2026

Blue Book on the Development of Generative AI Education Products

From Chatbots to Agents, Multimodality, and On-Device AI: Five Scenarios, Product Landscape, Benchmarks, and Roadmap

AI-SLI

2026 · AI-Assisted Research Edition

A Note on Method: How This Report Was Produced with AI Assistance

AI-SLI · Produced with AI Assistance

This blue book represents a systematic exploration by AI-SLI into AI-assisted knowledge production. Throughout its development, generative artificial intelligence served as a research assistant under the direction and oversight of the research team, undertaking four kinds of labour-intensive work. First, evidence-based research and data verification: fanning out multi-source searches across vendor public disclosures, primary policy texts, academic literature, and authoritative benchmarks; tracing every market figure, benchmark score, product specification, and policy-document number cited in this volume to its primary source, corroborating it with no fewer than two independent sources, and issuing corrections where warranted; and rigorously distinguishing "verified" from "yet to be verified" for frontier products and literature. Second, market sizing and literature synthesis: mapping the product landscape and technical lineage across the five scenarios of Empowering Teaching, Supporting Learning, Supporting Research, Intelligent Assessment, and Governance and Safety; triangulating market sizes and the competitive landscape across sources; and building on this basis an evaluation framework spanning the five scenarios. Third, drafting, figure generation, and citation management: composing the chapters and rendering the product-landscape and data figures to a unified specification, and maintaining a fully traceable citation apparatus with graded credibility annotations. Fourth, a parallel bilingual edition: producing semantically aligned Chinese and English versions under a controlled terminology glossary.

We state plainly that the role of AI here was to carry out the labour-intensive work of searching, sizing, drafting, illustrating, and managing citations; the choice of subject, the judgements of value, the industry assessments, and the final conclusions were directed and vouched for by the research team. Every datum and citation is required to be real and independently verifiable, and every market figure is required to be corroborated by no fewer than two independent sources; wherever verification was not possible, the text is marked [TBD] rather than invented. We offer this report as a forward-looking reference workflow for colleagues across the education industry to scrutinise and improve upon — a sincere experiment in a new paradigm of knowledge production, and in no way a substitute for expert judgement or professional due diligence.

Table of Contents

Chapter 1 Introduction: From Conversational AI to Agents, Multimodality, and On-Device Computing — the Paradigm Shift in Generative-AI Education Products

1.1 From "Report" to "Blue Book": Upgrading the Research Standard

This Blue Book is the annual flagship study in AI-SLI's Smart Education Blue Book series on generative-AI education products. It builds on the "teaching–learning–research" three-scenario tradition established by the earlier *Generative AI Product Development Report*, but it rebuilds the research object, methods, and evidentiary standard from the ground up. That rebuild is warranted by a real shift in product form over the past two years: what was being observed when the earlier report's scope was fixed is no longer the same category of thing being deployed on campuses today.

This report is written for three audiences at once. For **education policymakers and school administrators**, it offers a framework for judging whether a given product is worth adopting and what red lines must hold if it is. For **education researchers and classroom teachers**, it lays out the mechanics and limits of three technology threads, so readers can distinguish real capability from marketing claims. For **industry**, it offers an evidence-based comparative view of where the paradigm shift is genuinely furthest along and where the gaps remain. What unites these three audiences is a shared need for **restrained, verifiable analysis that discusses both capability and its limits** — a stance this report deliberately holds itself to, in contrast to product-showcase writing or techno-optimist narratives.

When the previous report was written, generative-AI education products were still dominated by a single paradigm: the conversational large language model. Natural-language dialogue was the interface, single- or multi-turn Q&A was the core capability, and a general-purpose cloud model — adapted to education through prompt engineering and light fine-tuning — sat underneath it all. That paradigm can be dated to a clear starting point: Khan Academy's launch of Khanmigo alongside GPT-4 in March 2023 — an AI tutor that answered questions and coached students through dialogue, and which represented the ceiling of product imagination at the time. This paradigm delivered real breakthroughs in concept explanation, content generation, and tutoring, but it also exposed structural limits in a domain as scenario-bound, constraint-heavy, and accountability-laden as education: it lacked controllable task orchestration (Q&A stopped after one turn, with no capacity to autonomously complete a multi-step workflow); it had no native understanding of multimodal learning materials (it couldn't read a blackboard, handwriting, or a lab demonstration); it had no awareness of the real physical context teachers and students were in (it wasn't present in the

classroom and had no idea what a student was doing at that moment); and it left no auditable trail of process or accountability (generation was the end of the interaction, with little to trace back to). Lu Yu and Tang Xiaoyu of Beijing Normal University, writing in *Modern Educational Technology* (issue 6, 2025), observe that generative AI's classroom applications today "mostly remain at a shallow, instrumental level of use ... confined to basic tasks such as material retrieval and resource generation, without yet tapping its deeper potential to empower teaching" — a diagnosis that is precisely the scholarly starting point for this Blue Book's choice to organize itself around "paradigm shift" rather than "product review." The three threads that follow can be read as direct responses to those four gaps: agentic capability answers the lack of task orchestration; multimodality answers the lack of multimodal understanding and situational awareness (especially once paired with on-device hardware); on-device computing answers the lack of a trustworthy, low-latency, compliant deployment form; and all three converge on the governance demand for an auditable chain of process and accountability.

By 2026, product form has shifted in substance. This Blue Book accordingly upgrades its research scope from "review of conversational products" to "industry and policy research on a paradigm shift," making two structural changes. First, it **adds two new scenarios — Intelligent Assessment and Governance and Safety — to the original three**, extending the analytical frame from "how AI teaches and learns" to "how AI judges learning outcomes" and "how AI can be made trustworthy and controllable in education." Second, it shifts the observational focus of the product landscape from the chat interface to three technology threads advancing in parallel: **agents, multimodality, and on-device hardware**. Neither change is an editorial preference; both respond to three verifiable developments between 2024 and 2026. The Ministry of Education's Teaching Guidance Committee for Basic Education issued the *Guidelines for General AI Education in Primary and Secondary Schools (2025 Edition)* and the *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* simultaneously on May 12, 2025, writing "supporting teacher instruction, fostering student growth, and advancing smart education management" into the same normative text as "safeguarding student privacy and data security." Mainstream education companies have been rolling out agentic and on-device product iterations in quick succession since early 2025. And the engineering conditions for multimodal and on-device small models matured over the same window. Together, these three developments form the empirical basis for this report's central claim: a shift from conversational AI to agents, multimodality, and on-device computing.

1.1.1 Why "Paradigm Shift" Rather Than "Product Iteration"

To be clear about our terminology: this Blue Book deliberately chooses "paradigm shift" over language like "product iteration" or "version upgrade," because what happened over the past two years was not the same class of product improving continuously along one capability axis — it was simultaneous change in **capability composition, interaction form, and accountability structure**. A

system that only answers questions inside a chat box, and a system that can decompose a goal, call tools, remember a student's learning history across sessions, and perceive handwriting and speech offline on a local device, do not differ in degree — they differ in kind, in what they can do at all. In Kuhnian terms, the yardstick used to evaluate the former — accuracy and fluency of a single-turn answer — is no longer adequate to evaluate the latter; the latter requires a new set of yardsticks built around task completion, process controllability, and traceable accountability. This is precisely why this Blue Book simultaneously rebuilds both its research object (the product landscape) and its evaluation instruments (the benchmarking framework): when the paradigm changes, the ruler used to measure it has to change too.

At the same time, we use the word "shift" advisedly: a shift denotes direction, not completion. In terms of actual penetration on campuses, what this Blue Book observes is closer to a picture of "old and new coexisting, at uneven depth" — most products currently on the market are still primarily conversational Q&A tools, with agentic, multimodal, and on-device capabilities layered on top as selling points whose reliability and usability have yet to be independently verified. This report therefore affirms the direction of travel while consistently keeping "vendor claims" and "reproducible capability" in two separate columns, rather than letting trend narrative substitute for verified fact.

1.1.2 Scope, Time Frame, and Division of Labor with Companion Volumes

This Blue Book defines its object of study as "products and systems built around generative AI as a core capability, aimed at education settings" — spanning software forms (apps, platforms, agent services) and hardware-software integrated forms (AI learning tablets, AI tutoring pens, AI glasses, education robots, and the like), covering K-12 and higher education and touching on some vocational and lifelong-learning scenarios. The study's time frame centers on 2024–2026, with the structural shift in product form during that window as the primary object of observation; earlier technical history is traced only as far as necessary, and longer-range trends are projected only where the evidence supports it.

Within the Blue Book series, this volume has a clear, non-overlapping division of labor with two companion volumes. This volume focuses on the overall landscape and five-scenario analysis of the generative-AI education product spectrum as a whole; the *AI Glasses Education Industry Blue Book 2026* goes deep on the on-device multimodal hardware cross-section — a dedicated volume on the "on-device" thread as it plays out in the glasses category; and the *Global Education Robotics Development White Paper 2026* goes deep on embodied intelligence — the point where the "agentic + multimodal + on-device" threads converge in robotic form. Readers can treat the three volumes as complementary: this volume supplies the framework and the panorama, and the two companion volumes supply category depth.

> This Blue Book's companion volumes — the *AI Glasses Education Industry Blue Book 2026* and the *Global Education Robotics Development White Paper 2026* — each develop the paradigm shift described

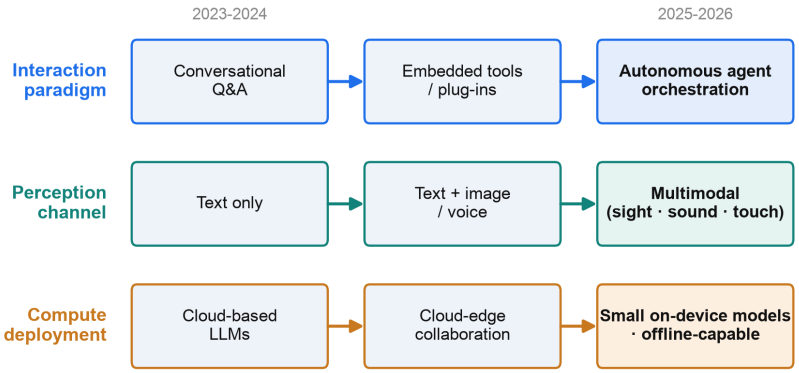
in this chapter in depth, from the on-device multimodal hardware and embodied-intelligence cross-sections respectively. Readers may consult them alongside this volume; see the two reports referenced above for details.

1.2 Three Threads: Agentic, Multimodal, On-Device

We summarize the evolution of generative-AI education products from 2024 to 2026 as three intertwined, mutually reinforcing technology threads. They are not a linear sequence of replacements but a layered, converging set of capabilities: agents supply the "can get things done" skeleton, multimodality supplies the "can perceive" input-output channels, and on-device computing supplies the "close at hand, trustworthy, low-latency" deployment form. The key to understanding these three threads is their **complementarity** — a system that can only converse, cannot read a blackboard, and must stay connected to the cloud is simply not in the same league, in terms of classroom usability and trustworthiness, as a system that can plan tasks, natively understand text, images, audio, and video, and respond instantly on a local device.

It's worth noting that none of these three threads appeared out of nowhere in 2024. The pursuit of "agent-like" capability in education technology has a long history: from early teaching machines, computer-assisted instruction (CAI), and intelligent tutoring systems (ITS) to adaptive-learning platforms, the field has always been searching for technology that can "perceive the student and teach autonomously." Gu Xiaoqing and Hao Xiangjun describe this lineage as "a century-long history of agent-like technology empowering education." Generative AI's real contribution isn't inventing the concept of an agent — it's the first time general-purpose language and multimodal understanding have been strong enough that planning, dialogue, perception, and generation can all run on a single model foundation, turning what used to require heavy rule engineering and domain customization into an intelligent tutor that can be built relatively generically and iterated quickly. The three threads are simply this qualitative shift projected onto three dimensions: capability (agents), perception (multimodality), and deployment (on-device).

Fig. 1 Three converging trajectories in generative-AI education products



Source: this report's analytical framework (see Chapter 1).

1.2.1 Agentic Capability: From "Answering" to "Getting Things Done"

The product logic of the conversational paradigm is "the user asks, the model answers" — the interaction loop closes after a single generation. The product logic of the agentic paradigm is "the user states a goal, and the system autonomously plans and calls tools to complete the task" — the interaction loop extends into the actual education workflow. The technical building blocks typically include task planning and decomposition, tool/plugin use, Retrieval-Augmented Generation (RAG), long-term memory and state maintenance, and multi-agent orchestration.

Mechanically, each of these five building blocks fixes a specific structural weakness of the conversational paradigm. **Planning** breaks a vague goal ("help me prepare a lesson on quadratic functions") into a sequence of executable sub-steps and adjusts it dynamically as intermediate results come in — moving the system from "one-shot generation" to "multi-step progress." **Tool use** lets the model step outside its own parameters to call external capabilities: looking up a question bank, computing values, generating a slide deck, writing into a learning management system (LMS) — turning "saying" into "doing." **RAG** retrieves relevant material from an external knowledge base before generation and injects it into context, constraining the model's answer to the teacher's course materials and authoritative references rather than letting it improvise purely from parametric memory. This mechanism matters especially in education, because it moves "what the answer is based on" from the model's uncontrollable internal memory to course material the teacher controls — researchers describe this as letting "the teacher retain control over the underlying literature the large model draws on." **Long-term memory** lets the system maintain continuous awareness of a given student's learning history across sessions, weaving fragmented question-and-answer exchanges into a traceable learning trajectory. **Multi-agent orchestration** splits different functions across multiple agents working in coordination, with an overseeing coordinator agent. Over the past two years, the literature on "Agentic RAG" and on memory in the age of AI agents has matured into a fairly

systematic body of review work, giving educational-agent engineering a reusable methodological skeleton to build on.

In an education context, agentic capability means a product can take on a more complete unit of work — for example, automatically running the full chain from "lesson prep → generate study guides → assign homework → grade → generate feedback," rather than offering piecemeal help at just one link in that chain. Gu Xiaoqing and Hao Xiangjun of East China Normal University, writing in the *Journal of East China Normal University (Educational Sciences)* (issue 5, 2025), use the metaphor of "Sun Wukong's hairs" to trace "a century-long history of agent-like technology empowering education," examining the digital-intelligence trajectory from teaching machines to intelligent learning companions, and systematically mapping how the shift "from standalone agents to multi-agent collaboration" is reshaping learning-technology systems — arguing that educational agents are evolving from single-point tools into collaborative "intelligent actors." Lu Yu and Tang Xiaoyu's four-tier framework offers a further calibration of this evolution: their intermediate tier, "human-AI collaboration and innovation activation," is explicitly defined as "generative AI functioning as a highly autonomous teaching agent, forming a multi-directional interactive relationship with teachers and students," while their advanced tier, "cognitive integration and thinking formation," points toward "integrated empowerment through the symbiosis of multiple agents" — a description that aligns closely with what this Blue Book calls "agentic capability."

Over the past two years, a body of academic work has emerged built around multi-agent education frameworks that divide labor among "teacher agents, student agents, and evaluator agents," offering engineering precedents for this form. Some proposed architectures unify three tiers at once — student-level personalization, educator-level automation, and institution-level intelligence. Others model themselves on a "professor–student" relationship, with a coordinator agent overseeing a team of "verifier" and "executor" agents that jointly plan, retrieve, design, and deliver lecture content. In writing-feedback frameworks, a teaching-assistant agent grades first and a teacher agent adjudicates, with students able to appeal their scores. What these designs share is a rewriting of "one model answering independently" into "multiple agents, each with a defined role, capable of mutual evaluation and negotiation, jointly completing an education task" — which is the technical substance of "from answering to getting things done."

One of the most scaled examples on the product side is Khan Academy's AI tutor, Khanmigo. According to the organization's own disclosures, the number of students and teachers using Khanmigo across partner districts grew from roughly 68,000 in the 2023–24 school year to more than 700,000 in 2024–25, with partner districts expanding from 45 to more than 380, and a target of crossing one million users in 2025–26. On the teacher side, Khanmigo is now completely free, used for differentiated instruction, generating lesson plans, writing questions, grouping students, and producing grading rubrics — the "administrative" side of teaching. Notably, Khanmigo's design deliberately distinguishes itself from a generic chatbot: it "doesn't hand over answers directly, but

patiently guides learners to work through problems themselves" — a design choice that is itself a product-level answer to the principle that agentic capability must not be conflated with automating learning away. Khan Academy ran a series of rigorous product experiments between October 2025 and April 2026 to test which changes actually improved the tutor's effectiveness — this practice of "testing agentic teaching efficacy through controlled experiments" is exactly the evidentiary posture this Blue Book advocates for.

To make the "from answering to getting things done" distinction concrete, consider a typical workflow. Under the conversational paradigm, a teacher preparing a lesson has to make a series of separate requests to the model: "Write me a set of learning objectives for quadratic functions," "Now give me three worked examples," "Make this explanation sound more conversational" — each one a separate, independent Q&A turn, with the teacher stitching the pieces together by hand. Under the agentic paradigm, the teacher states a single goal: "Help me prepare a 40-minute introductory lesson on quadratic functions for eighth graders, with five tiered practice problems and a short in-class quiz." The system then **plans** a sequence of sub-tasks, **retrieves** the teacher's uploaded textbook and curriculum standards (RAG), generates the lesson plan, worked examples, and quiz from that material, **calls tools** to format it into a downloadable courseware package, and **remembers** this teacher's past style preferences to keep the output consistent. The teacher's role shifts from issuing instructions line by line to reviewing and revising. The value here isn't that any single step is executed better — it's that scattered fragments get assembled into a deliverable finished product. That is the line between "getting things done" and "answering."

The industry-level picture points the same way. Gartner projected in 2025 that by 2028, 33% of enterprise software applications will have agentic AI built in — up from under 1% in 2024 — while simultaneously issuing a cautionary note: weighed down by rising costs, unclear business value, and inadequate risk controls, more than 40% of agentic AI projects could be canceled by the end of 2027. This combination of high expectations and a high failure rate carries a particular warning for a high-accountability domain like education: agentic capability is not automatic progress — it is a systems-engineering undertaking that demands rigorous evaluation and governance constraints. It raises the bar for product evaluation, shifting the object of assessment from "quality of a single-turn answer" to "completion rate, reliability, and controllability of a multi-step task" — a system that is 95% reliable at each of nine steps has an overall success rate of only about 63%, and this pattern of "reliability decaying with step count" makes multi-step robustness a make-or-break property for agentic products. It also raises new governance questions: when an education task is completed by multiple autonomous agents working together, and something goes wrong, how should responsibility be apportioned among teacher, product, and model — or even between one agent and another? This is a design problem that cannot be avoided. This Blue Book's inclusion of "traceable chain of accountability" as a core governance metric is a direct response to that problem.

1.2.2 Multimodality: From "Text Dialogue" to "Native Understanding of Text, Image, Voice, and Video"

Education is inherently multimodal: blackboard writing, textbook illustrations, lab demonstrations, spoken explanation, handwritten homework, classroom video — all of these carry learning information. The conversational paradigm is text-first and struggles to handle this material directly — it either hands images off to a separate recognition module to convert to text before feeding a language model, or it simply can't process them at all. "Omni-modal" or "natively multimodal" models — represented by OpenAI's GPT-4o and Google's Gemini series — can understand and reason over text, images, audio, and even video within a single architecture, without relying on separately bolted-together adapters for each modality. The distinction between "native multimodality" and "bolted-on multimodality" isn't just terminology: a native architecture preserves richer semantic connections across modalities (for instance, hearing "here" while simultaneously seeing where a finger is pointing on a diagram), enabling interaction that comes much closer to real teaching. This engineering leap has let AI "see" a geometry diagram, "hear" spoken language, and "read" a handwritten answer sheet. Lu Yu and Tang Xiaoyu likewise rank "multimodal information understanding, knowledge reasoning, and content generation" as the first of three core capabilities through which generative AI empowers education.

Multimodality does more than widen the input-output channel — it redraws the boundary of what education tasks AI can support at all. Oral-language assessment, photo-based homework grading, recognition of lab procedures, and accessibility support have all become productizable as a result. Take spoken-language assessment: a text-only model cannot evaluate pronunciation or intonation, while a model with speech understanding can directly score a student's reading aloud on pronunciation accuracy, fluency, and intonation. Take homework grading: a multimodal model can recognize handwritten answer sheets, locate the exact step where an error occurred, and generate targeted feedback — without requiring the student to manually transcribe the problem first.

Deployment on the product side has been dense, and heavily concentrated in 2025. One trigger was the release of the domestic open-source reasoning model DeepSeek-R1 in early 2025: NetEase Youdao announced integration of DeepSeek-R1 on February 6, 2025; TAL Education announced on February 8 that its learning tablets and smart hardware would integrate DeepSeek and build in a "deep-thinking mode"; Zuoyebang, Yuanfudao, and others followed suit shortly after, triggering an industry-wide upgrade of underlying models. Multimodal education hardware followed in quick succession: NetEase Youdao released hardware in 2025 — including an AI tutoring pen — built on its self-developed "Ziyue" education LLM, combining photo recognition, voice interaction, and reasoning-based tutoring into a single pen-shaped device. Yuanfudao Group unveiled its first AI paradigm for education, "Xiaoyuan AI," along with the companion Xiaoyuan AI learning tablet, on April 15, 2025; the model layer combined its self-developed Yuanli LLM with DeepSeek-R1 in a

matrix, and the product took the form of a study companion robot with "the tablet as eyes and ears, the intelligent core as hands," handling daily study plans, diagnostic assessment, knowledge graphs, and end-to-end tutoring. What these products share is that they couple multimodal perception (reading homework, listening to reading aloud) with agentic workflow (diagnose → plan → tutor) on the same device — confirming this chapter's opening argument that the three threads are not independent in real products but layered together.

The other half of multimodality's value lies on the "output" side, and it's often overlooked in product narratives. Once a model can not only understand but also generate images, speech, and video, the way educational content gets produced changes: an abstract physics principle can be rendered instantly as a visualized animation, a foreign-language text can be synthesized into a listening-comprehension recording with standard pronunciation, a geometry problem can come with a step-by-step diagram of auxiliary construction lines. This capacity to "generate multimodal learning material on demand" matters especially for personalized and accessible education — converting a chart into structured spoken description for a visually impaired student, converting spoken explanation into real-time captions for a hearing-impaired student, generating illustrated explanations at varying levels of difficulty for students at different levels — all of this shifts from dependence on manual production to scalable generation.

At the same time, the academic community is building more cautious benchmarks for how well multimodal models actually suit education. One study uses K-12 classroom video as a benchmark to test whether multimodal LLMs can genuinely "understand" science instruction and reason pedagogically about it, and its findings serve as a warning to the field: there remains a considerable gap between a model "being able to see an image" and "being able to understand a classroom." This gap has a direct implication for product evaluation: a vendor demo succeeding at "recognizing a photo of a homework page" is not the same as succeeding at "understanding pedagogical intent across a continuous real classroom video" — the latter requires an independent benchmark much closer to real conditions. Similarly, multimodal generation capability demands its own caution — generating an explanation that sounds fluent is not the same as generating one that is factually correct, and multimodal expressiveness can sometimes mask content errors, making hallucinations harder to detect rather than easier.

1.2.3 On-Device Computing: From "Centralized in the Cloud" to "Device-Cloud Collaboration"

The third thread is the migration of compute and deployment form. As on-device model compression, dedicated inference chips, and miniaturized multimodal models have advanced, a portion of inference has begun moving from the cloud onto devices such as learning tablets, tablets generally, AI glasses, and education robots. The technical foundation underpinning this shift is the rapid maturation, over

the past two years, of small language models (SLMs) and quantization/compression techniques. SLMs represented by Microsoft's Phi series, Google's Gemma series, and Alibaba's Qwen series typically range from a few hundred million to a few billion parameters; paired with low-bit quantization such as 4-bit, they can run on just a few gigabytes of memory on phones, tablets, and lightweight edge hardware, and on a number of text and multimodal tasks they approach — or in places even surpass — much larger cloud models. Some of these small models can already handle speech, vision, and text simultaneously within a single architecture, providing a ready-made model foundation for localized multimodal inference on education devices. The significance of this progress is that it turns "offline-capable AI tutoring" from an engineering concept into a mass-producible product premise.

The drivers of on-device computing are especially pronounced in education, and can be grouped into three. First, **data compliance and privacy**: minors' learning data is highly sensitive, and processing it locally means it never has to leave the device, cutting off the risk of leakage at the source — a requirement the *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* addresses head-on, with provisions such as "strictly prohibiting teachers and students from entering sensitive data such as exam questions or personal identifying information," "establishing a sound tool whitelist system," and "genuinely safeguarding student privacy and data security" — forming the policy rationale for on-device computing in China's education context. Second, **latency and availability**: scenarios like real-time in-class feedback and instant homework grading demand millisecond-to-second response and cannot tolerate interruption from network fluctuation; on-device inference naturally satisfies this constraint. Third, **cost**: education use cases involve frequent, repetitive calls, and routing every single one through a cloud API adds up to substantial long-term inference cost, whereas a one-time hardware investment on-device sharply dilutes the marginal cost of high-frequency use.

On-device computing does not mean the cloud disappears — it means moving toward "device-cloud collaboration": lightweight, privacy-sensitive, real-time tasks are handled on-device, while complex reasoning, knowledge updates, and large-parameter model capability remain the cloud's job, with the two coordinated through dynamic task routing. The division of labor broadly follows three criteria: **privacy sensitivity** (processing involving minors' identity or learning data stays on-device wherever possible), **latency requirements** (interactions needing instant feedback run on-device; deep reasoning that can tolerate delay runs in the cloud), and **task complexity** (simple intent recognition and common Q&A are handled directly by on-device small models; cross-disciplinary complex reasoning and long-form generation are routed up to large cloud models). This division has already become the mainstream architecture in real products — the learning-tablet products mentioned above commonly adopt a hybrid form in which "on-device small models handle real-time interaction and privacy-sensitive processing, while cloud-based large models handle deep reasoning." It's worth noting that device-cloud collaboration also introduces a new evaluation dimension: what a product

can still do while offline, and which functions degrade, becomes a real, measurable indicator of its "on-device authenticity" — one that cannot simply be taken on a vendor's claim that it "supports a local large model." Market data corroborates the industry heat around on-device computing: per RUNTO, China's learning-tablet sales across all channels reached about 5.923 million units in 2024, up 25.5% year-on-year, with sales revenue of about RMB 19.06 billion; separately, per IDC, China's learning-tablet shipments in Q2 2025 reached about 1.54 million units, up roughly 44.6% year-on-year, with the wave of DeepSeek integration and government subsidy policy widely seen as major drivers of this growth cycle in 2025. (Different research firms' figures for the same market vary in methodology and scope; this report cross-verifies and reconciles these figures in its industry chapters, and cites the data here only to illustrate the intensity of industry interest in on-device computing.)

This thread also has a more forward-looking projection on hardware product lines. NetDragon (HK:0777), an affiliated organization of this Blue Book, has strategically invested in AI-glasses and AR-interaction company Rokid since 2023, positioning itself in on-device multimodal hardware around education and interaction scenarios; according to public reporting, NetDragon's strategic investment in Rokid started at US$20 million with subsequent capital increases. Rokid's all-in-one AI+AR glasses, Rokid Glasses, won a Best of CES 2025 award at CES 2025 and moved toward mass-market launch within 2025. This kind of "on-device perception, device-cloud collaborative inference" hardware form is precisely the subject of deep-dive research in this Blue Book's companion volume, the *AI Glasses Education Industry Blue Book 2026*; this chapter does not expand on it further, and notes only its coordinates within the three threads: on-device computing gives agentic and multimodal capability a form factor that is close at hand, trustworthy, and low-latency, bringing AI back from "a chat box in the cloud" to "a device at the learner's side." (Specific figures on NetDragon's investment amounts and the pace of Rokid's product launches are subject to the cross-verified figures presented in the industry chapters.)

1.2.4 Where the Three Threads Meet: Convergence, Not Replacement

It bears repeating that agentic capability, multimodality, and on-device computing are not three independent technology curves — in real products they interlock and reinforce one another. Take away any one of them, and the value of the other two is discounted: an agent that can plan tasks but can't read a student's handwritten homework (missing multimodality) has "getting things done" capability that stops at the text layer; a natively multimodal model that has to upload every homework photo to the cloud for processing (missing on-device computing) carries the double burden of latency and minors' data-compliance risk; a small model running locally that can only do one-shot Q&A, with no tool use or memory of a student's learning history (missing agentic capability), falls right back into the old conversational paradigm.

Layer all three together, and an ideal form emerges: **the on-device multimodal agent** — one that natively sees, hears, and reads learning material on a device at the learner's side (on-device), while

autonomously planning, calling tools, and maintaining continuous memory to complete a full education task (agentic). This form has not yet become mainstream as of 2026, but product lines like learning tablets, AI tutoring pens, and AI glasses are already converging toward it. This Blue Book accordingly treats "the degree to which the three threads converge" as a yardstick for gauging product maturity: the more deeply a product fuses all three, rather than tacking on just one as a marketing label, the closer it sits to the real frontier of the paradigm shift.

1.3 The Five-Scenario Framework: This Blue Book's Analytical Axis

Building on the three threads above, this Blue Book examines generative-AI education products across **five scenarios**. The first three carry forward and update an existing tradition; the last two are new to the 2026 edition. This framework did not emerge out of thin air: the three application categories the Ministry of Education's Teaching Guidance Committee for Basic Education set out in its 2025 *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* — "supporting teacher instruction, fostering student growth, and advancing smart education management" — echo this framework's "Supporting Teaching, Supporting Learning, Supporting Research (Professional Development)/Governance" closely; and Lu Yu and Tang Xiaoyu's language of empowerment across "teaching, learning, assessment, and support" scenarios provides the scholarly basis for giving Intelligent Assessment its own, independent scenario.

Why five scenarios, rather than sticking with an existing division such as "teach–learn–assess–support" or "teach–learn–research"? This report's reasoning rests on three considerations. First, **separating "assessment" out from "teaching"**: under the conversational paradigm, assessment was largely treated as a subordinate part of teaching (writing questions, grading), foldable into "Supporting Teaching." But under the generative paradigm, AI has begun independently taking on scoring, diagnosis, and feedback, and the reliability and fairness of that work constitute a distinct set of technical and ethical questions that need their own scrutiny. Second, **adding "Governance and Safety" as a cross-cutting scenario**: governance isn't another category of "application" sitting alongside teaching, learning, assessment, and research — it's a constraint layer running through all of them. Listing it explicitly as its own scenario forces every product analysis in this report to answer whether it is safe, whether it is compliant, and whether responsibility can be traced — rather than treating governance as an optional appendix. Third, **retaining "Supporting Research (Professional Development)" to preserve institutional memory**: teaching research (jiaoyan) is a distinctive institution in China's basic education system and the primary channel for teacher professional development; generative AI's contribution to it shouldn't be absorbed into either "teaching" or "learning," so it remains its own category. The five scenarios thus form a grid of "three verticals (teaching, learning, research) plus two horizontals (assessment, governance)" — carrying tradition forward while responding to what's genuinely new.

Scenario	Core Question	Key Update from the Previous Edition	Representative Product Forms
Supporting Teaching	How does AI help teachers teach?	Expanded from lesson-prep/question-writing assistants to teaching agents and multimodal classroom sensing	Teaching agents, lesson-prep/question-writing assistants, classroom analytics tools (see product landscape in later chapters) [TBD: list of selected products and selection criteria]
Supporting Learning	How does AI help students learn?	Expanded from conversational Q&A to on-device learning companions, multimodal tutoring, and personalized pathways	AI learning tablets, AI tutoring pens, AI learning-companion apps (e.g., Xiaoyuan AI, Youdao-related products; see later chapters) [TBD: complete list]
Supporting Research (Professional Development)	How does AI support teacher professional development and teaching research?	Expanded from resource generation to research agents and evidence-based analysis	Research agents, evidence-based teaching-research platforms (see later chapters) [TBD: product list]
Intelligent Assessment (new)	How does AI judge learning and teaching outcomes?	Reliability of formative assessment, multimodal testing, AI-generated questions, and automated grading	Spoken-language assessment, homework grading, automated test assembly and scoring systems (see later chapters) [TBD: product list]
Governance and Safety (new)	How is AI made trustworthy and controllable in education?	Content safety, minor protection, data compliance, explainability, and chain of accountability	Tool whitelists, content-safety filtering, compliance auditing mechanisms (as required by the *Guidelines*) [TBD: list of institutional mechanisms/products]

The three carried-forward scenarios are not preserved unchanged in this edition — they are upgraded in step with the three threads. **Supporting Teaching** expands from early-stage lesson-prep/question-writing assistants to teaching agents capable of sensing the classroom and orchestrating instructional workflows, with its evaluative center of gravity shifting from "is the generated content usable" to "does it genuinely reduce teacher workload and improve instruction" — as the multi-agent research on reducing teacher workload cited above shows, workload reduction itself has become a quantifiable

product objective. **Supporting Learning** expands from conversational Q&A to on-device learning companions and personalized learning pathways; Khanmigo's design of "guiding rather than handing over answers directly," and the closed loop of "learning diagnosis → knowledge graph → daily practice" in domestic learning tablets, are both product-level expressions of this upgrade. At the same time, academic warnings that "over-automation may weaken students' metacognitive engagement and self-regulated learning" are a reminder that this scenario must guard against the paradox of "too much help producing less learning." **Supporting Research (Professional Development)** expands from scattered resource generation to research agents and evidence-based analysis, moving AI from "producing materials for teachers" to "helping teachers do research."

The new Intelligent Assessment scenario responds to generative AI's leap from "assisting generation" to "assisting judgment" — once AI begins participating in scoring, diagnosis, and feedback, its reliability, fairness, and bias control stop being merely technical concerns and become education concerns. The risk in that leap should not be underestimated: existing systematic reviews identify "technical reliability and hallucination" as the most-discussed challenge in generative AI's education applications, followed by over-reliance, the evolution of assessment and fairness, and privacy and safety; research specifically on the safety of LLM-based automated grading further shows that grading agents still have exploitable weak points in robustness and trustworthiness — for instance, specific manipulative text embedded in an answer can skew the resulting score. When a single score can affect a student's academic prospects and self-perception, "can the scoring model be gamed, and does it systematically favor certain kinds of students" stops being a technical footnote and becomes a matter of educational equity.

Precisely for this reason, intelligent-assessment products should not be designed as "AI adjudicates alone" but should move toward "human-AI collaborative assessment": AI handles the initial pass, flags anything suspicious, and provides evidence and confidence levels, while final judgment stays with the teacher. The multi-agent essay-feedback framework described earlier — where a teacher agent adjudicates a teaching-assistant agent's scoring and students are allowed to appeal — is one technical realization of exactly this approach. It also means that the evaluative priority for the Intelligent Assessment scenario is not just the agreement between AI scoring and human scoring (correlation coefficients, scoring gaps) but the **discoverability and correctability** of its errors — a system that occasionally errs but whose errors are easy for a teacher to spot and overturn may be a better fit for the classroom than a marginally more accurate "black-box scorer" that is hard to challenge. This evidence shows that giving assessment its own scenario, with independent evaluation and constraints, is not conceptual overreach but a necessary response to a real risk.

The new Governance and Safety scenario responds to the accountability questions that become unavoidable once products enter real campuses at scale: content safety for minor users, compliant use of learning data, protection against model hallucination and misinformation, and defining the boundary of human-AI accountability. The urgency here comes from the special nature of the

population involved — AI systems aimed at minors face additional risks that adult-facing products don't: children's particular vulnerability to unsafe content, and the cognitive risk that "misinformation delivered by AI in an authoritative tone is more readily mistaken for truth by students." Encouragingly, governance is moving from abstract advocacy toward something measurable and engineerable: dedicated benchmarks for content risk to minors (such as MinorBench), and evaluation frameworks that unify "safety, usefulness, and pedagogy" for education-focused LLMs, have both emerged, turning "safety" from a slogan into a scorable, comparable metric. At the policy level, the five application principles established by the *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* — education-oriented values, educational equity, values guidance, needs-driven design, bottom-line thinking — together with grade-differentiated usage norms (elementary students use open-ended content generation with teacher and parent assistance; middle-school students are guided to cross-verify the reasonableness of generated content; high-school students independently assess the social impact of generated content in light of the underlying technical principles), provide a localized institutional anchor for this scenario. This "the younger the student, the more constrained the use" design is itself a governance mechanism that can be built into a product — an ideal education AI product should be able to identify a user's grade level and adjust its autonomy and content openness accordingly, rather than treating every user the same.

Taken together, the five scenarios form a complete chain running from "building capability" to "verifying outcomes" to "controlling risk": Supporting Teaching, Supporting Learning, and Supporting Research (Professional Development) address what AI can do for education; Intelligent Assessment addresses how good and how accurate that work actually is; Governance and Safety addresses whether doing it is safe, compliant, and accountable. All three are indispensable: capability without assessment leaves a product's claims unverifiable; capability and assessment without governance leave risk unconstrained. These two new scenarios extend this Blue Book's framework from a "capability map" into a complete "capability–assessment–governance" chain — a product narrative that discusses only capability, without addressing how its outcomes are judged or its risks constrained, is incomplete, and irresponsible.

1.4 Research Method and Evidentiary Principles

As a flagship Blue Book positioned as industry and policy research, this report holds itself to a method of "research first, evidence first, conservative accuracy." This methodological stance follows directly from two features of what this report observes. First, generative-AI education products iterate extremely fast, and any static "capability checklist" risks going stale by the time this book is printed — so this report places more weight on offering a sustainably reusable **analytical framework** than on a perishable product ranking. Second, education is a high-accountability domain, and any overstated capability claim can end up being paid for by students — so this report's posture

toward fact is "better an acknowledged gap than an invented fact." This is expressed concretely in four ways.

- **Product mapping**: Using scenario as one axis and capability as the other, this report structures and cross-compares generative-AI education products currently on the market, aiming to show the real distribution of the paradigm shift rather than a list of individual cases. The value of mapping is that it reveals how the three threads' capabilities are distributed across products — which products only do conversational Q&A, which have layered on multimodality, which have genuinely achieved on-device agentic capability — turning a vague claim of "paradigm shift" into a checkable structural table. The baseline inclusion criteria are "aimed at K-12 or higher education, possessing generative capability, and on sale or publicly released between 2024 and 2026." [TBD: final count of included products and item-by-item screening criteria]
- **Capability benchmarking**: This report draws on education-vertical large language model (LLM) benchmarks, AI hardware evaluations (for learning tablets, AI glasses, and the like), and comparative product reviews, presented visually through capability radar charts and timelines, to avoid substituting vendor claims for independent observation. We place particular weight on distinguishing "vendor-claimed capability" from "third-party reproducible capability" — as the K-12 classroom-video benchmark cited earlier shows, the two can diverge systematically — and on distinguishing "general benchmark score" from "usability in an actual education setting," since a model that scores well on a general-purpose Q&A benchmark won't necessarily work stably, safely, and pedagogically soundly in a real classroom. [TBD: the evaluation dimensions, sample sizes, and data cutoff date used in this report]
- **Dissection of the new paradigms**: This report breaks down agent orchestration, RAG, long-term memory, and other supporting paradigms at the mechanism level, explaining the conditions under which they fit education settings and the boundaries at which they fail. For example, RAG's value in education lies in "constraining answers to course material the teacher controls," while its failure boundary lies in poor retrieval quality or an outdated knowledge base — a RAG system that retrieves the wrong material will deliver a wrong answer in a more confident tone, making the error harder to catch. Long-term memory improves personalized continuity, but it equally amplifies compliance risk around minors' data — every piece of a student's learning history that gets remembered is a piece of privacy that needs protecting. Multi-agent orchestration improves completion rates on complex tasks, but it makes attributing responsibility harder when something goes wrong. Presenting mechanism and boundary side by side is this report's deliberate method for avoiding techno-optimism: wherever this report describes a capability, it tries equally hard to describe the conditions under which that capability fails.
- **Verifiable sourcing**: This report strictly distinguishes among three types of statement — "verified fact," "trend judgment," and "analytical inference" — and marks that distinction both

in the text and in footnotes. Any factual claim involving market size, shipment volumes, market share, user counts, specific policy-document numbers, or literature citations is backed by a genuine public source noted in the footnote; where evidence is insufficient, we leave a placeholder rather than guess. On market size specifically, third-party research firms' estimates of the future size of the global and Chinese "AI + education" markets vary substantially by methodology and scope, and this report does not adopt any single uncross-verified figure in the main text — related figures are reconciled and cross-verified consistently in the industry chapters; figures involving currency strictly separate RMB, US dollars, and HK dollars, and are never mixed. [TBD: the market-data sources, methodology, currencies, and cutoff date used in this report]

On data boundaries, this report also holds to a principle of "caution with affiliated enterprises": for NetDragon (HK:0777), an enterprise affiliated with this institution, and its related products and investments, this report applies exactly the same evidentiary standard as for any other company — every specific figure is sourced to a public, verifiable reference, neither loosened nor tightened because of the affiliation, in order to preserve the independence and credibility of the research.

It should be specifically noted that this Blue Book was produced by AI-SLI, and that AI assistance tools were used in the writing process itself. This is, in one sense, this report putting its own research subject to the test; it also creates an added obligation of self-restraint. To keep the research rigorous and trustworthy, every factual claim generated by AI was traced back to its source and manually verified — any specific figure, product name, company name, benchmark score, or policy-document number that could not be traced to a credible primary source was left out of the main text entirely or marked as a placeholder. This practice is itself a self-application of the "defining the boundary of human-AI accountability" principle this report advocates in the Governance and Safety scenario; see this report's "Note on Preparation" page for further detail.

1.5 How This Blue Book Is Organized

The book as a whole is organized around the logic of "paradigm — scenario — evaluation — governance — outlook," forming a closed chain running from an overarching judgment through detailed analysis and back to policy recommendations. This chapter (Chapter 1) establishes the overarching judgment of a paradigm shift and the five-scenario analytical framework, and serves as the coordinate system for the rest of the book. The chapters that follow proceed, broadly, across three tiers. **The first tier is the overall industry and product landscape** — a structured classification and cross-comparison of generative-AI education products, answering "what products actually exist in the market, and how are they distributed across the three threads." **The second tier goes deep scenario by scenario** across the five scenarios — Supporting Teaching, Supporting Learning, Supporting Research (Professional Development), Intelligent Assessment, and Governance and

Safety — in turn, breaking down each one's mechanisms, surveying its representative products, and identifying its fit conditions and failure boundaries. **The third tier covers evaluation and outlook** — introducing education-vertical LLM benchmarks and AI hardware evaluations to independently check vendor claims, and building on that foundation to offer development recommendations and trend projections aimed at policymakers, education institutions, and industry. These three tiers build on one another, jointly supporting this report's unified "capability–assessment–governance" analytical thesis. [TBD: final correspondence of chapter titles, chapter order, and page numbers]

We hope this Blue Book serves not merely as a catalog-style inventory of products, but as an analytical framework that helps educators, policymakers, and industry understand what generative-AI education products are actually becoming. The paradigm shift is still underway — and as Gartner's warning about the high failure rate of agentic projects, and the academic community's ongoing concerns about hallucination and assessment fairness both suggest, this process is neither linear nor automatically benign: agentic capability can bring a "collapse in reliability across multi-step tasks," multimodality can bring the illusion of capability where a system "can see but can't understand," and on-device computing can create new tension between convenience and compliance. Acknowledging these tensions is precisely the precondition for advancing this shift responsibly. What this report presents is a structural picture of this process at the 2026 time cross-section, together with a set of evidence-based tools for observing it going forward; we hope it serves as a starting point that can be tested, corrected, and updated year after year — not a final, once-and-for-all conclusion.

Chapter References

1. Lu Yu, Tang Xiaoyu. Generative AI Empowering Classroom Teaching: Levels of Form and Pathways of Advancement [生成式人工智能赋能课堂教学的形态层级与进阶路径] [J]. Modern Educational Technology (电化教育研究), 2025, 46(6): 75-82+106. School of Educational Technology, Beijing Normal University. DOI:10.13811/j.cnki.eer.2025.06.010. (Original PDF: https://aver.nwnu.edu.cn/upload/formalarticle/202506/2025061005-生成式人工智能赋能课堂教学的%20 形态层级与进阶路径.pdf ; also at https://aic-fe.bnu.edu.cn/docs/2025-07/a4d9baa90d9e49d9920ee2c46dd283b7.pdf)
2. Gu Xiaoqing, Hao Xiangjun. Sun Wukong's Hairs: The Multi-Agent Systems Reshaping Learning Technology [悟空的毫毛：正在重塑学习技术系统的多智能体] [J]. Journal of East China Normal University (Educational Sciences) (华东师范大学学报（教育科学版）), issue 5, 2025. Faculty of Education, East China Normal University. URL: https://ercdee.ecnu.edu.cn/b8/9d/c16571a702621/page.htm

3. Teaching Guidance Committee for Basic Education, Ministry of Education. Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition) [中小学生成式人工智能使用指南（2025 年版）] and Guidelines for General AI Education in Primary and Secondary Schools (2025 Edition) [中小学人工智能通识教育指南（2025 年版）]. Issued 2025-05-12. (Full text via Beijing Daily client: https://xinwen.bjd.com.cn/content/s68217324e4b0ec1c3d96f32d.html ; interpretation via China Education and Research Network (CERNET): https://www.edu.cn/xxh/focus/zc/202505/t20250513_2667992.shtml)

4. Yuanfudao Group. "Xiaoyuan AI and Smart Hardware Strategy Launch Event" [小猿 AI 暨智能硬件战略发布会] (launching Xiaoyuan AI and the Xiaoyuan AI learning tablet, built on the Yuanli LLM plus DeepSeek-R1), 2025-04-15. China Daily (tech) report: https://tech.chinadaily.com.cn/a/202504/17/WS680071c2a310e29a7c4a9b3f.html (see also Sina Finance: https://finance.sina.com.cn/tech/roll/2025-04-15/doc-inetfxpu4284697.shtml)

5. DoNews (多知网). NetEase Youdao Launches New Education Smart Hardware [网易有道推出新款教育智能硬件] (Youdao AI Tutoring Pen Space X and others, built on the "Ziyue" education LLM), 2025. URL: http://www.duozhi.com/industry/insight/2025022317027.shtml

6. QbitAI (量子位) / Beijing Daily. Timeline of online education companies integrating DeepSeek-R1 [在线教育企业接入 DeepSeek-R1 时间线] (NetEase Youdao 2025-02-06, TAL Education 2025-02-08, and others). URL: https://www.qbitai.com/2025/02/254011.html ; https://xinwen.bjd.com.cn/content/s67b5d6bae4b068c68f1001dd.html

7. TechNode (动点科技). Targeting the Metaverse Opportunity: NetDragon Completes US$20 Million Investment in Rokid and Forms Strategic Partnership [瞄准元宇宙机会，网龙完成向 Rokid 投资 2000 万美元并达成战略合作], 2023-11-22. URL: https://cn.technode.com/post/2023-11-22/netdragon-rokid/ (Rokid Glasses winning Best of CES 2025 and moving to mass-market launch within 2025, see Securities Times (证券时报): https://www.stcn.com/article/detail/1579878.html)

8. Gartner. "Agentic AI to feature in 33% of enterprise software applications by 2028" (and "Over 40% of Agentic AI Projects Will Be Canceled by End of 2027"), 2025. URL:

https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

9. Khan Academy. Khanmigo AI tutor user scale and form [Khanmigo AI 导师用户规模与形态] (partner-district students/teachers grew to over 700,000 in the 2024-25 school year, teacher access now free). URL: https://blog.khanacademy.org/how-khan-academy-is-building-a-better-ai-tutor-our-most-recent-learnings/ ; https://www.khanmigo.ai/teachers
10. Systematic review: "Large language models in education: a systematic review of empirical applications, benefits, and challenges" [J]. ScienceDirect, 2025. URL: https://www.sciencedirect.com/science/article/pii/S2666920X25001699
11. Singh A. et al. "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG." arXiv:2501.09136, 2025. URL: https://arxiv.org/pdf/2501.09136 (education RAG tutoring applications also at ScienceDirect: https://www.sciencedirect.com/science/article/pii/S2590291125004796)
12. "MinorBench: A hand-built benchmark for content-based risks for children." arXiv:2503.10242, 2025. URL: https://arxiv.org/pdf/2503.10242 ; "GradingAttack: Exposing Security Vulnerabilities in LLM Based Educational Grading Agents," arXiv, 2026. URL: https://arxiv.org/pdf/2602.00979
13. Survey of on-device small language model deployment (Microsoft Phi, Google Gemma, Alibaba Qwen) on phones/tablets/edge devices [端侧小语言模型（Microsoft Phi、Google Gemma、Alibaba Qwen）在手机/平板/边缘设备的部署综述]. URL: https://www.digitalapplied.com/blog/small-language-models-business-guide-gemma-phi-qwen
14. "Can Multimodal LLMs 'See' Science Instruction? Benchmarking Pedagogical Reasoning in K–12 Classroom Videos." Springer Nature, 2025. URL: https://link.springer.com/chapter/10.1007/978-3-032-29755-6_28

Chapter 2 The Product Landscape, 2025–2026: Tracks, Form Factors, and Competitive Positioning

2.1 Why a "Landscape" at All: From Chat Products to Agentic Ecosystems

Around 2024, the dominant form of generative-AI education product was "a conversational LLM dressed up for the classroom": take a general-purpose or fine-tuned language model, wrap it in subject-specific prompts, a textbook knowledge base, and a chat front end, and the result was an "AI study buddy," an "AI teaching assistant," or an "intelligent grader." The capability ceiling of these products was essentially the capability ceiling of the underlying model — differentiation showed up in content operations and interface polish, not in any deeper paradigm shift. OpenAI's ChatGPT Edu, launched in May 2024 for higher education, is a clean example: it was pitched as "delivering enterprise ChatGPT to campus in a compliant, governable way," built on GPT-4o, spanning text and visual reasoning, supporting more than 50 languages, allowing institutions to build custom versions on their own data, and promising that conversation data would not be used for training. This is the archetypal "general model plus education-compliance wrapper" approach, and it grew directly out of early enterprise-tier deployments at Oxford, Wharton, UT Austin, Arizona State, and Columbia.

By 2025–2026, three technology threads converged to reshape the product landscape structurally — and in doing so, retired the old approach of "list every product on a single static spreadsheet":

- **Agentic**: Products moved from single-turn question-and-answer to agents that can plan, call tools, and execute multi-step tasks. The signal event here was Google's December 2024 release of Gemini 2.0, which the company explicitly positioned as built "for the agentic era," with native support for tool use — Google Search, code execution, custom third-party functions — plus multimodal input and output. Teaching, learning, and research-support tasks are now decomposed into orchestrable workflows, and product value has shifted from "answers accurately" to "closes the loop on a task."
- **Multimodal**: Beyond text, voice, images, whiteboard writing, screen content, physical objects, and classroom video have become first-class inputs. GPT-4o's native audio-to-audio interaction — capable of picking up hesitation, urgency, and irony in tone — together with the Gemini Live API and GPT Realtime-style interfaces pushing first-audio-chunk latency down to roughly 300–600 milliseconds, means products can now "see the classroom and understand the discussion." Assessment and feedback have correspondingly expanded from text-based homework to process-oriented, situated, multimodal evidence.

- **On-device**: Compute has moved from the cloud down onto AI learning tablets, tablets, AI glasses, and other endpoint devices, bringing low latency, offline capability, and data localization — and reshaping how value is split across hardware, model, and service. The 2024–2025 boom in China's AI learning tablet market (detailed in Section 2.5) is the direct projection of on-device computing onto the consumer education market.

These three threads reinforce rather than run parallel to one another: agentic behavior creates demand for "multi-step tasks that need cross-modal perception and tools"; multimodality supplies the input channel that lets a system "see the classroom, understand the discussion"; and on-device computing answers the deployment constraint that "perception happens in a real physical setting, so data shouldn't all go to the cloud." The net effect of all three together is to move the bar for product value up a level — from "is the single-turn answer correct" to "can a real, complete educational task be carried out reliably, controllably, and compliantly." That is precisely why the 2024-style static spreadsheet listing several dozen chat products broke down so quickly: once products are moving simultaneously along the paradigm, modality, and deployment axes, a list along any single axis can no longer capture the real competitive relationships between them.

This chapter therefore does not organize itself as a product directory. Instead, it builds a **competitive landscape** in more than two dimensions: the horizontal axis captures a product's **capability paradigm** (conversational → agentic orchestration), the vertical axis captures its **form factor and deployment mode** (pure cloud software / platform-and-middleware / on-device hardware), and a third overlay captures its **education-scenario assignment** (Teaching / Learning / Research (PD) / Assessment / Governance). This coordinate system is used both to place individual products and to track the direction of movement for the track as a whole. It is worth stressing that the coordinate system is an **analytical tool, not a leaderboard**: products sharing a quadrant do not form a simple better-or-worse ranking — their real competitiveness depends on scenario fit, compliance maturity, and the quality of their data loop, all of which are unpacked with evaluation data in Chapters 4 through 6.

> A deeper teardown of on-device hardware form factors — especially the supply chain, market sizing, and competitive structure for AI glasses and AI learning tablets — appears in this institute's *2026 AI Glasses Education Industry Blue Book* and *Global Education Robotics Development White Paper 2026*. This chapter includes only their coordinate position at the overview level and does not repeat that detail.

2.2 Carving Up the Track: Five Scenarios by Three Form Factors

Building on the previous edition's three scenarios — teaching, learning, and research support — this Blue Book adds two: **Intelligent Assessment** and **Governance and Safety**, for a total of **five application scenarios**. Orthogonal to these are the product's **three form factors**. Together they form this chapter's basic analytical grid. This five-scenario split is not an idiosyncratic classification of our

own invention — it lines up closely with both industry and policy framing. In September 2025, Huawei and iFlytek jointly released the "Spark Education and Healthcare Large-Model All-in-One Solution," which groups education-side capability into exactly five categories: teaching assistance, learning assistance, research assistance, administration assistance, and communication assistance. And the "Action Plan for AI + Education" jointly issued in April 2026 by the Ministry of Education and four other ministries explicitly calls, in its section on application integration, for "empowering student learning, empowering teacher instruction, empowering school governance, and empowering scientific research." This Blue Book's five scenarios track closely with both, while additionally carving Assessment out as its own category — a deliberate response to the core controversy over whether generative AI should be entrusted with an assessment function.

2.2.1 The Five Application Scenarios

Scenario	Core Need	Typical 2025–2026 Capability Shift	Representative Capability Keywords
Teaching	Lighten lesson-prep load, generate resources, support classroom delivery	From "generating lesson plans/slides" to "classroom-collaboration agents"	Lesson-prep agents, whiteboard understanding, AI teaching assistants
Learning	Personalized tutoring, Q&A, practice	From "conversational Q&A" to "Socratic guided learning + long-term-memory study companions"	Guided learning, error-pattern attribution, memory-enabled companions
Research (Professional Development)	Resource accumulation, teacher development, data-driven insight	From "material retrieval" to "PD RAG knowledge bases + lesson-observation analysis"	PD RAG, lesson-case analysis, teacher development
Assessment (newly added)	Process-oriented evaluation, multimodal evidence, competency profiling	From "grading objective questions" to "multimodal, process-oriented assessment"	Subjective-response scoring, process profiling, rubric alignment
Governance and Safety (newly added)	Compliance, content safety, data and ethics	From "content filtering" to "education-vertical safety alignment and auditing"	Content safety, data compliance, grade-band usage rules

The last two scenarios are the key additions in the 2026 landscape relative to the previous edition's *Generative AI Product Development Report*. **Intelligent Assessment** responds to the central controversy over whether generative AI can credibly take on an assessment role (see the research evidence in Section 2.4 and Chapter 5). **Governance and Safety** responds to the compliance, content-safety, and ethical risks that have become salient as products scale — China's *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* frames itself around "application first, governance as the foundation" and sets usage boundaries by grade band, while UNESCO's *Guidance for Generative AI in Education and

Research* (2023) recommends a minimum independent-use age of 13. Both signal that governance has moved from an external constraint on the product to a first-class design constraint of the product itself. Evaluation methods and detailed product analysis for these two scenarios appear in Chapters 5 and 6 respectively.

2.2.2 The Three Product Form Factors

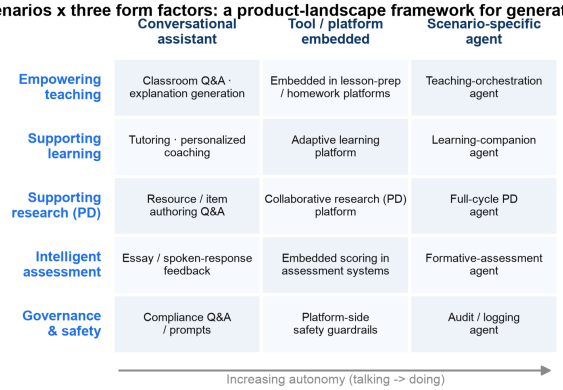
- **Software/App layer**: End-user applications and web/mini-program interfaces for teachers and students, directly carrying conversation, Q&A, and grading interactions. This layer has the largest number of products and the fastest iteration cycle, but also the highest degree of homogenization. Representative examples include OpenAI's ChatGPT Edu, Google's Gemini for Education, Anthropic's Claude for Education, and ByteDance's Doubao AI Xue.
- **Platform/middleware layer**: The underlying capability — model access, agent orchestration, RAG, memory, evaluation, and safety governance — that regions, schools, or product teams build on. Between 2025 and 2026, this layer evolved from "reselling model APIs" into "agent runtime plus education knowledge substrate," making it the highest-value ground in this round of product competition. Representative examples include iFlytek's Spark education LLM, the Huawei–iFlytek "all-in-one scenario appliance," and Alibaba's Qwen3-Learning education-vertical model.
- **On-device hardware layer**: AI learning tablets, tablets, AI glasses, classroom interaction terminals, and other hardware carrying on-device models. This layer physically embeds model capability into real teaching scenarios and is the direct carrier of the multimodal and on-device trends. Representative examples include iFlytek's AI learning tablets, Zuoyebang's and Step by Step's learning tablets, iFlytek's AI blackboard, and Rokid Glasses (a strategic investment of NetDragon (HK:0777)).

A clear value chain and lock-in relationship runs across the three layers: the platform/middleware layer supplies models, orchestration, memory, evaluation, and safety capability downward to both the app layer and the hardware layer, and once a given middleware platform becomes the "education knowledge substrate" for a region or a school system, it creates substantial switching costs and ecosystem lock-in for everything built on top of it. The app layer sits closest to teachers and students and iterates fastest, but because its underlying substrate is swappable, it has limited pricing power and faces heavy homogenization. The hardware layer embeds capability physically into the classroom, and its moat comes from supply chain, distribution channel (including the sales elasticity created by in-store demos combined with national subsidy policy), and the engineering discipline required for edge-cloud coordination. Understanding this value chain is a prerequisite for judging which layer is actually capturing value — and Section 2.5 corroborates it commercially, with data showing that "GenAI technical-capability value" is growing faster (45%) than the overall product-and-service market (37%) in China, evidence that value is concentrating in the middleware layer.

Crossing the five scenarios against the three form factors yields this chapter's **primary product-landscape grid**. The representative products filled into the table below all come from publicly available sources located for this chapter, and are meant to illustrate how each cell can be populated; the full product-by-product comparison appears in Chapter 6.

Form Factor \ Scenario	**Teaching**	**Learning**	**Research (PD)**	**Assessment**	**Governance and Safety**
Software/App layer	ChatGPT Edu; Gemini in Classroom	Claude for Education (Learning Mode); Doubao AI Xue	Khanmigo teacher tools (lesson plans/grouping/progress tracking)	LLM subjective-response scoring tools (mostly at research stage)	Grade-band "usage guideline" compliance modules [TBD: mature commercial products]
Platform/middleware layer	iFlytek Spark "teaching assistant"; Huawei–iFlytek all-in-one scenario appliance	Qwen3-Learning	PD RAG knowledge substrate [TBD: named platform]	School-based assessment engine [TBD: named platform]	Education-vertical safety-alignment/audit middleware [TBD: named platform]
On-device hardware layer	iFlytek AI Blackboard 2.0	iFlytek/Zuoyebang/Step by Step AI learning tablets	[TBD: PD-dedicated terminal]	[TBD: process-assessment terminal]	Rokid Glasses and other wearables (compliance and privacy as the core variables)

Fig. 2 Five scenarios x three form factors: a product-landscape framework for generative-AI education products



Source: this report's analytical framework; example products verified by research (see Chapters 2-7).

2.3 Four Drivers Behind the Form-Factor Shift

Before presenting the competitive coordinate map itself, it is worth spelling out the four underlying threads driving the shift in product form factors. They explain why products are migrating toward the upper-right of the map — toward agentic and on-device — and they also form the basis for the evaluation metrics used in later chapters.

1. **From model capability to task completion.** The locus of competition has moved from the quality of a single answer to whether a product can reliably complete an entire educational task end to end — for instance, "from curriculum standard, to a differentiated lesson plan, to classroom-implementation guidance." That requires three capabilities a purely conversational paradigm lacks: planning (breaking a task into ordered sub-steps), tool use (retrieving textbook content, generating charts, writing to a learning-record store), and error recovery (rolling back and redoing a step once a prior deviation is detected). Khanmigo's teacher-facing tools, which generate a lesson plan, a grading rubric, a grouping strategy, and a learning-progress summary in one pass, are a textbook case of "closing the task loop" rather than "answering one question." Gemini 2.0's official positioning as built "for the agentic era," together with its native tool chain for Google Search and code execution, is essentially this same closed-loop capability moving down from "an orchestration trick at the application layer" to "a native capability at the model layer." For product teams, this means the competitive moat has shifted upward — from prompt engineering to the stability of workflow design and task success rate, which demands sustained tuning against real-world scenario data and is accordingly a higher bar to clear.
2. **From general-purpose model to education-vertical substrate.** A general-purpose model struggles to satisfy three hard constraints at once — subject-matter accuracy, pedagogical fit, and safety for minors — which is what makes education-vertical LLMs and domain-specific knowledge bases (RAG) the key differentiator. Google's LearnLM family, built on Gemini and fine-tuned against the learning sciences, is designed around five learning-science principles:

active practice and feedback, cognitive-load management, personalization, curiosity, and metacognition. At I/O 2025, Google folded its capabilities into Gemini 2.5, and in its own controlled-experiment results reported that students who received a short tutoring session from LearnLM were about 5.5 percentage points more likely to solve a subsequent novel problem than a control group tutored only by human instructors. Alibaba's December 4, 2025 release of the Qwen3-Learning education-vertical model (as reported via iResearch's 2026 report) is China's move along the same logic. The value of going vertical is not just "more accurate answers" — it is baking pedagogy (Socratic guidance rather than handing over the answer directly) and safety boundaries (grade-band content red lines) directly into model behavior, which lowers the alignment burden for everything built downstream. Evaluation methods for this thread appear in Chapter 4.

3. **From stateless conversation to long-term memory.** Learning is a long process, and products are moving from single sessions toward cross-session, cross-semester learner memory and profiling — which unlocks a personalization dividend but also raises data-compliance and ethics stakes. iResearch's *2026 China GenAI + Education Industry Development Report* found that among parents tutoring their children, 37.1% stick to one or two GenAI tools consistently while 35.6% mix multiple tools, and that 40.9% of adult learners mix multiple AI tools. This fragmented, cross-product usage pattern is itself evidence that a unified learner memory spanning sessions and products remains an unmet need. But memory capability simultaneously pushes a governance question — whether collecting, storing, and reusing a minor's learner profile is compliant — from the margins to the center, which is one of the concrete reasons this edition breaks Governance and Safety out as its own scenario.
4. **From cloud-only software to edge-cloud co-designed hardware.** On-device computing moves part of inference and perception onto the terminal, trading for three advantages at once — latency, privacy, and situational awareness. Low latency serves real-time classroom interaction; local processing eases the compliance burden of minors' data leaving the device; and first-person or near-field sensing supplies situational signal the cloud alone cannot capture. Hardware has accordingly become a decisive variable in education-AI competition again. This is especially visible in China's consumer market: in 2024, China's AI learning tablets sold 5.923 million units across all channels, up 25.5% year-on-year (see Section 2.5 for detail). It is worth noting that on-device does not mean fully offline — the dominant form is edge-cloud coordination, where the device handles low-latency perception and privacy-sensitive processing while the cloud handles heavy inference and knowledge updates. That split of compute and data between edge and cloud is exactly what differentiates products at the hardware layer, and the full industry teardown is available in this institute's AI glasses and learning-tablet evaluations.

These four threads all point in the same direction: products moving "up and to the right" on the coordinate map — toward agentic behavior plus on-device/multimodal capability. They also

underpin the evaluation metrics used later in this Blue Book: Chapter 4's evaluation of education-vertical models corresponds to threads 1 and 2; Chapter 5's evaluation of assessment and governance corresponds to the compliance issues in thread 3; and Chapter 6's hardware and product comparison corresponds to thread 4.

2.4 The Competitive Coordinate Map: Two Axes and Clustering

Building on the framework above, this chapter constructs a competitive coordinate map with **capability paradigm (conversational ↔ agentic orchestration) as the horizontal axis and deployment form (pure cloud software ↔ on-device hardware) as the vertical axis**, clustering the major products into several groups. These two axes were chosen over more conventional dimensions such as vendor size or price because they map directly onto the two most structural of the four drivers in Section 2.3 — task completion and edge-cloud coordination — and are therefore the most sensitive indicators of where the track is heading. Placing any individual product precisely on this map requires real data support; what follows is the cluster structure and verified representative players, with precise coordinate values (the sizing behind the bubble metaphor) to be filled in by later evaluation chapters.

One caveat about "mobility" on this map: a single product may exist simultaneously in more than one form factor — Khanmigo, for instance, has both a learner-facing conversational companion and a teacher-facing agentic research/PD tool — so it occupies not a point on the map but a trajectory across it. Khanmigo's own scaling trajectory is itself a verifiable footnote to the "learning plus research support, twin engines" version of the A→B migration: according to Khan Academy's own public statements, its K-12 student user base grew substantially in the 2024–25 school year, and it expects to reach roughly a million users in the 2025–26 school year, expanding into more U.S. school districts and extending into classrooms in India, Brazil, and the Philippines; classroom access is provisioned only through a school or district-wide deployment. This kind of institution-mediated distribution itself bakes governance (quadrant E) directly into the product's path to deployment.

> Various figures for Khanmigo's user base circulate informally (from "tens of thousands to a million" to "795 school districts"), with inconsistent baselines and reference dates across sources. This Blue Book uses only statements traceable to Khan Academy's own official channels, with a clear baseline and date; other high-growth figures are marked **[TBD: needs verification against an official source]** rather than cited directly, to avoid perpetuating unverified claims.

Cluster (Quadrant)	Paradigm	Deployment	Verified Representative Players	Competitive Focus
A. Cloud conversational companions/assistants	Conversation-led	Pure cloud software	ChatGPT Edu; Gemini for Education; Doubao AI Xue	Content quality, subject coverage, compliance wrapping

B. Agentic research/teaching platforms	Agentic orchestration	Cloud + middleware	iFlytek Spark education LLM; Huawei–iFlytek all-in-one scenario appliance; Khanmigo teacher tools	Orchestration capability, RAG/memory, safety governance
C. On-device learning hardware (learning tablets, etc.)	Conversation → light agent	Edge-cloud coordination	iFlytek/Zuoyebang/Step by Step AI learning tablets; iFlytek AI Blackboard	On-device models, offline capability, parent-child/eye-health compliance
D. Multimodal wearables (AI glasses, etc.)	Multimodal agent	Primarily on-device	Rokid Glasses (strategic investment from NetDragon (HK:0777))	First-person perception, latency, privacy and ethics
E. Vertical assessment/governance engines	Agent + rules	Cloud/private deployment	LLM auto-grading and "usage guideline" compliance modules (mostly at research/pilot stage)	Assessment validity, compliance auditing, explainability

Reading the map (qualitative):

- The overall direction of migration runs from quadrant A (bottom-left: cloud conversational) toward quadrants B/D (top-right: agentic plus multimodal/on-device), reflecting a value upgrade toward "closing the task loop plus situational awareness." Claude for Education's and Gemini's "guided learning/Learning Mode" — using Socratic questioning instead of handing over answers directly — represents one branch of the A-to-B "learning-paradigm upgrade" migration; iFlytek's and Huawei's "all-in-one scenario appliances," spanning teaching, learning, research, administration, and communication assistance, represent the middleware consolidation happening within quadrant B.
- Quadrant B (agentic platforms/middleware) is the **highest-value ground and the most fiercely contested**, because it simultaneously serves teaching, research support, and governance needs, and because it locks in downstream apps and hardware. iResearch's estimate that "GenAI technical-capability value" is compounding at 45% annually, faster than the 37% for the overall product-and-service market (see Section 2.5), corroborates the sharpness of middleware-layer value commercially.
- Quadrant D (multimodal wearables) is the **zone of highest uncertainty**: the technology trajectory is compelling, but maturity, cost, ethics, and education-scenario fit are all still being worked out. NetDragon's 2023 Series C strategic investment in Rokid as lead investor — a round totaling US$112 million at a post-money valuation of US$1 billion — together with a

five-year strategic partnership the two companies signed, is this institute's directly affiliated industry position in this quadrant; industry detail appears in this institute's AI glasses Blue Book.

- Quadrant E (assessment/governance engines) is a **nascent but strategically important** cluster, corresponding directly to the two scenarios newly added in this edition. The current evidence suggests this quadrant's "validity is not yet stable and still requires human-in-the-loop oversight." On one hand, several 2025 peer-reviewed studies show LLM auto-grading already has considerable potential — a fine-tuned ChatGPT achieved fairly high agreement on rubric-based EFL essay scoring (Yavuz et al., 2025: intraclass correlation coefficient (ICC) of about 0.972 for the fine-tuned version and about 0.947 for the default version). But findings elsewhere are not consistent: Flodén (2025) found that ChatGPT's scoring tends to avoid extreme grade bands, achieves full agreement with human graders only about 30% of the time, and shows insufficient subject-level reliability. Together these findings point to uneven validity: the model handles surface-level dimensions — grammar, word choice, sentence structure — fairly well, but remains unstable on higher-order, construct-level dimensions such as coherence, argument clarity, and persuasiveness, and its reliability drifts across different prompts and different student populations. That is why researchers broadly insist that "human moderation and explicit validity argumentation" cannot be dispensed with. This "usable but not fully trustworthy" evidentiary state is precisely what will determine whether these products can scale into formal educational assessment settings, and it is why Chapter 5's evaluation treats "assessment validity plus the boundary of human-AI collaboration" as a core metric. In terms of industry maturity, quadrant E currently consists mostly of research and pilot activity with no named, mature commercial products, which is why most of the corresponding cells in this chapter are marked **[TBD]** — a research prototype is not represented here as a commercial product.

Figure 2-1 (planned): A bubble chart of the competitive coordinate map, with "conversational ↔ agentic orchestration" on the horizontal axis and "pure cloud software ↔ on-device hardware" on the vertical axis, showing the five clusters A–E, with bubble area representing **[TBD: market-size/installed-base basis, data to appear in the Chapter 4–6 evaluations]**. Precise coordinate values and the bubble-sizing basis are to be filled in once the evaluation chapters are complete; the placeholder here is not a fabricated estimate.

2.5 Market Size and Structure

The commercial foundation beneath this product landscape needs to rest on real market data. Every figure cited in this section is labeled with its source institution, year, and basis, and strictly observes a **currency firewall**: RMB and USD figures are kept in separate columns, with no exchange-rate conversion or cross-currency summation; where different institutions use different bases, cross-comparisons are offered only as directional color, never as like-for-like substitution.

2.5.1 The Global Market (USD Basis)

Per Research and Markets' *AI in Education Market Report 2026* (published April 2026): the global "AI in education" market is estimated to grow from about **US$7.52 billion** in 2025 to about **US$10.6 billion** in 2026 (a 2025→2026 CAGR of about 40.9%), and is projected to reach about **US$42.48 billion** by 2030 (a 2026→2030 CAGR of about 41.5%); by region, North America led in 2025, while Asia-Pacific is growing fastest. It bears noting that different institutions define the boundaries of "AI in education" very differently: Precedence Research gives a long-horizon estimate of about US$7.05 billion in 2025 rising to about US$136.79 billion by 2035, and MarketsandMarkets, Mordor Intelligence, Grand View Research, and Technavio each use their own base years and CAGRs that do not agree with one another. This Blue Book therefore **cites only a single institution's figures where the basis and year are both clearly stated**, and does not splice numbers together across institutions.

Dimension	Basis and Institution	Figures
Global AI-in-education market size	Research and Markets, published 2026, USD	2025 ≈ US$7.52 billion; 2026 ≈ US$10.6 billion
Global AI-in-education market size forecast	Research and Markets, published 2026, USD	2030 ≈ US$42.48 billion (2026→2030 CAGR ≈ 41.5%)
Long-horizon basis (for reference only, not to be summed with the above)	Precedence Research, published 2025, USD	2025 ≈ US$7.05 billion; 2035 ≈ US$136.79 billion

2.5.2 The China Market (RMB Basis)

Per iResearch's *2026 China GenAI + Education Industry Development Report* (published March 3, 2026):

Dimension	Basis (RMB)	Figures
Total GenAI + education product-and-service market	Full year, 2025 forecast	2025 ≈ RMB 344.2 billion; 2028 ≈ RMB 891 billion (CAGR ≈ 37%)
Of which, "GenAI technical-capability value"	Pure technical-capability basis	2025 ≈ RMB 66.4 billion; 2028 ≈ RMB 202.3 billion (CAGR ≈ 45%)
Technical-capability share of total	Average	About 20% in 2025 (i.e., roughly RMB 1 out of every RMB 5 in product-and-service spend goes toward purchasing technical capability)
Total consumer (C-end) education market (background context)	Consumer education spending	2025 ≈ RMB 1.3 trillion; 2028 ≈ RMB 1.5 trillion
Institutional (B-end) education-informatization/digitalization spending (background context)	School-side investment	2025 ≈ RMB 551.5 billion; 2028 ≈ RMB 680.2 billion
User-side penetration	As of December 2025	57.3% of parents use GenAI when

		tutoring their children; 25% of adult learners use it for exam prep/study

On the school-procurement side, iResearch reports that GenAI-related projects account for about 27% of spend at general universities (with typical individual budgets of RMB 1–4 million), about 35% at vocational colleges (RMB 1.5–4 million), and about 46% at primary and secondary schools (RMB 1–5 million). This pattern indicates that the institutional (B-end) market has entered a staged progression "from shallow hardware retrofits toward a digital substrate plus deep integration of large models" — first-tier cities have already reached the deep-integration stage, while second- and third-tier cities still rely mostly on combined software-hardware upgrades or hardware retrofits. Two further baseline facts on the user side are worth recording: as of December 2025, China's generative-AI user base stood at roughly 602 million people, a penetration rate of about 42.8%; and about 74.8% of children encounter GenAI for the first time through a general-purpose (rather than education-vertical) AI product — meaning "general-purpose entry point first, vertical product later" is the current real-world path by which users are reached, which places an urgent governance demand (content safety, grade-band rules) ahead of product maturity itself.

Policy is a demand-side variable that cannot be ignored in the China market. In May 2025, the Ministry of Education's Basic Education Teaching Guidance Committee released the *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* and the *Guidelines for General AI Education in Primary and Secondary Schools (2025 Edition)*. The former is framed around "application first, governance as the foundation" and sets usage boundaries by grade band: at the primary level, students are barred from independently using open-ended content-generation features, though teachers may use them appropriately in class; at the middle-school level, students may moderately explore the logical analysis of generated content; and at the high-school level, students are permitted to conduct inquiry-based learning that engages with the underlying technical principles. In April 2026, the Ministry of Education and four other bodies — the National Development and Reform Commission, the Ministry of Industry and Information Technology, the Ministry of Science and Technology, and the National Data Administration — jointly issued the *Action Plan for AI + Education* (Document No. Jiaokexin [2026] 1), setting an overall goal of "a basically established pattern of deep integration between AI and education" by 2030, and laying out four priority tasks: talent development and literacy building, deep educational integration, foundational infrastructure, and an open ecosystem; its application-integration section explicitly calls for "empowering student learning, empowering teacher instruction, empowering school governance, and empowering scientific research." The significance of this top-level document is that it converts GenAI from "a school's optional initiative" into "a systemic undertaking spanning every grade level," directly shaping where B-end budgets and scenario priorities will flow over the coming years — it is the policy basis for this chapter's five-scenario split, and a structural driver expanding demand in

both quadrant B (middleware) and quadrant E (assessment/governance). Internationally, UNESCO's *Guidance for Generative AI in Education and Research* (2023) sets a minimum independent-use age of 13, which — together with China's grade-banded red lines — forms a compliance baseline that products must now design around from the outset.

2.5.3 On-Device Hardware: AI Learning Tablets (RMB Basis)

AI learning tablets are the most mature carrier of on-device computing in China's consumer market, and figures from multiple institutions corroborate one another:

Dimension	Basis and Institution	Figures
All-channel unit sales	Full year 2024	5.923 million units, +25.5% year-on-year
All-channel sales revenue	Full year 2024	RMB 19.06 billion, +37.6% year-on-year
All-channel unit sales	Q1 2025	1.265 million units, +29.4% year-on-year; revenue RMB 4.02 billion, +15.8% year-on-year
Shipment volume	Q2 2025 (IDC basis, learning tablets)	1.54 million units, +44.6% year-on-year
Competitive standing	No. 1 by revenue, Q2 2025 (IDC)	iFlytek AI learning tablets (its first time topping revenue rankings since entering the category)
Competitive standing	No. 1 by both unit sales and revenue, Q1 2025	Zuoyebang (38.8% unit-sales share, 31.1% revenue share; 46.7% share in the mainstream RMB 2,001–4,000 price band)

Note: Learning-tablet figures are calculated separately by several institutions (IDC, iiMedia Research, industry half-year reports, etc.); unit sales, revenue, and shipment volume are three distinct bases and should not be mixed. Each row above is labeled by its own source. The full industry chain, installed-base estimates, and competitive landscape for both AI learning tablets and AI glasses appear in this institute's *2026 AI Glasses Education Industry Blue Book*.

2.5.4 Structural Takeaways

Taken together, these three data sets support three structural conclusions. First, **both the global and China markets are growing at a high 30–45% clip**, but their absolute size and measurement bases differ enormously, so any cross-comparison must hold the institution and basis constant. Second, **"technical-capability value" is growing faster than the overall product-and-service market** (45% versus 37% in China), meaning value is concentrating in the middleware layer of quadrant B (Section 2.4). Third, **on-device hardware is currently the most commercially certain landing spot** — AI

learning tablets have posted double-digit growth for several consecutive quarters — while wearables (quadrant D) and assessment/governance engines (quadrant E) remain early-stage, without enough scale data yet to support a precise bubble-sizing basis for Figure 2-1, which is accordingly marked **[TBD]**.

2.6 Timeline: Key Milestones, 2025–2026

To show the pace at which this track is evolving, this chapter lays out a timeline interweaving product releases and policy actions. Every entry below comes from a publicly available source located for this chapter; where an exact date could not be verified, the entry is marked with a range or **[TBD]**.

Date	Event Type	Event (verified source)	Source
2024-05	Product launch (International · App)	OpenAI launches ChatGPT Edu for higher education (built on GPT-4o)	Official OpenAI announcement
2024-12	Paradigm (International · Agentic)	Google launches Gemini 2.0, positioned "for the agentic era"	blog.google
Around 2025-04-02	Product launch (International · App)	Anthropic launches Claude for Education, including Learning Mode (Socratic guidance)	anthropic.com
2025-05	Paradigm (International · Learning Science)	Google I/O 2025: LearnLM folded into Gemini 2.5, Guided Learning launched	blog.google
Around 2025-05-12	Policy/Standard (China · Governance)	Ministry of Education's Basic Education Teaching Guidance Committee releases the *Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition)* and *Guidelines for General AI Education in Primary and Secondary Schools (2025 Edition)*	edu.cn/CERNET
2025-06-24	Hardware (China · On-device)	iFlytek's AI learning tablet summer launch event: 16 upgrades including AI one-on-one precision learning, tutoring Q&A, and	qbitai.com

		interactive lessons	
2025-07-15	Product launch (International · Channel)	Claude for Education listed on AWS Marketplace	aws.amazon.com
2025-09	Platform/middleware (China · B-end)	Huawei Connect: Huawei and iFlytek launch the "Spark Education and Healthcare Large-Model All-in-One Solution" (five scenarios: teaching, learning, research, administration, communication assistance)	e.huawei.com
2025-10	Hardware (China · On-device)	iFlytek launches its next-generation AI Blackboard 2.0	edu.iflytek.com
2025-12-04	Product launch (China · Vertical model)	Alibaba launches the Qwen3-Learning education-vertical model	As reported via the iResearch report
2026-03-03	Report (China · Market)	iResearch publishes the *2026 China GenAI + Education Industry Development Report*	iResearch
Issued 2026-04-02 / made public 2026-04-10	Policy (China · Top-level)	Ministry of Education and four other ministries jointly issue the *Action Plan for AI + Education* (Document No. Jiaokexin [2026] 1), setting a 2030 goal and four priority tasks	moe.gov.cn
2026-04-07	Report (International · Market)	Research and Markets publishes the *AI in Education Market Report 2026*	Research and Markets

Figure 2-2 (planned): A timeline of generative-AI education product evolution, 2025–2026, arranged across three lanes — "international products," "China products and hardware," and "policy and standards" — with nodes and sources as in the table above; cross-currency scale data is excluded from this figure to avoid mixed-basis calculation.

2.7 Summary

This chapter has established the analytical framework behind the 2026 product landscape: a primary grid crossing **five application scenarios** (Teaching, Learning, Research/PD, Intelligent Assessment, Governance and Safety) **against three form factors** (App, platform/middleware, on-device hardware), overlaid with a two-axis competitive coordinate map — capability paradigm by deployment form — that places the major players into five clusters (A: cloud conversational, B: agentic middleware, C: on-device learning hardware, D: multimodal wearables, E: assessment/governance engines), and identifies the dominant migration path as products moving toward "agentic plus multimodal/on-device."

Relative to the previous edition's largely list-based presentation of conversational products, this landscape emphasizes three shifts — **paradigm migration, scenario expansion, and evidence-based positioning** — and anchors its commercial foundation with three verifiable data sets: the global AI-in-education market (Research and Markets, USD basis) growing from about US$7.52 billion to about US$10.6 billion between 2025 and 2026; China's GenAI + education market (iResearch, RMB basis) at about RMB 344.2 billion in 2025 with a CAGR of about 37%, within which technical-capability value is concentrating in the middleware layer at a faster 45% growth rate; and on-device AI learning tablets selling 5.923 million units in 2024, continuing double-digit growth through the first half of 2025. All three data sets are presented separately by currency and by institution, never combined.

One boundary is worth stating candidly: quadrant D (wearables) and quadrant E (assessment/governance) remain early-stage, lacking the public scale data needed to support a precise bubble-sizing basis for Figure 2-1. The corresponding cells and figures in this chapter are accordingly left as **[TBD]**, to be filled in with real data by the education-vertical LLM evaluation, AI hardware evaluation, and product comparison in Chapters 4 through 6 — ensuring that the landscape has a skeleton, and that every number on it has a source.

Chapter References

1. OpenAI. Introducing ChatGPT Edu. OpenAI, 2024. https://openai.com/index/introducing-chatgpt-edu/

2. Google. Google introduces Gemini 2.0: A new AI model for the agentic era. Google Blog, 2024-12. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/

3. Google. I/O 2025: LearnLM in Gemini 2.5 and more AI updates to help people learn. Google Blog, 2025-05. https://blog.google/products-and-platforms/products/education/google-gemini-learnlm-update/

4. Anthropic. Introducing Claude for Education. Anthropic, 2025-04. https://www.anthropic.com/news/introducing-claude-for-education

5. AWS Public Sector Blog. Claude for Education now available in AWS Marketplace. Amazon Web Services, 2025-07. https://aws.amazon.com/blogs/publicsector/claude-for-education-now-available-in-aws-marketplace/

6. Khan Academy. Need-to-Know BTS 2025: AI-Powered Support, Classroom-Ready Updates, and District Features. Khan Academy Blog, 2025. https://blog.khanacademy.org/need-to-know-bts-2025/

7. iResearch. 2026 China GenAI + Education Industry Development Report [2026 年中国 GenAI+教育行业发展报告]. iResearch, 2026-03-03 (full text reprinted by Jiemian News). https://www.jiemian.com/article/14060378.html

8. Research and Markets. Global $10.6B AI in Education Market, 2026: Total Revenue Set to Quadruple During 2026-2030, Reaching $42.48 Billion. 2026-04-07 (reprinted by Yahoo Finance). https://finance.yahoo.com/sectors/technology/articles/global-10-6b-ai-education-163600564.html

9. Precedence Research. AI in Education Market Size to Surge USD 136.79 Bn by 2035. Precedence Research, 2025. https://www.precedenceresearch.com/ai-in-education-market

10. Huawei. Huawei and iFlytek Jointly Launch the "Spark Education and Healthcare Large-Model All-in-One Solution" [华为联合科大讯飞发布"星火教育、医疗大模型场景一体机解决方案"]. Huawei Enterprise, 2025-09. https://e.huawei.com/cn/news/2025/industries/education/spark-education-medical-large-model

11. QbitAI. iFlytek Accelerates "AI + Education" Again: Learning Tablet Upgrades Lead the Industry [科大讯飞"AI+教育"再提速：学习机功能升级引领行业发展]. QbitAI, 2025-06. https://www.qbitai.com/2025/06/300819.html

12. QbitAI. Q2 Learning Tablet Shipments Up 46%! IDC: iFlytek AI Learning Tablet Tops Market in Sales Revenue [Q2 学习机出货量增 46%！IDC：科大讯飞 AI 学习机登顶市场销售额第一]. QbitAI, 2025-09. https://www.qbitai.com/2025/09/328513.html

13. China.com. Learning Tablets Focus on Mainstream Price Bands — Zuoyebang Leads Strongly with Nearly Half of Market Share [学习机聚焦主流价格段　作业帮近五成份额强势领跑]. 2025-04. https://hea.china.com/articles/20250423/202504231664359.html

14. China Education and Research Network (CERNET). Full Text: Guidelines for Generative AI Use in Primary and Secondary Schools (2025 Edition) [《中小学生成式人工智能使用指南（2025年版）》全文]. edu.cn, 2025-05. https://www.edu.cn/xxh/focus/zc/202505/t20250513_2667992.shtml
15. Ministry of Education of the People's Republic of China. Notice on the Issuance of the "Action Plan for AI + Education" (Document No. Jiaokexin [2026] 1) [教育部等五部门关于印发《"人工智能+教育"行动计划》的通知（教科信〔2026〕1 号）]. moe.gov.cn, 2026-04. http://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html
16. UNESCO. Guidance for generative AI in education and research. UNESCO, 2023. https://www.unesco.org/en/articles/unesco-governments-must-quickly-regulate-generative-ai-schools
17. Zhitong Finance. Strategic Investment in Rokid Opens a New Chapter for "Pure AI Stock" NetDragon (00777) [战略投资 ROKID，"纯正 AI 股"网龙（00777）打开全新想象空间]. 2023-11. https://cn.investing.com/news/stock-market-news/article-2634061
18. Tencent News. AR Smart Glasses Company Rokid Completes US$112 Million Series C Round Led by NetDragon [AR 智能眼镜企业 Rokid 完成 1.12 亿美元 C 轮融资，由网龙网络领投]. 2023-11-21. https://news.qq.com/rain/a/20231121A08L2800
19. Yavuz F., et al. Utilizing large language models for EFL essay grading: reliability and validity in rubric-based assessments. British Journal of Educational Technology, 2025. https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13494
20. Flodén J. Grading exams using large language models: human vs AI grading with ChatGPT. British Educational Research Journal, 2025. https://bera-journals.onlinelibrary.wiley.com/doi/10.1002/berj.4069

Chapter 3 Empowering Teaching: Mechanisms, Product Forms, and Recommendations

3.1 Introduction: The Shift from "Conversational Assistance" to "Agentic Collaboration"

In the early days of generative AI in the classroom, "empowering teaching" typically meant one thing: a large language model providing conversational support for lesson prep, instruction, and classroom interaction. That framing made sense in 2023. It no longer captures the technical reality of 2026. On the teaching side, generative AI is evolving from a "question-answering tool the teacher invokes" into something closer to an **agent** — orchestrable, equipped with memory, capable of multimodal perception, and able to intervene proactively in the instructional workflow. This chapter keeps the "mechanism—product—recommendations" narrative structure of earlier editions, but widens the aperture from a single conversational LLM to three carriers: **education-vertical large language models (LLMs), teaching agents, and multimodal/on-device hardware**. It also introduces a two-axis "capability versus deployment maturity" lens, so that lab demos are not mistaken for classroom-scale reality.

Three intertwined technical drivers sit behind this shift. The first is **models moving from general-purpose question-answering toward education alignment**: continued pretraining and fine-tuning on educational corpora, plus alignment with learning-science principles — Google's LearnLM, for instance, is explicitly fine-tuned against five learning-science principles (active learning, cognitive-load management, personalization, curiosity, and metacognition) — so that outputs read as pedagogically grounded rather than generically correct. The second is **the move from single-turn generation to multi-tool orchestration**: agent frameworks let a model decompose a task, call tools, and chain steps together, upgrading "generate a piece of text" into "complete a piece of teaching work." The third is **the descent from cloud to on-device**: grading tablets, interactive whiteboards, and teacher-point-of-view hardware push inference to the classroom itself, addressing the twin constraints of real-time responsiveness and data compliance. Together, these three forces are pushing "conversational assistance" toward "agentic collaboration" — and that shift is precisely why this chapter reorganizes its narrative from mechanism through product.

One judgment discipline runs through the entire chapter and needs to be stated up front: the most reliably scalable classroom deployments remain concentrated in **teacher-productivity** scenarios — content generation and lesson-prep assistance, practice-problem and explanation generation, in-class language support, and grading-workload reduction. Several independent surveys now put numbers

behind that judgment. A Walton Family Foundation and Gallup survey of 2,232 U.S. public K-12 teachers, fielded March–April 2025, found that teachers using AI at least weekly saved an average of 5.9 hours per week — roughly six weeks over a school year. Those teachers reported that AI's most common uses were exactly the content-generation tasks one would expect: lesson planning (about 37%), building worksheets/activities (about 33%), and adapting materials to student needs (about 28%) — while more judgment-heavy uses such as grading (about 16%), one-on-one tutoring (about 14%), and student-data analysis (about 12%) trailed well behind. RAND's sampling of U.S. teachers across the 2024–2025 school year tells a similar story: roughly 53% of all K-12 teachers had used generative AI for instructional purposes that year, double the roughly 25% recorded in 2023–2024 (English and science teachers reported markedly higher use than math teachers or elementary teachers), a notably steep adoption curve. Over the same period, roughly 48% of U.S. school districts (as of fall 2024) reported having rolled out teacher AI training — up about 25 percentage points from the prior year — evidence that institutional support is catching up, but still lags front-line use. By contrast, scenarios such as "autonomous teaching agents," "real-time multimodal learning-state sensing," and "AI emotion recognition" remain largely at the pilot, demo, or regulation-constrained stage. This chapter flags a maturity band for every product category discussed, and leaves any specific claim unverifiable against a real source blank rather than guessing.

3.2 The Mechanisms Behind Empowering Teaching

3.2.1 A Three-Layer Mechanism Framework

The pathways through which generative AI empowers teaching can be organized into three progressive layers: **content, interaction, and decision-making**.

- **Content layer (generation and recombination)**: The model generates or recombines instructional content — lesson plans, slide decks, explanatory text, tiered practice problems, scenario-based cases, multilingual and accessibility materials — based on textbooks, curriculum standards, and teacher intent. In essence, this shifts the teacher's role from "creating from scratch" to "reviewing and refining," and it rests mechanically on the LLM's language-generation and knowledge-recombination capabilities. It is the most mature and reliable layer today. The Gallup survey's top teacher uses — lesson planning (37%), worksheets/activities (33%), adapting materials to student needs (28%) — all fall squarely within this layer.
- **Interaction layer (dialogue and multimodal perception)**: The model conducts multi-turn dialogue with teachers and students during class or lesson prep, progressively incorporating multimodal inputs — voice, images, handwriting on the board, on-screen content — to close the loop of "explain—ask—answer—correct." This layer's maturity hinges on multimodal recognition accuracy and real-time responsiveness, and on-device hardware (see this institute's

2026 Blue Book on the AI Smart Glasses Education Industry for an industry analysis of teacher-point-of-view classroom systems) is a key carrier here.

- **Decision-making layer (orchestration and intervention)**: The agent decides autonomously, based on teaching objectives, classroom state, and learning signals, when to prompt, when to supplement, and when to hand control back to the teacher — what might be called "teaching orchestration." This layer draws on Retrieval-Augmented Generation (RAG) to keep content anchored to school-specific knowledge bases, and on memory mechanisms to maintain continuity across class sessions. It is the frontier direction for 2026, but classroom-scale validation remains thin.

The three layers do not replace one another in sequence — they stack and reinforce. The content layer ensures the generation is correct; the interaction layer ensures the exchange flows naturally; the decision-making layer ensures intervention happens at the right moment. Earlier reports stopped mostly at conversational implementations of the content and interaction layers. This chapter treats the decision-making layer — agentic orchestration — as the core increment for the 2026 edition, while stressing that the further one moves toward decision-making, the more clearly human/machine responsibility must be delineated: instructional authority should always remain with the teacher.

Worth emphasizing: the maturity of each layer correlates closely with teacher acceptance. The content layer is widely embraced because its value is immediately visible and its risk is manageable; the decision-making layer, by touching on instructional judgment itself, more readily raises the question of whether the technology is overstepping. That acceptance gradient lines up neatly with the Gallup data cited above — heavy use of lesson-planning-type functions, light use of grading and learning-analytics functions — suggesting that teachers' actual choices are themselves a kind of vote on how mature each layer really is.

Viewed through the lens of cognitive load, the three layers correspond to three different ways of redistributing a teacher's workload. The content layer strips out extraneous cognitive load — offloading low-creativity mechanical labor such as formatting, searching, and drafting to the model, freeing teachers to focus their attention on germane load, i.e., pedagogical judgment. That is why the workload relief from lesson prep and grading is the easiest of the three to quantify and verify. The interaction layer reduces the working-memory pressure of real-time classroom response — when a teacher faces an unexpected question, needs to read handwriting on the board, or must respond to a multilingual need, the model's real-time supplementation functions as an instantly callable external memory. The decision-making layer, meanwhile, attempts to take on part of the "instructional monitoring" load — judging what should happen right now — which is precisely the most concentrated and least automatable core of teaching expertise. That is exactly why this layer has the lowest maturity and needs the tightest constraints on responsibility. This load-stratification view is

consistent with the chapter's overall judgment: moving from content to decision-making, certainty decreases and the need for constraint increases.

3.2.2 Mechanisms Across the Four Stages of Prep, Instruction, Assignments, and Feedback

Mapping the three-layer framework onto a teacher's actual workflow breaks it further into four stages — lesson prep, instruction, assignments, and feedback — each with its own mechanism and bottleneck.

- **Lesson prep (primarily content layer)**: The underlying mechanism is "intent alignment + knowledge recombination + curriculum-standard verification." The teacher supplies grade level, textbook edition, lesson objectives, and learning-context constraints; the model generates a lesson-plan outline, instructional design, tiered slide decks, and scenario-based cases. The bottleneck is alignment with curriculum standards and subject-matter accuracy — general-purpose models lack precise memory of specific textbook editions and curriculum-standard line items, and readily produce objectives and activities that look plausible but are actually misaligned. This is the fundamental reason school-specific RAG has become standard equipment in education-vertical products. The Gallup survey's finding that lesson planning (37%) is teachers' top AI use case is consistent with this being the most mature stage mechanistically.
- **Instruction (primarily interaction layer)**: The underlying mechanism is "real-time perception + instant generation + low-latency feedback." The model takes on explanatory supplementation, follow-up-question generation, board/screen-content recognition, and real-time multilingual transcription during class. The bottleneck is latency and reliability — the classroom is a high-real-time, low-tolerance-for-error environment, where even second-scale round-trip delays to the cloud, or the occasional hallucination, can break the flow of instruction. This is what is pushing on-device deployment, and it is also why "autonomous teaching" is unlikely to be viable for the foreseeable future.
- **Assignments (content layer plus decision-making layer)**: The underlying mechanism is "question generation + automated grading + learning-gap attribution." On the front end sits tiered question generation and item variation; on the back end sits grading of objective and subjective items plus error-cause analysis. Objective-item grading is already highly mature (essentially a rules-plus-recognition problem), whereas subjective-item grading enters the decision-making layer — it requires the model to score open-ended responses and justify the reasoning, and consistency and explainability are the core bottlenecks, making human review and an appeals channel non-negotiable. iFlytek's Spark Smart Grading Tablet, which integrates "grading—learning analytics—personalized assignments" into a single on-device pipeline, is a representative example of productization at this stage.

- **Feedback (primarily decision-making layer)**: The underlying mechanism is "learning-state modeling + differentiated recommendations + instructional redesign." Based on student responses and classroom data, the model generates recommendations for the teacher on what to teach next and who needs what kind of support, feeding back into the next round of lesson prep to close the loop. The bottleneck is the reliability of causal attribution and data compliance — learning-state signals are often correlational rather than causal, and the process depends heavily on collecting data from minors, where a single misstep can cross privacy or emotion-recognition red lines (see Section 3.4). This is also the stage with the blurriest product boundaries, most deeply entangled with Intelligent Assessment and Supporting Research (Professional Development).

These mechanism differences across the four stages directly shape the pace of product deployment: lesson prep and grading scale first because their bottlenecks are manageable; instruction depends on hardware evolution because of its real-time constraints; feedback remains the most cautious owing to compliance and causal-attribution challenges. The product analysis and maturity timeline later in this chapter can all be read back against this four-stage framework.

3.2.3 Mapping Key Technical Paradigms to Teaching

The table below maps 2026-era technical paradigms onto specific teaching stages, noting each one's underlying mechanism and current maturity. Maturity judgments synthesize the public cases and deployment figures this chapter was able to locate; items lacking public evaluation are left as placeholders.

Technical Paradigm	Teaching-Stage Mapping	Underlying Mechanism	Deployment Maturity (this chapter's assessment)
Education-vertical LLM	Lesson prep, question generation, explanation, grading	Aligns with curriculum standards and grade level via educational corpora, reducing general-model knowledge mismatches	Broad rollout (e.g., TAL's MathGPT, iFlytek Spark education models; evaluation basis in Section 3.5)
Teaching-agent orchestration	Classroom-workflow scheduling, multi-tool coordination	Decomposes tasks against teaching objectives, calls tools, decides when to intervene	Pilot (e.g., MagicSchool's "Raina," Khanmigo's tool-collection entry point)
RAG (Retrieval-Augmented Generation)	School-specific-knowledge-base explanation, Q&A	Constrains generation to trusted textbook/school-specific resources, suppressing hallucination	Pilot to broad rollout (school-corpus integration is now standard in education-vertical products)
Memory mechanisms	Continuity across class sessions	Maintains long-term context for a class/unit	Demo to pilot [TBD: classroom-scale validation of cross-session memory]

Multimodal perception	Board recognition, learning-state observation	Fuses voice/image/screen input to sense classroom state	Demo to pilot (constrained by emotion-recognition regulation; see Section 3.4)
On-device deployment	Offline/low-latency classrooms, grading tablets	Local inference reduces latency, protects data, and adapts to weak-network environments	Pilot (e.g., iFlytek Spark Smart Grading Tablet; hardware figures in the companion Blue Book)

3.2.4 Where General-Purpose and Education-Vertical Models Diverge on Teaching Capability

How credible a teacher-facing product is depends heavily on whether its underlying model is a general-purpose LLM or one aligned specifically for education. The divergence between the two can be understood along three dimensions.

The first is **knowledge alignment**. General-purpose models draw their knowledge primarily from internet-scale corpora and lack precise memory of specific textbook editions, curriculum-standard line items, or grade-level cognitive-development norms — they readily err on details such as which volume a knowledge point belongs to, whether the difficulty matches the grade level, or whether terminology conforms to the curriculum standard. Education-vertical models undergo continued training and alignment on educational corpora — textbooks, curriculum standards, item banks — which markedly reduces this kind of knowledge mismatch. China's Academy of Information and Communications Technology (CAICT) and other institutions have begun tiered evaluations specifically for education LLMs (TAL's MathGPT earning a Level 4+ rating comes from exactly this kind of evaluation), providing an early third-party anchor for distinguishing the two — though fine-grained evaluation targeting specific teaching tasks still has room to grow.

The second is **pedagogical alignment**. General-purpose models tend to give answers that are "correct but delivered flatly," whereas good teaching requires structure — sequencing, follow-up questions, deliberate pauses, error correction. Google's LearnLM, fine-tuned against five learning-science principles, and iFlytek Spark Teaching Assistant, which draws on "accumulated experience from excellent teachers," are both, at bottom, efforts to fill this gap — so the model doesn't just answer correctly, but teaches like a teacher.

The third is **safety and values alignment**. Educational settings serving minors demand far more from content safety, values guidance, and privacy protection than general-purpose settings do. Education-vertical products typically build in stricter content guardrails and independent, student-facing safe environments (such as MagicStudent), and must comply with education-specific regulation — China's Interim Measures for the Management of Generative AI Services and algorithm filing requirements, the U.S.'s FERPA/COPPA, and the EU's GDPR and AI Act. This is exactly the

risk teachers most need to watch for when "borrowing" general-purpose conversational products for classroom use.

One clarification is worth making: "vertical" does not mean "stronger" — it means "more correct for the task." On open-ended reasoning and cross-disciplinary synthesis, frontier general-purpose models often still have a higher capability ceiling. The ideal teacher-facing product uses a strong general-purpose model as its foundation, aligns it with educational corpora and RAG, and wraps it in task-specific tools and safety guardrails — capturing the best of both a high capability ceiling and educational trustworthiness.

3.3 A Map of Product Forms

3.3.1 Four Product Forms, from Conversational Tool to Agent

Teaching-side generative AI products in 2026 fall into four forms, forming a continuum from "light" to "heavy."

1. **Conversational teaching assistants**: Natural-language dialogue as the primary entry point, serving lesson-prep Q&A, content generation, and explanation polishing. This is the broadest-reach, lowest-barrier category, and it splits into two branches. One branch consists of general-purpose conversational products (ChatGPT, Ernie Bot, Tongyi, Doubao, and the like) that teachers "borrow" for instructional tasks — the advantage is a high capability ceiling and low cost, the drawback is a lack of curriculum-standard alignment and educational safety constraints, leaving prompt engineering and content review to the teacher. The other branch consists of education-dedicated conversational interfaces that split off from general-purpose products (such as Khanmigo, or Spark Teaching Assistant's dialogue entry point), constrained a second time by educational corpora and safety guardrails. The divide between the two branches is exactly the "generally usable" versus "educationally trustworthy" distinction this chapter keeps returning to.
2. **Embedded teaching plug-ins**: Generative functionality modules built directly into lesson-prep platforms, office suites, interactive whiteboards, and all-in-one teaching devices, letting teachers invoke AI within their existing workflow at low migration cost. A domestic example is NetDragon's 101 Education PPT, which integrates AI generation capability into lesson-prep-and-instruction software; international examples include Google embedding 30-plus AI tools into Google Classroom as "Gemini in Classroom," and Microsoft wiring Khanmigo and Copilot capabilities into its education office suite. This form's core value proposition — "doesn't change teachers' existing work habits" — tends to give it an edge in scaled adoption over standalone apps.
3. **Teaching agents and agent platforms**: Agents that can be configured with roles, bound to tools, connected to school-specific knowledge bases, and given a measure of autonomous orchestration, aimed at composite tasks such as "generate an entire unit's teaching package in one click," "select

tools conversationally," or "chain tasks across tools." MagicSchool's teacher-facing conversational agent "Raina" is a representative example: a user describes a need in a single sentence, and the agent decides which tool to call, or generates directly and iterates within the conversation. Khanmigo, for its part, organizes 25-plus tools into a collection scheduled by category — Plan, Create, Differentiate, Support, Learn. This form represents an early productization of "decision-layer orchestration," and mostly sits at the pilot-to-early-rollout stage; its autonomy currently extends only to "selecting and chaining tools," not to "instructional decision-making."

4. **Multimodal and on-device teaching hardware**: All-in-one teaching devices, interactive whiteboards, smart grading tablets, and teacher-point-of-view smart glasses that push generative capability down to the device itself. This form is the interface between "content/interaction capability" and the "physical classroom," addressing both real-time-responsiveness and weak-network problems while serving as an important lever for data compliance (data never leaves the device). For hardware forms, shipment figures, and evaluation data, see this institute's *2026 Blue Book on the AI Smart Glasses Education Industry* and *Global Educational Robotics Development Blue Book 2026*, which cover the supply chain and hardware evaluation in depth; this chapter does not repeat that market sizing and instead folds hardware in only from the angle of teacher-empowerment mechanisms.

The four forms are not mutually exclusive — they are the same underlying capability packaged at different "weight classes." The lighter the form, the easier it spreads but the harder it is to guarantee educational trustworthiness; the heavier the form, the better it can secure compliance and real-time performance, but at higher deployment cost. Competition among teacher-facing products is, in essence, a search for the point in the "usability—trustworthiness—deployment cost" triangle that best matches a given teaching scenario.

3.3.2 Where Products Land Across the Teacher's Full Instructional Workflow

Mapped onto a teacher's timeline, product capabilities land across three segments: before, during, and after class.

- **Before class (prep and question design)**: Lesson-plan generation, slide-deck generation, tiered practice problems and question design, cross-disciplinary scenario design, multilingual and accessibility material generation. This segment is the most mature and has the most clearly established value — it is also where product feature density is currently highest. China's Ministry of Education "AI+ Education" Action Plan explicitly calls for "empowering teacher instruction, advancing intelligent applications that cover the full before-, during-, and after-class cycle," and the before-class segment is the first to have been productized.

- **During class (instruction and interaction)**: Real-time explanatory supplementation, in-class question generation, board- and screen-content recognition, instant multilingual support, classroom-interaction design. This segment depends on multimodal and on-device hardware, maturity varies widely, and one must carefully distinguish "usable" from "demonstrable." Functions that involve biometric data or emotion inference are bound by regulatory red lines (see Section 3.4).
- **After class (feedback and redesign)**: Assignment grading and learning-state analysis, structured review of classroom recordings, teaching-behavior analysis, adaptive redesign for the next lesson. This segment already has mature, deployed-at-scale functions (such as smart grading), but is also deeply entangled with Intelligent Assessment and Supporting Research (PD), and product boundaries here are increasingly blurred.

A data loop runs across the three segments, and product design is steadily working to close it: after-class learning-state analysis should, in principle, flow back into the next round of before-class prep and question design, closing a "teach—learn—assess—prep" loop. Most current products remain single-point features, and this loop is not yet fully closed. Whether after-class feedback can be reliably converted into before-class redesign input is the key marker of whether a teacher-facing platform has truly entered the decision-making layer. The Ministry of Education's "AI+ Education" Action Plan calls for "intelligent applications covering the full before-, during-, and after-class cycle" — and the deeper requirement behind that phrase is precisely this closed loop, not a pile of isolated features.

3.3.3 Three Implementation Paradigms for Content Generation and Orchestration

Before examining specific products, it is worth laying out the three implementation paradigms shared across teacher-facing products — paradigms that determine both a product's capability ceiling and its floor for trustworthiness.

- **Task-based encapsulation (prompt-to-tool)**: Common teaching tasks are hard-coded into parameterized tools (e.g., "input grade level + topic → output a rubric"), so teachers can invoke them without needing prompt-engineering skills. MagicSchool's 60-to-80-plus tools, Khanmigo's five tool categories, and Google Classroom's 30-plus tools all fall into this category. The advantage is reliability, reusability, and evaluability; the cost is that flexibility is bounded by whatever the tool designer anticipated.
- **Retrieval-Augmented Generation (RAG) plus school-specific adaptation**: Before generating, the system retrieves relevant material from textbook libraries, curriculum-standard databases, and district/school-specific resources, then constrains generation accordingly — suppressing hallucination and improving alignment with curriculum standards. iFlytek Spark Teaching Assistant's retrieval integration across "self-built + district/school-specific + user-generated"

resources, and Google's emphasis on generating against learning objectives and textbook exemplars, are both instances of this paradigm. It is the technical watershed separating education-vertical products from general-purpose ones.

- **Conversational orchestration (agentic orchestration)**: An agent parses the teacher's intent and autonomously decides which tool to call, in what order to chain them, when to ask a clarifying question, and when to hand the result back to the teacher. MagicSchool's "Raina" is currently the most mature teacher-facing instance of this. This paradigm raises the one-shot completion rate for composite tasks, but it also hands part of the "what should happen" judgment to the system — which is precisely why it must rest on a clear delineation of responsibility and human final review.

In real products, the three paradigms are typically layered together: RAG ensures the generation is correct, task-based encapsulation ensures it's easy to use, and conversational orchestration ensures the steps chain together smoothly. Understanding these three paradigms will help separate superficially similar products by their real technical depth in the comparative evaluation in Section 3.5.

3.3.4 Profiles of Representative Teacher-Facing Products

The following profiles, organized into "international" and "domestic" groups, cover representative products this chapter was able to verify — products aimed directly at teachers as end users — with an emphasis on feature composition, genuine deployment scale, and evidence strength. Every figure is tagged with its source year; anything not independently verifiable is left blank.

(1) Khan Academy Khanmigo for Teachers (United States)

Khanmigo is Khan Academy's AI teaching assistant and tutoring system. Its teacher-facing edition, Khanmigo for Teachers, has been offered in partnership with Microsoft since May 2024, running on Azure OpenAI Service. According to Microsoft's Education Blog (August 2024), the teacher edition is provided free with 25-plus teacher-specific tools, organized into five categories — Plan, Create, Differentiate, Support, and Learn — with representative tools including lesson-opener design (Lesson Hook), question generation (Question Generator), report-card comment drafting (Report Card Comments), and class-newsletter composition (Class Newsletter). This edition is free in English across 49 countries. Its teaching value lies in repackaging "a general-purpose model that requires skilled prompting" into "task-based tools a teacher can invoke without learning prompt engineering."

Khanmigo's classroom evidence comes primarily from a regional pilot: at Newark Public Schools in New Jersey, following a 2023–2024 school-year pilot that was subsequently expanded, Chalkbeat reported that the district's math pass rate on standardized testing rose from about 15% in 2023 to 17.7% — though the district itself stated it could not attribute this quantitatively to Khanmigo tutoring specifically. By the time of the relevant reporting, the partnership covered dozens of schools in the

district and roughly 29,000 students. It bears emphasizing that Khanmigo serves a dual role — student tutoring and teacher assistance — and that causal evidence on the student-tutoring side is still accumulating, while the teacher-facing side's value currently shows up mainly as prep and grading efficiency rather than direct learning gains.

(2) MagicSchool AI (United States)

MagicSchool is one of the fastest-growing teacher-facing generative AI platforms. According to its own announcement, the company closed a $45 million Series B round on February 11, 2025, led by Valor Equity Partners with participation from Bain Capital Ventures, Adobe Ventures, and others; the announcement stated that the platform had more than 6 million registered educators, partnerships with more than 10,000 schools, and reach across roughly 160 countries. By early 2026, multiple sources indicated further growth in registered educators, with school partnerships exceeding 13,000. Its product is known for 60-to-80-plus task-based tools spanning tiered lesson plans, a Rubric Generator, test questions, family-communication tools, and IEP (Individualized Education Program) assistance; third-party commentary commonly cites a self-reported figure of teachers saving "7 to 10 hours per week."

The Rubric Generator offers a window into the mechanics of this task-based encapsulation: a teacher selects a grade level, fills in an assignment title, description, and objectives, and sets a scoring-band configuration, and the tool produces a tabulated rubric with labeled evaluation dimensions and per-band descriptors. This kind of tool compresses "a task that would take a teacher half an hour to work out" into a minutes-long operation — and is exactly the micro-level source of the "hours saved per week" figure. MagicSchool also offers MagicStudent, an independent, safety-controlled environment for students, where teachers can create supervised "student rooms," forming a "teacher generates, student uses" loop alongside the teacher-facing tools.

In 2025, MagicSchool introduced a teacher-facing conversational agent, "Raina," positioned not as a general-purpose chatbot but as an orchestration entry point "trained on pedagogy, able to understand a need in conversation and either route it to the right tool or generate content directly" — maintaining conversational context so results can be iterated on. This is a workable instance of the "decision-layer orchestration" discussed in Section 3.2.1, applied on the teacher-facing side. On compliance, MagicSchool states it complies with FERPA, COPPA, GDPR, and SOC 2 Type 2, and cites a favorable privacy rating from Common Sense Media. It is worth flagging that figures such as "hours saved" and "approval ratings" are mostly self-reported or vendor-sourced, without independent randomized-controlled verification; public information also indicates that in early 2026 MagicSchool redesigned its student-facing product and adjusted the personality design of its AI assistant, reflecting that the industry is still working out where to draw the line between "personified" and "tool-like" in education products. Selection decisions should be weighed against the evidence-based evaluation recommendations in Section 3.5.

(3) Google Gemini for Education / LearnLM (United States)

Google's education strategy can be summarized as "general-purpose model + learning-science fine-tuning + workflow embedding." Its Gemini for Education is powered by Gemini 2.5 Pro and incorporates LearnLM — a model family fine-tuned against learning-science principles (active learning and feedback, cognitive-load management, personalization, curiosity, and metacognition). The largest-scale teacher-facing deployment point is "Gemini in Classroom": Google has embedded 30-plus AI tools into Google Classroom, supporting lesson-plan generation against learning objectives, quiz-and-rubric generation from scratch or from exemplars, text-difficulty adjustment to match student reading levels, and guided prompts targeting common misconceptions — free for Workspace for Education users aged 18 and over. Google's 2025 year-end review stated that its education AI capabilities had reached more than a thousand U.S. higher-education institutions and tens of millions of students (this figure is primarily student-facing; teacher-facing scale [TBD: teacher-user figures]). One frequently cited internal LearnLM result states that "students who received a short LearnLM tutoring session had roughly a 5.5-percentage-point higher probability of solving a subsequent new problem correctly than students tutored only by a human tutor" — but this is student-tutoring-side evidence, self-evaluated by the vendor, and should be treated with appropriate caution. What is distinctive about Google's approach is its "ecosystem embedding" strategy: the overwhelming majority of K-12 teachers already use Google Classroom and Workspace, so adopting the AI tools carries essentially no additional migration cost — a structural reason Google can reach large user populations quickly. The flip side is that its educational trustworthiness depends heavily on the alignment quality of the underlying general-purpose model and the effectiveness of LearnLM's fine-tuning.

(4) TAL's "Jiuzhang" Large Model (MathGPT) (China)

TAL's self-developed Jiuzhang large model, branded MathGPT, is among the earliest domestically released hundred-billion-parameter-class large models focused on mathematics, with its core built around problem-solving and worked-explanation algorithms. It covers computation, word problems, and algebra problems spanning elementary, middle, and high school math, and has since expanded into functions such as essay grading. According to public information, MathGPT received a Level 4+ rating in CAICT's education-LLM evaluation — a relatively high tier as publicly announced at the time of that evaluation. On the teacher-facing side, TAL Smart Education offers a smart lesson-prep solution spanning "prep plus in-class delivery," supporting question lookup, test assembly, lecture-note and interactive-slide creation, and micro-lecture recording and editing. It is worth noting that MathGPT's strength lies in consistency of problem-solving and worked explanations within mathematics; independent evaluation of its cross-disciplinary generality and curriculum-standard alignment [TBD: third-party evaluation results by grade level/subject] is not yet available.

(5) iFlytek's "Spark Teaching Assistant" (China)

iFlytek, built on its Spark cognitive large model, has launched "Spark Teaching Assistant," positioned as "a professional AI partner that understands teaching." Through conversational interaction, it generates unit-level teaching plans, instructional designs, and context-appropriate slide decks for teachers, and covers scenarios including teaching reflection, lesson-topic inspiration, student comments, class-meeting design, and home-visit talking points. Its slide-deck generation draws on retrieval, analysis, and integration across iFlytek's self-built resources, district/school-specific resources, and user-generated content — a textbook example of the RAG-plus-school-specific-adaptation route. According to usage data disclosed by iFlytek Smart Education from front-line teachers: instructional-design efficiency improved by roughly 56.52%, resource-search convenience improved by roughly 56.22%, slide-deck-creation efficiency improved by roughly 64.18%, and the teacher approval rate stood at roughly 93%; the product is also reported to cover more than 4,000 schools nationwide and more than 200,000 teachers, with teachers averaging roughly 5.3 conversational generations per week. Beyond the assistant, iFlytek's Spark Smart Grading Tablet, released in June 2024, is a representative example of on-device hardware for teacher-workload reduction; the company states it can cut assignment-grading time from roughly 90 minutes to roughly 5 minutes, and learning-analytics time from roughly 60 minutes to roughly 1 minute. All of the above are vendor self-reported figures; independent third-party evaluation [TBD].

(6) NetDragon's 101 Education PPT (China; an affiliated enterprise of this institute)

101 Education PPT, from NetDragon (HK:0777), is an all-in-one lesson-prep-and-instruction tool aimed at teachers, covering different textbook editions across compulsory education, senior high school, and vocational secondary school, and offering free lesson plans, slide decks, micro-lectures, and other teaching resources along with subject- and classroom-interaction tools. According to publicly disclosed figures, 101 Education PPT has served roughly 9.7 million teachers nationwide, with cumulative installations exceeding roughly 35.76 million, and user coverage across 32 provincial-level administrative regions in China. NetDragon is among the earlier domestic enterprises to bring AI into education (it released an "AI Teaching Assistant" as early as 2018), and has continued integrating AI generation capability into its education product line to support lesson prep and slide-deck creation. It should be noted that this Blue Book applies the same source-verification standard to affiliated enterprises: for any dynamic data on the latest installation figures, teacher counts, or specific AI functionality, this chapter relies strictly on public disclosure and leaves unpublished figures blank [TBD: feature inventory and teacher-adoption rate for 101 Education PPT's generative AI module]. NetDragon's investments in the AI-glasses space (including industrial positioning related to Rokid) are not covered in this chapter; see this institute's *2026 Blue Book on the AI Smart Glasses Education Industry* for that analysis.

Across the six cases, three common evolutionary threads emerge for teacher-facing products. First, a move from "general-purpose conversation" toward "task-based toolsets," lowering the prompt-engineering barrier for teachers. Second, a move from "single-point tools" toward "conversational

orchestration entry points" (such as Raina), raising the one-shot completion rate for composite tasks. Third, a move from "cloud-based generation" toward "school-specific RAG plus on-device workload-reduction hardware," simultaneously addressing content trustworthiness and data-compliance constraints. Cross-comparable data on subject coverage, deployment scale, and teacher-adoption rates across products remains insufficient; this chapter recommends filling that gap with the evidence-based tools proposed in Section 3.5.

3.3.5 Comparing the International and Domestic Deployment Paths

Placing the cases above side by side reveals two deployment paths for teacher-facing generative AI — related, but distinct — in the international and domestic markets.

- **The international path centers on "platform toolsets plus ecosystem embedding."** MagicSchool, Khanmigo, and Google all follow a route of packaging large numbers of teaching tasks into tools and then embedding them into teachers' existing ecosystems — office suites, Classroom, LMS platforms — scaling extremely quickly (MagicSchool reaching millions to tens of millions of educators; Khanmigo, via its Microsoft partnership, reaching dozens of countries). The strength here is speed of adoption and breadth of tooling; the weak point is relatively thin localization to domestic textbooks and curriculum standards outside their home market, along with heavy dependence on underlying general-purpose LLMs (mostly from the OpenAI or Google families). Compliance is anchored to established frameworks such as FERPA, COPPA, and GDPR.
- **The domestic path centers on "vertical models plus integrated prep-and-instruction plus on-device hardware."** TAL's MathGPT, iFlytek's Spark Teaching Assistant and Grading Tablet, and NetDragon's 101 Education PPT place greater emphasis on deep adaptation to domestic textbook editions, curriculum standards, and examinations, and are frequently bundled with hardware — interactive whiteboards, grading tablets, AI learning tablets — forming an integrated "software plus hardware plus school-specific resources" solution. The strength here is curriculum-standard alignment and closed-loop scenario coverage; the weak point is relatively scarce cross-disciplinary generality and internationally comparable evaluation data. Compliance is anchored to the Interim Measures for the Management of Generative AI Services, algorithm filing, data classification and grading requirements, and the recently issued Guidelines for Teachers' Use of Generative AI.

Both paths face the same underlying challenge: how to convert self-reported "efficiency gain" figures into independent evidence of improved teaching quality, and how to hold the line on minors' privacy and emotion-recognition red lines when handling multimodal and learning-state data. These are exactly the questions Sections 3.4 and 3.5 take up next.

3.4 Typical Scenarios and Risk Boundaries

3.4.1 High-Value, High-Certainty Scenarios

- **Lesson-prep workload reduction**: Freeing teachers from repetitive content production is the easiest value to verify. The Gallup–Walton finding of "5.9 hours saved per week, roughly six weeks per school year" and iFlytek's claimed "instructional-design efficiency improvement of more than 56%" point in the same direction. The risk lies in content accuracy and curriculum-standard fit, which requires teacher final review.
- **Assignment grading and learning-state feedback**: Smart grading tablets and grading tools have already scaled in some regions and represent the highest-certainty workload-reduction point in the after-class stage. The risk lies in the consistency and explainability of subjective-item scoring, which requires preserving a human-review and appeals channel.
- **Tiered and personalized question design**: Rapidly generating tiered practice sets and rubrics for differentiated learning needs. The risk lies in controllability of difficulty calibration and knowledge-point coverage, which calls for RAG constraints anchored to school-specific item banks and curriculum standards.
- **In-class language and accessibility support**: Real-time text/speech support for multilingual classrooms and for students who are deaf, hard of hearing, blind, or low-vision. This scenario is strongly tied to on-device hardware; see this institute's AI-glasses Blue Book for the relevant industry and hardware-evaluation figures.
- **Family communication and administrative writing**: Parent newsletters, student comments, home-visit talking points, notices, and summaries — administrative writing that consumes a substantial share of teachers' hidden work hours, and a scenario heavily covered by multiple products (Khanmigo's Class Newsletter, iFlytek's home-visit talking points, MagicSchool's family-communication tools). Its value is well established and its risk is low, since the content is largely formulaic — though teachers still need to verify accuracy and appropriateness for individual students, to avoid mechanical phrasing that damages home-school relationships.

The common feature across these scenarios is clear task boundaries, manageable error costs, and low cost of human final review. They have scaled first precisely because they can create certain value at the content layer — the most mature layer — without needing to rely on the still-immature layers of multimodal perception and autonomous orchestration.

3.4.2 Scenarios Requiring Caution, and Red Lines

- **Claims of "autonomous teaching agents"**: Current instances are mostly demos or tightly controlled pilots, and it is strictly impermissible to represent them as capable of replacing teacher instruction without supervision. China's Ministry of Education, in its 2025 Guidelines

for Teachers' Use of Generative AI (First Edition), explicitly requires that teachers "consistently play the leading role in student development, using generative AI only as an auxiliary tool," and stipulates that in key areas of student development — values guidance, moral education, emotional cultivation, psychological support — work "must be led by the teacher and may not be delegated to technology." This draws a firm boundary of responsibility that product positioning cannot cross.

- **Classroom multimodal perception and emotion recognition**: Because these touch minors' biometric and emotional data, they must strictly observe regulatory red lines. The EU AI Act, Article 5(1)(f), has prohibited — since February 2025 — inferring a natural person's emotions in the workplace or in educational institutions based on biometric data (except for medical or safety purposes), with violations subject to fines of up to 7% of global annual turnover; its accompanying rationale (Recital 44) points directly to such systems' insufficient scientific basis and limited reliability and generalizability. In practice, this means that "camera-based attention/emotion monitoring" classroom features are, in principle, prohibited within EU jurisdictions, and should be designed cautiously elsewhere too, on a basis of minimal collection, clear notice, and an opt-out.
- **Content quality, bias, and hallucination**: Multiple 2024–2025 studies show that AI-generated lesson plans, absent scaffolding such as role constraints, exemplars, and rubrics, readily suffer from vague objectives, shallow cognitive levels, and superficial differentiation — and can even "fabricate facts" (hallucinate) when textbook information is insufficient. In blind evaluations, human-written lesson plans still generally outscore AI-generated ones on quality dimensions. On this basis, generated content must pass professional teacher review before entering the classroom, and products should build in rubric-based constraints and source-traceability mechanisms.
- **Data leaving the device and on-device compliance**: The collection and transmission of sensitive data — classroom audio/video, student responses — makes on-device deployment a compliance path as well as a latency solution, but it cannot substitute for institutional data-governance arrangements. China's Guidelines for Teachers' Use of Generative AI also requires relevant enterprises to implement data classification and grading, security assessment, and algorithm filing in accordance with law.
- **The "de-skilling" concern for teacher professional capability**: Over-reliance on generative tools risks weakening young teachers' core professional capacity to independently design instruction and diagnose learning states — a hollowing-out risk of "knowing how to use the tool but not how to teach." This risk is hard to quantify, yet it bears on the long-term professionalism of the teaching workforce, and should be counterbalanced in teacher training and product design under the principle of "assisting rather than replacing thought" — for example, by having tools

output "an editable draft plus a rationale for the design" rather than "an unquestionable finished product."

3.4.3 Grading Evidence Strength: Distinguishing Self-Reported, Piloted, and Causal Claims

The "evidence" this chapter located varies enormously in credibility and must be graded accordingly, to avoid using low-strength evidence to support high-strength conclusions.

- **Vendor self-reported figures (weakest)**: Examples include "saves 7–10 hours per week," "efficiency improved 56%/64%," "93% approval rating." This kind of data comes from vendors or their partners, with sample composition, controls, and measurement methods typically undisclosed — usable only as a directional reference, never as a decisive basis for procurement decisions.
- **Third-party survey data (moderate)**: Examples include the Gallup–Walton survey of 2,232 teachers and RAND's national teacher sample. This kind of data has clearly defined samples and methods and carries higher credibility, but it still reflects teachers' self-reported use and perceptions rather than a causal measure of learning outcomes.
- **Quasi-experimental/regional pilot data (moderate to strong)**: Examples include the change in math pass rates following the Khanmigo pilot in Newark. This kind of data is closer to real classrooms, but commonly lacks a control group and struggles to rule out confounding factors — and the districts themselves often state they "cannot attribute this quantitatively."
- **Randomized-controlled/controlled-experiment data (strongest)**: Rigorous randomized-controlled evidence specifically for "teacher-facing generative AI improving teaching quality" remains scarce. The controlled experiments that do exist mostly concentrate on the student-tutoring side (such as LearnLM's problem-solving accuracy figures, or AI-tutoring comparison studies), and cannot be directly transposed into conclusions about teacher empowerment.

This grading points to the field's core evidence gap: workload reduction from lesson prep and grading already has moderate-strength evidence behind it, but whether that workload reduction translates into better teaching quality and student learning outcomes still lacks strong causal evidence. Product selection and policymaking should proceed with cautious optimism accordingly, and should prioritize investment in evaluation designs capable of producing strong evidence.

3.5 Recommendations for Evidence-Based Evaluation: Making "Empowering Teaching" Comparable and Verifiable

Vendor-disclosed figures on "time saved," "efficiency-improvement percentages," and "approval ratings" are mostly self-reported or measured under controlled conditions, without a unified method

or independent review. To keep product capability claims from resting on marketing language alone, this chapter recommends introducing three categories of evidence-based tools into the Blue Book going forward.

- **Teaching-task evaluation for education-vertical models**: Design evaluation sets specific to teaching tasks — lesson-plan generation, question-design quality, explanation accuracy, curriculum-standard alignment, grading consistency — reporting capability by grade level and subject, with scoring criteria and sample composition disclosed publicly [TBD: unified evaluation method and results]. Existing industry-level evaluations (such as CAICT's tiered education-LLM evaluation) can serve as a reference anchor, but need to be supplemented with fine-grained evidence targeting specific teaching tasks.
- **Product comparison and capability radar charts**: Score products of the same form on a unified basis across six dimensions — content generation, interaction experience, orchestration capability, school-specific adaptation, compliance and safety, and on-device performance — and present the results as radar charts, clearly distinguishing vendor self-reported figures from independently reviewed ones [TBD: comparison sample and scoring].
- **A deployment-maturity timeline**: Use a timeline to mark the transition points for each product form as it moves from demo to pilot to broad rollout to full scale, distinguishing "capability achieved" from "actually deployed in the classroom." For example: lesson-prep and grading tools have already reached "broad rollout to full scale"; conversational orchestration entry points sit at "pilot to early rollout"; cross-session memory and multimodal learning-state sensing sit at "demo to pilot"; and regulation-constrained emotion recognition does not enter productization in most jurisdictions.

On evaluation methodology, this chapter further recommends three operational principles to put the tools above into practice.

- **Anchor to specific tasks rather than vague capability claims**: Evaluation should be tied to concrete teaching tasks (such as "lesson-plan generation for a specific unit of eighth-grade physics under the Renmin Education Press edition," or "grading-consistency for open-ended middle-school math problems"), blind-scored by subject-matter teachers against a shared rubric — rather than a generic judgment of whether "the model is strong." Lesson-plan quality can be scored along dimensions such as curriculum-standard alignment, measurability of objectives, distribution of cognitive levels, quality of differentiated design, and feasibility of activities; grading quality can be measured by agreement coefficients with teacher scores (such as quadratic-weighted Kappa) and the accuracy of error-cause explanations.
- **Human-versus-AI comparison and incremental measurement**: Beyond comparing different products against each other, compare the quality and time difference between "teacher working alone" and "teacher plus AI," to measure AI's genuine marginal contribution — ideally

introducing random assignment to approximate causal evidence and help close the evidence gap identified in Section 3.4.3.

- **Transparency of measurement basis and disclosure of interests**: Any cited efficiency or effectiveness figure should be tagged with sample size, control-group setup, measurement method, and data source (vendor / third party / academic), with charts visually distinguishing vendor self-reported figures from independently reviewed ones, so marketing language never gets folded into evidence-based conclusions.

The evaluations above must adhere to a principle of source verifiability: any score, ranking, or scale figure cited should carry a footnote with its genuine source and basis, and remain a placeholder pending empirical support; particular caution is warranted against misattributing "student-tutoring-side evidence" as "teacher-empowerment-side evidence" (as with some of the learning-outcome figures from LearnLM and Khanmigo, which are actually evidence from the student-tutoring pathway).

3.6 Recommendations for Development

For the development of teaching-side generative AI products, this chapter offers six recommendations, directed respectively at product developers, schools and education authorities, and evaluation and governance bodies.

1. **Anchor product positioning to "teacher productivity" rather than "teacher replacement."** Prioritize making real gains in high-certainty, already-quantified scenarios: before-class prep, after-class grading, and in-class language support. Treat decision-layer agentic orchestration as a gradual increment rather than a marketing hook — do not claim maturity that has not been reached — and make explicit, in both the product and any contract, the red line that "key areas of student development are led by the teacher," consistent with the Ministry of Education's Guidelines for Teachers' Use of Generative AI. Product output should take the form of "an editable draft plus a rationale for the design," rather than "an unquestionable finished product," to protect rather than erode teachers' professional judgment.
2. **Make RAG and school-specific knowledge bases the default foundation for teaching agents.** Constraining generation against school-specific textbooks and curriculum standards to suppress hallucination and knowledge mismatch is the key differentiator between education-vertical and general-purpose products (iFlytek Spark Teaching Assistant's school-specific resource retrieval and Google's curriculum-standard-aligned generation are both instances of this route), and it is the precondition for trustworthy deployment. Generated content should be traceable to its source and reviewable by teachers against a rubric; for requests that exceed the knowledge base's coverage, the product should flag the uncertainty honestly rather than fabricate an answer.

3. **Build compliance and safety into the product-design process from the start.** For classroom multimodal data, minors' data, and emotion-recognition red lines, implement minimal collection, on-device-first processing, and auditability at the architecture level. For EU and similar jurisdictions, biometric-based emotion-inference functions should be off by default (the EU AI Act's Article 5 already lists this as a prohibited practice); for China, implement data classification and grading, security assessment, and algorithm filing. See this Blue Book's governance-and-safety-related chapter for the detailed governance framework.
4. **Establish verifiable teaching-evaluation and comparative-benchmarking mechanisms.** Use education-specific evaluations, capability radar charts, and maturity timelines to support selection decisions; do not let demos substitute for empirical evidence or self-reported figures substitute for independent review, and rigorously distinguish evidence from the teacher-empowerment side from evidence on the student-tutoring side. Related hardware forms' industry and evaluation figures can be cross-referenced against this institute's *2026 Blue Book on the AI Smart Glasses Education Industry* and *Global Educational Robotics Development Blue Book 2026*.
5. **Advance teacher digital-literacy training in step with product deployment.** RAND data shows that U.S. school-district teacher AI training, while growing fast, still lags front-line use; domestically, the Ministry of Education's "AI+ Education" Action Plan likewise lists "comprehensively raising teacher digital literacy" as a key priority. Schools introducing teacher-facing products should simultaneously roll out training on "how to review AI output" and "how to hold the line on data and student-development boundaries," to avoid the misuse and de-skilling that result when tools arrive ahead of literacy.
6. **Use the "teach—learn—assess—prep" closed loop, not isolated features, as the platform-selection standard.** Real value lies not in the number of tools but in whether after-class learning-state data can reliably flow back into before-class redesign. When procuring, schools and education authorities should focus on whether a platform can close the full-cycle data loop and connect with the existing teaching ecosystem and school-specific resources, rather than being dazzled by the length of a tool list.

In sum, the central question for teaching-side empowerment in 2026 is no longer "can generative AI assist teaching" — multiple independent surveys have already confirmed its workload-reduction value in lesson prep and grading — but rather "in what form, at what maturity level, and under what governance constraints does it enter the classroom." This chapter has offered a structured answer through its three-layer mechanism framework and four-stage breakdown, its four product forms and three implementation paradigms, six real-product profiles, an evidence-strength grading scheme, and evidence-based evaluation recommendations. One thing should be kept clearly in view: the evidence for workload reduction is now relatively solid, but strong causal evidence that workload reduction translates into better teaching quality and learning outcomes remains scarce. Preserving teachers'

leading role in student development, and holding the line on minors' data, are the preconditions for any of this entering the classroom. Specific figures that remain uncertain or lack public verification are marked [TBD], pending further evidence.

Chapter References

1. Walton Family Foundation & Gallup. "Teaching for Tomorrow: Unlocking Six Weeks a Year With AI" (the "AI Dividend" / 5.9-hours–six-weeks figure; 2,232 U.S. public K-12 teachers; survey fielded 2025-03-18 to 2025-04-11). 2025. https://www.waltonfamilyfoundation.org/the-ai-dividend-new-survey-shows-ai-is-helping-teachers-reclaim-valuable-time ; Gallup version: "Three in 10 Teachers Use AI Weekly, Saving Six Weeks a Year." https://news.gallup.com/poll/691967/three-teachers-weekly-saving-six-weeks-year.aspx
2. RAND Corporation. "Uneven Adoption of Artificial Intelligence Tools Among U.S. Teachers and Principals in the 2023–2024 School Year" (RR-A134-25). 2024. https://www.rand.org/pubs/research_reports/RRA134-25.html
3. Microsoft Education Blog. "Khanmigo for Teachers: Your free AI-powered teaching tool" (25+ tools, five categories, Azure OpenAI, free in 49 countries). 2024-08-13. https://www.microsoft.com/en-us/education/blog/2024/08/khanmigo-for-teachers-your-free-ai-powered-teaching-tool/
4. Khan Academy. "Khanmigo: Free, AI-powered teacher assistant." https://www.khanmigo.ai/teachers
5. Chalkbeat Newark. Reporting on the Khanmigo pilot at Newark Public Schools and the math pass-rate change (15%→17.7%), and on district-wide expansion. 2024. https://www.chalkbeat.org/newark/2024/05/13/artificial-intelligence-khanmigo-chatbot-tutor-pilot-testing-districtwide-expansion/ ; https://www.chalkbeat.org/newark/2024/11/15/newark-receives-25k-gates-foundation-grant-to-expand-khanmigo-ai-tutor-chatbot/
6. MagicSchool. "Announcing MagicSchool's $45M Series B Fundraise" ($45 million Series B, 2025-02-11, led by Valor Equity Partners, 6M+ educators, 10,000+ schools, 160 countries). 2025. https://www.magicschool.ai/blog-posts/series-b-fundraise-for-teacher-ai
7. MagicSchool. "Rubric Generator" and "AI Tools for Teachers / Raina" (teacher-facing conversational agent; FERPA/COPPA/GDPR/SOC2 compliance; Common Sense Media privacy rating). 2025—2026. https://www.magicschool.ai/tools/rubric-generator ; https://www.magicschool.ai/magic-tools
8. Google for Education / Google Blog. "New Gemini tools for students and educators" (Gemini in Classroom 30+ tools, Gemini 2.5 Pro, LearnLM's five learning-science principles). 2025. https://blog.google/products-and-platforms/products/education/gemini-iste-2025/ ; year-end

review: https://blog.google/outreach-initiatives/education/google-for-education-year-in-review-2025/

9. iFlytek Smart Education. "Spark Teaching Assistant" (instructional-design efficiency +56.52%, slide-deck creation +64.18%, 93% approval rating, coverage across 4,000+ schools and 200,000+ teachers). 2023—2025. https://edu.iflytek.com/solution/school/teachers-assistant ; company news: https://edu.iflytek.com/about-us/news/company-news/795.html
10. iFlytek Smart Education. "iFlytek Spark V4.0 / Spark Smart Grading Tablet" (grading time 90→5 minutes, learning analytics 60→1 minute). 2024-06-27. https://edu.iflytek.com/about-us/news/company-news/1095.html
11. TAL's Jiuzhang large model (MathGPT), official and introductory materials (hundred-billion-parameter-class math model; problem-solving/worked explanation; CAICT education-LLM evaluation Level 4+). https://www.mathgpt.com/ ; https://www.100tal.com/
12. 101 Education PPT official site (integrated lesson-prep-and-instruction tool, covering multiple textbook editions, free resources); public reporting on NetDragon's education-AI deployment (9.7 million teachers, 35.76 million installations figures). https://ppt.101.com/
13. EU AI Act, Article 5 (prohibiting emotion inference based on biometric data in the workplace and in educational institutions, applicable from 2025-02, maximum fine of 7% of global annual turnover) and Recital 44. https://artificialintelligenceact.eu/article/5/ ; Future of Privacy Forum analysis: https://fpf.org/blog/red-lines-under-eu-ai-act-unpacking-the-prohibition-of-emotion-recognition-in-the-workplace-and-education-institutions/
14. Ministry of Education of the People's Republic of China. Guidelines for Teachers' Use of Generative AI (First Edition) [教师生成式人工智能应用指引（第一版）] (the first generative-AI usage standard aimed at teachers; key areas of student development led by teachers; enterprises to implement data classification/grading, security assessment, and algorithm filing in accordance with law). 2025-11. Expert commentary: https://www.edu.cn/xxh/focus/li_lun_yj/202512/t20251201_2704690.shtml
15. Ministry of Education and four other departments. "AI+ Education" Action Plan [《"人工智能+教育"行动计划》] (empowering teacher instruction, covering the full before-, during-, and after-class cycle; four major priority tasks). 2026. http://www.moe.gov.cn/jyb_xwfb/s271/202604/t20260410_1433232.html
16. Ministry of Education. Guidelines for Generative AI Use by Primary and Secondary School Students (2025 Edition) and Guidelines for General AI Literacy Education in Primary and Secondary Schools (2025 Edition) [《中小学生成式人工智能使用指南（2025 年版）》《中

小学人工智能通识教育指南（2025 年版）》]. 2025-05-12.

https://www.edu.cn/xxh/focus/zc/202505/t20250513_2667990.shtml

17. A review of empirical studies on the quality, bias, and hallucination of AI-generated lesson plans (arXiv 2510.19866, evaluation of high-school physics lesson plans; socialinnovationsjournal on pedagogical bias in lesson-plan generators; MIT Sloan guidance on AI hallucination and bias in teaching; findings that human-authored lesson plans score higher in blind evaluation). 2024—2025. https://arxiv.org/pdf/2510.19866 ; https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

Chapter 4 Supporting Learning: Mechanisms, Product Forms, and Recommendations

4.1 Introduction: From "Q&A Tool" to "Learning Companion"

In the 2024 edition of the *Generative AI Product Development Report*, the "Supporting Learning" scenario centered on conversational LLMs answering questions on demand, explaining concepts, and helping with homework — with a chat window as the primary interface. By 2026, as agent orchestration, Retrieval-Augmented Generation (RAG), long-term memory, and on-device multimodal capability have all matured, both the technical foundation and the product logic behind learning support have shifted structurally. Generative AI is no longer just a passive Q&A tool; it is evolving into a "learning companion" capable of managing goals, accompanying learners through a process, adapting to the individual, and remembering across sessions.

This shift traces back to a well-known theoretical touchstone: educational psychologist Benjamin Bloom's 1984 "2 Sigma Problem." Bloom found that the average student who received one-on-one mastery-based tutoring outperformed roughly 98% of peers in a conventional classroom — an effect size of about two standard deviations. The catch was cost: hiring enough tutors to give every student that kind of attention was never going to scale. For forty years, ed-tech's central puzzle has been how to deliver the benefits of one-on-one tutoring at a price everyone can afford. Generative AI has captured the field's attention in 2024–2026 precisely because it is the first technology to make "every learner gets an on-call, personalized tutor" economically plausible. That said — and this chapter returns to the point repeatedly — being technically feasible is a long way from being pedagogically effective and safe.

This chapter keeps the book's "mechanism–product–evidence–recommendations" structure but rebuilds each section for the current moment. First, on mechanisms, the analysis expands from single-turn "knowledge retrieval" to multi-turn, cross-modal, memory-enabled modeling of the whole learning process. Second, on products, the scope widens from conversational apps to agentic learning applications, on-device learning hardware (AI learning tablets, practice devices, AI glasses, and the like), and multimodal learning environments, illustrated throughout with real, shipping products. Third, this edition adds an entirely new section on tutoring-effectiveness evidence, systematically reviewing the 2024–2026 crop of randomized controlled trials (RCTs) and meta-analyses on generative-AI tutoring, and carefully separating vendor claims from verifiable evidence. Fourth, the recommendations move beyond "how to use the tools well" to the harder question of how to protect learner agency and cognitive development within human–AI collaboration. One scoping note: this chapter focuses on the learner's side of the equation (the student perspective), which naturally

overlaps with Chapter 3 ("Empowering Teaching," the teacher's perspective) and Chapter 6 ("Intelligent Assessment") — readers may want to cross-reference both.

4.2 The Mechanisms Behind Supporting Learning

4.2.1 Cognitive Scaffolding: A Zone of Proximal Development That Adjusts Itself

The core mechanism by which generative AI supports learning is best understood through Vygotsky's "Zone of Proximal Development" and the related concept of "scaffolding." Traditional learning support is bottlenecked by teacher availability and resources, which makes it hard to give every learner guidance calibrated precisely to where they currently stand. Generative models, by contrast, can read a learner's input in real time and dynamically produce explanations, examples, follow-up questions, and hints at whatever difficulty and granularity fit — building a scaffold that stretches between what the learner already knows and what they don't.

Compared with traditional adaptive-learning systems such as Intelligent Tutoring Systems (ITS), generative models change the nature of scaffolding altogether. ITS relies on knowledge components, rule bases, and feedback templates that experts write in advance — scaffolding that is discrete and pre-built. Generative models, on the other hand, produce natural-language explanations, analogies, and follow-up questions on the fly at inference time — scaffolding that is continuous and generated, and that can in principle cover a near-infinite range of problem variants and phrasings. This "continuous scaffolding" capability is the real increment over prior ed-tech: it turns what used to be spotty coverage of a handful of high-frequency problems into blanket coverage that responds instantly to any input.

The crux of this mechanism is **fading** — the deliberate, gradual withdrawal of scaffolding. Good learning support doesn't keep handing out answers; it tapers the intensity of its hints as the learner's ability grows, until the learner can go it alone. Tellingly, nearly every tutoring system that performed well between 2024 and 2026 explicitly encoded a specific pedagogical principle into its interaction strategy: don't give the answer directly — guide the learner toward it through heuristic questioning. Khan Academy's Khanmigo is explicitly designed to withhold direct answers in favor of Socratic questions that nudge learners toward their own solutions. Harvard's homegrown physics-tutoring agent, PS2 Pal, made "step-by-step, active-learning-style guidance rather than direct answers" its first design principle (see Section 4.5 for the effectiveness evidence). And Google DeepMind's LearnLM, fine-tuned specifically for teaching, was praised by supervising teachers for being "good at drafting Socratic questions that provoke deep reflection." All of this marks a shift in how learning-support products are designed at the mechanism level — from the 2023-era "answer anything asked" toward "guidance constrained by pedagogy."

It's worth being explicit about the precondition for "continuous scaffolding" to actually work: the quality of a generative model's scaffolding depends heavily on prompt engineering and on how explicitly pedagogical strategy gets encoded into the system. The very same underlying model, given a system prompt that says "answer directly" versus one that says "guide Socratically," can produce wildly different educational value. That explains a counterintuitive finding that shows up again and again in the evidence: **model capability alone does not predict tutoring effectiveness. What predicts it is the engineering quality of how pedagogical principles get built into the interaction strategy.** This point runs through the rest of the chapter.

4.2.2 Personalization: From Adapting Content to Adapting Process

Personalization is the second mechanism through which generative AI supports learning, and it unfolds in two stages:

- **Content adaptation**: adjusting how an explanation is phrased, which examples are used, and which modality is used to present it, based on a learner's knowledge state, interests, and language proficiency. For the same math problem, a struggling student might get a more granular step-by-step breakdown, while a stronger student gets a terser hint at the underlying approach.
- **Process adaptation**: using long-term memory to track a learner's recurring errors, preferences, and progress across multiple sessions, and then adjusting the pace and sequencing of subsequent learning accordingly.

The jump from content adaptation to process adaptation depends on **memory and user modeling** capabilities that have gradually come online in 2025–2026. On the student-facing side, this is no longer theoretical: Yuanfudao's "Xiaoyuan AI" publicly claims to have built a "memory mode" that tracks a user's grade level, practice history, and results — so that "when different students practice or ask about the same problem, the system matches the explanation style to their individual usage history." This kind of design pushes personalization from tweaking tone and difficulty within a single session to something closer to a learning profile that persists across sessions.

A second pillar of process adaptation is structuring learning data into a computable "knowledge profile." Chinese AI learning tablets typically build on a knowledge-graph backbone: a subject gets broken into fine-grained knowledge points, and the system estimates a learner's mastery probability for each point based on performance in diagnostic assessments and routine practice. That estimate then drives a closed loop of locating weak points, recommending targeted problems, explaining, and retesting. Plugging in a generative LLM upgrades the "explaining" step in that loop — from playing a pre-recorded video to generating a step-by-step explanation tailored to the learner's specific mistake in real time, marrying structured diagnostic capability with generative explanation. This is the

technical core of how AI learning tablets have repositioned themselves between 2024 and 2026, from "question-bank machines" to "AI tutoring machines."

But process adaptation introduces new mechanism-level risks of its own. The accuracy, freshness, and privacy boundaries of memory become critical variables in the quality of learning support. If the user model misjudges a learner's level, personalization can backfire — locking the learner into the wrong difficulty band and creating a "the more you practice, the narrower it gets" feedback loop. And retaining practice data and emotional-state signals from minors over the long term runs straight into data-minimization and privacy-compliance red lines (see Chapter 5, "Governance and Safety"). The stronger a product's personalization, the more it needs matching mechanisms for explainability and correction, so that learners and parents can see — and fix — how the system is judging the learner.

4.2.3 Knowledge Anchoring: Using Retrieval to Suppress Hallucination

The third mechanism concerns the reliability floor beneath learning support: **knowledge anchoring**. Generative models have a well-known Achilles' heel — hallucination, the tendency to generate factually wrong content in a fluent, confident tone. In casual chat, hallucination is a mild annoyance; in a learning context, it can propagate incorrect knowledge directly into a learner's head, and the harm is compounded by the fact that learners are often the least equipped to catch the error.

Retrieval-Augmented Generation (RAG) is the leading engineering approach to this problem today. The mechanism: before generating a response, the system first retrieves relevant passages from a controlled knowledge base — textbooks, course materials, vetted worked solutions — and then asks the model to answer based on what it retrieved, anchoring the response in verifiable external knowledge rather than leaning on the model's internal parametric memory. Research shows that anchoring responses in retrieved text meaningfully lowers hallucination rates relative to pure generation. In education specifically, RAG brings two further benefits: it lets the system inject up-to-date, domain-specific knowledge that teachers have vetted, and it allows the model to cite its sources, making answers traceable and open to challenge. That combination fits neatly with how demanding learning contexts are about theoretical and mathematical correctness.

For student-facing products, RAG's value isn't just fewer errors — it's making **the boundary of what the system knows, and who's responsible for it, explicit**. A tutoring system anchored to an authoritative textbook has an answer scope and evidentiary basis that can be defined and audited, which gives content-compliance work aimed at minors something concrete to hold onto. It's no accident that Chinese education-vertical LLMs typically draw their knowledge sources from proprietary subject-specific question banks and solution-process data — that's this same mechanism taking shape in the Chinese market. RAG is no silver bullet, of course: when nothing relevant turns up in retrieval, the model can still improvise, and retrieval quality, knowledge-base freshness, and resolving conflicting sources are themselves new engineering headaches.

4.2.4 Multimodal Embodiment: Getting Closer to Real Learning Contexts

The fourth mechanism comes from multimodality and on-device processing. Learning doesn't happen only in text chat — it's embedded in reading, writing, observing, hands-on manipulation, and speaking. Multimodal LLMs let a system "see" a learner's handwriting, a photographed problem, or a lab procedure, and "hear" their reading aloud or spoken language, providing support in contexts much closer to how learning actually happens.

Two categories of student-facing products make this mechanism especially visible. The first is **photo-based problem-solving**: Photomath (acquired by Google in 2023) is the archetype, using optical character recognition paired with a computer algebra system to read a photographed math problem and generate a step-by-step solution with animated explanation; Chinese AI learning tablets commonly build in similar "photograph the problem, get a step-by-step explanation" capability. The second is **spoken-language practice**: Duolingo Max's Roleplay and Video Call features let learners hold open-ended spoken conversations with an AI character, bringing listening-and-speaking — a highly embodied language skill — squarely into AI-supported territory.

Spoken practice is a particularly high-value application of the multimodal mechanism because it targets a real structural pain point in traditional learning: the "output" language skills (speaking, writing) depend far more on frequent, immediate, judgment-free practice opportunities than the "input" skills (reading, listening) do — and those opportunities have always been scarce in a traditional classroom, thanks to student-teacher ratios and the anxiety of speaking up. Generative AI's spoken-dialogue capability drives the marginal cost of speaking practice to near zero, and an AI partner never tires and never judges — which happens to fill exactly the gap that has been hardest to scale. That's the deeper reason language learning became one of the earliest and most mature applications of student-facing generative AI.

On-device processing, meanwhile, cuts latency, protects privacy, and removes the need for a network connection, letting learning support follow learners into the classroom, to the desk at home, and even outdoors. AI learning tablets typically run a hybrid setup — a small local model paired with a larger cloud model — to balance offline availability against compute demands. AI glasses and other wearables push multimodal sensing further still, toward a "see it, learn it" hands-free interaction model. For a deeper look at how on-device multimodal hardware is evolving and how it holds up under safety evaluation, see our institute's *Blue Book on AI Glasses in Education 2026*.

4.2.5 Two Sides of the Same Coin: The Tension Between Empowerment and Dependence

A caution is in order here. The mechanisms above cut both ways. The very same instant, fluent, humanlike support that can accelerate learning can also induce **cognitive offloading** and intellectual

passivity — learners leaning on the model to do their thinking for them, which erodes their own knowledge construction and metacognitive skills.

Learning science offers a clear explanation for why this tension runs so deep. Effective learning generally requires **desirable difficulties** — retrieval practice, spaced repetition, a healthy dose of productive struggle and trial and error — which are exactly what builds durable long-term memory and the ability to transfer knowledge to new situations. Generative AI's core selling point, though, is reducing difficulty and eliminating friction: it swaps out the retrieval, organizing, and self-correction a learner would otherwise have to do themselves for one smooth question-and-answer exchange. When the "friction" being removed is precisely the "desirable difficulty" that makes learning happen in the first place, convenience collides head-on with deep learning. This is the theoretical explanation behind the "faster completion, less learning" phenomenon covered in Section 4.5.3.

This tension isn't just theoretical — there's empirical backing for it. A 2025 Microsoft Research and Carnegie Mellon University survey of 319 knowledge workers, covering 936 real-world AI usage episodes, found that **higher trust in AI correlated with less critical thinking applied by the user, while higher self-confidence in one's own ability correlated with more critical thinking applied.** In other words, AI can shift cognitive effort away from "solving the problem" and toward "verifying the AI's output" — and if that verification step gets skipped too, critical thinking gets outsourced wholesale. In a learning context, this risk is amplified, because learners are still in a critical stage of cognitive development; some research uses the term "metacognitive laziness" to describe learners offloading metacognitive effort onto AI and, in the process, doing less of their own reflection and self-regulation.

Notably, this risk isn't evenly distributed. The survey above, along with related research, suggests that younger users and those with lower self-confidence are more prone to over-reliance, while more education and stronger AI literacy have a buffering effect against cognitive offloading. That finding carries two policy implications. First, products aimed at younger learners need built-in protections for learner agency, precisely because their self-regulation skills aren't yet fully formed. Second, AI-literacy education is itself a protective intervention — not merely a skill for "knowing how to use the tool."

Taken together, the mechanism design behind learning-support products cannot be judged solely on "quality of response" — it must also build in protections for learner agency. That's the deeper reason fading matters so much as a design principle. An ideal learning companion should dynamically balance "helping the learner past an obstacle" against "preserving the difficulty that's necessary for growth," rather than chasing the single easiest possible answer every time. This tension between empowerment and dependence is the central concern running through both the effectiveness-evidence review and the recommendations later in this chapter.

4.3 A Map of Product Forms (2026)

Compared with the 2024 edition's map, which centered on conversational apps, learning-support products in 2026 show a clear pattern of diverging into agentic applications, more varied hardware, and multimodal capability. The table below offers a broad categorization; only products verified in this chapter are listed as representative examples.

Product Category	**Core Form**	**Typical Capabilities**	**Representative Products (Verified)**
Conversational/Socratic tutoring apps	App/web chat interface	Heuristic Q&A, step-by-step guidance, explanation	Khanmigo (Khan Academy)
Language learning and spoken practice	App + voice/video dialogue	Roleplay, spoken practice, answer explanation	Duolingo Max (Roleplay/Video Call)
Photo-based problem solving	Photo recognition + computation engine	Step-by-step math solutions, animated explanation	Photomath (Google), photo-search features in various AI learning tablets
On-device learning hardware (AI learning tablets/practice devices)	Dedicated terminal + local/cloud hybrid model	Photo search, step-by-step explanation, learning diagnostics, parental controls	TAL xPad, Zuoyebang learning tablet, Yuanfudao practice tablet, iFLYTEK AI learning tablet
Education-vertical LLMs	Underlying model capability	Subject problem-solving, essay grading, spoken dialogue	TAL's "Jiuzhang," Zuoyebang's "Yinhe" (Galaxy), Yuanfudao's "Yuanli/Kanyun"
AI glasses and other wearables	On-device multimodal wearable	Contextual prompts, real-time translation, hands-free operation	See *Blue Book on AI Glasses in Education 2026*

4.3.1 From "Chat" to "Agent": Orchestrating Learning Applications

The most visible product shift in 2026 is the move from single-turn chat to agentic orchestration. An agentic learning application no longer just sits and waits for a question — it can take on a learning goal (say, "master this unit within two weeks"), break it into tasks on its own, call tools (retrieval, computation, code execution), track progress across many sessions, and proactively initiate follow-up questions and reviews. RAG lets it draw on an authoritative, controlled knowledge base to keep hallucination down; memory lets it carry a coherent learning thread across sessions.

Khanmigo is the most studied and most discussed student-facing example of this approach. It's built on OpenAI's GPT-4 family, but wrapped in education-specific constraints — the most central of which is that it doesn't give direct answers, guiding students toward their own reasoning through Socratic questioning instead, alongside features for writing practice and spoken-language practice. In

May 2024, Microsoft announced it would donate Azure OpenAI infrastructure to make "Khanmigo for Teachers" free for every teacher in the US (the teacher version had previously cost about US$4 a month). After migrating to Azure OpenAI, Khanmigo gained access to GPT-4o, GPT-4, Whisper, and DALL·E 3, among other models. On adoption: as of the end of 2023, roughly 45 school districts and 40,000 students were piloting it; by the end of the 2024–2025 school year, about 795 districts and 770,000 students in the US were using it. Student accounts are still billed to school districts — reportedly around US$35 per student per year — and Khan Academy has said it is working to bring that down into the US$10–20 range. One important caveat: **adoption scale is not the same thing as learning effectiveness**. Causal evidence for Khanmigo's own impact is still being built (see Section 4.5).

At the architecture level, an agentic learning application is roughly four layers deep. At the base sits the LLM itself (general-purpose or education-vertical). Above that is the RAG knowledge layer, anchoring responses to controlled textbooks and question banks. Above that is the memory layer, maintaining a learner profile across sessions. At the top is the orchestration and strategy layer, handling task decomposition, tool calls (retrieval, computation, code execution, drawing), and pedagogical strategy — when to hint, when to ask a follow-up question, when to withdraw the scaffold. Of these four layers, the memory and strategy layers are what actually separate a genuinely agentic learning application from a chatbot with a fresh coat of paint: memory determines whether it can carry a learning thread across sessions, and strategy determines whether it acts as a guide or as an answer machine. This is also why this chapter keeps insisting that "pedagogical design," not "model capability," is the deciding factor: the model itself is the layer that's easiest to obtain and most commoditized across products — the real room for differentiation lies in the three layers built on top of it.

In China, agentic capability on the student-facing side shows up mostly through "AI tutoring built into a learning tablet or practice device" and "education-vertical LLMs," rather than as a standalone general-purpose conversational agent — a pattern tied to China's cautious regulatory stance on letting minors use open-ended general-purpose LLMs. There's a coherent logic to this approach: since minors' self-regulation skills are still developing (see the risk distribution discussed in Section 4.2.5), constraining AI support to content-compliant, controllable dedicated devices is a more effective way to manage risk than leaving minors free to use an open general-purpose assistant. The tradeoff is some loss of flexibility and a lower ceiling on capability.

4.3.2 Going On-Device and Into Hardware: Learning Support Steps Off the Screen

The second major trend is hardware diversification and on-device processing — a pattern especially pronounced in the Chinese market. AI learning tablets and practice devices carve out a distinctive niche in home learning thanks to offline availability, parental controls, eye-health features, and content compliance. Adding a generative LLM upgrades them from "search a question bank, play a

video" to "AI-generated step-by-step explanations, learning diagnostics, and personalized problem recommendations."

On market size: according to RUNTO data, China's learning-tablet market (including AI learning tablets) sold about 5.923 million units across all channels in 2024, up roughly 25.5% year over year, with sales revenue of about RMB 19.06 billion, up roughly 37.6%. Growth continued into 2025: first-quarter unit sales across all channels came in at about 1.265 million, up roughly 29.4% year over year; second-quarter shipments reached about 1.54 million units, up roughly 44.6%. According to IDC data, iFLYTEK took the top spot in AI-learning-tablet sales revenue for the first time in Q2 2025, having already held the top spot in the high-end segment for several consecutive years. On overall market share, one research firm's tally from early 2025 put Zuoyebang at roughly 31.8%, TAL at roughly 20.9%, and Yuanfudao at roughly 11.7% (different firms report meaningfully different figures; consult each firm's original report for specifics). National trade-in and education-hardware subsidy programs are widely credited with stimulating demand in 2024–2025.

Representative products and their underlying LLM capabilities:

- **TAL's xPad series**: runs on TAL's homegrown "Jiuzhang" model (English name: MathGPT). Jiuzhang started in mathematics, built around problem-solving and explanation algorithms, and was among the first education-vertical LLMs in China to complete official registration. It has since expanded to cover all grade levels and subjects, supporting single-problem grading, essay assistance and grading, and spoken-dialogue practice. The xPad's Jiuzhang-powered "Ask Math Anytime" feature is claimed by the company to answer roughly 80% of K-12 math questions instantly; for the remainder, a human-tutor explanation is uploaded within an hour at the fastest, with an AI-generated video explanation following within 20 minutes.
- **Zuoyebang's learning tablet**: runs on the company's proprietary "Yinhe" (Galaxy) model, covering multi-subject problem-solving and multilingual dialogue. Reportedly, its smart learning tablet has a built-in question bank of more than 850 million items spanning the full K-12 range, including synchronized-curriculum courses and AI-driven interactive topic sessions.
- **Yuanfudao's practice tablet**: the full product line integrates the DeepSeek reasoning model and is deeply merged with the company's own "Yuanli" model; Yuanfudao's proprietary Kanyun model has completed official registration. Xiaoyuan AI's headline feature is personalization through "memory mode," and it was the first to introduce a "mental-health companion" feature — using the LLM to detect a user's emotional state and offer support through an "empathize, comfort, act" pathway. Public reporting indicates the Yuanfudao practice tablet passed one million units sold within roughly 16 months of launch.
- **iFLYTEK's AI learning tablet**: built on the iFLYTEK Spark (Xinghuo) model, it has held a long-standing position in the high-end market and, in Q2 2025, took the top spot in industry sales revenue for the first time.

For head-to-head benchmarking, radar-chart comparisons, and safety evaluation of this hardware, see our institute's *Blue Book on AI Glasses in Education 2026* and Chapter 7 of this Blue Book. A caution: figures such as "roughly 80% instant answers" and "850 million-item question bank" above are vendor-reported figures and should be read alongside third-party evaluation results, not treated as independently verified evidence of effectiveness.

4.3.3 Subject-Specific and Education LLMs: From General-Purpose to Purpose-Built

The third major trend is verticalization. General-purpose LLMs still have real limits in tasks that demand strong subject-specific logic — step-by-step math reasoning, code debugging — and carry hallucination and content-compliance risks, which has driven demand for subject-specific tools and education-vertical LLMs. All three leading Chinese education companies have launched, and officially registered, their own education LLMs — TAL's "Jiuzhang," Zuoyebang's "Yinhe," and Yuanfudao's "Yuanli/Kanyun" — and all commonly layer in general-purpose reasoning models such as DeepSeek to boost capability. The shared approach: use domain-specific data, process-level annotation (labeling not just the answer but the solution steps), and purpose-built evaluation to chase higher subject-matter accuracy and, just as important, get the pedagogical act of "explaining a problem" right — not merely "getting the answer right."

In language learning, verticalization has taken a different form: **scaling up content production itself**. Using generative AI, Duolingo launched 148 new language courses in a single push in 2025, more than doubling its course catalog. By the company's own account, "the first 100 courses took about 12 years to build; with generative AI, a shared content system, and internal tooling, nearly 150 courses went from development to launch in about a year." The lesson here: generative AI's impact on learning support isn't confined to the learner-facing interaction layer — it has also reshaped how educational content gets produced on the supply side.

Worth noting: this supply-side transformation came with real organizational and public turbulence. In April 2025, Duolingo's CEO announced an "AI-first" strategy — gradually replacing automatable outsourced work with AI and requiring teams to demonstrate that a task was fully automated before adding new headcount. The announcement triggered a sharp user backlash over perceived "quality decline" and "replacing people with AI," with a wave of cancellations across social media. The CEO later clarified that no full-time staff would be cut — the change affected only a small number of hourly contractors doing repetitive tasks — and that hiring pace overall would be unchanged. Although the controversy briefly dented user growth, the company's revenue that quarter still beat expectations. The episode carries a warning that reaches beyond this one company: **learning products carry an implicit promise of human care, and if an "AI-first" efficiency narrative overshadows a "learner-first" values narrative, it can cost a company real trust.**

From an equity standpoint, generative AI's potential to broaden access to learning support coexists with the risk of opening new divides. On one hand, Khanmigo's teacher version being free nationwide (thanks to Microsoft's funding) and Duolingo's continued free tier both lower the barrier to quality learning support, echoing the original goal of scaling one-on-one tutoring affordably. On the other hand, the most capable features — Duolingo Max's video call, the per-student annual fee districts pay for Khanmigo accounts, the price of premium learning tablets — still sit behind a paywall, which risks creating a new form of stratification in who actually gets access to the best capability. The upshot: judging the social value of a student-facing product means looking not just at its technical ceiling, but at how well it extends genuinely effective capability to learners with fewer resources.

For the evaluation methodology and results for education-vertical LLMs, see Chapter 7, "Evaluating Education-Vertical LLMs."

4.4 Photo-Based Problem Solving, Language Practice, and General-Purpose Assistants: Three Student-Facing Interaction Patterns

Cutting across the agent and hardware trends above, student-facing generative AI support falls into three high-frequency interaction patterns, each with its own mechanism emphasis and risk profile.

First, photo-based problem solving with step-by-step explanation. Photomath is the prototype: it uses enhanced optical character recognition paired with a computer algebra system so that scanning a problem with a phone camera immediately produces a step-by-step solution on screen; its paid tier adds textbook-matched solutions and animated walkthroughs. Google announced its acquisition of Photomath in 2022, closed the deal in 2023 after EU review, and folded it into the Google Play distribution system in early 2024 — with its recognition and problem-solving capability expected to feed back into products like Google Lens and Search. This pattern lines up closely with the genuine, everyday need for homework help, but it's also the pattern most prone to sliding into "just copy the answer." Whether the mechanism actually helps depends on whether the product guides understanding or simply hands over the result — which is exactly why Chinese AI learning tablets lead with "step-by-step explanation" and "AI video walkthrough" as their core selling points, rather than just delivering a final answer.

Second, language learning and spoken-language practice. Duolingo Max is the leading example. Its Roleplay feature lets learners hold open-ended, situational conversations with an AI character — ordering at a Paris café, discussing travel plans — while Video Call (a video conversation with the virtual character Lily) pushes practice further, toward real-time-paced spoken exchange; "Explain My Answer" gives targeted feedback on the accuracy and sophistication of a learner's response. What's distinctive about this pattern is that it makes "speaking out loud" — the hardest skill to practice cheaply in a traditional app — accessible at low cost.

Third, general-purpose conversational assistants used for learning. A large number of learners simply use ChatGPT, Gemini, DeepSeek, and other general-purpose assistants to look things up, revise essays, and write code. According to a 2024 Common Sense Media survey, about 70% of US teens have used at least one AI tool, and roughly four in ten of them used it for school-assignment tasks; Pew Research Center surveys from 2023–2024 similarly found that about 26% of US middle and high schoolers had used ChatGPT, and that most surveyed teens considered "using it to look things up" acceptable but "using it to write an essay for you" unacceptable. This split in what teens consider acceptable is itself a valuable teaching moment: learners already have an intuitive sense of where "assistance" ends and "doing it for you" begins — the job for schools is to translate that intuition into clear, actionable usage norms.

One structural tension worth flagging: **purpose-built tutoring apps and general-purpose assistants pull in opposite design directions.** Khanmigo, PS2 Pal, and similar dedicated products deliberately withhold answers in favor of guidance; a general-purpose assistant's whole product goal, by contrast, is satisfying the user's request as efficiently as possible — so a learner who types "just solve this problem for me" typically gets a complete answer immediately. How well learning support actually works, then, depends heavily on which category of tool a learner actually reaches for, and how they use it. General-purpose assistants are the most flexible and have the highest capability ceiling, but they also carry the fewest pedagogical constraints and content guardrails, putting them in the sharpest tension with academic integrity and cognitive-offloading concerns — which makes them the primary focus of student-facing governance (see Chapter 5). This also explains why, for minors specifically, China has leaned toward "learning tablets and practice devices with built-in educational constraints" over "open general-purpose conversational assistants" as the main delivery vehicle.

4.5 Personalization and Tutoring Effectiveness: The Range and Limits of the Evidence

This is a substantial new section in this edition. Given how widespread vendor claims of effectiveness have become, this section insists on separating evidence by strength: **(a) correlational evidence linking platform usage to outcomes; (b) evidence from rigorously designed randomized controlled trials (RCTs); (c) meta-analytic evidence pooling multiple studies; and (d) counter-evidence exposing potential risks.** These four categories differ sharply in strength and in what they can and can't support — together, they form the real picture of generative-AI tutoring effectiveness in 2024–2026.

4.5.1 Correlational Evidence: More Use, More Progress — But Not Necessarily Causal

Take Khan Academy as an example. Its study of roughly 350,000 students in grades 3–8 across the 2022–2023 school year found that students who used the platform at the recommended rate — at least 30 minutes a week, or roughly 18-plus hours over the full year — showed academic progress about 20% above what would otherwise be expected, an effect size of roughly 0.36 (as measured by MAP Growth). A related method was later reinforced by a study published in the *Proceedings of the National Academy of Sciences* (PNAS), which controlled for unobserved factors such as teacher and student characteristics. **Two boundaries need to be stated clearly here.** First, this effect is a "more use, more progress" correlation (bolstered by quasi-experimental controls), not a clean causal attribution to any single feature. Second — and this is the more important point — these gains largely trace back to Khan Academy's decade-plus of accumulated **practice problems and content**, not to its generative-AI component, Khanmigo. Attributing platform-level effects directly to Khanmigo is a common inference, but it doesn't hold up. As of this writing, no peer-reviewed causal-effectiveness study of Khanmigo itself has been published. A Khanmigo RCT run by J-PAL and the University of Toronto in Canadian grade 6–8 classrooms is reportedly expected to report results around mid-2026, and is worth watching.

4.5.2 Causal Evidence: Well-Designed AI Tutoring Can Significantly Outperform Good Classroom Teaching

The evidence that actually supports causal claims comes from a handful of small but carefully designed RCTs.

The Harvard physics-tutoring RCT (Kestin et al., published in *Scientific Reports*, June 2025) **is by far the most widely cited study to date. Using a crossover randomized design, 194 Harvard physics undergraduates received either "AI tutoring" or "high-quality active-learning classroom instruction." The homegrown AI tutoring system, PS2 Pal, was built to rigorously embed research-backed teaching principles: guided, step-by-step answering rather than direct answers; deliberate management of cognitive load in manageable chunks; growth-mindset language; personalized feedback targeted at a student's specific misconception; and richly engineered prompts designed to suppress hallucination. The results: students in the AI group saw learning gains more than double** those in the classroom group (post-test median of 4.5 versus 3.5), with a reported effect size of roughly 0.73–1.3 standard deviations. The AI group also finished faster (median 49 minutes versus 60), and reported higher engagement (4.1 out of 5.0 versus 3.6) and motivation (3.4 out of 5.0 versus 3.1). The finding has been held up as strong evidence that "AI tutoring can match or exceed a good classroom in less time."

That said, this study demands a very cautious reading. Limitations flagged by the authors and by commentators include: the intervention was short — a single two-week session — which can't rule out a novelty effect fading over time, and long-term retention wasn't measured at all; the sample was highly selective Harvard undergraduates, raising real doubts about whether the results generalize to community colleges, younger students, or groups with less access to technology; the study focused on the "understand-analyze" mid-tier of Bloom's taxonomy of cognitive skills, without touching higher-order thinking, collaboration, or social-emotional competencies; and the comparison group was "high-quality active-learning instruction," not mediocre teaching. In other words, what this study demonstrates is that "AI tutoring, when rigorously designed around sound pedagogical principles, can be highly effective in the short term" — not that "any AI tutoring beats classroom teaching." **The quality of the pedagogical design, not the AI itself, is the variable that determines the outcome.**

A UK secondary-school math RCT (December 2025, arXiv 2512.23633) offers a useful complement, one closer to a real classroom setting. This exploratory RCT spanned five UK secondary schools and 165 students, connecting LearnLM — a generative model fine-tuned specifically for teaching — to chat-based tutoring on the Eedi math platform. The study used a deliberately conservative design in which an expert teacher supervised the AI: the teacher could edit any message the AI proposed to send before it went out, and in practice roughly 76.4% of the AI's draft messages were sent with no edits or only minimal ones. Students who received LearnLM support solved subsequent new problems at a 66.2% correct rate, versus 60.7% for students tutored by a human teacher alone — a gap of roughly 5.5 percentage points. Teachers specifically praised the system's ability to generate Socratic questions that provoked deep reflection, with some reporting that they picked up new teaching techniques from it themselves.

Taken together, the causal evidence supports a qualified but real conclusion: **AI tutoring that has been pedagogically fine-tuned, and that follows a principle of guiding rather than directly answering, can deliver genuine and often substantial learning gains under controlled conditions.** Both RCTs point to the same underlying lesson: what actually drove the gap in outcomes wasn't "a more powerful model" — it was "better instructional design." Harvard's PS2 Pal owed its results to embedded pedagogical principles; the UK's LearnLM owed its results to teaching-specific fine-tuning plus teacher supervision. This reconfirms the judgment from Section 4.2.1: the quality of pedagogical design is the deciding variable behind effectiveness.

4.5.3 Meta-Analytic Evidence: A Significant Overall Effect, With Systematic Biases to Watch For

Beyond individual studies, meta-analyses offer a more robust — though still carefully qualified — overall picture. Ma et al. (2025, *Journal of Computer Assisted Learning*) pooled 34 (quasi-)experimental studies and reported a combined effect size of roughly 0.68 for generative AI's impact

on overall learning outcomes, with stronger effects in the cognitive dimension ($g \approx 0.80$) and skills dimension ($g \approx 0.71$), and a moderate effect in the affective dimension ($g \approx 0.51$). Subject area emerged as a significant moderating variable, while intervention duration, type of knowledge, and prior-knowledge level showed no significant moderating effect. Other meta-analyses point in the same direction — several reviews report medium-to-large effects of generative AI on academic achievement and learning motivation, with a positive lift on the motivation dimension as well.

On the question that matters most to educators — higher-order thinking — the evidence gets more nuanced. One meta-analysis pooling 29 experimental and quasi-experimental studies found that generative AI has an overall positive effect on higher-order thinking skills (critical thinking, creative thinking, problem-solving), while cautioning that "over-reliance on AI-generated content may weaken independent learning and self-regulation" — meaning positive and negative effects may well coexist, depending on how the tool is used. Especially worth policymakers' attention is an **inverted-U relationship between intervention duration and effect size**: one study found that interventions lasting roughly 8–16 weeks produced the best outcomes, while shorter interventions weren't enough to build capability and longer ones risked accumulating the negative consequences of over-reliance. This has a direct design implication: AI tutoring is not simply "the more, the better" — there appears to be an optimal dosage range that needs to be deliberately engineered into product and curriculum design.

Three caveats apply to this meta-analytic evidence. First, **publication bias**: positive results are more likely to get published, which may inflate the overall effect estimate. Second, **intervention-quality bias**: most studies feeding into these meta-analyses involve carefully designed, short-term interventions implemented directly by researchers — a far cry from the scattered, unsupervised way learners actually use these tools day to day. Extrapolating a lab-measured effect size (like 0.68) straight onto everyday usage is a common misreading. Third, **measurement scope**: most studies rely on short-term post-test scores of knowledge or skill, with far less measurement of long-term retention, transfer, or lasting effects on learning habits. Meta-analytic evidence, then, can support the conclusion that "well-designed generative AI has a positive effect on learning," but it cannot support the stronger claims that "any use is beneficial" or that "the effect holds up over the long term."

4.5.4 Counter-Evidence: Convenience May Come at the Cost of Learning Depth

Just as important as the causal evidence is the counter-evidence pointing to real risks. The Microsoft/CMU survey cited earlier (Section 4.2.5), on higher AI trust correlating with less critical thinking, already flags the risk at the mechanism level. On actual learning tasks, there's direct experimental evidence of a "convenience-for-depth" tradeoff. One experimental study on math-problem practice (arXiv 2605.21629, aptly titled "Faster Completion, Less Learning") found that when generative AI was available, students finished math problems significantly faster — but built up less knowledge in the process. In other words, AI improved "completion efficiency" without

necessarily improving, and possibly even harming, "learning outcomes." This lines up neatly with the "desirable difficulty" explanation from Section 4.2.5: when AI smooths over the retrieval and trial-and-error a learner should be going through themselves, the gain in efficiency comes directly at the expense of learning.

This counter-evidence has a real-world behavioral counterpart. Per the 2024 Common Sense Media survey cited earlier, about 70% of US teens have used at least one AI tool, with roughly four in ten using it for school-assignment tasks; related statistics also show that among college students who admit to using AI, the share using it specifically for assignments is very high. When "using AI to complete an assignment" slides from "aiding understanding" to "replacing thinking," the result is exactly the "faster completion, less learning" pattern the counter-evidence describes. The upshot: product form (guiding versus directly answering) and mode of use (process-oriented versus outcome-oriented) together determine whether the very same AI capability ends up helping or undermining learning — the technology itself is neutral on this question.

The evidence on language learning is more positive, though it, too, calls for caution. A pair of pre/post surveys of Duolingo Max users (385 learners total, studying French or Spanish) found that after a month of using Roleplay and Explain My Answer, learners' self-reported self-efficacy rose significantly. In the first study, a significant number of learners who had not previously agreed with the statement "I feel ready to use this language in a real-world setting" agreed with it at post-test. In the second study, six of seven self-efficacy measures showed significant gains; 86%–99% of learners felt the AI features meaningfully supported their learning, and 63%–73% reported applying what they'd learned in real-world situations outside the app. But the researchers themselves openly acknowledge the limitations: no control group, reliance on self-report, participant attrition, and no direct measurement of actual language proficiency. This evidence, then, can only support a claim about "improved confidence and subjective engagement," not one about "improved objective language ability" — self-efficacy and learning outcomes may correlate, but they are not the same thing. This case is a textbook illustration of why the evidence-grading approach at the start of Section 4.5 matters: satisfaction and self-efficacy surveys are the weakest of the three evidence tiers.

A methodological checklist for evaluating effectiveness claims. To evaluate any effectiveness claim about a student-facing AI tutoring product, ask six questions in sequence: (1) **Type of evidence** — is it correlational, an RCT, or just a satisfaction/self-efficacy survey? The three decrease in strength in that order. (2) **Source of evidence** — is it a vendor's own self-assessment, or independent/peer-reviewed research? (3) **Comparison condition** — is the comparison good teaching, ordinary teaching, or no intervention at all? A stronger comparison makes the conclusion more credible, but usually shrinks the measured effect. (4) **Measurement window** — is it an immediate post-test score, or retention and transfer measured weeks or a semester later? (5) **Duration and ecological validity** — was it a short, carefully controlled intervention run by researchers, or genuine day-to-day use by learners? (6) **Attribution boundary** — is the effect properly attributed to the

specific AI feature being evaluated, or vaguely credited to the whole platform? Mapped onto this chapter's own examples: Khan Academy's 0.36 is "correlational, and attributed to the platform rather than to Khanmigo specifically"; the Harvard and UK RCTs are "causal, but short-term and limited in sample"; and the Duolingo Max self-efficacy gains are "the weakest tier — satisfaction-based, with no measurement of objective ability." Vendor marketing heavy on phrases like "raises scores" or "highly efficient" deserves real scrutiny if it doesn't meet these conditions.

4.6 Recommendations

4.6.1 For Product Developers: Build Pedagogy Into the System

- **Turn scaffold fading into an actual product feature**: design interaction strategies around fostering independent learning, not around answering every question asked. Khanmigo's Socratic guidance, PS2 Pal's step-by-step-answering principle, and LearnLM's guided questioning have all been shown by evidence to be key drivers of effectiveness, and should be the default mode rather than an optional setting. Products should offer step-by-step hints, heuristic follow-up questions, and "try first, reveal later" patterns, and should treat "whether the learner is prompted to attempt the problem first" as a measurable product metric rather than an invisible internal policy.
- **Hold the reliability line with knowledge anchoring**: products aimed at minors should anchor responses to vetted textbook and solution knowledge bases via RAG or similar methods, reducing hallucination and supporting source traceability, and should clearly define a behavioral boundary — declining to answer, or explicitly flagging uncertainty, when a question falls outside what the knowledge base covers — rather than letting the model improvise when it lacks grounding.
- **Design memory and user modeling with care**: while building out process-adaptation capability (as with Xiaoyuan's "memory mode"), clearly define the boundaries of what memory is visible, editable, and deletable, protect learner privacy, and guard against a mistaken user model locking a learner into the wrong difficulty band (see Chapter 5). The stronger a product's personalization, the more it needs an explainable, correctable interface that lets learners and parents see how the system is "reading" them.
- **Design for "optimal dosage," not "maximum time on task"**: since meta-analytic evidence suggests intervention duration and effectiveness may follow an inverted-U curve, products should not treat average daily usage time as their sole growth metric — they should instead balance learning outcomes against dependence risk, building in usage-pacing features and "step away" reminders where appropriate.

- **Drive iteration with real, properly graded evidence**: subject-matter accuracy, hallucination rate, and pedagogical appropriateness should be measured through third-party or publicly auditable evaluation. When disclosing effectiveness claims externally, clearly distinguish correlational, RCT, and satisfaction-survey evidence, and disclose intervention duration, comparison group, and measurement approach — rather than letting platform-wide or short-term experimental effects be mistaken for a product's own causal effectiveness. Organizations with the resources to do so should invest in, or support, independent RCTs, building genuine differentiation through real causal evidence. See Chapter 7 of this Blue Book for relevant methodology.

4.6.2 For Schools and Teachers: Protect Learner Agency Within Human–AI Collaboration

- **Set clear boundaries and norms for human–AI collaboration**: be explicit about which parts of the learning process are suitable for AI support and which must be completed independently by the learner, to guard against cognitive offloading eroding deep learning. The "teacher supervises AI output" model used in the UK LearnLM RCT offers a workable, conservative template for responsibly introducing AI tutoring into the classroom.
- **Build learners' AI literacy**: fold "how to ask good questions, how to question the answer, how to verify it" into everyday teaching. Research suggests that people confident in their own ability are more likely to sustain critical thinking, and that education level buffers against cognitive offloading — meaning AI-literacy education is itself a protective factor against over-reliance, and should aim to make learners masters of the tool rather than dependents on it.

4.6.3 For Learners and Parents: From "Get the Answer" to "Grow"

- Favor process-oriented use over outcome-oriented use — put AI to work checking your work, practicing with you, and helping you review, rather than asking it for the answer outright. Be especially wary of high-risk, convenience-for-depth uses such as photographing a problem for the solution or having a general-purpose assistant write an essay for you.
- Pay attention to eye health, screen time, and content compliance on on-device hardware. Given the feature homogeneity and pricing controversies in China's AI-learning-tablet market, parents shopping for one should look past marketing phrases like "AI one-on-one tutoring" and "hundred-million-item question bank" and focus on whether a product genuinely offers guided explanation rather than direct answers, cross-checking with third-party evaluations where possible. Families should also consider setting explicit house rules on when and for what purpose these devices get used.

4.6.4 For Policy and Governance: Set a Safety Floor for Learning Support

Generative products aimed at minors should face more careful governance around content safety, data minimization, algorithmic transparency, and anti-addiction measures. China's registration system for education LLMs — under which TAL's "Jiuzhang," Zuoyebang's "Yinhe," and Yuanfudao's relevant models have all completed registration — sets a baseline for market entry, but registration is focused on compliance at the entry point; it is not equivalent to ongoing evaluation of teaching effectiveness or child cognitive safety. The latter requires independent, longitudinal evaluation mechanisms as a complement. For the specifics of governance frameworks, standards, and compliance requirements, see Chapter 5, "Governance and Safety," as well as our institute's related standardization research.

4.7 Summary

By 2026, "Supporting Learning" is undergoing a profound shift — from conversational Q&A toward an agentic, multimodal, on-device learning companion. Its underlying mechanism has expanded from single-turn knowledge retrieval to modeling the whole learning process, with memory, adaptivity, and closeness to embodied context. Its product form has splintered from the single chat window into Socratic tutoring apps, spoken-language practice, photo-based problem solving, AI learning tablets and practice devices, and education-vertical LLMs.

On effectiveness, this chapter's central judgment is this: **the evidence supports the claim that "well-designed AI tutoring can meaningfully advance learning," but the deciding variable is the quality of pedagogical design, not the AI technology itself.** The Harvard and UK RCTs demonstrate the value of guided, pedagogically fine-tuned tutoring; the "faster completion, less learning" counter-evidence and the research on cognitive offloading are a reminder that every leap in technical capability carries a matching tension around eroding learner agency. The Khan Academy case is its own warning: a platform-level correlational effect cannot simply be credited to its AI component, and adoption scale is not the same thing as causal effectiveness.

In the Chinese market, student-facing deployment has mostly taken the compliant path of "AI learning tablets/practice devices plus education-vertical LLMs": in 2024, learning tablets sold roughly 5.923 million units across all channels for about RMB 19.06 billion in revenue, with growth continuing into 2025; TAL, Zuoyebang, Yuanfudao, and iFLYTEK are driving a closed loop of "photograph the problem, explain it step by step, diagnose learning gaps" using their own registered education LLMs. This path is distinguished by content compliance and parental controls, but it also faces shared challenges — feature homogeneity, pricing controversy, and a persistent shortfall of solid effectiveness evidence. Real differentiation, in the end, will come down to a single pedagogical question: does the product genuinely guide learners toward growth?

This chapter's conclusion, then, is this: the value of a learning-support product is ultimately not measured by how much it can do for the learner, but by what kind of independent learner it helps that learner become. Building scaffold fading into the system, holding the line on knowledge anchoring, grading evidence honestly, and treating protection of learner agency as a hard boundary — these are the preconditions for this transformation to proceed steadily and sustainably. This judgment also sets up the two chapters that follow, on "Intelligent Assessment" and "Governance and Safety."

Chapter References

1. Kestin, G., et al. "AI tutoring outperforms in-class active learning: a randomized controlled trial." (Harvard physics-tutoring RCT) · Scientific Reports · 2025 · https://www.nature.com/articles/s41598-025-97652-6 ; review coverage at Educational Technology and Change Journal, 2025-11 · https://etcjournal.com/2025/11/10/review-of-kestin-et-al-s-june-2025-harvard-study-on-ai-tutoring/
2. "AI tutoring can safely and effectively support students: An exploratory RCT in UK classrooms" (LearnLM / Eedi UK secondary-school math RCT) · arXiv:2512.23633 · 2025 · https://arxiv.org/abs/2512.23633
3. Ma, X., et al. "A Meta-Analysis of the Impact of Generative Artificial Intelligence on Learning Outcomes." · Journal of Computer Assisted Learning, 41(5): e70117 · 2025 · https://onlinelibrary.wiley.com/doi/10.1111/jcal.70117
4. Khan Academy. "Khan Academy Efficacy Results, November 2024" (approx. 350,000 students, effect size 0.36) · 2024 · https://blog.khanacademy.org/khan-academy-efficacy-results-november-2024/
5. "Computer-assisted learning in the real world: How Khan Academy influences student math learning" · PNAS · 2025 · https://www.pnas.org/doi/10.1073/pnas.2507708123
6. CNBC. "Microsoft, Khan Academy provide free AI assistant for all educators in US" (Khanmigo free access and adoption scale) · 2024-05 · https://www.cnbc.com/2024/05/21/microsoft-khan-academy-launch-free-ai-assistant-for-all-us-teachers.html
7. GovTech. "Microsoft Deal Makes Khan Academy's AI Assistant Free for Teachers" (district-level pricing, 795 districts/770,000 students) · 2024 · https://www.govtech.com/education/k-12/microsoft-deal-makes-khan-academys-ai-assistant-free-for-teachers
8. Duolingo. "Duolingo Max Uses OpenAI's GPT-4 For New Learning Features" (Roleplay/Video Call/Explain My Answer) · Duolingo Blog · 2023 · https://blog.duolingo.com/duolingo-max/

9. Duolingo. "Duolingo Launches 148 New Language Courses" (generative AI at-scale course production) · Investor Relations · 2025 · https://investors.duolingo.com/news-releases/news-release-details/duolingo-launches-148-new-language-courses

10. TechCrunch. "The backlash against Duolingo going 'AI-first' didn't even matter" · 2025-08 · https://techcrunch.com/2025/08/07/the-backlash-against-duolingo-going-ai-first-didnt-even-matter/

11. Frontiers in Education. "Mobile language app learners' self-efficacy increases after using generative AI" (Duolingo Max self-efficacy pre/post surveys, 385 participants) · 2025 · https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1499497/full

12. Wikipedia. "Photomath" (Google acquisition and features) · 2024 · https://en.wikipedia.org/wiki/Photomath ; also Sammy Fans, 2024-03 · https://www.sammyfans.com/2024/03/01/google-acquires-photomath-a-popular-math-solver-app/

13. Microsoft Research. "The Impact of Generative AI on Critical Thinking / Tools for Thought at CHI 2025" (AI trust and critical thinking, 319 knowledge workers) · 2025 · https://www.microsoft.com/en-us/research/blog/the-future-of-ai-in-knowledge-work-tools-for-thought-at-chi-2025/

14. "Faster Completion, Less Learning: Generative AI Reduced Study Time on Math Problems and the Knowledge They Build" · arXiv:2605.21629 · https://arxiv.org/pdf/2605.21629

15. Qbitai (量子位). "Q2 Learning-Tablet Shipments Up 46%! IDC: iFLYTEK AI Learning Tablet Tops Market in Sales Revenue" [Q2 学习机出货量增 46%！IDC：科大讯飞 AI 学习机登顶市场销售额第一] · 2025-09 · https://www.qbitai.com/2025/09/328513.html

16. Zhiyan Consulting (智研咨询). "2025 China AI Learning Tablet Industry: Supply Chain, Sales Volume, Revenue, Competitive Landscape and Outlook" [2025 年中国 AI 学习机行业产业链、销量、销售额、竞争格局及前景展望] (RUNTO/洛图 data, 2024: 5.923 million units / RMB 19.06 billion) · 2025 · https://www.chyxx.com/industry/1224843.html

17. Sina Finance (新浪财经). "Learning Tablets Compete on AI — Who's Winning?" [学习机拼 AI，谁是赢家？] (Zuoyebang/TAL/Yuanfudao market share) · 2025-07 · https://finance.sina.com.cn/tech/roll/2025-07-01/doc-infcyazn3425896.shtml

18. GeekPark (极客公园). "TAL Showcases Jiuzhang LLM and Learning Tablet at the World Artificial Intelligence Conference" [学而思携九章大模型、学而思学习机亮相世界人工智能大会] (Jiuzhang/MathGPT, xPad) · https://www.geekpark.net/news/337563 ; also Tencent News (腾讯新闻). "2024 CIFTIS: Learning Tablets Boosted by LLMs" [2024 服贸会｜大模型加持学习机] · 2024-09 · https://news.qq.com/rain/a/20240913A00E1O00

19. China Daily (中国日报网). "Fully Embracing the AI Era: Yuanfudao Education Products Deeply Integrate DeepSeek and the Yuanli LLM" [全面拥抱 AI 时代 小猿教育产品深度融合 DeepSeek 和猿力大模型] (Xiaoyuan memory mode, mental-health companion feature, registration) · 2025-02 · https://bj.chinadaily.com.cn/a/202502/19/WS67b5d47ba310510f19ee7ea5.html

20. Southern Metropolis Daily (南方都市报). "What Do Vertical Education-AI Products Actually Compete On?" [垂类教育 AI 比拼什么？猿辅导王向东：瞄准三重 AI 布局] and "TAL CTO Tian Mi: As General-Purpose Model Technology Races Ahead, Vertical Models Must Know the Industry Better" [好未来 CTO 田密：通用模型技术狂飙，垂直模型必须更懂行业] (approaches to education-vertical LLMs) · 2025 · https://m.mp.oeeee.com/a/BAAFRD0000202504161069982.html

21. Common Sense Media / Pew Research (2024 survey on teen generative-AI usage and academic-integrity attitudes) · cited via Nerdynav, "ChatGPT Cheating Statistics (2025)" · https://nerdynav.com/chatgpt-cheating-statistics/

22. TechCrunch. "The backlash against Duolingo going 'AI-first' didn't even matter" (Duolingo's "AI-first" strategy and public backlash) · 2025-08 · https://techcrunch.com/2025/08/07/the-backlash-against-duolingo-going-ai-first-didnt-even-matter/

23. "Exploring the use of retrieval-augmented generation models in higher education: A pilot study on AI-based tutoring" (RAG suppressing hallucination, anchoring to teacher-vetted material) · ScienceDirect · 2025 · https://www.sciencedirect.com/science/article/pii/S2590291125004796

24. "Does Generative Artificial Intelligence Improve Students' Higher-Order Thinking? A Meta-Analysis Based on 29 Experiments and Quasi-Experiments" (positive effect on higher-order thinking, over-reliance caution, inverted-U intervention duration) · MDPI Journal of Intelligence, 13(12):160 · 2025 · https://www.mdpi.com/2079-3200/13/12/160

25. "A systematic review and meta-analysis of the effectiveness of Generative Artificial Intelligence (GenAI) on students' motivation and engagement" (meta-analysis of motivation and engagement) · ScienceDirect · 2025 · https://www.sciencedirect.com/science/article/pii/S2666920X25000955

26. Sciendo/etc: "Learners' AI dependence and critical thinking: the psychological mechanism of fatigue and the social buffering role of AI literacy" (AI dependence, critical thinking, and the buffering role of AI literacy) · ScienceDirect · 2025 · https://www.sciencedirect.com/science/article/pii/S0001691825010388

Chapter 5 Supporting Research (Professional Development): Agentic Restructuring, Product Forms, and Recommendations Across the PD Cycle

5.1 Redefining the PD Scenario: From "Lesson-Prep Assistance" to the "Research Agent"

Educational research and professional development (jiaoyan, 教研) is the hub connecting curriculum standards, classroom implementation, and teacher growth. Within China's education governance structure, it spans everything from school-level collective lesson planning, peer observation and evaluation, lesson-case studies, item and assignment design, and school-based curriculum building, to district- and province-level regional PD, evidence-based guidance from subject PD specialists, and PD data governance. Where "Empowering Teaching" (Chapter 3) serves classroom delivery directly and "Supporting Learning" (Chapter 4) serves individual students directly, the core object of PD is **the collective professional practice of teachers** — and the accumulation, transfer, and reproduction of instructional knowledge. This is what determines whether front-line classrooms keep improving, and whether good teaching practices can spread across the boundaries of individual, school, and region.

Generative AI's entry into PD didn't begin in 2026, but its form has shifted substantially over the past two years. In the earlier, conversational-LLM phase, "supporting PD" mostly meant using general-purpose models to help teachers with discrete tasks — drafting lesson plans, searching for materials, drafting test items — essentially a stronger search-and-writing tool. The *Blue Book on AI-Empowered Applications in Basic Education (2025)*, published in July 2025 by the Ministry of Education's Engineering Research Center for Intelligent Technology and Educational Applications together with the Beijing Digital Education Center, characterized this stage as "using intelligence to aid research," noting its value in "integrating teaching and research data to support precise analysis and scientific decision-making, and easing teachers' burden in tasks such as organizing materials and evaluating outcomes," and judged that the teacher's role was shifting "from experience-driven to data-driven." That judgment got the direction right, but the emphasis still fell on task assistance.

By 2026, as agents, multimodal understanding, and on-device compute matured along three parallel tracks, PD support has been evolving from a single-purpose Q&A tool into what might be called a Research-oriented Teaching Agent spanning the entire PD cycle. Lu Yu and Tang Xiaoyu of Beijing Normal University, writing in *e-Education Research* (Issue 6, 2025), proposed a "generative AI

empowering classroom teaching" framework built on "four progressively advancing, interconnected tiers": a foundational tier of "labor substitution and task assistance," a beginner tier of "capability enhancement and boundary expansion," an intermediate tier of "human-machine collaboration and innovation activation," and an advanced tier of "cognitive integration and mindset shaping." Although that framework was built for classroom teaching, its underlying logic — moving from tool-based empowerment to agentic collaboration — applies equally well to PD. A research agent, on this view, is distinguished by three things:

- **A shift from reactive answering to proactive orchestration.** Agents can decompose PD tasks on their own against a stated goal, invoke tools (retrieval, grading, visualization, classroom-data analysis), and self-check and iterate on intermediate outputs — the same capability profile Lu and Tang describe for "teaching agents with a high degree of autonomy."
- **A shift from single-modality text to multimodal evidence.** Classroom video, whiteboard writing, student work, and spoken interaction are now material that can be understood and structured for PD purposes. This is the single biggest capability gain in 2026 relative to earlier reports, and it's the technical precondition for moving peer observation "from subjective impression to evidence-based diagnosis."
- **A shift from generic knowledge to traceable domain memory.** Retrieval-Augmented Generation (RAG) and long-term memory turn school-based teaching resources, curriculum standards, and PD history into a reusable, citable knowledge base. The whole point of RAG is to "anchor the generation process in retrieved documents, so the model answers on the basis of real evidence, with results that are more traceable."

It's worth stressing that this shift is not just a technology narrative — it has entered top-level national policy. In April 2026, five ministries — the Ministry of Education, the National Development and Reform Commission, the Ministry of Industry and Information Technology, the Ministry of Science and Technology, and the National Data Administration — jointly issued the *"AI+ Education" Action Plan* (Jiaokexin [2026] No. 1), which explicitly calls for "using intelligent technology to analyze classroom teaching behavior, carrying out evidence-based PD practice enabled by AI, and building a teacher-training model suited to the intelligent era." On the governance side, it also calls for "advancing applications such as intelligent item generation, intelligent test assembly, intelligent proctoring, and intelligent grading," while requiring the "development of intelligent tools for education evaluation, exploring longitudinal evaluation of students' entire learning process and horizontal evaluation across moral, intellectual, physical, aesthetic, and labor dimensions." "Evidence-based PD" has thereby graduated from an academic concept to a policy term — giving the product forms discussed in this chapter an institutional anchor.

To be clear: the "research agent" discussed in this chapter is not meant to replace the collective wisdom of a PD group with an automated system, but to embed AI capability — steerable by

teachers — into every stage of the PD process. PD is a high-value scenario for generative AI precisely because of three characteristics. First, **task density is high and much of the work is repetitive** — teachers spend enormous amounts of time on lesson prep, finding materials, writing test items, drafting observation notes, and compiling PD briefs, all of which are exactly the kind of "drafting" work generative AI is good at. Second, **evidence that could be structured has instead stayed impressionistic** — judgments like "this class had good interaction" or "the questions skewed shallow" have long lacked quantifiable, comparable backing, and multimodal understanding is exactly what turns the classroom scene into a statistical evidence stream. Third, **knowledge that could be preserved instead walks out the door with people** — the tacit expertise of excellent teachers is hard to make explicit and hard to transfer across schools, and RAG plus long-term memory offer a new technical vehicle for "capitalizing" school-based PD assets. These three characteristics are what make PD one of the few scenarios where all four capability layers — generation, multimodal understanding, retrieval/memory, and agentic orchestration — can operate simultaneously and form a genuine closed loop.

> Note: This chapter focuses on the productization mechanisms and forms of generative AI in PD. Hardware-side, first-person classroom capture capability (such as AI glasses used in peer observation) is covered in this institute's *2026 Blue Book on the AI Glasses Education Industry*.

5.2 The Empowerment Mechanism: How the Four Capability Layers Act Across the PD Cycle

5.2.1 Decomposing PD Work into Structured Tasks

PD work can be broken down along an "input–processing–output–feedback" chain into subtasks that AI can plausibly assist with. The table below maps typical PD stages to the corresponding AI capabilities. The maturity column is a qualitative judgment, grounded in the product evidence gathered for this chapter and in authoritative reports (such as the discussion of intelligent lesson-prep, item generation, and test assembly maturity in the *Blue Book on AI-Empowered Applications in Basic Education (2025)*); quantitative evaluation criteria are discussed separately in Section 5.4.

PD Stage	Core Task	Primary AI Capability	2026 Typical Paradigm	Maturity (Qualitative)
Collective lesson planning	Aligning teaching objectives, generating lesson plans, tiered design	Generation, structuring, knowledge alignment	RAG + curriculum-standards knowledge base	Fairly mature (resource retrieval/content generation already at scale)
Classroom observation	Capturing classroom process, tagging	Multimodal understanding, speech	On-device capture + event extraction	Rapidly maturing (hardware + LLM

	behavior	transcription		integration already piloted in demonstration zones)
Post-observation review	Synthesizing evidence, dimension-based diagnosis	Synthesis, visualization, rubric mapping	Classroom-analysis agent	In active use (multi-metric reports already generated; standards still unsettled)
Lesson-case study	Iterative improvement, accumulating experience	Long-term memory, comparative analysis	PD memory repository	Early stage (cross-cycle memory mechanisms still immature)
Item and assignment design	Generating test items, difficulty/reliability reference	Generation, constraint satisfaction	Domain-vertical model	In active use (generation is efficient, but quality needs teacher verification)
Regional PD	Aggregating resources, PD data governance	Retrieval, aggregation, de-identification	PD platform + agentic orchestration	Early stage (governance and compliance demands are high; standardization is lacking)

5.2.2 Four Technical Layers Underpinning PD

· **Generation and restructuring layer.** Education-vertical or general-purpose LLMs draft and restyle lesson plans, rubrics, test items, and PD reports. The mechanism is translating a teacher's tacit expertise into a structured product via natural-language instructions, cutting the cognitive load of drafting from scratch. The *Blue Book on AI-Empowered Applications in Basic Education (2025)* notes that this kind of "instructional content generation" can "automatically generate content matching instructional elements — such as test items, lesson plans, and courseware — based on basic information the teacher inputs, like topic, grade, and subject," calling it key to "easing teachers' lesson-prep burden." NetEase Youdao's "Ziyue" education LLM illustrates the point: its Ziyue 3 math model, open-sourced in August 2025, claims to "cover high-frequency needs across all subjects, enabling end-to-end, multi-role empowerment across lesson prep, item generation, grading, and Q&A" — pushing generation capability across the entire upstream PD workflow.

· **Multimodal understanding layer.** This layer recognizes, transcribes, and semantically tags classroom video, whiteboard images, student work, and teacher–student dialogue, turning the unstructured classroom scene into searchable, statistically tractable PD evidence. This is the

single largest capability gain relative to earlier reports. Seewo's classroom intelligent-feedback system, for instance, "captures multidimensional data in real time — classroom teaching behavior, teacher–student interaction, and more — via smart devices," generating a learning-profile and a visual diagnostic report that moves peer observation from subjective impression toward discussable evidence.

- **Retrieval-augmentation and memory layer.** RAG brings school-based resources, curriculum standards, and PD history into the generation process so outputs are "grounded in evidence"; long-term memory carries teacher preferences, class-specific learning profiles, and prior improvement items across PD cycles. Several 2025 RAG surveys describe the technique as "enhancing the accuracy and trustworthiness of generation by retrieving relevant knowledge from external databases," and note that it can "incorporate new knowledge simply by updating the knowledge base, enable rapid iteration, and strengthen the traceability of results" — three properties that map directly onto PD's hard requirements for traceability, updatability, and auditability. For PD specifically, RAG matters in three ways: first, it **grounds** the model — connecting a general-purpose model to local curriculum standards and school-based resources solves the "generic but not locally grounded" problem; second, it makes output **traceable** — every generated statement can point back to a specific document, so PD conclusions can be audited; third, it makes the system **updatable** — when textbooks are revised or curriculum standards change, only the knowledge-base documents need to be swapped, with no retraining required. Long-term memory then upgrades a single session into a cross-cycle asset: a teacher's preferences for tiered instruction, a class's learning-profile baseline, and the improvement items from the last round of lesson-case study can all be retained and reused in the next cycle. Lu and Tang likewise note that current teaching agents have "achieved an initial capability in multimodal perception, retrieval-augmented generation, reasoning, and planning," but that their "ability to precisely parse the attributes, relationships, and semantic information of teaching resources still needs improvement" — a reminder that how well RAG performs in PD depends heavily on the structural quality and annotation granularity of the school-based knowledge base itself.
- **Agentic orchestration layer.** Using a plan–execute–reflect loop, this layer organizes the capabilities above into a PD workflow that can advance automatically, handing off to teachers for review at key checkpoints (human-in-the-loop). The industry commonly describes education agents as systems "built around an LLM as the core reasoning engine, combined with RAG for processing an education knowledge base, and supporting multimodal interaction across text, speech, and images," emphasizing their ability to "track learning trajectories, understand individual differences, and continuously optimize strategy based on historical data." Take a complete lesson-case study as an example: the orchestration layer can organize "settling on a research topic → capturing and transcribing classroom video → extracting diagnostic

dimensions against a rubric → comparing with historical lesson cases → generating improvement recommendations → writing back to the PD memory repository" into a pipeline that advances automatically but stays controllable at each node — the agent handles "planning tasks, invoking tools, synthesizing evidence, and drafting," while the teacher handles "confirming the topic, reviewing the diagnosis, and finalizing recommendations." It's important to stay clear-eyed here: Lu and Tang explicitly judge that "the construction of LLM-based teaching agents is still at an early stage," with existing agents needing improvement in their "ability to precisely parse the attributes, relationships, and semantic information of teaching resources," their "precision in recognizing behavioral patterns, linguistic features, and latent intent in teaching subjects," and their "depth of understanding of interactive behaviors, activity design, and goal attainment across the teaching process." These three gaps are exactly where the orchestration layer is most prone to overstepping and drawing conclusions beyond its warrant in PD contexts — so deployment should follow a conservative path: start with high-value, low-risk stages, and keep mandatory human review at critical checkpoints.

5.2.3 Maturity by Form: An Uneven Distribution of the Four Layers Across the PD Cycle

Mapping Lu and Tang's four-tier framework onto the full PD cycle reveals a clear maturity gradient. In **lesson prep and resource retrieval**, capability has already advanced to the beginner tier of "capability enhancement and boundary expansion" — generation and retrieval are deployed at scale, and the *Blue Book on AI-Empowered Applications in Basic Education (2025)* describes this as cracking the "difficulty of lesson prep" for front-line teachers, by modeling "a teacher's lesson-prep needs, teaching objectives, teaching context, and student learning profile" and "dynamically ranking to form a resource-recommendation list matched to the teacher's lesson-prep needs." In **peer observation and classroom analysis**, capability sits right at the threshold of "human-machine collaboration" — multimodal capture and diagnostic reports are already generated at scale in demonstration zones, but the "evidence-to-conclusion" reasoning still depends heavily on preset rubrics and teacher interpretation. In **lesson-case study and regional PD governance**, capability remains stuck between "task assistance" and "capability enhancement" — cross-cycle long-term memory, de-identified aggregation of multi-school data, and autonomous identification of common problems have not yet coalesced into a stable, reliable product loop. This gradient suggests that the research agent's value will be realized **stage by stage, in phases** — attempting to build a fully automated "PD brain" in one leap is neither realistic nor sound.

5.2.4 The Boundaries and Preconditions of the Mechanism

It bears emphasizing that the mechanism above only holds under three preconditions: **teacher-led, evidence-traceable, and privacy-compliant.** Generative AI's role in PD is to "amplify professional

judgment," not "replace professional judgment." This positioning aligns exactly with Lu and Tang's call to "establish a principle and code of conduct of 'human-machine collaboration' rather than 'machine replacement,' giving full play to [the teacher's] own leading role," and echoes the bottom line set out in the *"AI+ Education" Action Plan* to "effectively guard against problems such as AI-enabled fraud, academic misconduct, exam-oriented over-optimization, and privacy leaks."

On the boundary question, at least three preconditions are non-negotiable. **First, teacher leadership cannot be ceded.** Generated diagnostic conclusions, item difficulty levels, and improvement recommendations all require teacher review — the *Blue Book on AI-Empowered Applications in Basic Education (2025)* states plainly that intelligent item generation may produce "items, explanations, and answers [that] may contain errors, requiring teacher verification before use in actual instruction." The professionalism of PD lies precisely in value judgments about "why teach it this way" and "where exactly this lesson went wrong" — a layer that cannot be outsourced to a model. **Second, the evidence chain cannot be broken.** PD's persuasive power comes from conclusions being traceable to evidence. A system that offers only the conclusion "this lesson relied mainly on closed-ended questions," with no way to locate the specific timestamp and dialogue segment behind it, produces a conclusion that is neither easy to challenge nor easy to trust — which is exactly the point of RAG and evidence anchoring. **Third, privacy compliance cannot be a fix applied after the fact.** The classroom multimodal data these systems draw on touches the privacy of teachers and students (especially minors), and must follow data-minimization and de-identification principles at the point of capture — not after data has already been aggregated (see Chapter 7, "Governance and Safety," for the relevant governance requirements). Together, these three preconditions form the foundation that makes a research agent usable, trustworthy, and sustainable.

5.3 A Map of Product Forms: From Chat Tools to Research Agents

Products supporting PD in 2026 fall into four broad forms, showing an overall shift from "general-purpose conversation" toward "domain-specific agents." A note on method: the products below are described strictly from vendor-disclosed information and credible media reporting; this section states only disclosed capabilities and scale, and makes no claims about unverified market share or benchmark rankings.

5.3.1 Conversational PD Assistants (General-Purpose Foundation)

These are chat-style assistants for teachers, built on general-purpose LLMs, covering high-frequency discrete tasks like lesson-plan generation, material search, Q&A, and copy polishing. Their strength is zero barrier to entry and broad coverage; their limitation is a lack of school-specific context and classroom evidence.

Internationally, Khan Academy's Khanmigo offers teachers free generation of "standards-aligned lesson design, learning objectives and rubrics, exit tickets, and differentiated instruction and quiz items," with Khan Academy stating that "what used to take hours — differentiated designs, lesson plans, quiz questions, student groupings, warm-up activities — can now be done in minutes." Its "Lesson Planner" is guided by "a research-based structure created by Khan Academy educators," and its "Create a Hook" feature offers contextualized options such as stories, activities, poems, and examples, and can also draft multilingual-family emails and class newsletters. Khanmigo is worth treating as a benchmark not because of technical leadership, but because its product philosophy — freeing teachers from transactional work so they can focus on instructional design — is executed down to the level of being free, open, and standards-aligned.

Domestically, teacher assistants built on general-purpose LLMs (such as iFlytek Spark, Doubao, ERNIE Bot, Tongyi Qianwen, and DeepSeek) fall into the same category, characterized by strong general capability, fast iteration, and low cost — teachers can get lesson-plan frameworks, draft test items, and PD-brief drafts directly through natural-language prompts. Their shared weakness: **alignment with national curriculum standards and school-based resources depends heavily on prompt engineering and bolt-on knowledge bases** — a general-purpose model has no inherent knowledge of, say, a specific region's textbook edition's unit pacing or a specific class's learning-profile baseline, and without access to local resources its output tends to be "generic but not locally grounded." This is exactly why the product form naturally evolves toward the next category — domain-vertical models paired with a PD platform.

5.3.2 Education-Vertical LLMs and PD Platforms

These are education-vertical LLMs trained or fine-tuned for K-12 and teacher professional development, emphasizing curriculum-standards alignment, subject-matter accuracy, and PD compliance, typically deployed as a "PD platform" at the regional or school level, integrated with resource repositories and PD management systems. Representative products include:

- **iFlytek Spark Education LLM.** Iterated through 2025 to versions such as iFlytek Spark X1.5, positioned around "understanding education." Its "Spark Teacher Assistant" is designed to "provide suggested curriculum outlines and lesson plans, helping teachers reduce time spent searching for materials, organizing information, and producing teaching aids," and is deployed on hardware like the iFlytek AI Blackboard to deliver "end-to-end intelligence across lesson prep, instruction, and assessment." In September 2025, Huawei and iFlytek jointly released the "Spark Education and Healthcare LLM Scenario All-in-One Solution," combining Ascend compute with the Spark LLM to point toward localized, controllable deployment.
- **NetEase Youdao's "Ziyue" education LLM.** On August 20, 2025, Youdao released multiple new AI products and proposed an L1–L5 grading scale for education-AI application capability

— Xinhua reported that the industry is currently "accelerating from L3 active learning tutoring toward L4 virtual teacher (already possessing near-human-teacher thinking ability)." This L1–L5 grading echoes the four-tier framework this chapter borrows elsewhere, both attempts to give a common coordinate system to the question "how far has education AI actually gotten." On upstream PD capability, its Ziyue 3 math model, open-sourced in 2025, claims to "cover high-frequency needs across all subjects, enabling end-to-end, multi-role empowerment across lesson prep, item generation, grading, and Q&A"; the Ziyue 3.0 small-language-model variant supports real-time mutual translation across 38 languages. On the compliance dimension, the Ziyue education LLM passed the China Academy of Information and Communications Technology (CAICT)'s trustworthy-AI education-LLM evaluation, earning the top rating available at the time (Level 5) — one of the few education-vertical models to enter third-party compliance review.

- **TAL Education's "Chapter Nine" model (MathGPT).** Positioned for global math enthusiasts and research institutions, with problem-solving and explanation algorithms at its core, MathGPT was among the first domestic education LLMs to complete government filing. TAL describes it as a tool supporting teachers in "designing curricula, assigning homework, and conducting student assessment," trained on TAL's years of accumulated PD and teaching data along with user data. Math's demanding requirements for derivational rigor and solution-set correctness make MathGPT a useful test case for whether deep vertical specialization actually pays off in item and explanation quality — an area where general-purpose models have a well-known weakness on complex STEM problems.

The compliance foundation for vertical models is also filling in: by the end of 2024, roughly 302 generative AI services had completed filing with the Cyberspace Administration of China (CAC) (with roughly 238 newly added that year); by 2025, the number of filings continued to grow (per CAC announcements, the cumulative count of filed generative AI services expanded further), with education LLMs an active category within it. Compliance filing is a prerequisite gate for any product entering K-12 schools, and it is the first thing this chapter checks before describing any education LLM. One caveat: parameter scale, training-corpus composition, and third-party evaluation methodology for each product should be taken from vendor and evaluator public disclosures only; where such information is not public, this chapter draws no inference, and vendor-reported figures for "accuracy" or "coverage" are never treated as equivalent to third-party conclusions.

5.3.3 Multimodal Classroom Analysis and Peer-Observation Products

These products take classroom video and teacher–student interaction as input and output structured diagnostics — the ratio of teacher-to-student talk, the cognitive level of questions asked, the sequencing of instructional segments — serving peer observation and lesson-case study. This

category has evolved rapidly in 2026 thanks to gains in on-device compute and multimodal models, and it is, in this chapter's assessment, the most solidly deployed category of PD-support product.

Seewo's classroom intelligent-feedback system is representative, built on a "hardware plus software" architecture: an AI camera (reportedly a four-lens setup) captures classroom audio and video unobtrusively, and a teaching LLM performs in-depth data analysis, generating a "precise learning-profile" and a visual diagnostic report. On the teacher side, the system precisely tallies "questioning, follow-up questions, and guidance," and "offers personalized questioning suggestions based on individual teacher differences"; on the student side, it tracks head-up rate, hand-raising rate, and engagement, "objectively presenting the level of classroom interaction and participation." This diagnostic-dimension system essentially hands the manual work of classic classroom-observation coding schemes, such as the Flanders Interaction Analysis System (FIAS), over to a multimodal model to perform automatically — compressing classroom analysis that once took a PD specialist hours of frame-by-frame annotation into a report generated automatically after class.

At scale: per Seewo's disclosure in January 2026, as of December 31, 2025, the system had been deployed across 19 key demonstration zones, covering more than 5,600 schools and over 17,000 classrooms, with more than 97,000 teachers having generated a cumulative total of over 650,000 feedback reports — roughly double the figures Seewo disclosed in June 2025 (over 3,000 schools, over 7,000 classrooms, roughly 360,000 reports), a near-doubling within six months. This growth rate is itself indirect evidence that multimodal classroom analysis is currently the category of PD AI with the most rigid demand and the clearest repeat usage. Related practice was also selected for the Ministry of Education's Center for Educational Technology and Resource Development (China Educational Television and Audio-visual Center) 2025 list of teacher AI-application cases (see Scenario 2 in Section 5.5).

It should be noted that while this category's diagnostic dimensions are already fairly rich (some deployments in Zhejiang claim "200-plus data metrics per lesson"), their value depends heavily on two things: first, **alignment between metrics and PD goals** — a statistic like "62% of questions were closed-ended" only matters if it points toward an actionable recommendation like "increase the share of open-ended questions"; second, **traceability of evidence** — a diagnostic conclusion needs to be locatable to a specific timestamp and dialogue segment to support evidence-based review. First-person capture can complement this fixed-camera approach with AI glasses, supplying supplementary evidence for the teacher's own viewpoint, whiteboard detail, and one-on-one tutoring moments that fixed cameras struggle to cover (see this institute's *2026 Blue Book on the AI Glasses Education Industry* for detail).

5.3.4 Research Agents and Orchestration Platforms

This new generation of products, built around agentic orchestration, RAG, and long-term memory, can autonomously drive the PD loop of "topic selection → evidence capture → diagnosis → improvement → accumulation," handing off to teachers for review at checkpoints. Most products in this category are still at an early stage of deployment, and their reliability, explainability, and teacher acceptance still require systematic evaluation (see Section 5.4).

The industry's typical architectural description of an education agent is "an LLM as the reasoning engine + RAG for processing an education knowledge base + multimodal interaction + strategy optimization driven by historical data," with the value proposition repeatedly framed as "not replacing the teacher, but supporting the teacher — offloading repetitive work so the teacher can focus on instructional design and emotional care." In their framework, Lu and Tang further sketch a picture of "multi-agent collaboration": in classroom teaching, a "teaching-assistant agent" handles knowledge Q&A and personalized guidance, a "learning-companion agent" simulates diverse roles within collaborative learning, and a "metacognitive-mentor agent" monitors project progress in real time, analyzes each participant's engagement, and provides feedback. This same division of labor across multiple agents can transfer to PD — for example, a "retrieval agent" handling calls to school-based resources and curriculum standards, an "analysis agent" extracting diagnostic dimensions from classroom evidence, and an "accumulation agent" writing finalized improvement plans back to the memory repository. On the standards side, the "General Reference Framework for Education LLMs" consortium standard, released at the 2025 World Digital Education Conference, offers authoritative guidance for moving education agents from the lab to at-scale deployment, and is regarded as a key inflection point for education agents "moving from the laboratory to large-scale application."

Domestic ed-tech companies are also layering agentic capability onto existing lesson-prep-and-delivery platforms. NetDragon (HK:0777)'s 101 Education PPT is a case in point: on its integrated lesson-prep-and-delivery platform, it has added an "AI Teaching Assistant" (letting teachers choose or customize an assistant persona, add source material to create personalized content, and execute in-class commands) and an upgraded "Lesson Prep Studio" (letting teachers create courseware directly and pull lesson-prep resources and templates). NetDragon states that 101 Education PPT already provides more than 770,000 free lesson plans, courseware, and micro-lecture resources to roughly 9.7 million teachers nationwide. This enormous installed base of teachers and this resource repository provide a real foundation for a research agent's resource accumulation, memory-repository construction, and personalized recommendation — for a research agent, "how much searchable, reusable, high-quality school-based resource exists" often matters more to real-world usefulness than "how strong the underlying model is." Overall, the research agent is the category with the largest imaginative upside but the lowest maturity of the four forms discussed here: the more autonomous its orchestration becomes, the higher the bar for evidence traceability and human review — the key to

sound deployment is "automate the high-certainty stages first, and leave the uncertainty to the teacher."

Product capability radar (proposed dimensions): curriculum-standards alignment, multimodal support, evidence traceability, orchestration autonomy, privacy compliance, teacher controllability. Scores for each product against these dimensions need to be grounded in third-party benchmark data — see [TBD: benchmark data/source]; the capability radar chart appears as Figure 5-1.

5.4 Evaluation and Benchmarking: Making "Supporting PD" Testable

Consistent with this Blue Book's overall commitment to evidence-based evaluation, this section proposes an evaluation framework for the PD scenario and cites publicly available evaluation criteria and standards wherever possible. Evaluating PD-oriented products cannot simply borrow a general-purpose LLM leaderboard; dimensions need to be designed around the question "has PD actually gotten better," covering at minimum: **curriculum-standards alignment** (whether generated content matches national curriculum standards and the specific textbook edition), **subject-matter factual accuracy** (particularly the correctness of test items and their explanations), **evidence traceability** (whether diagnostic conclusions can be traced back to source evidence), **diagnostic consistency** (agreement between AI classroom analysis and expert human annotation), **orchestration reliability** (agent task-completion rate, correctness of tool invocation, hallucination rate, and frequency of human intervention required), and **teacher controllability and acceptance** (whether teachers can understand, correct, and trust the output).

The standardized foundation for evaluation is taking shape. The 2025 national standard GB/T 45288.2—2025 *Artificial Intelligence — Large Models — Part 2: Evaluation Metrics and Methods*, together with the *General LLM Evaluation System 2.0* released by the National Key Laboratory of Cognitive Intelligence (led by iFlytek), provide a general framework of metrics and methods for LLM capability evaluation; evaluation aimed specifically at the education sector further "covers multi-subject K-12 knowledge and competency, aligns with China's education system, and assesses model performance in scenarios such as intelligent lesson-prep content generation and personalized learning-path planning." CAICT's trustworthy-AI education-LLM evaluation (as with Ziyue's Level 5 rating mentioned above) provides a third-party reference point for the compliance and trustworthiness dimension.

- **Evaluation of education-vertical LLMs.** This involves designing item banks and scoring rubrics for tasks such as curriculum-standards alignment, subject-matter factual accuracy, item quality, and PD-report usability, reporting scores, sample sizes, and evaluation methodology. Particular caution is warranted around claims of "high accuracy": the curriculum-comprehension accuracy and Q&A accuracy figures various vendors report are typically based on their own

private test sets, lacking cross-vendor comparability; specific figures [TBD: third-party evaluation data/source].

- **Evaluation of multimodal classroom analysis.** This means benchmarking against annotated classroom video to assess the accuracy and consistency of behavior recognition, dialogue attribution, and segment splitting. Classic coding schemes such as the Flanders Interaction Analysis System (FIAS) can serve as an annotation baseline, assessing the agreement between AI-generated distributions of teacher–student talk ratio and question cognitive level against human annotation [TBD: evaluation data/source].
- **Evaluation of research agents.** This means assessing task-completion rate, correctness of tool invocation, hallucination rate, and frequency of human intervention required. Given Lu and Tang's finding that agents still need improvement in their "depth of understanding of interactive behaviors, activity design, and goal attainment across the teaching process," this kind of evaluation should focus specifically on whether an agent draws conclusions beyond its warrant when evidence is insufficient, and whether its diagnostic conclusions can be traced back to source evidence [TBD: evaluation data/source].

A special case: evaluating intelligent item generation and test assembly demands particular caution. Item design is the PD stage with the highest bar for factual correctness, and the one most easily masked by an appearance of "efficiency" hiding a real quality gap. The *Blue Book on AI-Empowered Applications in Basic Education (2025)* offers a fairly sober assessment here: current intelligent item-generation systems already support template-based generation by "selecting grade level, subject, item type, difficulty, and knowledge-point coverage," and some even "connect to an LLM to support automatic document analysis, extracting core knowledge points and generating test items" — but "the items, explanations, and answers generated may contain errors, requiring teacher verification before use in actual instruction," and current systems "have not yet truly matched real teaching scenarios." The report identifies three areas needing improvement: first, on generation type, the focus should shift toward "higher-order thinking, contextualized, and cross-disciplinary item generation," rather than staying at low-order item types like arithmetic drills and fill-in-the-blank; second, on explanation generation, systems should work to "conquer complex, difficult problems in subjects like math and physics that involve formula derivation, special solution methods, and geometric figures," building "a scientifically rigorous, logically sound standard for explanation generation"; third, on difficulty control, systems should generate, based on a student's cognitive trajectory, "items spanning a range of difficulty levels that emphasize testing personalized higher-order thinking ability." Intelligent test assembly, meanwhile, can draw on "multi-objective optimization techniques such as genetic algorithms and ant-colony algorithms" plus knowledge-graph techniques to balance the multiple constraints of "difficulty, knowledge-point coverage, and item-type distribution" — but "scientific test assembly" in the sense of "constraint satisfaction" is not the same thing as "test items with real educational value." Evaluation aimed at item design should

therefore not report only "generation speed" and "knowledge-point coverage," but should also introduce educational-measurement metrics like **difficulty, discrimination, and reliability**, with human review by subject-matter education experts serving as the quality gold standard. Related benchmark data: [TBD: third-party evaluation of item quality/source].

Evaluation-results timeline. The iteration milestones and version evolution of different products on key capabilities appear in the Figure 5-2 timeline. All evaluation results should be tagged with test date, model version, and dataset, to avoid treating a single result as an absolute conclusion — this is also why this Blue Book stays cautious about any claim of "top of the leaderboard" or "99% accuracy." In fact, during the research for this chapter, individual channels were seen claiming figures such as "98.5% curriculum-comprehension accuracy" or "99.8% Q&A accuracy" for particular products — but such figures are typically based on vendors' private test sets, lack public methodology, and lack cross-vendor comparability, so this chapter treats them uniformly as unverified vendor claims and excludes them from any quantitative conclusion.

5.5 Illustrative Scenarios (Evidence-Based Narratives)

This section relies primarily on public, verifiable real-world cases, supplemented by framework-level narrative, in an effort to avoid any fabrication.

- **Scenario 1: School-based collective lesson planning and resource accumulation.** A PD group connects a PD assistant to the school's historical lesson plans and curriculum standards, generating a tiered draft lesson plan for a given unit; teachers revise it and write it back, building a reusable school-based PD memory. An idealized workflow looks like this: a lesson-prep agent first uses RAG to retrieve the school's historical lesson plans, high-quality courseware, and corresponding curriculum-standard requirements for the unit, generating a structured draft covering "teaching objectives — key/difficult points — tiered activities — matching exercises"; the teacher then edits and localizes the draft based on the specific class's learning profile; the system writes the finalized version and its revision trail back to the memory repository, so that "the next time this unit comes up for prep, the system understands this school better." In this process, the AI handles "drafting and alignment," while the teacher handles "judgment and finalization" — a clean division of labor.

The *Blue Book on AI-Empowered Applications in Basic Education (2025)* describes this kind of capability as efficient resource retrieval and precise recommendation that cracks the "difficulty of lesson prep" for front-line teachers, with the core technique being multidimensional modeling of "a teacher's lesson-prep needs, teaching objectives, teaching context, and student learning profile," then "dynamically ranking to form a resource-recommendation list matched to the teacher's lesson-prep needs"; the report specifically stresses the need to integrate public resources, such as the National Smart Education Platform for Primary and Secondary Schools, with school-based repositories,

achieving a "dual-drive" of public and school-based resources. On the product side, iFlytek's "Spark Teacher Assistant" is sold on "providing suggested curriculum outlines and lesson plans for teachers, reducing time spent searching for materials and producing teaching aids," with its Spark education LLM (iterated through 2025 to versions such as X1.5) built around "understanding education"; NetDragon's 101 Education PPT, as a lesson-prep-and-delivery platform serving roughly 9.7 million teachers with a vast shared library of lesson plans and courseware, provides the real-world resource base for this scenario.

- **Scenario 2: Multimodal peer observation and evidence-based improvement.** This is the scenario with the most solid deployment in 2026, and the one that best embodies the meaning of "evidence-based PD." Huachuan Elementary School in Yongkang, Jinhua, Zhejiang Province, adopted Seewo's AI classroom intelligent-feedback system, generating more than 200 data metrics per lesson; the system records every question a teacher asks, tracks their movement around the room, and captures interaction moments, generating a real-time feedback report with instructional-improvement suggestions — teachers affectionately call it their "AI PD specialist."

A verifiable string of details completes the causal chain of "evidence-based improvement": the system detected that a given teacher asked a question on average once every 3 minutes, with 62% of those questions closed-ended, and automatically generated the recommendation to "increase the proportion of open-ended questions"; after the teacher adjusted their questioning strategy accordingly, the rate of students voluntarily speaking up in that class rose from 35% to 68%. This chain — evidence → diagnosis → recommendation → improvement → re-measurement — is exactly what turns the old style of PD, where "review was based on impression and discussion was based on experience," into evidence-based PD where "data does the talking." The school built out a full PD-cycle loop of "evidence collection — data analysis — problem diagnosis — instructional improvement — iterative refinement," producing a new pattern of teacher development that is "data-driven, problem-oriented, and tiered in progression," and moved roughly 40 teachers from an "experience-based" to an "evidence-based" mode of practice. This lines up closely with the shift the *Blue Book on AI-Empowered Applications in Basic Education (2025)* describes from "experience-driven" to "data-driven" teachers, and offers a vivid, concrete instance of the *"AI+ Education" Action Plan*'s call to "carry out evidence-based PD practice enabled by AI."

Seewo's related practice was also selected for the Ministry of Education's Center for Educational Technology and Resource Development (China Educational Television and Audio-visual Center) 2025 call for teacher AI-application cases, under case titles such as *Building a New Paradigm for Smart PD: Using AI to Drive Evidence-Based Improvement in Teachers' Instructional Behavior*, *A Practical Study of GenAI-Driven Evidence-Based PD in Diagnosing and Improving Classroom Problems*, and *Seewo's AI Classroom Feedback System Empowering Integrated Teaching–Research–Assessment Practice in Elementary Chinese Language Instruction*. These case titles alone sketch out three main lines along which multimodal classroom analysis is landing in PD practice:

evidence-based improvement of teacher instructional behavior, diagnosis and improvement of classroom problems, and integration of teaching, research, and assessment.

- **Scenario 3: Regional PD data governance.** A regional PD platform aggregates de-identified PD data across multiple schools, using an agent to cluster themes and identify common problems, providing a basis for regional PD topic selection and evidence-based guidance from PD specialists. For instance, when classroom-analysis reports from multiple schools in a district all point to common problems like "insufficient higher-order questioning" or "group collaboration that's superficial in form only," the regional platform can use this pattern to set the semester's PD theme and route high-quality lesson cases directly to the teachers who need them most — scaling up from individual-case improvement to district-wide quality gains. Seewo's 19 key demonstration zones nationwide embody exactly this "single school to region" organizational form. The *"AI+ Education" Action Plan*'s call for "evidence-based PD practice enabled by AI" and "building a teacher-training model suited to the intelligent era" creates the policy space for this scenario; but regional-level data aggregation carries a higher bar for privacy compliance and security, given that it involves data flows across schools and stakeholders, requiring that security-protection standards be "determined by classification and grading" and that "privacy leaks" be guarded against (see Chapter 7 for detail). This scenario remains largely at the demonstration-zone pilot stage; quantified outcomes for specific regional deployments are [TBD: regional case/source] — this chapter makes no fabricated claims here.

5.6 Challenges and Risks

- **Factuality and hallucination.** Errors in vertical models' subject-matter knowledge and test items can be subtle and easily missed, potentially misleading PD judgments. The *Blue Book on AI-Empowered Applications in Basic Education (2025)* states plainly that intelligent item generation may produce "items, explanations, and answers [that] may contain errors, requiring teacher verification before use in actual instruction," and acknowledges that items generated by existing systems "have not yet truly matched real teaching scenarios," still lagging on higher-order-thinking, contextualized, and cross-disciplinary items. Factuality risk requires a double check: both evaluation and teacher review.
- **Insufficient evidence traceability.** Some products' diagnostic conclusions lack a traceable chain back to source evidence, weakening PD's persuasiveness and auditability. RAG and evidence anchoring are the mitigation path, but not every product offers a traceable mapping from conclusion back to evidence.
- **Item quality and the risk of exam-oriented over-optimization.** Intelligent item generation and test assembly boost efficiency, but if pursued purely for item volume and coverage, they risk reinforcing rote drilling. The *"AI+ Education" Action Plan* specifically lists "guarding against

exam-oriented over-optimization" as a bottom line; intelligent test assembly can use multi-objective optimization techniques like genetic and ant-colony algorithms to balance difficulty, knowledge-point coverage, and item-type distribution — but "scientifically balanced test assembly" is not the same as "test items with real educational value," and item design must stay oriented around core competencies.

- **Privacy and compliance.** Classroom multimodal data touches the privacy of minors and teachers; its capture, storage, and use must comply with data-minimization principles and "effectively guard against ... privacy leaks and similar problems" (see Chapter 7 for detail).
- **Rebalancing teacher workload.** Introducing a tool without redesigning the underlying workflow risks creating "use for the sake of using it" busywork that defeats the purpose of reducing teacher burden. Real workload reduction comes from a closed-loop process (like Huachuan Elementary's five-stage PD loop), not from stacking tools on top of one another.
- **Lack of evaluation standards and incomparable methodologies.** AI products aimed at PD still lack a widely accepted capability benchmark and evaluation methodology. Although general standards like GB/T 45288.2—2025 *Artificial Intelligence — Large Models — Part 2: Evaluation Metrics and Methods* exist, and the World Digital Education Conference's "General Reference Framework for Education LLMs" consortium standard is a start, a benchmark specific to the PD scenario — curriculum-standards alignment, item quality, classroom-diagnosis consistency — is still needed, and vendor-self-reported accuracy figures have weak cross-comparability, leaving schools and districts without a trustworthy basis for comparison when selecting products.
- **Model-version drift and reproducibility of conclusions.** Education LLMs iterate frequently — during this chapter's research window alone, models like Spark and Ziyue underwent multiple version updates within a few months — and the same query can produce different diagnoses or item-generation results across versions. This poses a challenge to PD's "reproducibility": if a PD conclusion cannot be tagged with the model version and dataset it depended on, its academic value and traceability are significantly diminished. Evaluation and PD records alike should therefore log model version, timestamp, and methodology.
- **Over-reliance and the erosion of professional skill.** If teachers habitually "outsource" lesson prep, item design, and observation-based diagnosis to AI while neglecting their own independent judgment, their instructional-design ability and PD sensitivity risk atrophying. Lu and Tang stress the need to "develop innovative awareness and capability, encouraging teachers and students to explore human-machine collaborative teaching models suited to their discipline" — the implication being that AI should function as scaffolding for a teacher's professional growth, not a substitute for it; the ultimate purpose of PD remains the teacher's own professional development.

5.7 Recommendations

1. **Make "teacher-led, evidence-traceable" a red line.** Clarify AI's supporting role in PD; every diagnosis and improvement recommendation must retain a traceable evidence chain and undergo teacher review. This is consistent with Lu and Tang's principle of "human-machine collaboration rather than machine replacement" and with the bottom-line requirements of the *"AI+ Education" Action Plan*.
2. **Push education-vertical models toward curriculum-standards alignment and open evaluation.** Encourage vendors to publish their evaluation methodology and sample data, and build a public benchmark for PD — grounded in national standards like GB/T 45288.2—2025 and the General Reference Framework for Education LLMs — for third parties (such as the China Educational Television and Audio-visual Center, CAICT, and relevant professional bodies) to organize evaluations, breaking the current deadlock where "no vendor's accuracy figures are comparable to any other's."
3. **Prioritize multimodal classroom-evidence capability.** Move peer observation from subjective impression to evidence-based diagnosis (as demonstrated by the Seewo feedback system and the "AI PD specialist" practice), coordinating with first-person capture hardware (see this institute's *2026 Blue Book on the AI Glasses Education Industry*); use classic coding schemes such as FIAS as an annotation baseline to safeguard the scientific rigor and consistency of AI diagnosis.
4. **Accumulate PD assets through a school-based memory repository.** Use RAG and long-term memory to turn lesson plans, lesson cases, and PD history into reusable, citable knowledge assets, preventing expertise from walking out the door with departing staff; make full use of public resources, such as the National Smart Education Platform for Primary and Secondary Schools, together with school-based repositories, so the system "understands the school better the more it's used."
5. **Build governance and safety into product design from the start.** Embed compliance capability into data capture, de-identification, retention, and authorization, implementing security-protection standards "determined by classification and grading" (see Chapter 7, "Governance and Safety," for detail).
6. **Keep item design and test assembly oriented around core competencies.** Push intelligent item generation from mere "knowledge-point coverage" toward items with real educational value — higher-order thinking, contextualized, cross-disciplinary — strengthening collaboration with subject-matter education experts to prevent efficiency gains from curdling into exam-oriented over-optimization.
7. **Roll out agentic orchestration deployment cautiously.** Pilot agentic automation first in high-value, low-risk PD stages, retaining human review at critical checkpoints before gradually

expanding; stay realistic about the fact that "teaching agents are still at an early stage," and avoid overpromising on orchestration autonomy.

8. **Make teacher professional growth the final measure.** The success of any PD AI should ultimately be judged by whether teachers have genuinely grown and classrooms have genuinely improved — not by tool-usage frequency or report-generation volume. Evidence-based PD enabled by AI should be built into teacher-training systems and AI-literacy programs (the *"AI+ Education" Action Plan* has already called for "developing teacher AI-literacy standards" and "building a contextualized assessment system"), so teachers both use the tools well and are never replaced by them — ultimately realizing a new PD model of "human-machine collaboration, teacher-led."

Chapter References

1. Lu Yu, Tang Xiaoyu. Form Hierarchy and Progression Pathway of Generative AI Empowering Classroom Teaching [生成式人工智能赋能课堂教学的形态层级与进阶路径]. *e-Education Research*, Issue 6, 2025 (No. 386 overall). School of Educational Technology, Beijing Normal University. DOI:10.13811/j.cnki.eer.2025.06.010. https://aic-fe.bnu.edu.cn/docs/2025-07/a4d9baa90d9e49d9920ee2c46dd283b7.pdf

2. Engineering Research Center for Intelligent Technology and Educational Applications, Ministry of Education; Beijing Digital Education Center (Beijing Educational Television Station). Blue Book on AI-Empowered Applications in Basic Education (2025) [人工智能赋能基础教育应用蓝皮书（2025 年）]. July 2025. https://vmcl.bnu.edu.cn/docs/2025-07/3a416f18f01b42f998062babd78b9ca8.pdf

3. Ministry of Education, National Development and Reform Commission, Ministry of Industry and Information Technology, Ministry of Science and Technology, National Data Administration. "AI+ Education" Action Plan [“人工智能+教育”行动计划] (Jiaokexin [2026] No. 1). April 2026. Ministry of Education of the People's Republic of China official portal. http://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html ; full text also at China Education Online https://www.eol.cn/zhengce/wenjian/202604/t20260410_2727386.shtml

4. Xinhuanet. NetEase Youdao Releases Multiple New AI Products for Its Ziyue Education LLM, Defines an L1-L5 Grading Scale for Education AI Application Capability [网易有道发布子曰教育大模型多款 AI 新品 定义教育 AI 应用能力 L1-L5 分级]. August 21, 2025. http://www.news.cn/tech/20250821/11dbed39f03b4b3aa0723e0a2ce910d9/c.html

5. iFlytek Smart Education. iFlytek Spark X1.5 Released: Understanding Education, Understanding You Better [讯飞星火 X1.5 发布：懂教育、更懂你] (company news). 2025. https://edu.iflytek.com/about-us/news/company-news/2535

6. Huawei. Huawei and iFlytek Jointly Release the "Spark Education and Healthcare LLM Scenario All-in-One Solution" [华为联合科大讯飞发布“星火教育、医疗大模型场景一体机解决方案”]. September 2025. https://e.huawei.com/cn/news/2025/industries/education/spark-education-medical-large-model

7. Seewo. Ministry of Education Publishes 2025 List of Teacher AI Application Cases; Seewo Classroom Intelligent Feedback System Selected for Multiple Cases [教育部 2025 年教师 AI 应用案例名单公布，希沃课堂智能反馈系统多案例入选]. January 2026. https://www.seewo.com/article/detail/2889

8. Huicong Education Network. 200+ Data Metrics Per Lesson? Zhejiang's "AI PD Specialist" Keeps Every Class Improving [每节课 200+项数据指标？浙江“AI 教研员”让每一节课都有进步]. August 2025. https://edu.hczyw.com/2025/0801/77901.html (also see Seewo https://www.seewo.com/article/detail/2797)

9. Khan Academy. Khanmigo for Teachers: Free, AI-Powered Teacher Assistant. 2025. https://www.khanmigo.ai/teachers ; related lesson-planning capability at Khan Academy Blog, "10 Creative Ways to Leverage Lesson Planning with AI," https://blog.khanacademy.org/10-creative-ways-to-leverage-lesson-plan-ai/

10. TAL Education "Chapter Nine" model (MathGPT) product materials. Launched 2023, iterated 2024–2025. (Among the first domestic math-education LLMs to complete government filing.) Related coverage at https://www.aihub.cn/tools/llm/mathgpt/ and China.com https://m.tech.china.com/digi/digi/20240419/202404191508210.html

11. NetDragon Websoft (HK:0777). 101 Education PPT official site and product materials. https://ppt.101.com/ ; platform-scale figures from Tencent Software Center https://pc.qq.com/detail/8/detail_22548.html and Zhitong Finance https://m.zhitongcaijing.com/contentnew/appcontentdetail.html?content_id=384035

12. State Administration for Market Regulation, Standardization Administration of China. GB/T 45288.2—2025, Artificial Intelligence — Large Models — Part 2: Evaluation Metrics and Methods [人工智能 大模型 第 2 部分：评测指标与方法]. 2025. https://lib.scu.edu.cn/genai/static/wenjian/GBT45288.2-2025-genaim.pdf

13. National Key Laboratory of Cognitive Intelligence. General LLM Evaluation System 2.0 Officially Released [《通用大模型评测体系 2.0》正式发布] (led by the National Key Laboratory of Cognitive Intelligence). https://cogskl.iflytek.com/archives/3296
14. Cyberspace Administration of China (CAC). Announcement on the Publication of 2024 Filing Information for Generative AI Services Already on Record [关于发布 2024 年生成式人工智能服务已备案信息的公告]. January 2025. https://www.cac.gov.cn/2025-01/08/c_1738034725920930.htm (filing statistics also reported at C114 https://m.c114.com.cn/w5339-1259288.html)

Chapter 6 Intelligent Assessment (A New Scenario): Mechanisms, Product Forms, and Recommendations

6.1 Scope and Rationale for a New Scenario

Building on the three scenarios this Blue Book has used to organize generative-AI education products so far — Empowering Teaching, Supporting Learning, and Supporting Research (Professional Development) — Blue Book 2026 breaks out Intelligent Assessment as a fourth, independent scenario. Two parallel trends drove this decision.

The first is that generative AI's capabilities are pushing outward from "producing content" toward "understanding, judging, and giving feedback." Once a model can read an essay, check it against a scoring rubric, and produce both a score and sentence-by-sentence commentary, it has functionally crossed the threshold from a content-generation tool into a judgment-making one — it now has the technical footing to take on assessment, a fundamentally evaluative task. Compared with earlier-generation auto-grading built on keyword matching, regex rules, or shallow statistical features, generative models represent a qualitative leap in handling semantics, structure, and lines of reasoning. That leap is the direct reason this chapter pulls assessment out as its own scenario rather than folding it into the others.

The second is that assessment has long been the most labor-intensive, slowest-to-return, and hardest-to-standardize link in the education chain. A single provincial spoken-English exam for the high school entrance test (zhongkao) can involve hundreds of thousands of oral responses; grading a set of midterm essays can eat an entire weekend for a frontline teacher; and feedback routinely arrives only after students have already moved on to the next unit. This is exactly where the demand for scale, immediacy, and personalization is most acute — and exactly where generative technology is best positioned to deliver value. iFlytek's intelligent speech-assessment technology offers a instructive reference point from Chinese practice: it has been deployed at scale in the spoken-English components of the gaokao (China's national college entrance exam) across 29 provinces and municipalities and the zhongkao across 132 cities, serving more than 45 million test-takers cumulatively[1]. High-stakes, large-scale, heavily standardized scenarios like these are precisely where assessment intelligence is landing first — and where the most caution is warranted.

It's worth stressing that assessment is tightly coupled with teaching, learning, and research support: intelligent assessment is simultaneously an "output" of teaching and an "input" to learning diagnostics and instructional improvement. Treating it as a separate scenario doesn't sever it from the assess-teach-learn loop; it highlights the distinct mechanisms and product logic that assessment

carries in its own right. The moment a machine is allowed into scoring, a tension immediately opens up between technical feasibility and educational appropriateness — a model that *can* assign a grade is not necessarily a model that *should be trusted* to assign one. This chapter is organized around that tension: Section 6.2 lays out the underlying mechanisms, Section 6.3 surveys product forms, Section 6.4 takes a hard look at validity and fairness risks, and Section 6.5 closes with recommendations for building trustworthy systems.

There is also a clear policy backdrop for treating assessment as its own scenario. In October 2020, the CPC Central Committee and the State Council issued the *Overall Plan for Deepening the Reform of Education Evaluation in the New Era* — the first systematic reform document on education evaluation in the history of the People's Republic. It called for eliminating the entrenched over-reliance on test scores, admission rates, diplomas, published papers, and titles ("唯分数、唯升学、唯文凭、唯论文、唯帽子"), and for "making full use of information technology to improve the scientific rigor, professionalism, and objectivity of education evaluation" by building evaluation mechanisms supported by big data and AI, encompassing multidimensional process-based, value-added, and comprehensive evaluation[19]. In April 2025, the Ministry of Education and eight other departments jointly issued the *Opinions on Accelerating the Advancement of Education Digitalization*, folding "AI-empowered education evaluation" into the broader digital-transformation agenda[19]. In other words, intelligent assessment is being pulled forward by both a strong technology push and a clear policy pull — it is not an optional efficiency add-on but a load-bearing piece of the national evaluation-reform roadmap. Precisely because of that, this chapter's treatment of risk has to be unusually careful: when a technology is granted institutional legitimacy to intervene in high-stakes evaluation, any validity or fairness shortfall gets amplified into a systemic educational consequence.

One definitional note up front: "Intelligent Assessment" in this chapter refers to products and practices in which generative AI — often working alongside discriminative models and purpose-built scoring engines — intervenes in educational measurement and feedback. It spans all three purposes of assessment (diagnostic, formative, and summative) and the full chain from item writing and response collection through scoring judgment to feedback intervention. It is neither synonymous with narrow "automated grading" nor confined to exam settings — routine homework grading, writing-process coaching, spoken-language practice, in-class learning analytics, and even holistic-competency profiling all fall within its scope. It is precisely this "spans the entire assessment chain, cuts across high- and low-stakes settings" character that sets this scenario apart from the other three and earns it a chapter of its own.

6.2 Mechanisms: How Generative AI Intervenes in Assessment

At its core, generative AI's intervention in assessment converts the judgment task of "scoring" into a language- and multimodal-processing pipeline: understand the response, check it against a standard, and generate a judgment plus feedback. To see how this differs from earlier technology, it helps to first distinguish the two technical lineages of automated scoring.

The first is the **explicit-feature approach**, exemplified by ETS's e-rater. It extracts a set of interpretable, explicit microfeatures from response text — grammar error rate, usage, mechanics (spelling, punctuation, capitalization), style, vocabulary sophistication, organization and development — then aggregates those features into a score via regression or machine learning[2][5]. Its strength is traceability: every point can be traced back to a specific feature, making it explainable and auditable. Its cost is that the features are manually engineered by humans and struggle to capture "below the surface" constructs like argument quality or depth of thought, leaving them exploitable by formulaic writing strategies.

The second is the **generative-understanding approach**, exemplified by large language models. Rather than extracting explicit features, it feeds the entire response in as context and, guided by a scoring rubric, understands and evaluates it holistically at the semantic level, producing a score and sentence-level commentary directly in natural language. Its strength is a qualitative leap in handling synonymous phrasing, partial correctness, and lines of reasoning, with feedback that is readable and improvement-oriented. Its cost is that the "reason for the score" is itself a generated artifact — making the process more of a black box — and scores can drift with prompt design and decoding temperature.

It's worth emphasizing that these two approaches are not simply substitutes for one another. Industrial-grade products today are typically hybrids of "explicit features plus generative understanding": a discriminative model or purpose-built engine anchors reliability and auditability, while a generative model fills in semantic understanding and feedback generation. Grasping this technical divide is the prerequisite for understanding the chapter's central tension later on — that high agreement does not imply high validity. Functionally, generative assessment breaks down into six categories.

6.2.1 Automated Scoring (Objective and Semi-Structured Items)

For objective items and semi-structured items with a relatively bounded answer space — fill-in-the-blank, short answer, formula problems, true/false with justification — models score by matching semantics against a reference answer. Compared with traditional keyword- or regex-based auto-grading, generative models can recognize synonymous phrasing, partial correctness, and valid reasoning paths, covering a much wider range of response forms. Take a short-answer item that asks students to "explain photosynthesis in your own words": a regex rule can only match pre-set

keywords and will miss a student who expresses the same idea in different but equivalent phrasing, whereas a generative model can judge semantic equivalence and award partial credit — with a specific pointer — when a student "gets the mechanism right but uses imprecise terminology."

In large-scale Chinese practice, iFlytek treats intelligent grading for the gaokao and zhongkao, oral-language assessment, and essay grading as "technologically the same family." Its spoken-English assessment system automates both testing and scoring and has been rolled out to spoken-English exams for the gaokao and zhongkao in more than 20 provinces and municipalities[1]; its smart-exam solution claims to "grade a full test paper in under 15 seconds," compressing the marginal time cost of grading subjective items to near zero[13]. The value of such systems lies in freeing enormous volumes of repetitive, well-defined scoring work from human hands.

But automated scoring of objective items is not a zero-risk, low-hanging fruit. Its risk concentrates in two places. First is the **"boundary response" problem** — misjudging answers that are reasonable but deviate from the reference answer, or cases where the reference answer itself is incomplete; this is especially acute in open-ended short answers and mathematical proofs, where a student's non-standard but correct solution path can be marked wrong. Second is **propagation of reference-answer quality** — automated scoring treats the item-writer's preset reference answer as the sole arbiter, so if that reference answer itself contains an error or gap, the error gets systematically amplified across every single response. Even in automated scoring — the relatively mature end of the spectrum — human spot-checks and anomaly review need to stay in the loop rather than handing the whole process over to the machine.

6.2.2 Essay and Open-Response Grading

This is where generative AI represents the sharpest capability jump over prior technology. Guided by a preset rubric, models can score open-ended responses — essays, short answers, argumentative pieces — across multiple dimensions such as content, structure, language, and logic, while pinpointing the specific sentences or passages where problems occur. Its value lies not just in assigning a score but in generating feedback that is readable and improvement-oriented.

On validity and reliability, the past two years have produced a body of verifiable empirical findings, and the overall picture is one of cautious optimism with strings attached:

- **Agreement with human raters can match or exceed inter-rater levels, but it hinges on prompt engineering and temperature settings.** One study comparing GPT-4, GPT-3.5, and Claude 2 on English-learner writing found GPT-4 performed best, showing excellent intra-rater reliability and good validity; under a "rubric plus anchor examples" prompting strategy, GPT-4's scoring accuracy improved by roughly 112% and 114% over GPT-3.5 and Claude 2, respectively[3]. The same study noted that lowering the temperature setting makes the model

more deterministic and significantly improves repeated-measures consistency — evidence that prompting strategy and decoding parameters are key variables governing reliability[3].

- **Reliability is very high in higher-education EFL essay grading — provided scoring is anchored to a rubric.** Yavuz et al. (2025), writing in the *British Journal of Educational Technology*, had 15 EFL teacher-raters plus ChatGPT and Bard score the same set of essays against a five-dimension rubric (grammar, content, organization, style and expression, mechanics), and found that a fine-tuned ChatGPT model achieved extremely high reliability across 10 repeated measurements (intraclass correlation coefficient ICC = 0.972, standard deviation 0.00)[4]. The takeaway: when assessment is firmly anchored to an explicit rubric and randomness is controlled, generative models can deliver highly stable scores.
- **Alignment with human judgment in multidimensional scoring varies by dimension.** A study using LLMs for multidimensional writing assessment found that models agree closely with human raters on surface-level linguistic dimensions (grammar, vocabulary, mechanics) but agreement drops noticeably on higher-order dimensions like content, argumentation, and depth of thought[3] — corroborating a judgment that runs throughout this chapter: models are better at judging the surface of language than the thinking underneath it. The design implication is that products shouldn't just hand back a single composite score; they should report scores and confidence levels by dimension, separating what the machine can be trusted to judge from what still needs a human check.
- **Legacy purpose-built engines have already established a comparison baseline.** As a reference point, ETS's e-rater engine correlates with human raters at roughly r≈0.78–0.79 on GRE argument/issue essay tasks, with weighted kappa between 0.73 and 0.77; on TOEFL independent and integrated writing tasks the correlation is roughly r≈0.75/0.73 with weighted kappa≈0.70[5]. A frequently cited finding is that on TOEFL independent and GRE Issue tasks, e-rater's agreement with a single human rater actually exceeds the agreement between two independent human raters[2]. But ETS's own research reports also carry an explicit warning: agreement statistics alone cannot prove "what is actually being measured" — high agreement can still coexist with construct underrepresentation or construct-irrelevant variance[5]. That warning applies equally to generative models and is the key to understanding the risks discussed in Section 6.4.

One methodological caveat is worth adding: the reliability and validity figures cited above (QWK, ICC, correlation coefficients, weighted kappa) were mostly obtained on controlled research datasets, and are highly sensitive to prompt design, rubric clarity, and the distribution of the response sample. Extrapolating them directly to how a specific product will perform in a real classroom is not rigorous — the same model can see its agreement drop significantly with a different rubric, a different grade band, or a student population with a different native-language background. This chapter cites these

numbers to characterize the ceiling on capability and its conditionality, not to endorse any particular product. Taken together, essay and open-response grading is where generative assessment's capability is strongest — and also where the need for human-machine collaboration constraints is greatest: it can score reliably and generate readable feedback, but scoring consistency is not the same thing as scoring validity.

6.2.3 Spoken-Language and Multimodal Assessment

Spoken-language assessment has the longest technical pipeline in assessment intelligence: automatic speech recognition (ASR) must first transcribe speech to text, after which pronunciation, fluency, vocabulary, grammar, content, and discourse coherence are each scored as separate sub-constructs. Internationally, multiple testing organizations have deployed purpose-built engines — ETS's SpeechRater, Cambridge's CASE, Pearson's (PTE) Ordinate, and the Duolingo English Test's (DET) Duo Speaking Grader[6]. Take DET's public scoring documentation (2024) as an example: its automated scoring covers all six sub-constructs found in human rubrics (content, discourse coherence, vocabulary, grammar, fluency, pronunciation), with the speaking sub-score aggregated from "classically scored speaking" items and "IRT-based (item response theory) read-aloud" items[7]. On reliability, SpeechRater's correlation with human raters is roughly r=0.57–0.68[6] — noticeably lower than the agreement written-essay engines achieve with human raters, reflecting the fact that spoken-language constructs (especially "content/communicative effectiveness") are harder for machines to capture reliably than written language.

The validity risks in spoken-language assessment have their own distinctive character. First, **error propagates down the pipeline**: ASR transcription errors directly contaminate every downstream scoring dimension, and ASR error rates run higher for students with heavy accents or unusual native-language backgrounds — turning a technical error into a fairness problem. Second, **uneven construct coverage**: machines score "acoustically quantifiable" dimensions like pronunciation, fluency, and speech rate fairly accurately, but score higher-order dimensions — whether the content is on-topic, whether communication is effective, whether the argument is persuasive — much less reliably. This is the underlying reason SpeechRater's correlation with human raters tops out around r≈0.57–0.68[6]. Third, **the gap between read-aloud and open-response items**: the very reason DET uses IRT for read-aloud items and classical scoring for open speaking is that read-aloud answers converge to a bounded space and are easy to automate, while open expression items involve complex constructs that are hard to score stably[7].

In China, spoken-language assessment is among the most mature at-scale deployments. iFlytek already provides spoken-language assessment services for zhongkai (zhongkao) exams in 64 cities nationwide and for gaokao spoken-language scoring or supplementary tests in 13 provinces and municipalities[1]. Guangdong offers a case in point: as part of the gaokao English exam, the Computer-based English Listening and Speaking Test (CELST) uses a machine-learning-based

automated scoring engine trained on a human-rated calibration set[8]. Notably, high-stakes spoken-language exams like this typically still retain a human re-scoring or spot-check step — machine scoring is not the sole arbiter — echoing the "machine plus human" dual-scoring practice common in high-stakes international writing exams. As multimodal terminals such as AI glasses and educational robots mature, spoken-language assessment is expanding from static "read a passage, answer a question" testing toward previously hard-to-assess dimensions — narrating while running an experiment, oral contribution within group collaboration, contextualized role-play — drawing on speech, vision, motion, and collaboration traces simultaneously as evidence sources (see this institute's companion volumes, the *AI Glasses in Education Blue Book 2026* and the *Global Education Robotics Development White Paper 2026*, for more detail). This expansion broadens the constructs assessment can cover, but it also compounds risks around ASR error, multimodal alignment, and privacy in data collection — the more complex the construct, the more essential it is to keep humans in the loop rather than ceding the call to the machine.

6.2.4 Process-Based Assessment

By modeling learning-process data — response trajectories, revision history, interaction dialogue, collaboration records, keystroke logs — models can convert previously hard-to-quantify process behaviors into observable, feedback-ready indicators, helping formative assessment shift from "scoring the outcome" to "characterizing the process." Recent research treats keystroke logging, qualitatively analyzed via S-notation, as a form of learning analytics that makes the scope, sequence, and integration of revisions during writing visible, thereby supporting process-oriented formative feedback[9]. Products like Khan Academy's Writing Coach are moving in this direction: it emphasizes that "process matters as much as product," delivers guided feedback at the outlining, drafting, and revision stages, and gives teachers class- and individual-level feedback summaries, time-on-stage reports, and conversation logs — so a teacher can grasp a student's progress and flag suspected plagiarism without paging through hundreds of draft histories one by one[10].

Process-based assessment is where generative technology holds the most promise relative to the traditional "one exam, one score" summative model, because it extends the evidentiary basis of assessment from the "final product" to the "entire process." Take writing as an example: a final draft only shows the outcome, while keystroke logs and draft history can reveal whether a student wrote in one continuous burst or revised repeatedly, whether edits clustered around word choice or overall structure, and how the student sought help when stuck. This kind of process evidence is often more valuable for formative feedback than a single score on the final draft. Khan Academy Writing Coach's product logic — combining time-on-stage reports, conversation logs, and stage-by-stage feedback as the core of teacher insight — is built on exactly this premise, letting teachers track progress without manually reviewing hundreds of draft histories[10].

That said, it has to be said plainly that the underlying science of process-based assessment is still immature. Several 2025 review articles note that the evidence base for using learning analytics to support formative assessment remains thin, that formative assessment resists being reliably operationalized, and that teachers tend not to trust — or act on — analytics results when they don't align well with a formative-assessment model[11]. Three bottlenecks stand out: first, **data completeness** — process data is often incomplete and noisy, and hard to stitch into a coherent trajectory across devices and platforms; second, **inferential uncertainty** — the chain of reasoning from "revised many times" to "diligent" or "struggling" is fragile and easy to over-interpret; third, **privacy and compliance** — process data collection is the most intensive and closest to student behavior of any assessment type, and crossing a line here amounts to over-surveillance of learners. In other words, there is still real distance between "being able to collect process data" and "being able to render a valid process-based judgment." For now, process-based assessment is better suited as a supporting cue for teacher judgment than as a standalone basis for scoring.

6.2.5 Learning Diagnostics and Feedback

Beyond scoring, models can attribute errors, pinpoint weak knowledge points, infer the likely source of a misconception, and generate personalized remediation and a follow-on learning path — moving assessment from "measurement" to "intervention" and feeding directly back into the learning scenario. This is the most direct link in the assess-teach-learn loop where assessment feeds learning: a score only tells a student "how many points you got," while a good diagnosis tells a student "which specific knowledge point you got wrong, why you got it wrong, and what to practice next."

Learning hardware from China's leading ed-tech companies is where this mechanism is most concentrated. TAL Education's Nine Chapters LLM (MathGPT) is the first education LLM in China purpose-built for mathematics and among the first to clear official filing (备案); centered on problem-solving and step-by-step tutoring algorithms, it offers automated math problem-solving and complex word-problem grading, Chinese- and English-essay grading, and AI step-by-step tutoring, and has already been deployed in the "Xuexiaoban" app and the Xueersi AI learning tablet Xpad. Its "Math Anytime" feature targets instant Q&A for primary- and middle-school problems (this is a vendor claim; the actual coverage rate [TBD: needs third-party verification])[12][13]. NetDragon's (HK:0777) 101 Education PPT AI Teaching Assistant approaches this from the classroom side, assigning students exercises calibrated to their individual level based on in-class performance data and grading them online — embedding diagnosis-and-feedback directly into the daily teaching workflow[18]. The common logic across these products is translating assessment results immediately into the next learning action, shortening the "test, feedback, practice" cycle.

The core limitation for diagnosis-and-feedback products is **attribution accuracy**. A wrong score affects one grade; a wrong attribution affects a student's entire subsequent learning path. If a model

misreads "a careless computational slip" as "a conceptual misunderstanding," it may push the student a heavy load of unnecessary remedial concept review — wasting time and undermining confidence — and the reverse error is just as costly. What makes this more insidious is that a generative model's attribution tends to be phrased with confidence and apparent evidence, which makes it easy for students and parents to over-trust. Diagnosis-and-feedback products therefore need to transparently surface "what is this attribution based on, and how confident is it," and preserve teachers' right to review any high-impact judgment.

6.2.6 AI-Assisted Item Writing

Upstream of scoring, models can help generate test items, distractors, and scoring rubrics based on knowledge points, difficulty level, and item-type requirements, and provide an initial quality check on items (discrimination, coverage, wording clarity) — shortening the item-writing cycle. This link puts generative technology's home-turf strength — content production — to direct use against assessment's professional constraints, with an immediate payoff: item writing has long been a time-consuming, high-skill professional task, and a good set of test items requires repeated polishing of stems, careful design of effective distractors, and controlled calibration of difficulty and discrimination.

Solid empirical support has recently emerged for this piece. One of the larger field studies to date covered 91 university classes in the United States and roughly 1,700 students (71 classes used AI-generated exams; 20 statistics classes used standardized AP items), evaluated with a two-parameter logistic (2PL) IRT model. It found that AI-generated items matched high-stakes standardized test items on difficulty and discrimination — AI-generated items averaged a discrimination parameter of $\bar{\alpha}\approx1.3$ versus $\bar{\alpha}\approx1.2$ for the standardized comparison set, and the AI-generated exam achieved a maximum test information of $I_{max}\approx3.85$ (reliability $R\approx0.79$), outperforming the standardized comparison's $I_{max}\approx2.61$ ($R\approx0.72$)[14]. A separate comparative study found that expert reviewers sometimes even preferred LLM-generated items over human-authored ones, but both studies consistently emphasized that **human review remains indispensable for final quality**: LLM-generated items occasionally produce multiple correct answers, non-functional distractors, or factual errors, and items still need to pass through multi-stage automated screening — evaluation, deduplication, difficulty and appropriateness checks — before entering an item bank[14][15].

For China's basic education system, AI-assisted item writing overlaps with the direction of item-writing reform: the 2026 zhongkao item-writing guidelines call for sharply reducing the share of rote-memorization items in favor of contextualized, competency-oriented ones[19] — and it is exactly this "context-heavy, transfer-heavy, memorization-light" style of item that is most labor-intensive to write by hand, and best suited to generative technology producing candidate items in bulk for subject teachers to curate and polish. But the bias risk in item generation shouldn't be dismissed: models may introduce stereotypes or regional bias through the scenarios, character names, or cultural background

they generate, or produce items that look reasonable but are actually beyond the curriculum scope or off the standards. "Machine generates candidates, teacher makes the final professional call" should be the basic working norm for this link.

The table below summarizes the input, output, and key limitations of the six mechanism categories.

Mechanism	**Primary Input**	**Primary Output**	**Key Limitation**	**Verifiable Evidence (example)**
Automated scoring	Objective/semi-structured responses	Score, correct/incorrect judgment	Boundary-response misjudgment	iFlytek at-scale gaokao/zhongkao assessment[1]
Essay and open-response grading	Open-ended text responses	Score, multidimensional commentary	Scoring agreement ≠ validity; weak explainability	GPT-4 high intra-rater reliability, prompt-dependent[3]; ICC=0.972[4]; e-rater κ≈0.70–0.77[5]
Spoken-language and multimodal assessment	Speech/multimodal responses	Sub-construct scores, pronunciation feedback	ASR error; content constructs hard to measure	SpeechRater r≈0.57–0.68[6]; DET six sub-constructs[7]
Process-based assessment	Learning-process/keystroke data	Process indicators, profile	Data completeness, privacy, immature operationalization	Keystroke-log learning analytics[9]; evidence still thin[11]
Learning diagnostics and feedback	Response and error data	Attribution, remediation, path	Attribution accuracy	MathGPT grading + step-by-step tutoring[12][13]
AI-assisted item writing	Knowledge points, rubrics	Test items, distractors, rubrics	Item quality and bias; needs human review	91-class/1,700-student field study[14][15]

6.3 Product Forms

From 2024 to 2026, the technical foundation of assessment products has been shifting from single-turn conversational tools toward a combined "agentic + multimodal + on-device" form. Three observable threads run through this shift. First, a move from single-turn Q&A-style grading toward an agent architecture that can plan, invoke tools, and reason across multiple steps — letting assessment autonomously orchestrate a sequence from diagnosis to tutoring to re-practice. Second, a move from pure text toward multimodal handling — handwriting recognition, formula parsing, image understanding, spoken-language assessment — letting assessment cover the full range of real-world response formats. Third, some inference is moving on-device (AI learning tablets, AI glasses) to cut latency, protect privacy, and support offline use.

On market size, estimates of China's overall AI-in-education market diverge substantially and should be cited with caution. Grand View Research puts China's AI-in-education market at roughly US$509 million in 2024, growing to roughly US$2.846 billion by 2030 (a 2025–2030 CAGR of about 31.6%)[16]; domestic brokerage and consultancy figures tend to run much higher — for example, estimates that put China's "AI+education" market above RMB 70 billion in 2025, approaching nearly RMB 300 billion by 2030[17]. These two sets of figures differ by more than an order of magnitude, rooted in inconsistent statistical boundaries, currency, and scope for what counts as "AI+education" — the former counts narrowly defined AI software in US dollars, while the latter often folds in hardware, content, services, and even broader education-informatization budgets, denominated in RMB. **To avoid conflating incompatible figures, this Blue Book does not convert or merge the two.** One further note of caution: as a sub-segment within this space, intelligent assessment still lacks an independent, authoritative breakdown of its own market size [TBD: an authoritative independent source for the intelligent-assessment sub-market's size/product count]. In terms of product structure, current assessment products fall broadly into four categories, each discussed below with real-world examples and key metrics.

6.3.1 Intelligent Grading

Aimed at homework, exam, and essay-grading scenarios, this category delivers automated scoring, comment generation, and error-cause annotation, with teachers as the primary user and a value proposition of stripping repetitive grading out of teacher workload. Representative examples include: iFlytek's intelligent grading and spoken-English assessment system for the gaokao and zhongkao, which claims to grade a full test paper in under 15 seconds[1][13]; TAL Education's Nine Chapters LLM (MathGPT), which powers math homework/exam grading, complex word-problem grading, and Chinese/English essay grading, and is already built into the Xueersi AI learning tablet Xpad[12][13]; NetDragon's 101 Education PPT AI Teaching Assistant, which offers personalized homework customization, online grading, and spoken-language assessment, and has cumulatively served 9.7 million teachers nationwide with an install base of over 35.76 million, reaching 32 provincial-level administrative regions across China[18]; and, internationally, Khan Academy Writing Coach, which covers argumentative, expository, and literary-analysis essays and provides immediate, specific, actionable feedback on structure, argument support, introductions and conclusions, and style and tone — made free for teachers starting in February 2025[10].

Multimodal capability is fast becoming the dividing line within this category. A large share of student responses in China's basic education system exist in handwritten form — paper homework, handwritten exams, scratch-work calculations — so whether a product can reliably recognize handwritten answer sheets, parse mathematical formulas and geometric figures, and process photographed images directly determines its usable range in a real classroom. A grading tool that only handles clean, typed text has quite limited applicability in K-12 settings. The key metrics for

this category are **scoring agreement rate** (agreement with human scores) and **error-cause localization accuracy** (whether an error can be pinpointed to a specific knowledge point and passage); but publicly disclosed, third-party-verified agreement-rate data from vendors remains scarce [TBD: independent evaluation data on public scoring-agreement rates/error-cause localization accuracy for mainstream intelligent-grading products].

6.3.2 Learning Analytics and Diagnosis

Aggregating response data at the class and individual level, this category outputs knowledge-mastery levels, weak spots, and learning-status reports to support teachers' targeted instruction and professional development. It differs from intelligent grading in scope: grading concerns the "right or wrong, good or bad" of a single response, while learning analytics concerns "group- and individual-level patterns" aggregated across responses and over time. iFlytek's smart-exam solution provides post-exam learning-status feedback on top of grading; NetDragon's 101 Education PPT AI Teaching Assistant delivers real-time in-class data feedback, tracks after-class data, and assigns personalized exercises and online grading accordingly[18]; TAL Education's learning hardware is built around a "diagnose, tutor, re-practice" loop as its core[12].

This category aligns closely with the direction of national assessment reform: the *Overall Plan for Deepening the Reform of Education Evaluation in the New Era* explicitly calls for building evaluation mechanisms supported by big data and AI, encompassing multidimensional process-based, value-added, and comprehensive evaluation[19] — and learning analytics is exactly where "value-added evaluation" (which tracks a student's improvement over a period rather than an absolute score) lands as a product. Its key metrics are **diagnostic coverage** (whether it covers the knowledge and competency points required by the curriculum standards) and **robustness of mastery estimation** (whether mastery judgments for the same student are consistent and credible across different items). The risk is that once a learning-status report is presented as precise numbers and charts, it can create an "illusion of precision" — packaging a rough inference drawn from limited responses as a firm conclusion, which can mislead a teacher's instructional decisions. Learning-analytics products should therefore clearly flag the uncertainty range around their estimates rather than presenting only point estimates.

6.3.3 Adaptive Testing

This category adjusts item difficulty and path dynamically based on real-time responses, estimating ability with fewer items and providing immediate feedback. It combines the mature paradigm of computer-adaptive testing (CAT) and item response theory (IRT) with generative item-writing's capacity to supply items: CAT/IRT handles "how hard the next item should be, and how to converge on a true ability estimate with the fewest items," while generative item writing handles "supplying

enough qualified candidate items on demand" — the two complement each other and directly address the item-bank-size and exposure-control bottleneck that has long constrained adaptive testing.

The Duolingo English Test is a mature exemplar: its speaking sub-score aggregates classically scored items with IRT-based read-aloud items, scoring across six sub-constructs — content, discourse coherence, vocabulary, grammar, fluency, pronunciation[7]. The field study cited earlier, covering 91 classes and roughly 1,700 students, further demonstrated that a generative item bank can match standardized-test-level difficulty and discrimination (the AI-generated exam achieved $I_{max}\approx3.85$, reliability $R\approx0.79$)[14], lending empirical credibility to the "generative item supply plus adaptive item selection" combination. Its key metrics are **measurement efficiency** (number of items needed to reach a target reliability) and **standard error of ability estimation** (precision of the estimate). The risk is that if generative item supply lacks rigorous item-parameter calibration and exposure control, it can introduce miscalibrated difficulty, item leakage, or content bias — undermining the very item-bank quality that adaptive testing depends on. In China, adaptive-testing capability is mostly embedded within learning hardware and online practice systems rather than sold as standalone products, and a list of independent products with disclosed reliability metrics remains insufficient [TBD: a source listing independent domestic adaptive-testing products and their reliability metrics].

6.3.4 Competency Profiling

Aggregating assessment data across subjects and over time, this category builds learner profiles oriented around core competencies or ability frameworks, supporting comprehensive-quality evaluation and career development. It is the most highly aggregated of the four product categories and sits closest to high-stakes decisions: intelligent grading evaluates a single response, learning analytics evaluates a single subject, adaptive testing estimates a single ability — while competency profiling attempts to characterize a student's overall competency profile across subjects and school stages.

This form is tightly bound to national policy: the *Overall Plan for Deepening the Reform of Education Evaluation in the New Era* calls for using information technology to objectively record students' day-to-day and standout conduct as an important component of comprehensive-quality evaluation, and for folding digital literacy into that same comprehensive-quality evaluation[19]. As a result, competency-profiling products are mostly deployed at the regional/school level rather than sold to individuals, and their data often feeds into consequential processes like university admissions and merit recognition. Precisely because of that, its risks are the sharpest of the four: **first, construct representativeness** — constructs like "core competencies" and "comprehensive quality" are themselves abstract and fuzzy at the edges, and approximating them with quantifiable indicators easily slides into "measuring what's measurable as a stand-in for what isn't" (substituting attendance, competition results, and gradable behaviors for genuine character and creativity); **second, fairness** — if a profile incorporates resource-dependent indicators like extracurricular activities or competition

experience, it will systematically favor students from resource-rich families; **third, explainability and appeal rights** — when a profile factors into university admissions, students and parents have a right to know what the evaluation is based on and whether it can be appealed. Its key metrics are indicator validity, profile explainability, and data compliance. Representative products and deployment details [TBD: a source listing comprehensive-quality-evaluation/competency-profiling platform products and validity evidence].

The table below compares the four product forms side by side.

Product Form	Core Function	Primary Users	Representative Products/Vendors	Key Metrics
Intelligent grading	Automated scoring, comments, error causes	Teachers	iFlytek intelligent grading[1]; MathGPT/Xueersi AI learning tablet[12]; 101 Education PPT AI Teaching Assistant[18]; Khan Writing Coach[10]	Scoring agreement rate, error-cause localization accuracy [TBD: public agreement-rate data]
Learning analytics and diagnosis	Mastery level, learning-status reports	Teachers/instructional researchers	iFlytek smart-exam[1]; 101 Education PPT[18]	Diagnostic coverage, mastery-estimation robustness [TBD]
Adaptive testing	Dynamic item selection, ability estimation	Students/teachers	Duolingo English Test (international reference)[7]	Measurement efficiency, standard error of ability estimation [TBD]
Competency profiling	Competency aggregation, profiling	Schools/regions	[TBD: product list]	Indicator validity, explainability, compliance

Across these four product forms runs a clear spectrum — aggregation level rises, stakes rise, and verification difficulty rises in lockstep: from intelligent grading (single response, low stakes, relatively verifiable), to learning analytics and adaptive testing (subject-level, medium stakes), to competency profiling (cross-subject, cross-time, high stakes, hardest to verify). This spectrum maps directly onto the chapter's central tension — the higher the aggregation level and the greater the stakes, the more abstract the construct becomes and the harder validity and fairness are to guarantee, which is exactly where the most caution is needed. Product maturity and deployment stakes should therefore be matched: scale up boldly in relatively mature, verifiable areas like intelligent grading, but rely primarily on humans — with machines in a supporting role — in areas like competency profiling that directly affect university admissions, treating explainability and the right to appeal as

hard prerequisites. This judgment is the direct basis for the risk discussion in the next section and the recommendations in Section 6.5.

6.4 Validity and Fairness Risks

The technical feasibility of intelligent assessment is not the same thing as its educational appropriateness. A contrast that has come up repeatedly in the previous section — agreement can be very high, but agreement is not the same as validity — is the crux of this section. Five categories of risk deserve clear-eyed attention, especially in high-stakes settings.

First, construct validity: agreement does not mean measuring the right thing. This is the most fundamental risk in this section, and the one most easily masked by high-agreement numbers. Construct validity asks whether an assessment actually measures what it claims to measure (say, "writing ability" or "mathematical reasoning") rather than some related but off-target proxy. ETS's research reports have long made clear that high agreement between automated and human scoring can still coexist with construct underrepresentation (measuring only part of the target ability) or construct-irrelevant variance (counting things toward the score that shouldn't count)[5]. This risk is more insidious with generative models: they can produce fluent, seemingly well-supported commentary while actually scoring surface linguistic features — length, syntactic complexity, vocabulary richness — rather than whether an argument holds up, whether the content is accurate, or whether the thinking is deep. This is the origin of the classic critique of automated essay scoring — critics such as Perelman, analyzing NAPLAN's automated scoring, have pointed out that such systems correlate strongly with essay length, cannot evaluate meaning or argumentation, cannot verify factual accuracy, and can be fooled into awarding high scores to structurally complete but semantically nonsensical text (such as gibberish essays deliberately assembled with a generator)[20]. Handing scoring authority to a system that "scores on form" erodes an assessment's construct validity — an erosion that typically goes unnoticed in everyday low-stakes use and only surfaces its cost when the system is deliberately attacked or put to high-stakes use.

Second, fairness: systematic bias against particular linguistic backgrounds. Bias in training data and stylistic preference can produce systematic disadvantage for particular dialects, native-language backgrounds, expressive styles, or disadvantaged groups, amplifying existing educational inequity. Recent empirical work shows that automated essay scoring risks introducing bias against subgroups such as English-language learners (ELLs), and that when certain populations are underrepresented or absent in training data, representational bias interacts with pre-existing historical and measurement bias in that data to amplify unfairness toward disadvantaged groups[21]. Comparative studies on German-learner essays, and validity studies of automated scoring for elementary-age ELLs, have both documented cross-group scoring disparities[22].

For China's basic education system, this risk takes a distinctly local shape: accent and dialect affect ASR recognition quality, and urban-rural and regional differences shape students' expressive style and speech habits — both of which could systematically depress automated spoken-language and essay scores for rural students, students from ethnic-minority regions, or students with heavy dialectal accents. A comparative study of AI scoring across four Guangdong cities points to exactly this kind of regional disparity in automated-scoring results, and flags it as worth further stratified verification[8]. This means that "acceptable average agreement" can mask a problem where certain subgroups are systematically scored too low or too high — a system with a high overall QWK can still be unfair to specific subgroups. Fairness therefore needs to be evaluated by subgroup (native language, dialect, urban/rural, socioeconomic background), and a "subgroup fairness diagnostic" should be a mandatory pre-launch check for any high-stakes assessment system, not an afterthought behind an aggregate metric.

Third, explainability and the right to appeal. The process by which a score and comments are generated is hard to fully trace, leaving "why this score" without verifiable grounding — which weakens an assessment's credibility and undercuts the ability to appeal it. The black-box nature of generative models makes them harder to explain than explicit-feature models like e-rater — the latter can at least be traced back to specific grammatical/structural microfeatures[2] — whereas the former's stated "reasoning" is itself a generated artifact that may not match the model's actual internal computation. A lack of explainability directly erodes due process in high-stakes assessment.

Fourth, gaming the system and test-driven distortion (the behavioral counterpart of construct-irrelevant variance). Once scoring rules can be reverse-engineered, learners may optimize their responses to what the model prefers rather than to genuine ability; when a scoring system correlates strongly with essay length or particular sentence templates, "test-taking strategy" displaces "genuine ability," producing the distortion where "piling on long sentences, following a template, and dropping in advanced vocabulary" is enough to earn a high score. The NAPLAN case mentioned earlier — where critics used deliberately generated, structurally complete but semantically nonsensical text to obtain a high score — is the sharpest demonstration of exactly this risk[20]. Even trickier is the prospect of **bidirectional cheating**: a student uses a generative tool to draft the response, and the grading side uses a generative tool to score it — assessment can degenerate into an empty loop of "AI writes, AI judges," turning "assessment for learning" into "assessment to get through." Khan Academy Writing Coach's built-in plagiarism flag, which uses process data as corroborating evidence[10], is a product-level response to exactly this risk — but detection on the tooling side is no substitute for robustness in assessment design itself. The real fix is making the assessment task itself harder to "perform" — for instance, by emphasizing contextualization, process, and oral explanation.

Fifth, data privacy and compliance, and over-displacement of teacher professional judgment. Process-based assessment collects the most intensive learning-behavior data of any category (keystrokes, time-on-task, revision trails, conversation logs), so it carries the highest compliance risk;

and once a competency profile is used in a high-stakes decision like university admissions, the risk of misuse scales up accordingly. Beyond that, if a product defaults to treating the machine's conclusion as authoritative — presenting a machine score as the default starting point — it can quietly displace teacher professional judgment. That is both an ethical problem and a validity problem: a teacher's judgment about context, intent, and individual difference is precisely the part of the construct that current models are worst at replacing. Automation bias — people's tendency to over-trust the output of an automated system — compounds this displacement, turning what was meant to be "assistance" into de facto "control."

In sum, these five risk categories share a common technical root: generative models are good at rendering highly consistent judgments about the surface of language, while what education assessment actually needs to measure is usually the thinking, content, and competency underneath that surface. That gap underlies a basic principle: **the more consistently a technology can produce a score, the more vigilance is warranted about whether it is measuring the right thing.** This becomes especially consequential once an assessment is institutionally granted a high-stakes purpose, at which point the gap gets amplified into a systemic educational consequence. This gap between technical capability and assessment goals is the premise every recommendation below has to reckon with.

6.5 Recommendations

Toward building trustworthy systems for the assess-teach-learn loop, and drawing on the mechanisms and risks discussed above, we offer the following recommendations. These point to different responsible parties: tiered risk classification and human-machine collaboration are principles for product design and school-level deployment; assessment quality benchmarks and third-party evaluation need to be led by government and professional bodies; and data governance requires regulators and vendors to act together, upfront. The overarching principle is this: the pace of intelligent-assessment adoption should track how verifiable its validity and fairness are — move boldly where it can be verified, and proceed cautiously where it cannot yet be.

- **Maintain human-machine collaboration with teachers as the final arbiter, and classify applications by stakes.** Application boundaries should be drawn by stakes: low-stakes settings (routine homework feedback, practice-oriented spoken-language assessment, writing-process coaching) should be encouraged to scale up to reduce workload and improve efficiency; high-stakes settings (admissions exams, competency certification) should strictly confine the model's role to an "advisory party," keeping final determination with teachers and retaining human re-scoring. International high-stakes practice is generally "machine plus human" dual scoring rather than machine-only scoring — e-rater, for instance, is used jointly with human raters on GRE and TOEFL iBT writing[2][5] — and can serve as a reference for tiered design.

- **Establish assessment quality benchmarks and independent third-party evaluation.** Reliability, fairness, and explainability benchmarks and public datasets for education assessment should be developed so product quality can be compared and monitored. Benchmarks need to move beyond a single "overall agreement" metric to include: subgroup fairness diagnostics (stratified by native language, dialect, socioeconomic background)[21][22]; robustness against "gaming" (resistance to structurally complete but semantically invalid text)[20]; and construct-validity evidence (the relationship between scores and external criteria)[5]. An independent benchmark and authoritative body for China's basic-education assessment landscape is currently lacking [TBD: a source for domestic assessment-quality benchmarks/evaluation bodies/public datasets].
- **Strengthen explainability and appeal mechanisms.** Scores should come with supporting grounds attached — tied to specific rubric items and located to specific passages — with a channel for human review and appeal. Comments should be traceable back to the rubric rather than presented as a stand-alone block of generated text; for high-stakes results, the original response, model version, and scoring parameters should be retained to support review.
- **Anchor to rubrics and control randomness to improve reliability.** Empirical evidence shows that firmly anchoring scoring to an explicit rubric and lowering decoding temperature can significantly improve scoring repeatability (a fine-tuned ChatGPT model reached ICC=0.972)[4]; prompting strategy (rubric plus anchor examples) has a large effect on accuracy[3]. Products should treat "rubric alignment plus low temperature plus fixed model version" as the engineering default for high-stakes scoring, and publish their reliability metrics.
- **Put data governance and privacy protection in place upfront.** Collection, storage, and use of process-based data should follow the principle of minimal necessity and applicable compliance requirements; when competency profiles feed into comprehensive-quality evaluation, they must comply with the *Overall Plan for Deepening the Reform of Education Evaluation in the New Era*'s requirements around objective recording and misuse prevention[19], with detailed norms picked up in Chapter 7's Governance and Safety scenario.
- **Coordinate with multimodal and on-device advances to extend coverage into hard-to-assess dimensions.** Products should draw on the multimodal sensing capabilities of new terminals like AI glasses and educational robots (see this institute's companion Blue Books for detail) to bring long-standing "hard to assess" dimensions — spoken language, hands-on performance, collaboration, expression — into assessment. But the machine reliability figures for constructs like spoken language (SpeechRater correlates with humans at only $r \approx 0.57–0.68$[6]) are a reminder that the more complex the construct, the more essential human-machine collaboration is — technical feasibility should not be mistaken for assessment validity.

- **Build AI-assessment literacy among teachers and item writers, and guard against automation bias.** Intelligent assessment is not meant to remove teachers from the loop — it asks them to take on a higher-order role, shifting from "person who grades one response at a time" to "editor-in-chief of assessment evidence": able to use tools to get machine judgments efficiently, but also able to recognize where machine judgment breaks down and resist the automation bias of over-trusting automated systems. AI-assessment literacy should be built into teacher professional development and item-writing training, with a clear collaborative norm of "machine proposes candidates, teacher makes the final call." For AI-assisted item writing specifically, subject teachers must have final professional say over an item's scientific soundness, alignment with curriculum standards, and cultural appropriateness[14][15]. This is consistent with the Ministry of Education's push to use digitalization to empower teacher development[19].

In sum, the value of intelligent assessment lies not in replacing teachers' professional judgment with machines, but in freeing teachers from repetitive scoring so that generative AI's capacity for feedback at scale can be converted into timely, personalized learning support for every learner. A thread running through this entire chapter is that generative models can produce highly consistent scores, but "consistent" does not mean "valid," and "can be assessed" does not mean "should be assessed" this way. The healthy development of intelligent assessment depends on coordinated calibration across three things — technical capability, the realities of education, and governance constraints — and in high-stakes settings especially, caution and verifiability should always take priority over efficiency and scale.

Chapter References

1. iFlytek Smart Education. *Smart Exam* and *AI Listening-and-Speaking Classroom* solution pages and case studies (including claims of "gaokao listening-and-speaking across 29 provinces/municipalities, zhongkao listening-and-speaking across 132 cities, cumulative service to over 45 million test-takers," "spoken-language assessment across 64 cities for zhongkao and 13 provinces/municipalities for gaokao," "rolled out to spoken-English gaokao/zhongkao exams in 20+ provinces/municipalities") [科大讯飞智慧教育《智慧考试》《AI 听说课堂》解决方案页及案例]. iFlytek. 2024–2025. https://edu.iflytek.com/solution/examination ; https://edu.iflytek.com/solution/school/listening-and-speaking ; http://ex.chinadaily.com.cn/exchange/partners/82/rss/channel/cn/columns/snl9a7/stories/WS609b84a6a3101e7ce974ecbb.html

2. Attali, Y., Bridgeman, B., & Trapani, C. "Performance of a Generic Approach in Automated Essay Scoring." *Journal of Technology, Learning, and Assessment*, 10(3). 2010 (finds e-rater's agreement with a single human rater exceeds agreement between two human raters on TOEFL

Independent and GRE Issue tasks). https://ejournals.bc.edu/index.php/jtla ; ETS, *How the e-rater Scoring Engine Works*. https://www.ets.org/erater/how.html

3. "Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability." *Computers and Education: Artificial Intelligence*. 2024 (GPT-4 vs. GPT-3.5 vs. Claude 2; GPT-4 shows superior intra-rater reliability, ~112%/114% accuracy improvement; temperature and prompting strategy identified as key variables). https://www.sciencedirect.com/science/article/pii/S2666920X24000353

4. Yavuz, F., Çelik, Ö., & Yavaş Çelik, G. "Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments." *British Journal of Educational Technology*, 56, 150–166. 2025 (15 human raters plus ChatGPT/Bard, five-dimension rubric; fine-tuned ChatGPT ICC=0.972). https://bera-journals.onlinelibrary.wiley.com/doi/10.1111/bjet.13494

5. ETS Research Report, *Evaluation of the e-rater Scoring Engine for the TOEFL Independent and Integrated Prompts*, and related e-rater validity studies (GRE r≈0.78/0.79, κw 0.73–0.77; TOEFL r≈0.75/0.73, κw≈0.70; agreement statistics alone do not establish construct validity). ETS. 2007–2014. https://files.eric.ed.gov/fulltext/EJ1109838.pdf ; https://onlinelibrary.wiley.com/doi/full/10.1002/ets2.12005

6. "Evaluating automated evaluation systems for spoken English proficiency: An exploratory comparative study with human raters." *PLOS One*. 2025 (overview of SpeechRater/CASE/Ordinate/Duo Speaking Grader; SpeechRater correlates with humans at r≈0.57–0.68). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0320811 ; https://pmc.ncbi.nlm.nih.gov/articles/PMC11952242/

7. Duolingo English Test, *An Overview of Duolingo English Test Administration and Scoring*. Duolingo. 2024 (six spoken-language sub-constructs: content/discourse coherence/vocabulary/grammar/fluency/pronunciation; speaking sub-score = classical scoring plus IRT read-aloud aggregation). https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_test_scoring_2024_v1.pdf

8. "A Comparative Analysis of AI-Scored Results in Computer-Based English Listening and Speaking Test (CELST) Across Four Cities in Guangdong, China." *2024 International Conference on Artificial Intelligence and Future Education*. 2024 (CELST is a component of the gaokao English exam, using a machine-learning automated scoring engine trained on a human-rated calibration set). https://dl.acm.org/doi/10.1145/3708394.3708436

9. "Integrating qualitative learning analytics and retrospective reflections to support formative, process-oriented feedback in EFL writing." *Frontiers in Education*. 2026 (keystroke logging plus S-notation as learning analytics, making the writing-revision process visible). https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2026.1773621/full

10. Khan Academy. *Meet Khan Academy Writing Coach* and *Writing Coach* product and help pages. Khan Academy. 2024–2025 (process valued equally with product, stage-by-stage guided feedback, class/individual feedback summaries, time-on-stage reports, plagiarism flags; made free for teachers as of 2025-02-27). https://blog.khanacademy.org/meet-khanmigo-writing-coach-helping-learners-become-better-writers/ ; https://www.khanmigo.ai/writingcoach
11. Banihashem, S. K., et al. "A Critical Review of Using Learning Analytics for Formative Assessment: Progress, Pitfalls and Path Forward." *Journal of Computer Assisted Learning*. 2025 (evidence for using learning analytics to support formative assessment remains thin, operationalization is difficult, and teachers are reluctant to trust results that don't align with the model). https://onlinelibrary.wiley.com/doi/full/10.1111/jcal.70056
12. TAL Education Nine Chapters LLM (MathGPT) introduction and official-filing information (first education LLM in China purpose-built for mathematics and among the first to clear official filing; automated math problem-solving/complex word-problem grading, Chinese/English essay grading, AI step-by-step tutoring; deployed in the Xuexiaoban app and the Xueersi AI learning tablet Xpad; "Math Anytime" targets instant Q&A for primary/middle school, coverage rate [TBD]) [好未来九章大模型（MathGPT）介绍与备案信息]. TAL Education / China.com / GeekPark. 2023–2024. https://m.tech.china.com/digi/digi/20240419/202404191508210.html ; https://www.geekpark.net/news/337563
13. "AI grades a test paper in under 15 seconds: education empowered by AI breaks through the 'impossible triangle'" [批改一张试卷不超过 15 秒，AI 赋能教育突破"不可能三角"]. *Southern Metropolis Daily* [南方都市报]. 2024. https://m.mp.oeeee.com/a/BAAFRD0000202411251028303.html
14. Isley et al. "Assessing the Quality of AI-Generated Exams: A Large-Scale Field Study." *arXiv*. 2025 (91 classes/~1,700 students; 2PL-IRT; AI-generated items discrimination $\bar{\alpha}$≈1.3 vs. 1.2 standardized; AI exam I_{max}≈3.85/R≈0.79 outperforms standardized 2.61/0.72). https://arxiv.org/abs/2508.08314
15. "Harnessing Generative AI for Assessment Item Development: Comparing AI-Generated and Human-Authored Items." *International Journal of Selection and Assessment*. 2025 (LLM-generated items occasionally preferred by expert review, but human review remains indispensable for final quality; issues include multiple correct answers/non-functional distractors). https://onlinelibrary.wiley.com/doi/10.1111/ijsa.70021
16. Grand View Research, *China AI in Education Market Size & Outlook, 2025–2030* (approx. US$509 million in 2024, approx. US$2.846 billion in 2030, CAGR≈31.6%). Grand View

Research. 2025. https://www.grandviewresearch.com/horizon/outlook/ai-in-education-market/china

17. Duojing/brokerage-house figures on the AI+education market size (China's AI+education market exceeding RMB 70 billion in 2025, approaching nearly RMB 300 billion by 2030, among other figures; scope and boundaries are inconsistent and should be compared with caution) [多鲸/券商口径 AI+教育市场规模]. Reposted via TopMarketing / Zhihu industry research. 2025. https://itopmarketing.com/info20459
18. NetDragon 101 Education PPT and AI Teaching Assistant product information (AI Teaching Assistant: personalized homework customization/online grading/spoken-language assessment; real-time in-class and after-class data feedback with personalized exercise assignment; cumulative service to 9.7 million teachers, install base over 35.76 million) [网龙 101 教育 PPT 与 AI 助教产品信息]. NetDragon 101 Education PPT official site / Duozhi.com. 2018–2024. https://ppt.101.com/news/06202018/20180620153929242.shtml ; http://www.duozhi.com/company/201804247108.shtml
19. CPC Central Committee and State Council, *Overall Plan for Deepening the Reform of Education Evaluation in the New Era* [深化新时代教育评价改革总体方案], and Ministry of Education press briefing (calls for building evaluation mechanisms supported by big data and AI; multidimensional process-based/value-added/comprehensive evaluation; using information technology to record and incorporate conduct into comprehensive-quality evaluation, folding digital literacy into comprehensive-quality evaluation). Issued 2020 / interpreted 2024. https://www.eol.cn/zhengce/jiedu/202410/t20241013_2636778.shtml
20. Perelman, L., analysis of automated essay scoring and NAPLAN (AES correlates strongly with essay length, cannot evaluate meaning or argumentation, cannot verify factual accuracy, can be fooled by structurally complete but semantically nonsensical text into awarding high scores). Australian Education Union (AEU) and related sources. Approx. 2017–2018. Summary confirmed via WebSearch retrieval (original PDF link no longer accessible; conclusions based on search results and existing literature) [待补：稳定可访问的原文链接] [TBD: a stable, accessible link to the original text].
21. "Assessing fairness in finetuned scoring models with demographically restricted training data" and related reviews (risk of introduced bias against subgroups such as ELLs; when populations are underrepresented or absent in training data, representational bias interacts with pre-existing historical/measurement bias to amplify unfairness). *Studies in Educational Evaluation /

ScienceDirect*. 2024–2026. https://www.sciencedirect.com/science/article/pii/S1075293526000206

22. "Fairness in Automated Essay Scoring: A Comparative Analysis of Algorithms on German Learner Essays" (ACL BEA 2024), and "Validity of automated essay scores for elementary-age English language learners: Evidence of bias?" (evidence of cross-group scoring disparities). ACL Anthology / *Studies in Educational Evaluation*. 2024. https://aclanthology.org/2024.bea-1.18/ ; https://www.sciencedirect.com/science/article/abs/pii/S1075293524000084

Chapter 7 Governance and Safety (A New Frame): Compliance, Privacy, and Academic Integrity

7.1 Framing the Problem: From a Capability Narrative to a Responsibility Narrative

Layered on top of the four scenarios established in the first six chapters of this Blue Book — Empowering Teaching, Supporting Learning, Supporting Research (Professional Development), and Intelligent Assessment — governance and safety form a fifth scenario, one that runs through all the others rather than sitting alongside them. It is not another category of application; it is the **constraint layer and precondition layer** that determines whether the first four categories can be deployed legally, credibly, and sustainably. The 2024 edition of this report treated conversational LLMs as the main storyline and tucked compliance and ethics into a single paragraph under "recommendations for development." By 2026, product form factors have shifted decisively toward agentic, multimodal, and on-device architectures, and the governance challenge has scaled up accordingly. An education agent that can plan autonomously, invoke tools, retain memory across sessions, and drive wearable hardware (see our companion report, the *AI Glasses in Education Industry Blue Book 2026*) generates data flows, decision chains, and lines of accountability that a compliance framework built for simple question-and-answer exchanges simply cannot cover. That is why this chapter treats governance and safety as a stand-alone subject — and why it functions as a **veto-level dimension** in product evaluation and procurement decisions.

The chapter's central claim is this: by 2026, the primary bottleneck constraining generative AI education products at scale is shifting from insufficient model capability to insufficient governance capability. We call this the **capability–compliance gap** — the extent to which a product's technical capabilities have outrun the compliance, privacy, and integrity safeguards that education settings require — and it is the analytical thread running through the entire chapter. The data bear this out directly: a nationally representative survey by the Center for Democracy & Technology (CDT) covering the 2024–2025 school year found that 85% of teachers and 86% of students reported using AI in the past school year, yet fewer than half of either group said their school had provided any AI-related training or guidance; only one in ten teachers reported having received training on how to respond when a student's AI use appears to be harming their well-being (CDT, 2025). Capability is spreading far faster than governance mechanisms can be built — that gap is the starting point for everything that follows in this chapter.

7.1.1 Three Layers of What Governance Must Cover

- **Data layer**: the collection, storage, circulation, and reuse of data on learner identity, behavior, biometric traits (voice, face, gaze), and inferred psychological states.
- **Model layer**: the training-data provenance, bias, hallucination, and explainability of foundation models and education-vertical models, plus the traces left by agent memory and tool invocation.
- **Application layer**: the boundaries of real classroom use, including special protections for minors, the informed consent and choice afforded to teachers and students, and how responsibility is divided between humans and machines.

These three layers are coupled to one another. Moving processing on-device eases privacy pressure at the data layer — less raw data travels to the cloud — but it introduces a new problem at the model layer: local models that are hard for anyone to audit. Multimodal interaction enriches what happens at the application layer, but it substantially widens the surface for sensitive-data collection at the data layer. The four themes that structure the rest of this chapter — multi-jurisdictional compliance (7.2), privacy and data protection (7.3), academic integrity (7.4), and accountability and human–machine collaboration (7.5) — cut across these three layers, and Section 7.6 pulls them together into a governance-maturity scale that is procurable and comparable across products.

7.1.2 Why Governance Earns Its Own Scenario in 2026

Elevating governance and safety from a footnote under "recommendations" to a scenario in its own right reflects three structural shifts, each of which has exposed the inadequacy of the old framework.

First, **the product itself has moved from tool to actor**. A conversational product is a passive question-answering tool: its data flow is a single closed loop of "user input, model output," and the chain of responsibility is straightforward. An agent — once it can plan autonomously, invoke tools, and retain memory across sessions — begins acting as a quasi-independent party on behalf of teachers and students: querying a learning-analytics database, calling a grading API, revising a study plan, pushing updates to parents. Every tool call is its own discrete act of data processing and decision-making. The notice-and-consent mechanisms the old framework designed for a single question-and-answer exchange cannot cover this kind of unfolding, dynamic chain of actions.

Second, **the collection boundary has expanded from the input box to physical space**. A text-only product processes only what a user actively types. Multimodal and on-device hardware — wearables, always-on cameras and microphones — passively capture voice, images, faces, and gaze from the physical environment, and that capture reaches third parties who are not the product's user at all. Data collection shifts from something actively provided to something passively captured, and the validity of consent erodes accordingly.

Third, **the cost of governance has moved from after-the-fact remediation to a precondition for market entry**. Between 2023 and 2025, China's Interim Measures and Labeling Measures, the EU

AI Act, and the revised COPPA rule in the United States all took effect in succession. Compliance has shifted from a soft constraint that could be fixed after launch to a hard threshold that must be cleared before launch: filing, safety assessment, content labeling, whitelist review — missing any one of these can keep a product out of the market or out of schools altogether. This is the change that turns governance from a "nice to have" into a "pass/fail," and it is the fundamental reason this chapter designates it a veto-level dimension in procurement.

7.2 The Compliance Landscape: Product Obligations Under Overlapping Jurisdictions

Generative AI education products face a rule-making environment that is multi-jurisdictional, multi-layered, and evolving quickly. A product that serves China's K12 and higher-education markets while also expanding into the EU or North America must satisfy several independent — and at times conflicting — compliance chains simultaneously. This section works through that landscape in the order China, the EU, the United States, and international bodies; every document number, year, and effective date is drawn from official or authoritative sources located in the course of this research.

7.2.1 The Layered Structure of China's Rules

China's regulation of generative AI education products follows a three-tier structure: overarching law, dedicated rules, and industry-specific access requirements.

Overarching law. The Personal Information Protection Law of the People's Republic of China (adopted at the 30th session of the Standing Committee of the 13th National People's Congress on August 20, 2021, and effective November 1, 2021 — PIPL) establishes the basic framework for processing personal information. Several provisions bear directly on education products. Article 28 explicitly classifies "personal information of minors under the age of 14" as **sensitive personal information**, alongside biometric data, medical and health information, and location tracking. Article 29 requires that processing sensitive personal information obtain the individual's **separate consent**, or written consent where laws and administrative regulations so require. Article 31 requires that processing the personal information of a minor under 14 obtain the consent of a parent or other guardian and that the processor establish **dedicated rules for handling such information** (Standing Committee of the National People's Congress, 2021). In practice, this means that any education product serving primary school and lower-secondary students that collects biometric traits such as face, voice, or gaze — or that makes psychological inferences about minors — will almost always trigger the combined, high-intensity compliance obligations that apply to sensitive information and to minors at once.

At the level of administrative regulation for minor protection, the Regulation on the Protection of Minors in Cyberspace (State Council Order No. 766, adopted at the 15th executive meeting of the

State Council on September 20, 2023, effective January 1, 2024) is China's first dedicated, comprehensive piece of legislation on protecting minors online. It sets out systematic rules on online content standards, personal information protection, and prevention of internet addiction; it requires personal information processors to strictly control access to minors' personal information and to conduct compliance audits; and it folds internet-literacy education into schools' quality-education curricula (State Council, 2023). The regulation complements PIPL Article 31: PIPL sets the baseline for processing personal information — guardian consent plus dedicated processing rules — while the regulation adds scenario-specific requirements around age-appropriate content, addiction prevention, and cyberbullying. Education products must satisfy both. In addition, the Data Security Law (effective September 1, 2021) provides an overarching basis for data classification and tiered security obligations; the portion of education data involving personal information of minors at scale may trigger a higher tier of data-security protection [TBD: whether education data has been specifically designated as "important data" under any local catalog].

Dedicated rules. The Interim Measures for the Administration of Generative Artificial Intelligence Services (jointly issued by seven agencies — the Cyberspace Administration of China, the National Development and Reform Commission, the Ministry of Education, the Ministry of Science and Technology, the Ministry of Industry and Information Technology, the Ministry of Public Security, and the National Radio and Television Administration — on July 13, 2023, effective August 15, 2023) is China's core dedicated rule for generative AI. Its key obligations for education products include: training data must come from lawful sources, respect intellectual property, and take measures to improve data quality; services with public-opinion attributes or the capacity for social mobilization must undergo a safety assessment and complete algorithm filing; generated images, video, and similar content must be labeled in accordance with the Provisions on the Administration of Deep Synthesis Internet Information Services; and providers must "take effective measures to prevent minors from over-relying on or becoming addicted to generative AI services" (Cyberspace Administration of China et al., 2023). As of December 31, 2025, a cumulative total of 748 generative AI services had completed filing with the Cyberspace Administration of China, and 435 generative AI applications or features had completed registration; 446 new service filings were added in 2025 alone (Cyberspace Administration of China, 2026) — the sheer growth in filing volume is itself direct evidence of how forcefully this rule is being enforced.

Content-labeling rules have been further refined in recent years. The Measures for Labeling AI-Generated and Synthesized Content, jointly issued by the Cyberspace Administration of China, the Ministry of Industry and Information Technology, the Ministry of Public Security, and the National Radio and Television Administration, took effect September 1, 2025; the accompanying mandatory national standard, **GB 45438-2025, "Cybersecurity Technology — Methods for Labeling AI-Generated and Synthesized Content"** (published February 28, 2025, implemented September 1, 2025 alongside the Labeling Measures), establishes the technical baseline for explicit and implicit

labeling. Explicit labels are human-perceptible text prompts or standard symbols placed at the start, end, or middle of text, or embedded in audio, images, and video; implicit labels include information written into file metadata and digital watermarks embedded in generated content (Cyberspace Administration of China et al., 2025; GB 45438-2025). For education products, this means AI-generated homework explanations, model essays, practice questions, and read-aloud audio must all be labeled as required by law, and any platform that also functions as a distribution channel — a community forum, a homework-showcase feature — is separately obligated to verify implicit metadata labels and detect explicit labels or generation traces in third-party content.

Education-sector access requirements. Layered on top of the general AI rules are dedicated access requirements specific to schools. In May 2025, the Ministry of Education's Committee for Guiding Basic Education Teaching issued the Guidelines on the Use of Generative Artificial Intelligence by Primary and Secondary School Students (2025 Edition) and the Guidelines on General AI Education for Primary and Secondary Schools (2025 Edition). The former establishes **grade-band-specific rules of use**: primary school students are barred from independently using open-ended content-generation features, though teachers may make appropriate use of them to assist instruction in class; lower-secondary students may explore the logical structure of generated content to a limited degree; upper-secondary students are permitted to undertake inquiry-based learning that engages with the underlying technical principles. The guidelines require schools to build a **"whitelist" system** for generative AI tools — subjecting each tool to rigorous review and evaluation and admitting to campus only those tools that meet pedagogical needs and satisfy data-security requirements — and to build an end-to-end safeguard mechanism spanning data security, ethical review, content oversight, and risk prevention (Committee for Guiding Basic Education Teaching, Ministry of Education, 2025).

Tier	Type of Rule	Core Obligation for Education Products	Representative Instrument (Document No. / Year)
Overarching law	Data and personal information	Lawful basis, data minimization, separate consent for sensitive information	PIPL (effective 2021.11.1), Articles 28/29
Overarching law / administrative regulation	Minor protection	Special protection for minors' information, guardian consent, addiction prevention	PIPL Article 31; Regulation on the Protection of Minors in Cyberspace (State Council Order No. 766, effective 2024.1.1)
Dedicated rule	Generative AI services	Algorithm filing, safety assessment, training-data compliance, addiction prevention for minors	Interim Measures for the Administration of Generative Artificial Intelligence Services (effective 2023.8.15)
Dedicated rule	Labeling of	Explicit + implicit labeling	Measures for Labeling AI-

	generated/synthesized content	(metadata / digital watermark)	Generated and Synthesized Content + GB 45438-2025 (effective 2025.9.1)
Education sector	K12 campus deployment	Whitelist admission, grade-band use rules, pedagogical fit, end-to-end safeguards	Guidelines on the Use of Generative Artificial Intelligence by Primary and Secondary School Students (2025 Edition)

Table 7-1 The Tiered Structure of China's Compliance Rules for Generative AI Education Products

It is worth stressing that education-sector rules are typically layered on top of general AI rules, producing a dual threshold of "**general compliance plus industry admission**." A conversational product that is fully compliant in the general consumer market still has to clear whitelist review, age-appropriateness screening, pedagogical-fit review, and parent-and-school notification before it can enter a K12 campus. The evolving practice of campus admission governance can usefully be read alongside the discussion of hardware admission to campuses in our companion report, the *Global Blue Book on the Development of Educational Robots 2026*.

This layered structure carries three practical implications for product teams. First, **filing and registration are a precondition for entering the domestic market, not an option**. The Interim Measures bring generative services with "public-opinion attributes or the capacity for social mobilization" within the scope of safety assessment and algorithm filing, and public-facing conversational education products typically fall within that scope. Filing volume grew from a few dozen products in early 2024 to 748 by the end of 2025 — a trajectory that reflects how substantively the regulator is enforcing this requirement. Product teams must complete filing before a feature launches and must build the filing number into their compliance disclosures. Second, **the content-labeling obligation runs through every step of the generation chain**. Any AI-generated model essay, worked solution, spoken narration, or diagram must carry both explicit and implicit labels under GB 45438-2025, and a product that also functions as a distribution platform — a community feature, a homework showcase — additionally has to verify labels on content it did not itself generate; this is a compliance blind spot for many education products today. Third, **the whitelist system pushes part of the review burden down to schools**. That means a vendor cannot simply assert its own compliance — it has to hand schools a review package they can actually evaluate: a data-flow description, a privacy policy, proof of filing, and an age-appropriateness rationale. Without that package, getting onto a school's procurement list is difficult.

7.2.2 The European Union: Horizontal Legislation Built Around Risk Tiers

The EU AI Act (Regulation (EU) 2024/1689, hereafter the EU AI Act) is the world's first comprehensive, horizontal piece of AI legislation, and it takes a risk-tiered regulatory approach. Its

timeline matters a great deal for education products: the Act **formally entered into force on August 1, 2024**; the **prohibited AI practices** provision (Article 5) and the AI-literacy obligation apply from **February 2, 2025**; obligations and governance rules for general-purpose AI (GPAI) models apply from August 2, 2025; and the Article 50 transparency obligations, along with most high-risk obligations, were originally scheduled to apply from August 2, 2026 (European Commission, 2024).

Two provisions of the Act constitute hard constraints for education products.

First, **the prohibited-practices provision names education explicitly**. Article 5(1)(f) prohibits the use of emotion-recognition systems in **workplaces and educational institutions**, carving out only a very narrow safety or medical exception (artificialintelligenceact.eu, Article 5). In practical terms, any feature that infers a student's "attentiveness" or "emotional state" from facial expression or vocal tone and then acts on that inference is, in principle, prohibited in EU education settings — a bright line for the large number of multimodal education products currently marketed around "emotion sensing" or "attention detection."

Second, **several categories of education AI are classified as high-risk**. Annex III, point 3, designates the following education AI systems as high-risk: systems used to determine access or admission, or to assign individuals to educational institutions or vocational-training programs at any level; systems used to evaluate learning outcomes, including outcomes used to steer the learning process; systems used to assess the appropriate level of education an individual can access; and systems used to monitor and detect prohibited student behavior during tests (artificialintelligenceact.eu, Annex III). High-risk systems carry a full slate of obligations covering risk management, data governance, technical documentation, log retention, human oversight, and accuracy and robustness. Put simply, any education AI that touches the "admissions sorting, learning evaluation, or exam proctoring" chain falls into the EU's heaviest compliance tier.

It is worth noting that in 2026, the EU adjusted this timeline via its **Digital Omnibus** simplification package, pushing several deadlines back. Under the provisional political agreement reached by the Council and the European Parliament on May 7, 2026 (subsequently confirmed by the Parliament on June 16 and the Council on June 29), obligations for independently deployed Annex III high-risk systems are postponed to **December 2, 2027**, and obligations for Annex I systems embedded in regulated products are postponed to **August 2, 2028**. AI systems already placed on the market before August 2, 2026 get a four-month grace period, until December 2, 2026, on the Article 50(2) watermarking obligation, but the broader Article 50 transparency obligation — disclosing to users that they are interacting with AI — remains on its original schedule of August 2, 2026 (Gibson Dunn, 2026; Inside Privacy, 2026). This postponement does not relax the substantive obligations that apply to education; it simply extends the runway for compliance preparation. The Article 5 red lines, including the emotion-recognition ban, are unaffected by the delay and have been in force since February 2025. The Digital Omnibus also adds a new prohibition to Article 5 covering AI-generated

non-consensual intimate imagery ("nudifiers") and child sexual abuse material — an additional red line that any education product with image-generation capability aimed at minors needs to build filtering against.

For Chinese products expanding into the EU, the combined effect of these two provisions matters most: first, **the emotion-recognition ban runs directly counter to the "emotion sensing / attention detection" selling point common in the domestic market** — the same multimodal product that may need only separate consent to collect this data domestically will find the entire feature category prohibited in EU education settings, requiring region-specific feature gating; second, **the compliance cost of the high-risk obligations is substantial** — once a product enters the admissions-sorting, learning-outcome-evaluation, or exam-proctoring chain, it must stand up a risk-management system, produce technical documentation, retain operational logs, implement human oversight, and pass conformity assessment. This cost structure is far heavier than domestic filing and is a variable that has to be evaluated up front in any decision to expand into the EU.

7.2.3 The United States: A Patchwork of Rules in the Absence of a Unified Federal AI Law

At the federal level, the United States **has no unified AI statute**. Compliance for education AI rests mainly on existing privacy law, layered with state legislation and executive orders — a genuinely fragmented picture.

Existing federal privacy law. Two statutes form the foundation for education data. The first is the Family Educational Rights and Privacy Act (FERPA, 20 U.S.C. § 1232g; implementing regulations at 34 CFR Part 99), which protects students' "education records" and generally prohibits disclosing them to a third party without written consent from a parent or an adult student, absent the "school official exception" (34 CFR § 99.31(a)(1)). In practice, the most common FERPA violation occurs when a teacher pastes an education record containing personally identifiable student information directly into a general-purpose tool like ChatGPT, Claude, or Gemini that has not signed an enterprise agreement with the school (Future of Privacy Forum, 2024). The second is the Children's Online Privacy Protection Act (COPPA, 15 U.S.C. § 6501 et seq.), which applies to online services directed at, or that knowingly collect personal information from, children under 13. The Federal Trade Commission (FTC) finalized amendments to the COPPA Rule in January 2025 that strengthen protections for children's data — providers can no longer obtain default consent for advertising uses and must obtain and document a parent's affirmative consent for each separate use (FTC, 2025). Under COPPA, consent given by a school on a student's behalf covers only the school-authorized educational purposes.

State legislation. At the state level, the pattern is rapid legislating followed by repeated revision. Colorado is a case in point: its Artificial Intelligence Act (SB24-205, the Colorado AI Act — the

nation's first comprehensive state-level AI law), enacted in May 2024, originally centered on "high-risk AI systems" and preventing algorithmic discrimination. The law was substantially revised before it ever took effect: the governor signed an amended version in May 2026 that pushes the effective date back to January 1, 2027, and reframes the law around transparency for "automated decision-making technology" (ADMT) in "consequential decisions," dropping the original "high-risk AI system" language altogether (Skadden, 2026; EPIC, 2026). There is friction between the federal government and the states as well: as of 2025, the federal government has taken a position opposed to state-level AI regulation, and a December 2025 executive order names Colorado's law specifically, directing the Department of Justice to establish an "AI Litigation Task Force" to identify and challenge state AI laws (Cooley, 2026).

Executive orders and federal action plans. In April 2025, the United States signed Executive Order 14277, "Advancing Artificial Intelligence Education for American Youth," establishing a White House AI Education Task Force to advance public–private partnership in K12 AI education. In July 2025, the administration released "Winning the AI Race: America's AI Action Plan," proposing more than 90 federal actions organized around three pillars — accelerating innovation, building infrastructure, and leading internationally (The White House, 2025). These executive initiatives are oriented toward promoting adoption, and together with the privacy-law constraints described above, they make up the regulatory environment for education AI in the United States.

The boundaries of the school official exception. FERPA permits schools to disclose education records — without obtaining consent case by case — to a "school official" performing an institutional function, and an ed-tech vendor designated as a school official may process data on that basis. But the exception applies only under strict conditions: the vendor must (1) perform a function the school would otherwise perform itself, (2) remain under the school's direct control — particularly regarding the use and re-disclosure of the data — and (3) use the data solely for the authorized educational purpose, with no secondary use permitted. On this basis, a teacher who pastes a student's education record into a general-purpose tool that has not signed an enterprise agreement — one whose terms permit the vendor to use the data to train its models — satisfies neither the "direct control" nor the "purpose limitation" condition, and this is a textbook FERPA violation (Future of Privacy Forum, 2024). The direct implication for product teams: a product aimed at U.S. schools must offer enterprise-grade contract terms that limit use to the educational purpose, exclude use in model training, and place the vendor under the school's control — otherwise the school has no lawful way to adopt it through the school official exception.

For products expanding abroad, the practical takeaway from the U.S. landscape is this: **federal privacy law (FERPA/COPPA) is the most stable part of the compliance foundation, while state-level AI law remains in a state of considerable flux**, and a product's data processing agreement (DPA) and contract terms need to be adaptable across states. A sound strategy is to set the baseline at the strictest jurisdiction: align directly with FERPA/COPPA and the EU's high standards on data

minimization, purpose limitation, exclusion from training, and deletability, then apply feature gating only for the handful of jurisdiction-specific prohibitions (such as the EU's emotion-recognition ban) — avoiding the need to rebuild a separate compliance architecture for every state or country.

7.2.4 International Bodies: Non-Binding but Widely Influential Governance References

UNESCO published the world's first Guidance for Generative AI in Education and Research in September 2023, arguing for a human-centered approach and recommending that the minimum age for students to use generative AI tools independently be set at **13**, with younger students permitted to use such tools only under teacher or parent supervision (UNESCO, 2023). In September 2024, UNESCO followed with the AI Competency Framework for Students and the AI Competency Framework for Teachers, building out a literacy framework around human-centered thinking, AI ethics, technical application, and system design (UNESCO, 2024). Although these documents are not legally binding, they serve as widely cited governance references for education authorities and product teams around the world, and their principles — a minimum age, humans in the loop, literacy before deployment — have already been absorbed into national policies, including the grade-banded Chinese guidelines discussed in this section.

7.2.5 A Cross-Jurisdictional Minimum Compliance Checklist for Product Teams

Drawing all of the above together, education products aimed at the domestic market with potential for international expansion can distill the following "minimum intersection" of obligations:

1. **Algorithm filing and safety assessment**: complete the algorithm filing — and, where applicable, the safety assessment — required by the Interim Measures in China.
2. **Training-data compliance**: source training data lawfully, avoid infringing intellectual property, exclude non-compliant content, and retain evidence of provenance.
3. **Generated-content labeling**: apply both explicit and implicit labels to generated content as required by the Labeling Measures and GB 45438-2025.
4. **Special protection for minors**: obtain guardian consent for users under 14 (under China's PIPL) or under 13 (under COPPA/UNESCO guidance); adopt dedicated processing rules, age-appropriate content tiers, and addiction-prevention design.
5. **Sensitive information and the emotion-recognition red line**: obtain separate consent for biometric data and psychological inference; disable emotion recognition in EU education settings.
6. **Auditable trace records**: ensure an agent's decision chain, tool calls, and memory updates are traceable, providing the evidentiary basis needed for high-risk scenarios under the EU's Annex III and for after-the-fact accountability more generally.

7.3 Privacy and Data Protection: New Problems Introduced by Multimodal, On-Device, and Agentic Design

7.3.1 The Expanding Collection Surface: The Privacy Cost of Going Multimodal

As products move from plain text to multimodal interaction, the type of data they collect expands from "text a user actively types" to "voice, images, faces, and gaze passively captured from the environment." Wearables and always-on microphones and cameras raise this concern most acutely — what they capture is rarely limited to the intended user; it typically sweeps in third parties who happen to share the room, whether other students, teachers, or family members. This is a "**bystander privacy**" problem. The compliance basis, disclosure method, and data boundaries for this class of hardware need to be resolved up front, at the design stage; a fuller analysis of these hardware form factors appears in our companion report, the *AI Glasses in Education Industry Blue Book 2026*.

Under Chinese rules, voice and facial data qualify as biometric information, while gaze and emotional inference may fall under sensitive personal information; collection must satisfy the minimum-necessity standard and obtain separate consent, with guardian consent and dedicated processing rules layered on top for minors. In the EU, emotion recognition in education settings is directly prohibited by Article 5 of the AI Act. There is, in other words, a direct tension between showcasing multimodal capability and the compliance cost it carries: each additional passively collected modality raises the compliance burden by something close to a full step, not a marginal increment.

The bystander-privacy problem is especially hard to solve in education settings. The traditional notice-and-consent mechanism presumes the user is actively choosing to engage with the product, but classmates, hallway passersby, and family members swept up by an always-on device have no legal relationship with the product at all — there is no one to notify and no consent to obtain. Three mitigation paths are available. The first is **technical suppression at the point of collection**: for instance, blurring faces on-device, suppressing non-target voiceprints, and retaining only structured features rather than raw imagery. The second is **boundary-setting at the scenario level**: physical indicator lights, recording notices, and defined capture zones that reduce the likelihood a bystander is captured at all. The third is **institutional consent at the venue level**: a unified notice and public posting, issued by school or classroom administrators, describing the scope of collection. No single one of these measures eliminates the risk on its own; product design needs to combine them, and bystander-protection measures should be included in the admission package schools use for review.

7.3.2 Sensitive Inference: From "What Was Collected" to "What Was Inferred"

The privacy risk of generative products extends beyond raw data collection into **secondary inference**: inferring cognitive level from answering behavior, inferring emotional state from vocal tone,

inferring mental-health risk from interaction patterns. These inferences are themselves highly sensitive personal information, and they are frequently generated, stored, and used to drive personalized decisions without the user's knowledge. The CDT 2024–2025 survey offers direct evidence that this risk is already materializing: nearly a third of students (31%) reported having had **non-academic, back-and-forth personal conversations** with AI on a device or tool provided by their school (CDT, 2025). Once conversations of this kind are used for emotional or psychological inference, the sensitivity compounds with the fact that the user is a minor, creating a very high compliance and ethical risk. Emotional and psychological inference about minors demands particular caution, and several jurisdictions have already placed direct limits or outright bans on emotion recognition in education settings (see Section 7.2.2).

7.3.3 The Double-Edged Effect of Moving On-Device

On-device inference (discussed as a trend in the product landscape in Chapter 3 of this Blue Book) keeps part of the data processing local to the device, which objectively reduces how much raw data is exposed by moving to the cloud — a genuine step forward for privacy. Researchers have recently proposed architectures combining federated learning with differential privacy that let "data stay on the device while only encrypted, noised model updates are uploaded" — an approach that shows promise for keeping raw student video and behavioral data from being centrally aggregated (Frontiers in Computer Science, 2025).

But moving on-device introduces two new problems of its own. First, local models and local data are hard for third parties to audit — "**privacy**" can shade into "**invisibility**": regulators, schools, and even parents have no good way to verify what an on-device model has actually inferred or retained. This is a directional paradox: going on-device trades away cloud exposure in exchange for privacy, but "data that cannot be externally verified" is a side effect of that same trade, and the privacy gain and the audit loss come from the same source. Researchers are exploring blockchain-based tamper-proof audit trails to close this "audit gap," but the work remains at the research stage and is far from a standard product feature (Preprints.org, 2025). Second, in hybrid edge–cloud architectures, exactly what data moves to the cloud, and when, and how it is governed once it gets there, is often left vague; the "on-device" label is frequently used as a marketing claim rather than a substantive privacy commitment. In a hybrid architecture where raw audio is preprocessed on-device and inference then runs on a large cloud model, the raw speech may be converted to text or features locally, but the data that actually carries semantic meaning still travels to the cloud — the privacy benefit is far smaller than the "on-device" label implies. When evaluating an on-device product, then, **"on-device" should not be treated as synonymous with "private and secure."** A product should be required to state three things explicitly: the edge–cloud data boundary (which data is processed locally, which moves to the cloud), the local retention policy (what is stored locally, for how long, and how it is deleted), and the audit mechanism (how an external party can verify the first two). Only when all three are in

place does "on-device" carry substantive privacy value; absent that, it should be treated as marketing language.

7.3.4 A Practical Checklist for Data Minimization and Purpose Limitation

- **Minimize collection**: collect only the data types and granularity necessary to deliver the educational function; justify necessity explicitly for every additional modality.
- **Limit purpose**: education data must not be used for ad targeting, cross-product tracking, or other non-educational purposes (consistent with the 2025 COPPA amendments' tight restriction on advertising uses and with China's purpose-limitation principle).
- **Set retention limits**: specify a retention period for each data category and a mechanism for deletion once it expires.
- **Enable portability and deletion**: give teachers and students an operable path to view, export, and delete their own data.
- **Govern agent memory**: writes, reads, and forgetting of long-term memory must be controllable, explainable, and clearable (see Section 7.5).
- **Sign a data processing agreement (DPA) up front**: put a DPA in place with the school before deployment, clarifying data ownership and responsibility — CDT-style surveys repeatedly reveal that many school districts are using AI tools without a DPA in place with the vendor, which is itself a concrete symptom of the governance gap.

7.4 Academic Integrity: From a Detection Arms Race to Reshaping Assessment

7.4.1 Reframing the Problem

The impact of generative AI on academic integrity was initially framed narrowly as "how do we detect AI-written text." But AI text detectors suffer **structural limitations** in accuracy, false-positive rate, and resistance to evasion, which makes betting integrity governance on detection tools an unstable strategy. The most compelling evidence comes from a Stanford team's 2023 study published in *Patterns*: seven leading AI detectors misclassified 61.3% of human-written text by non-native English speakers (TOEFL test-takers) as AI-generated; 97.8% of these texts were misjudged by at least one detector, and 19.8% were unanimously misjudged by all seven (Liang et al., 2023, *Patterns*). This "systematic bias against non-native writers" has produced real consequences: Turnitin's own published research has documented higher false-positive rates for non-native English speakers, and multiple universities — including Vanderbilt, Yale, Johns Hopkins, and Northwestern — have **disabled Turnitin's AI-detection feature entirely** as a result. Detectors that are unreliable,

easy to evade through paraphrasing, and prone to disproportionately harming disadvantaged groups make a "prohibit and detect" strategy hard to sustain.

The reason detectors fail is structural, not something engineering can tune away. The output distribution of a generative model is, by design, an approximation of the distribution of human language, and as models keep improving, the two distributions increasingly overlap — any classifier built on statistical features like "perplexity" or "burstiness" faces declining separability over time. Meanwhile, the spread of "paraphrase and humanize" tools has pushed the cost of evasion toward zero. Worse, the false positives are not random: detectors tend to flag text with tidy word choice, simple sentence structure, and few unusual expressions as AI-generated — and that is precisely the linguistic profile of many non-native writers and some neurodivergent students, which means the harm falls systematically on already-disadvantaged groups. Building academic sanctions on a tool that is "biased against disadvantaged groups, easily evaded, and getting less accurate as models improve" is hard to justify on either procedural-fairness or educational-equity grounds.

The deeper question is this: when writing, coding, and problem-solving can all be produced to a high standard by AI, what should education even be **assessing**, and **how**? This connects directly to the Intelligent Assessment scenario discussed in Chapter 6 — integrity governance is, at bottom, a question about assessment design. The direction for reshaping assessment is to move the center of gravity away from "final products that AI can generate in one pass" and toward "the process and understanding that AI cannot easily replace": process-based assessment that tracks how a draft evolves, the trail of revisions, and periodic reflection; performance-based assessment that brings in live oral defense, on-the-spot problem-solving, and hands-on demonstration; and inquiry-based assessment that treats AI as an openly acknowledged collaborator, requiring students to show the full arc of questioning, follow-up, critique, and re-creation in their finished work. None of this excludes AI — it folds "using AI in a way that is evidence-based, transparent, and responsible" into the assessment target itself, turning integrity from something that gets detected into something that gets cultivated.

7.4.2 Comparing Three Governance Orientations

Orientation	Core Mechanism	Strengths	Limitations
Prohibit and detect	AI detectors, use bans	Clear boundary in the short term	Unreliable detection, disproportionate harm to non-native speakers, evadable, not sustainable
Permit and disclose	Require disclosure of how and how much AI was used; retain conversation logs	Builds transparent habits of use; verifiable	Depends on good-faith disclosure; verification is costly
Redesign for immunity	Redesign assessment so AI is hard to substitute for	Resolves the problem at its	High cost to design and

	(process-based / performance-based / inquiry-based)	root	administer

Table 7-2 Three Governance Orientations for Academic Integrity, Compared

International practice is moving quickly toward a combination of "disclosure plus redesign." Australia's Tertiary Education Quality and Standards Agency (TEQSA) has made "assessment reform" its core strategy for responding to generative AI, arguing for a shift away from easily-copied outputs and toward authentic, process-based assessment. A growing number of universities now require students to attach an **AI use statement** (an AI acknowledgement or disclosure) at the point of submission, describing where and to what extent AI was involved, and some courses require students to retain their AI conversation logs for possible verification, treating undisclosed use as academic misconduct (TEQSA, 2025). This chapter favors the combination of "disclosure" and "redesign": build transparency through an **AI-use disclosure standard**, and reduce the weight placed on final text output through **process-based, performance-based, and inquiry-based assessment** — shifting the center of gravity from the product to the process and the understanding behind it.

7.4.3 Integrity-by-Design Recommendations for Product Teams

- **Transparent, not concealed**: products should log and allow export of where and how much AI was involved, supporting disclosure by teachers and students rather than helping to evade it (this aligns with the direction of the explicit/implicit labeling obligations under the Labeling Measures).
- **Scaffold, don't do the work**: products aimed at learners should default to guiding, probing, and step-by-step prompting rather than handing over a finished answer (consistent with the mechanism design for "Supporting Learning" in Chapter 5, and with the grade-banded Chinese guidance that bars independent use of open-ended generation in primary school).
- **Keep the teacher in the loop**: final judgment on integrity questions should remain with the teacher, with the product supplying evidence rather than a verdict — especially given how high detector false-positive rates run, no "AI probability score" should ever serve as the sole basis for a disciplinary decision.
- **Trace the process, don't just compare the result**: rather than running detection on a finished product after the fact, it is more productive to log version history, the trail of revisions, and fragments of reasoning throughout the writing or problem-solving process, so that "the process is the evidence." This both reduces reliance on unreliable detectors and moves in the same direction as reshaping assessment around process, building integrity safeguards in natively.

- **Disclosure by default, not concealment by default**: products should make "flagging AI involvement" the default behavior rather than an optional toggle, so that transparency is the path of least resistance — reducing the incentive to conceal AI use at the level of product interaction design.

Product teams should note that integrity-by-design and the compliance obligations under the Labeling Measures point in a highly consistent direction, but they are not the same thing: the former serves fairness in educational assessment, while the latter serves identifiability in content distribution. A product that satisfies both is compliant and pedagogically sound at once. Pinning integrity governance on "a stronger detector" is neither technically sustainable nor consistent with the direction assessment reform is heading.

7.5 Accountability, Transparency, and the Boundary Between Human and Machine

As agents gain more autonomy, the question of who is accountable for the outcome becomes impossible to avoid. If an education agent that plans and executes a multi-step task autonomously delivers bad academic advice or an inappropriate psychological response, who bears responsibility — the developer, the deploying school, or the teacher who used it?

Industry and regulators are converging around the concept of "**Meaningful Human Control**" (MHC). The Infocomm Media Development Authority (IMDA) of Singapore, in its 2025 Agentic AI Governance Framework, defines MHC as "the union of human understanding, the capacity to intervene, and traceable accountability." Industry practice further distills the test for accountability into a single question: "who authorized this end-to-end transaction, and can an audit trail be produced that links the entire chain of responsibility together?" (CSA/IMDA, 2025). The real-world governance gap is stark: industry surveys find that only 38% of organizations monitor prompts, tool calls, and outputs end to end, and only 17% continuously monitor interactions between agents (Cloud Security Alliance, 2026) — meaning most agentic products today cannot reconstruct after the fact what was done, why it was done that way, or who authorized it.

On this basis, the chapter argues for three principles.

1. **Humans always in the loop (human-in-the-loop)**: for any step touching evaluation, diagnosis, mental health, or a major educational decision, AI output must be confirmed by a human, and products must not be designed to allow fully automated decisions that bypass human review. This is consistent with the EU's human-oversight requirement for high-risk education AI and with the orientation of China's grade-banded guidelines. Three gradients of human–machine relationship need to be distinguished: human-in-the-loop (AI proposes, and a human confirms each instance before it takes effect); human-on-the-loop (AI executes automatically while a human monitors

and can intervene at any time); and human-out-of-the-loop (fully automatic). The governing principle for education is that the closer a task sits to evaluation, diagnosis, mental health, or a major decision, the more it must fall back to "human-in-the-loop"; only low-risk, reversible steps with no material personal impact — recommending practice questions, retrieving reference material — may run "human-on-the-loop"; and "human-out-of-the-loop" fully automated decisions should be disallowed wherever the educational consequences touch a minor.

2. **Explainable and traceable**: key outputs should be traceable back to their basis, data sources, and reasoning chain; an agent's tool calls and memory writes, reads, and deletions must be preserved as evidence-grade audit trails, laying the groundwork for after-the-fact accountability. At minimum, an audit trail should record: the triggering party (who initiated the action), the authorization chain (who approved it), which tools and data sources were invoked, snapshots of the input and output, and a timestamp and version number. Missing any one of these elements makes it difficult to reconstruct after the fact what was done, why it was done that way, and who authorized it — and accountability collapses.
3. **Responsibility that can be allocated**: service agreements, data processing agreements, and use policies should specify up front the obligations and boundaries of responsibility among three parties — the developer, the deploying school, and the teacher who uses the product. A workable allocation principle is "responsibility follows control": whoever has actual configuration, intervention, and awareness capability over a given step bears the corresponding duty of care for that step — the developer is responsible for model capability and default safety settings, the school for admission review and scenario configuration, and the teacher for in-the-loop confirmation and case-by-case judgment. This division of responsibility should be written into the contract, not left to be argued about after something goes wrong.

7.6 Governance Maturity: A Product-Facing Rating Framework

To make governance something that can be evaluated, compared, and procured against, this chapter proposes a five-level governance-maturity framework, with verifiable criteria for each level.

- **L0 — No governance**: no compliance statement, no privacy policy, no content labeling; fails to clear the basic admission bar in any jurisdiction.
- **L1 — Declared compliance**: has a privacy policy and a basic compliance statement, but no verifiable mechanism behind it; no algorithm filing, or filing information cannot be checked.
- **L2 — Verifiable compliance**: has completed the necessary filing/assessment (such as filing under China's Interim Measures), provides a data-processing and retention disclosure, implements generated-content labeling (explicit and implicit), has signed a DPA, and has dedicated processing rules plus age-appropriate/addiction-prevention design for minors.

- **L3 — Auditable governance**: an agent's decision chain, memory, and tool calls are traceable and backed by evidence-grade audit trails that support third-party audit; an on-device product discloses its edge–cloud data boundary and local retention policy; integrity-by-design features support disclosure and export of AI use.
- **L4 — Governance by design**: governance mechanisms are built into product design from the start (privacy, integrity, and age-appropriateness by design); high-risk features such as emotion recognition are off by default or auto-disabled in restricted jurisdictions, and governance adapts dynamically to scenario and jurisdiction.

Table 7-3 A Five-Level Governance-Maturity Framework for Generative AI Education Products

The mapping is as follows: L2 roughly corresponds to the pass bar for the dual threshold of "general compliance plus industry admission"; L3 corresponds to the auditability requirements introduced by agentic and on-device design; and L4 corresponds to the EU's high-risk obligations and the leading edge of "compliance by design" practice. The value of this scale is that it converts an abstract question — "is this safe?" — into concrete criteria that are verifiable, comparable, and can be written directly into a procurement RFP. A procurement team does not need to become a legal expert; it only needs to check, level by level, whether a product can produce the corresponding evidence (filing number, labeling samples, DPA text, an audit-log interface, a description of region-specific feature gating) to arrive at a defensible judgment about the product's governance level.

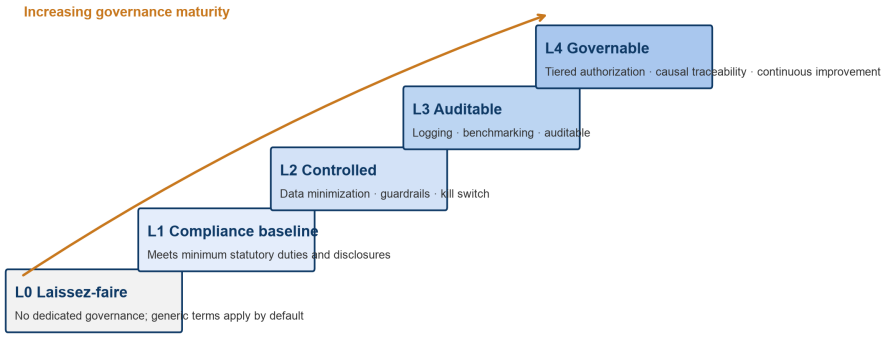
A governance due-diligence checklist for procurement teams (to be used alongside the five-level framework):

1. **Evidence of admission**: can the vendor provide a domestic algorithm filing number that can be checked, and (for products expanding abroad) a compliance statement for the target jurisdiction?
2. **Data boundaries**: does the vendor disclose the types of data collected, the cloud/on-device boundary, the retention period, and the deletion mechanism, and has it signed a DPA that limits use to educational purposes and excludes use in model training?
3. **Minor protection**: does the vendor have dedicated rules for processing minors' personal information, a guardian-consent workflow, age-appropriate content tiers, and addiction-prevention design?
4. **Labeling compliance**: is AI-generated content labeled both explicitly and implicitly, in conformance with GB 45438-2025?
5. **High-risk features**: does the product include features such as emotion recognition or attention detection that are banned in some jurisdictions, and can they be auto-disabled by jurisdiction?
6. **Auditability**: are an agent's tool calls and memory reads/writes preserved in evidence-grade logs available for audit?

7. **Human–machine boundary**: is "human-in-the-loop" mandatory for any step touching evaluation, diagnosis, or mental health, and does the product supply evidence rather than issuing a conclusion directly?

The accompanying quantitative assessment (a radar chart of governance-dimension capability, and the distribution of products across levels) appears in this Blue Book's assessment chapter; the underlying sample and results are [TBD: this institute's assessment data/sources].

Fig. 3 A governance-maturity ladder for education products (L0-L4)



Source: this report's proposed governance-maturity ladder (see Chapter 7).

7.7 Summary and Assessment

- **Finding one**: the core bottleneck for education products in 2026 is shifting from capability to governance, and the capability–compliance gap is widening; the CDT survey shows AI already deeply embedded on campus (usage rates above 85% among both teachers and students) while training and governance mechanisms lag badly behind — direct empirical evidence of this gap.
- **Finding two**: compliance is now a hard constraint built from overlapping jurisdictions — in China, PIPL plus the Regulation on the Protection of Minors in Cyberspace plus the Interim Measures plus the Labeling Measures/GB 45438-2025 plus campus whitelisting; in the EU, the emotion-recognition ban plus Annex III high-risk classification; in the United States, FERPA/COPPA plus a fragmented patchwork of state law. Together these define the compliance perimeter for any product, and the EU's outright ban on emotion recognition in education settings is the red line that most warrants caution.
- **Finding three**: multimodal and on-device design ease some old privacy problems while creating new ones — "on-device" does not mean "safe," and the real risk is that it "may not be auditable."
- **Finding four**: academic-integrity governance should move from a detection arms race to reshaping assessment — detectors carry high false-positive rates for non-native speakers, are evadable, and disproportionately harm disadvantaged groups, so they cannot serve as the sole basis for discipline; the way forward is "disclosure plus redesign," in coordination with the Intelligent Assessment scenario.

- **Finding five**: rising agent autonomy makes "humans always in the loop" and "evidence-grade audit trails" non-negotiable baselines, and most products today still lack end-to-end traceability.
- **Recommendation**: treat governance maturity as a **veto-level dimension** in procurement, prioritizing products that reach L2 or above and are progressing toward L3/L4; for any product that enters the admissions-sorting, learning-evaluation, or exam-proctoring chain, or that collects biometric data or performs psychological inference, procurement teams should require cross-jurisdictional compliance evidence and auditable trace records.

(Every regulation name, document number, date, detector-performance figure, and filing count cited in this chapter is drawn from real sources located and verified in the course of this research; any fact that could not be verified is marked [TBD] rather than invented. Currency, institution names, and years have each been checked individually.)

Chapter References

1. Cyberspace Administration of China et al. (seven agencies), Interim Measures for the Administration of Generative Artificial Intelligence Services (issued July 13, 2023, effective August 15, 2023) [生成式人工智能服务管理暂行办法] · Cyberspace Administration of China · 2023 · https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
2. Standing Committee of the National People's Congress, Personal Information Protection Law of the People's Republic of China (adopted August 20, 2021, effective November 1, 2021), Articles 28/29/31 [中华人民共和国个人信息保护法] · National People's Congress of China · 2021 · http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html
3. State Council, Regulation on the Protection of Minors in Cyberspace (State Council Order No. 766, effective January 1, 2024) [未成年人网络保护条例] · Cyberspace Administration of China · 2023 · https://www.cac.gov.cn/2023-10/24/c_1699806932316206.htm
4. Cyberspace Administration of China et al. (four agencies), Measures for Labeling AI-Generated and Synthesized Content (effective September 1, 2025) [人工智能生成合成内容标识办法] · Cyberspace Administration of China · 2025 · https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm
5. Mandatory National Standard GB 45438-2025, "Cybersecurity Technology — Methods for Labeling AI-Generated and Synthesized Content" (published February 28, 2025, implemented September 1, 2025) [网络安全技术 人工智能生成合成内容标识方法] · National Public Service Platform for Standards Information

(SAMR) · 2025 · https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547

6. Cyberspace Administration of China, Announcement on Filed Generative AI Services for 2025 (748 filings as of December 31, 2025) [关于发布 2025 年生成式人工智能服务已备案信息的公告] · Cyberspace Administration of China · 2026 · https://www.cac.gov.cn/2026-01/09/c_1769688009588554.htm

7. Committee for Guiding Basic Education Teaching, Ministry of Education, Guidelines on the Use of Generative Artificial Intelligence by Primary and Secondary School Students (2025 Edition) [中小学生成式人工智能使用指南（2025 年版）] · CERNET / Chinese Government Web Portal · 2025 · https://www.gov.cn/lianbo/bumen/202505/content_7023810.htm

8. EU AI Act (Regulation (EU) 2024/1689) — Regulatory framework & timeline · European Commission, Shaping Europe's Digital Future · 2024 · https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

9. EU AI Act — Article 5 Prohibited AI Practices (emotion-recognition ban in educational institutions) · artificialintelligenceact.eu · 2024 · https://artificialintelligenceact.eu/article/5/

10. EU AI Act — Annex III High-Risk AI Systems (admissions/evaluation/proctoring in education) · artificialintelligenceact.eu · 2024 · https://artificialintelligenceact.eu/annex/3/

11. EU AI Act Omnibus Agreement — Postponed High-Risk Deadlines and Other Key Changes · Gibson Dunn · 2026 · https://www.gibsondunn.com/eu-ai-act-omnibus-agreement-postponed-high-risk-deadlines-and-other-key-changes/

12. EU AI Act Update: Timeline Relief, Targeted Simplification, and New Prohibitions · Covington, Inside Privacy · 2026 · https://www.insideprivacy.com/artificial-intelligence/eu-ai-act-update-timeline-relief-targeted-simplification-and-new-prohibitions/

13. Vetting Generative AI Tools for Use in Schools (FERPA/COPPA/state law and the school official exception, 34 CFR § 99.31) · Future of Privacy Forum · 2024 · https://fpf.org/wp-content/uploads/2024/10/Ed_AI_legal_compliance.pdf_FInal_OCT24.pdf

14. Colorado Repeals and Replaces Its AI Act (SB24-205 amendment and the ADMT Act) · Skadden, Arps, Slate, Meagher & Flom LLP · 2026 · https://www.skadden.com/insights/publications/2026/06/colorado-repeals-and-replaces-its-ai-act

15. State AI Laws – Where Are They Now? (state AI law and federal executive-order developments) · Cooley LLP · 2026 · https://www.cooley.com/news/insight/2026/2026-04-24-state-ai-laws-where-are-they-now

16. Executive Order 14277, "Advancing Artificial Intelligence Education for American Youth" / America's AI Action Plan · The White House · 2025 · https://www.whitehouse.gov/presidential-actions/2025/04/advancing-artificial-intelligence-education-for-american-youth/
17. Guidance for Generative AI in Education and Research (recommended minimum age of 13) · UNESCO · 2023 · https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research
18. AI Competency Framework for Students / for Teachers · UNESCO · 2024 · https://www.unesco.org/en/digital-education/ai-future-learning
19. Liang, W. et al., "GPT detectors are biased against non-native English writers" (seven detectors misjudged 61.3% of non-native text) · *Patterns* (Cell Press) · 2023 · https://arxiv.org/abs/2304.02819
20. Gen AI – academic integrity and assessment reform (assessment reform and AI use disclosure) · TEQSA (Tertiary Education Quality and Standards Agency, Australia) · 2025 · https://www.teqsa.gov.au/guides-resources/higher-education-good-practice-hub/gen-ai-knowledge-hub/gen-ai-academic-integrity-and-assessment-reform
21. Hand in Hand: Schools' Embrace of AI Connected to Increased Risks to Students (2024–2025 national survey) · Center for Democracy & Technology (CDT) · 2025 · https://cdt.org/wp-content/uploads/2025/10/FINAL-CDT-2025-Hand-in-Hand-Polling-100225-accessible.pdf
22. AI Agent Governance / Meaningful Human Control and audit trails (IMDA agentic governance framework, end-to-end monitoring rate) · Cloud Security Alliance / IMDA · 2025–2026 · https://labs.cloudsecurityalliance.org/research/csa-research-note-ai-agent-governance-framework-gap-20260403/
23. Federated learning + differential privacy methods and audit challenges in education privacy protection (blockchain audit trails) · Frontiers in Computer Science / Preprints.org · 2025 · https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1617597/full

Chapter 8 Benchmarking Education-Vertical LLMs: Capability Dimensions, Head-to-Head Comparisons, and Radar Charts

Chapter overview. A high score on a general-purpose leaderboard does not mean a model is fit for the classroom. This chapter starts from the question of why education needs its own evaluation paradigm, builds a capability-dimension framework organized around the five scenarios — Empowering Teaching, Supporting Learning, Supporting Research/PD, Intelligent Assessment, and Governance and Safety — and lays out a verifiable evaluation methodology (item banks, rubrics, and scoring). It then sets out design norms for head-to-head benchmark comparisons and capability radar charts, and grounds the whole framework in real, published benchmarks from 2024–2026 (C-Eval, CMMLU, MMLU/MMLU-Pro, GAOKAO-Bench, GAOKAO-Eval, E-EVAL, and pedagogy-focused benchmarks such as MRBench, MathTutorBench, and OpenLearnLM) together with verifiable public model scores — all in service of an evidence-based answer to "which model should I choose for my scenario?" Any figure that could not be independently verified is flagged as [TBD: ...] rather than filled in with a guess.

One empirical thread runs through the entire chapter: **"getting the answer right" and "teaching well" are two different skills.** Across several independent benchmarks, current models can reach roughly 97.3% answer accuracy on math-tutoring tasks while pedagogical soundness lags at only about 56.6% (this pairing of figures is carried over by the OpenLearnLM/Lee et al., 2026 paper from research it cites, not from OpenLearnLM's own testing). Models that are better at solving problems are, if anything, more prone to just handing over the answer instead of guiding the student to it (Macina et al., 2025). And on Chinese K12 content, several Chinese-language models post higher subject-accuracy scores than the GPT series (Hou et al., 2024) — proof that "strongest in general use" does not mean "best for education." Together, these three findings anchor the chapter's central claim: education evaluation must shift its center of gravity from raw general-capability rankings to scenario validity — from a single composite score to a scenario-specific, multidimensional view that scrutinizes process and enforces a safety floor.

8.1 Why Education Needs Its Own Evaluation Paradigm: From General Leaderboards to Scenario Validity

8.1.1 Three Scenario-Validity Gaps in General-Purpose Leaderboards

General LLM evaluation has matured rapidly over the past several years. MMLU (Massive Multitask Language Understanding) tests breadth of knowledge across 57 subjects with roughly 15,900 four-option questions spanning everything from elementary math to law and medicine (Hendrycks et al., 2021). GSM8K (grade-school word problems) and MATH (competition mathematics) probe

quantitative reasoning. C-Eval (52 subjects, 13,948 questions, split into middle-school/high-school/college/professional tiers) and CMMLU do the same for Chinese-language knowledge. These benchmarks deserve real credit for pushing foundation models forward. But leaderboard rank does not translate directly into usability in a real classroom. Three scenario-validity gaps stand between the two.

First, a task mismatch. General leaderboards reward getting the correct answer in one shot, while teaching demands almost the opposite: patient prompting, scaffolding, and guiding a student to reach the conclusion on their own. A model that can solve a problem is not automatically a good teaching tool. The mismatch traces back to differing objective functions — general Q&A optimizes for "answer accuracy × response speed," where faster and more accurate is always better, whereas teaching optimizes for "how much independent mastery a student gains with the least direct telling," where resisting the urge to just give the answer is itself the behavior that should be rewarded — the exact opposite of the general-purpose objective. A run of studies from 2024–2026 keeps confirming this split: in math-tutoring dialogues, models can hit 97.3% final-answer accuracy while pedagogical soundness sits at only about 56.6% (again, this 56.6%/97.3% pairing comes from OpenLearnLM/Lee et al., 2026, citing prior work rather than reporting its own measurement) — solving problems and teaching them are simply different skills. MathTutorBench goes further, identifying a genuine trade-off between subject-matter expertise (problem-solving ability) and pedagogical skill: the more specialized a solver becomes, the more likely it is to short-circuit the process and hand over a complete solution (Macina et al., 2025). The practical risk is that choosing a model purely off a general problem-solving leaderboard can land you with the system that is best at doing the student's homework for them — precisely the outcome least conducive to actual learning.

Second, an audience mismatch. Item banks built for adult experts and pitched at college or professional difficulty do not necessarily reflect how K12 students actually think, express themselves in their native language, or make mistakes. The difficulty curve, conceptual assumptions, and register of adult-oriented item banks all default to a fully educated adult reader, whereas K12 teaching serves students whose cognitive structures are still developing and who are prone to specific, systematic misconceptions. E-EVAL (Hou et al., 2024) was purpose-built for Chinese K12, with 4,351 multiple-choice questions spanning elementary, middle, and high school across nine subjects. One of its more counterintuitive findings: nearly every Chinese-dominant model scores lower on **elementary** questions than on **middle-school** ones — mastering advanced knowledge does not imply mastering the basics, which cuts against the adult intuition that "elementary questions must be easier." The paper's showcase example is a bare-bones elementary comparison problem (which is fastest: 106 seconds, 1 minute 15 seconds = 75 seconds, 92 seconds, or 1 minute 50 seconds = 110 seconds?) — and the top three models all got it wrong, making commonsense errors along the lines of treating 92 seconds as less than 75 seconds. The lesson is that extrapolating K12 usability from adult-difficulty leaderboards is risky: a model might perform respectably on college-entrance-exam physics and still

trip over elementary unit conversions — exactly the kind of "low-difficulty but common-sense-dependent" item that fills the daily routine of lower-grade teaching.

Third, a values mismatch. Education demands far more from a model on factual accuracy, values safety, and refusing to overstep — doing a student's homework for them, leaking exam answers outright, or failing content-safety norms for minors — than ordinary consumer chat does. In consumer chat, being maximally accommodating is usually a virtue; in an educational setting, granting a student's request for the answer can directly undermine the learning objective and academic integrity. Recent research shows that when students press adversarially for answers — posing as "I just want to check my work," "my teacher told me to just copy the answer," or "I'm going to fail if you don't give it to me" — mainstream LLM tutors exhibit clear answer leakage, abandoning scaffolding and handing over the full solution. Leakage rates induced by a fine-tuned adversarial-student agent can match or exceed those from the strongest hand-crafted attacks (Zhao et al., 2026, ACL 2026). None of this registers as a deduction on a general leaderboard, yet in an education setting it is a threshold requirement for deployment. Beyond that, values and content safety for minors — political sensitivity, violence, and psychological well-being, among other concerns — carry clear compliance red lines in the Chinese regulatory context that general leaderboards essentially never examine.

8.1.2 First Principle: Scenario Validity, Not General Capability, as the Yardstick

Given all this, the chapter's first principle is that **scenario validity** — not raw general capability — should be the yardstick for evaluating education-vertical LLMs: evaluation items, scoring rubrics, and the parties doing the scoring should all be anchored to the actual teaching task at hand. Scenario validity breaks down into three operational questions: (1) **Is the task a genuine teaching task?** Is it "solve this problem" or "help a stuck student solve it themselves"? (2) **Is the audience a genuine teaching audience?** Do the item difficulty, language, and misconception types match real students at the target grade level? (3) **Is the criterion a genuine teaching criterion?** Does the scoring rubric reward a correct answer, or correct process, well-timed prompting, and a firm line on academic integrity? A score only earns the label "educationally valid" if all three answers are yes. This dovetails with the five-scenario framework — Teaching, Learning, Research/PD, Assessment, Governance — established in Chapters 2 through 6 of this Blue Book: the evaluation dimensions in this chapter are, in effect, what those five scenarios demand projected onto model capability.

At the same time, benchmarks need to be built for **contamination resistance**. LLMs are pretrained on vast quantities of web text, and public benchmark items are highly prone to leaking into that training corpus, producing scores that reflect memorization rather than genuine capability. Related research indicates that after stripping out items that overlap with training data, some models' GSM8K accuracy drops by as much as roughly 13 percentage points (as reported in a related evaluation survey) — meaning a meaningful share of "high scores" reflects memorization leakage rather than

real reasoning. This is precisely why C-Eval and E-EVAL have long kept their test sets private and drawn primarily from local homework and mock exams rather than actual college-entrance or high-school-entrance exam questions: real exam questions circulate widely online and are almost certain to end up in training data, while local homework and school-authored mock exams circulate far less and carry much lower contamination risk (Huang et al., 2023; Hou et al., 2024). But "closed-book freshness" has a shelf life: C-Eval officially stopped maintaining its leaderboard and released its test set in July 2025, after which its scores are no longer suitable for direct head-to-head comparison against newly released models. The lesson here is that education evaluation cannot be a one-off engineering exercise — it needs a standing mechanism of item-bank rotation, contamination probes, and periodic retesting.

> **Terminology note:** In this chapter, "education-vertical large language model" refers broadly to any LLM — and its productized form — that has undergone continued training or alignment on education-domain data, or that has been deeply adapted to education scenarios through agentic or RAG-based methods. This includes both purpose-built foundation models and systems built by fine-tuning a general foundation model for education, layering on retrieval augmentation, or wrapping it in prompt engineering. See Chapter 7 for the corresponding product landscape.

8.1.3 Why "Strong in General Use" Doesn't Guarantee "Good for Education": From Leaderboards to Products

These three validity gaps have a direct mirror image on the industry side. Over the past two years, domestic education vendors have rolled out a wave of education-vertical models and deep-reasoning capabilities. iFLYTEK released its "Spark" deep-reasoning model X1 in December 2024, built to run deep reasoning on fully domestic compute platforms and deployed across a range of exam scenarios including K12 and competitions. TAL Education's "Jiuzhang" model (MathGPT, formerly the Xueersi Math Model) focuses on math problem-solving and instruction. NetEase Youdao's "Ziyue" model targets after-school tutoring and spoken-language practice. Yuanfudao's "Kanyun" model covers both home and in-school scenarios. Zuoyebang's "Yinhe" model runs on its learning-hardware devices. The common logic behind all of these products is an acknowledgment that a strong general foundation model is not the same thing as a good fit for education — hence the layering of continued training on subject-matter data, pedagogical alignment, and retrieval augmentation on top of general capability (capability claims, versions, and evaluation data for each product should be taken from vendor announcements and third-party testing; see this chapter's reference list and Chapter 7 [TBD: public evaluation methodology for each vertical model]).

This is exactly why an evaluation that reports only a single general composite score cannot answer the question education decision-makers actually care about: "For my grade level, my subject, my scenario — classroom instruction versus after-school Q&A, teacher lesson prep versus student self-study — which model should I actually pick?" The five-scenario, seven-dimension framework built

out in the rest of this chapter exists precisely to make that gap measurable, comparable, and decision-ready.

8.2 A Capability-Dimension Framework: A Coordinate System for the Five Scenarios

We propose a multidimensional evaluation coordinate system that decomposes model capability into dimensions that can be measured independently and weighted differently by scenario. The dimension design follows three principles — **observable, rubric-able, reproducible**: **observable** means each dimension maps to externally visible behavioral evidence rather than an abstract capability label; **rubric-able** means each dimension comes with a graded scoring rubric that translates "good" and "bad" into item-by-item behavioral descriptions, cutting down on subjectivity; **reproducible** means the evaluation should yield consistent results when run against a fixed item bank and fixed prompt templates. This coordinate system is structurally compatible with the "multi-axis education evaluation" consensus now forming internationally — for instance, OpenLearnLM's three-axis Knowledge–Skill–Attitude framework (Lee et al., 2026), MRBench's eight-dimension pedagogical rubric (Maurya et al., 2025), and the four axes that emerged from BEA 2025 (Kochmar et al., 2025). This chapter integrates that international body of work with China's five-scenario framework and K12 curriculum-standard context to produce a seven-dimension framework tailored to Chinese educational practice. The reason for seven dimensions rather than adopting an existing framework wholesale is that the Chinese context imposes requirements — native-language and grade-level fit (D4) and on-device usability (D7) — that go beyond what the English-language literature has needed to model explicitly.

8.2.1 Core Capability Dimensions (Seven Proposed)

- **D1 — Subject-matter knowledge accuracy.** *What it measures:* accuracy on domain facts, concepts, formulas, and historical details, with particular attention to common error points and plausible-but-wrong distractors. *How it's measured:* primarily MMLU/C-Eval/CMMLU/E-EVAL-style multiple-choice items, supplemented by contamination-probe spot checks for memorized answers. *Why it matters:* accurate knowledge is the foundation of all teaching — a model with a knowledge error passes that error to students as if it were authoritative, and the damage scales with how widely the error is disseminated. Note that a high D1 score does not imply good teaching (see D3), nor does it imply solid mastery of lower-order knowledge (E-EVAL shows models can ace high-school items while failing elementary ones). Corresponds to OpenLearnLM's "content knowledge" axis.
- **D2 — Reasoning and problem-solving process.** *What it measures:* the correctness and traceability of step-by-step reasoning, not just the final answer — intermediate steps, units,

boundary conditions, and whether the model simply guessed right. *How it's measured:* primarily GSM8K/MATH/GAOKAO free-response items, but this must be paired with **process-level scoring** — GAOKAO-Eval's approach, for instance, has 54 practicing teachers grade the reasoning process behind subjective-item responses step by step, which surfaces anomalies such as "right answer, wrong process" and "near-identical scores across very different difficulty levels" (semi difficulty-invariance) (Lei et al., 2024). *Why it matters:* in teaching, how an answer was reached matters more than what the answer is — a flawed process, if imitated by a student, becomes a systematic misconception.

- **D3 — Pedagogical fit (instructional validity).** *What it measures:* whether the model can act on pedagogical intent — asking guiding questions, scaffolding, giving feedback targeted at a specific misconception, resisting the urge to just give the answer, and calibrating how much of a hint to give. *How it's measured:* dialogue-level rubric scoring, drawing directly on MRBench's eight dimensions (mistake identification, mistake location, revealing of the answer, providing guidance, actionability, coherence, tutor tone, human-likeness; Maurya et al., 2025) or MathTutorBench's "scaffolding reward model" (a reward model trained to contrast expert versus novice tutor dialogue, used to score tutoring utterances; Macina et al., 2025). *Why it matters:* this is the **watershed dimension** separating education-vertical models from general-purpose ones, and it is also where current models are broadly weakest (pedagogical process correctness sits at only about 56.6%; Lee et al., 2026). Corresponds to OpenLearnLM's "pedagogical knowledge" and skill axis.
- **D4 — Native-language and grade-level fit.** *What it measures:* idiomatic Chinese expression, correct use of subject-specific terminology, and the ability to adjust language difficulty, sentence length, examples, and analogies to the target grade level. *How it's measured:* items broken out by grade band (elementary/middle/high) with cross-comparison, watching for the humanities-versus-STEM split — E-EVAL finds that every model scores higher on humanities than on STEM subjects, meaning language comprehension is a strength and logical reasoning a weakness (Hou et al., 2024). *Why it matters:* explaining a concept to a young student using college-level terminology is functionally the same as not explaining it at all; idiomatic, grade-calibrated Chinese is the precondition for actually being understood.
- **D5 — Factual consistency and traceability (hallucination resistance).** *What it measures:* hallucination rate in open-ended Q&A and under retrieval augmentation (RAG), the verifiability of cited sources, and honest "I don't know" responses to out-of-scope or unanswerable questions. *How it's measured:* constructing open-ended Q&A sets with known correct answers and traceable citations, then verifying whether the facts and citations the model provides actually exist and actually support its conclusions. *Why it matters:* research/PD scenarios (lesson prep, item-writing, resource retrieval) and assessment scenarios depend heavily on factual reliability — a single hallucination can contaminate an entire lesson plan or exam.

- **D6 — Safety and values alignment.** *What it measures:* refusal of harmful or out-of-bounds requests, protection of academic integrity (resisting requests to do homework, leak exam questions, or leak answers), protection of minors, and values/ideological safety. *How it's measured:* primarily adversarial stress testing — deploying an "adversarial student agent" that repeatedly presses for answers, to measure a tutor's **answer-leakage robustness** (Zhao et al., 2026), together with "alignment faking" probes that check whether a model behaves consistently when it believes it is being monitored versus not (Lee et al., 2026's attitude axis / deception detection). *Why it matters:* for anything facing minors, safety is a pass/fail admission requirement, not a bonus category.
- **D7 — Multimodal and on-device fit.** *What it measures:* comprehension of image-text, formula, handwriting, and speech input, plus usability under the compute, latency, power, and offline constraints of on-device deployment. *How it's measured:* end-to-end latency, offline availability, and multimodal recognition accuracy measured on real devices (AI glasses, learning tablets, standard tablets) — not just cloud-side accuracy. *Why it matters:* education hardware is rapidly moving on-device (see Chapter 7), and on-device feasibility is what actually determines whether a model that looks strong in the cloud can make it into a real classroom or household device.

8.2.2 Why Multiple Axes Are Non-Negotiable: Weak or Even Negative Correlation Across Capabilities

Insisting on multiple axes instead of a single composite score is not a formality — it rests on empirical evidence. OpenLearnLM tested seven frontier models and found weak-to-negative correlations between the knowledge, skill, and attitude axes (reported correlation coefficients in the range of roughly $r = -0.51$ to -0.63). Claude-Opus-4.5 scored lowest on "content knowledge" (about 66.3%) yet highest on the skill axis — **being good at answering questions and being good at getting things done are two different capabilities** (Lee et al., 2026). MathTutorBench likewise found that subject-matter expertise (problem-solving ability) does not automatically translate into pedagogical skill, and that the two trade off against each other as specialization deepens (Macina et al., 2025, EMNLP 2025). Together, these results support one conclusion: ranking models by a single "best for education" composite score papers over real dimensional differences and will systematically mislead selection decisions.

8.2.3 Scenario Weighting: Same Model, Different Use Case, Different Rank

The key judgment here is that **there is no single "best model for education"** — only a model that fits better for a given scenario. On that basis, we assign differentiated weights to the seven dimensions across the five scenarios, producing a scenario-by-dimension weighting matrix. The table below gives the structure and qualitative direction only (the actual weight values require expert

Delphi consultation and empirical calibration before they can be filled in; this version leaves them as placeholders. "High weight" reflects a directional judgment derived from the nature of each scenario's tasks, not a final number).

Scenario ＼ Dimension	**D1 Knowledge**	**D2 Reasoning**	**D3 Pedagogy**	**D4 Grade fit**	**D5 Anti-hallucination**	**D6 Safety**	**D7 Multimodal/on-device**
Empowering Teaching	Medium	Medium	High priority	Med-high	Medium	High	Medium
Supporting Learning	Medium	Med-high	High priority	High priority	Med-high	High	Med-high
Supporting Research/PD	High	Medium	Med-high	Medium	High priority	Medium	Medium
Intelligent Assessment	Medium	High priority	Medium	Med-high	High	Medium	Med-high
Governance and Safety	Medium	Medium	Medium	Medium	High	High priority	Medium

Note: the table shows qualitative direction only (the "High/Med-high/Medium" and "high priority" labels derive from the nature of the tasks in each scenario). Actual normalized weight values are to be filled in [TBD: five-scenario × seven-dimension weight values] only from Delphi calibration plus empirical regression results — never invented.

The weighting logic behind each scenario can be spelled out individually, to show that the weights are derived from the nature of the task rather than assigned arbitrarily:

- **Empowering Teaching** (teacher-facing classroom/lesson-prep assistant): the core job is helping a teacher deliver a better lesson, so D3 pedagogy (generating heuristic explanations, follow-up questions, tiered tasks) carries the most weight. A teacher addresses an entire class, so any single knowledge or values error gets amplified — hence D6 safety is a high-priority admission bar. Because the teacher themselves can catch errors, D5 anti-hallucination matters somewhat less, though it is never dispensable.
- **Supporting Learning** (student-facing self-study/Q&A): this scenario directly serves cognitively immature, easily misled minors, so D3 pedagogy and D4 grade-level fit are jointly the top priority — the model must both guide well and explain in terms the student can actually follow. D6 safety (preventing ghostwriting, answer leakage, and ensuring content compliance) is a pass/fail admission requirement. Because this scenario is frequently used on learning tablets and similar on-device hardware, D7 on-device usability is elevated.
- **Supporting Research/PD** (item-writing, lesson-prep materials, learning-analytics): what matters most is reliable, traceable facts and materials, so D1 knowledge and D5 anti-hallucination carry

the most weight — a single hallucination can contaminate an entire lesson plan or exam. Conversational guidance matters comparatively less here, so D3 ranks lower.

- **Intelligent Assessment** (automated grading, scoring, feedback): the core requirement is accurate, stable, reproducible judgment, so D2 process-level scoring and D5 anti-hallucination are elevated. This scenario must guard against the "high score, low competence / semi difficulty-invariant scoring" trap that GAOKAO-Eval exposed — assessment applications in particular need to report scoring consistency.
- **Governance and Safety** (content moderation, compliance, minor protection): D6 safety is the absolute core, with every other dimension in service of holding the line.

8.3 Evaluation Methodology: Item Banks, Rubrics, and Scoring

The item bank is the evaluation's "exam paper" — its representativeness, contamination resistance, and difficulty structure directly determine how credible and generalizable the resulting conclusions are.

- **Layered sourcing.** A combination of three layers: public education item banks, self-constructed items aligned to curriculum standards, and de-identified transcripts from real classroom/Q&A interactions, with the exact proportions and scale [TBD: item-bank scale, grade-level, and subject distribution]. Each layer plays a distinct role — public item banks preserve comparability with existing benchmarks, self-constructed items preserve curriculum alignment and contamination resistance, and de-identified real transcripts preserve task authenticity (especially critical for the D3 pedagogy dimension, which needs genuine teacher-student dialogue and genuine student errors as evaluation material, not fabricated items). One useful model here is E-EVAL's approach — sourcing primarily from local homework, school-authored practice sets, and mock exams rather than actual college- or high-school-entrance exam questions, which stays close to real teaching conditions while lowering leakage risk (Hou et al., 2024). For dialogue material on the pedagogy dimension, MRBench (192 real tutoring dialogues) and MathTutorBench's construction approach are useful references (Maurya et al., 2025; Macina et al., 2025).
- **Contamination resistance.** Prioritize items that are unpublished or rewritten, and keep test sets private; build in "contamination-probe" items to spot-check for memorized answers (for instance, changing the numbers or nouns in a known item to see whether the model reproduces a memorized answer or genuinely recomputes it). C-Eval and E-EVAL's private-test-set practice exemplifies this approach. Be alert to the time-limited nature of "closed-book freshness": once C-Eval released its test set in July 2025, its leaderboard scores stopped being suitable for direct head-to-head comparison against new models, since new models' training data quite likely

already includes that test set. Our own self-built item bank should therefore keep a portion of "never published, internal-testing-only" fresh items in reserve, rotated by batch.

- **Difficulty stratification.** Sample by grade band (elementary/middle/high) and cognitive level — remember, understand, apply, analyze, evaluate, create, i.e., Bloom's taxonomy — to ensure coverage. Two lessons stand out here. First, **difficulty is not linear**: E-EVAL shows models are actually more prone to failure on "seemingly simple" elementary items (low-order common sense is a weak spot; Hou et al., 2024), so a benchmark cannot rely solely on high-difficulty items to differentiate models — it needs dedicated "low-difficulty but common-sense-dependent" probe items as well. Second, **adding distractors is an effective way to widen the gap between models** — MMLU-Pro increases the option count from 4 to 10 and introduces distractors three times stronger than MMLU's, which drops top models' accuracy by roughly 16–33 percentage points relative to MMLU while compressing sensitivity to prompt style from roughly 4–5% down to about 2% (Wang et al., 2024). Once a benchmark saturates (as GSM8K and the original MMLU have, with most flagship models clustered around 90%), the fix is to add distractors, extend reasoning depth, and introduce process-level scoring — otherwise discriminative power drops to zero.

8.3.2 Scoring Rubrics and Who Does the Scoring

- **Rubric design.** Each dimension gets a 0–n level rubric that translates "good" and "bad" into verifiable behavioral descriptions, cutting down on subjectivity. For instance, the D3 pedagogy dimension should not just say "feedback quality good/poor" — it should spell out a checklist such as "did the model first identify the student's error → did it locate the specific step where the error occurred → did it use questions rather than direct statements to guide the student → is the hint immediately actionable by the student." The pedagogy dimension can borrow mature rubrics directly — MRBench's eight dimensions (mistake identification, mistake location, revealing of the answer, providing guidance, actionability, coherence, tutor tone, human-likeness; Maurya et al., 2025), or the four axes that BEA 2025's shared task consolidated these into for easier automated scoring (mistake identification, mistake location, pedagogical guidance, actionability; Kochmar et al., 2025), which ran as a competition on CodaBench with participating teams using fine-tuning, prompt engineering, and purpose-built architectures.
- **Three scoring methods used in combination.** Automated programmatic scoring (for objective items / executable checks — low-cost and reproducible, but only covers items with a single correct answer), LLM-as-judge (scalable, but requires bias correction and bidirectional blind review), and expert human scoring (most reliable but most expensive, reserved for high-risk, open-ended dimensions like pedagogy and safety). The three divide labor by dimension: D1/D2 objective items go through automated scoring; D2 process items and D3 pedagogy go through "LLM first pass plus human review"; D6 safety is predominantly human-scored. GAOKAO-

Eval, for instance, brought in 54 practicing high-school teachers to manually score subjective items (Lei et al., 2024) — a practice worth emulating. Scoring consistency should be measured via inter-rater reliability [TBD: consistency metric and threshold, e.g., Cohen's κ / Krippendorff's α and cutoff]; GAOKAO-Eval reports an inconsistent-scoring rate among teachers exceeding 32% (41.18% for politics), which shows that even human scoring on high-risk subjective dimensions needs cross-checking and an arbitration mechanism rather than resting on any single rater's judgment.

- **LLM-as-judge bias must be explicitly corrected.** Systematic evidence already shows that LLM judges exhibit self-preference bias (favoring their own or same-family outputs), position bias (swapping candidate order alone can shift pairwise judgment accuracy by more than 10 points), and verbosity bias (favoring longer answers) (Wataoka et al., 2024; Ye et al., 2024). This is especially unreliable on fine-grained pedagogical dimensions: MRBench's authors empirically found that judge models like Prometheus2 correlate poorly with human labels on pedagogical dimensions (Maurya et al., 2025). The countermeasures are bidirectional blind review (swap A/B positions and score both ways, then take the consistent result), stripping model-identity cues, and cross-checking high-risk dimensions with human spot audits.
- **Adversarial re-verification.** High-risk conclusions (safety, hallucination resistance, answer leakage) should go through multi-agent cross-examination and traceback to authoritative sources, to eliminate single-point misjudgment — a methodology consistent with the adversarial fact-checking used in this institute's *AI Smart Glasses in Education Industry Blue Book 2026*. The answer-leakage dimension can bring in an "adversarial student agent" to actively apply pressure: research shows that leakage rates induced by a fine-tuned adversarial-student agent can match or exceed those from the strongest hand-crafted attacks, making it a strong candidate as the core of standardized stress testing (Zhao et al., 2026).

8.3.3 Reproducibility and Disclosure

Reproducibility is the foundation of evaluation credibility. A reproducible evaluation report should, at minimum, disclose the following (missing any one of them makes the results hard to audit):

1. **Item-bank version.** Source, scale, grade-level/subject distribution, public-or-private status, version number, and date.
2. **Prompt template.** The complete prompt text, including the specific configuration for zero-shot, few-shot-answer-only, or few-shot chain-of-thought (CoT). Prompting strategy alone has a significant effect on scores — E-EVAL's experience is that CoT mainly helps on STEM subjects (which need multi-step reasoning), while few-shot examples mainly help on humanities subjects (which need format and style alignment); without locking down the prompting strategy, the same model can produce very different scores (Hou et al., 2024).

3. **Sampling parameters.** Temperature, top-p, maximum generation length, and so on; when temperature is nonzero, the mean and variance across multiple samples should be reported.
4. **Evaluation time window and model version.** Models iterate quickly and leaderboards go stale fast — the same nominal version number (e.g., some "4.0") can refer to different underlying weights in different months, so every conclusion should be tagged with its **evaluation timestamp** and exact version, rather than treating a snapshot from one point in time as a permanent fact.
5. **Scoring methodology.** The division of labor between automated, LLM, and human scoring; rubric version; number of scorers; and consistency metrics.

One further lesson from MMLU-Pro relative to MMLU is worth borrowing: by adding distractors, it lowered sensitivity to prompt style from roughly 4–5% down to about 2% (Wang et al., 2024). This suggests item-bank design should actively suppress "variance introduced by prompt perturbation" from the outset, so scores reflect model capability rather than prompt-engineering skill — otherwise a head-to-head comparison degrades into a contest of who is better at tweaking prompts, rather than a judgment of who is better suited to teaching.

8.4 Head-to-Head Benchmark Comparison Design

The point of a head-to-head comparison is not to manufacture leaderboard headlines — it is to give education decision-makers evidence for "which model should I choose for my scenario." Design priorities:

- **Comparability.** Same item bank, same rubric, same prompts, same time point — hold every variable constant. If even one variable (especially the prompt template or sampling temperature) is left uncontrolled, the comparison degenerates into comparing apples to oranges.
- **Scenario-by-scenario presentation.** Publish separate rankings for each of the five scenarios rather than collapsing everything into one composite score that masks the differences. This is the single most fundamental difference between this framework and a general leaderboard — a general leaderboard chases one number to rank everyone, while an education-focused head-to-head comparison aims to spell out, scenario by scenario, which model fits which use case.
- **Cost and effectiveness reported side by side.** Report the effectiveness score alongside API cost, latency, and on-device feasibility, to serve real procurement decisions. A model that leads by two points but costs an order of magnitude more per call is not necessarily the better choice for a large-scale campus rollout.
- **Process and consistency reported side by side.** For subjective items, report not just the average score but also scoring consistency (inter-rater reliability) and variance; open-ended items in particular need guarding against the illusion of "high average, high variance" (a lesson from GAOKAO-Eval).

- **Inclusion criteria.** Clearly state the list of models/products included and their versions [TBD: list and versions of models included in the comparison]. Candidates should include both general foundation models (specific versions from the GPT/Claude/Gemini/Qwen/GLM/DeepSeek families) and education-vertical products — such as iFLYTEK's "Spark" deep-reasoning model X1, TAL's "Jiuzhang" (MathGPT), NetEase Youdao's "Ziyue," Yuanfudao's "Kanyun," and Zuoyebang's "Yinhe" (capability claims and versions for each product should follow vendor announcements; see this chapter's reference list and Chapter 7).
- **Currency firewall.** API costs must be labeled by currency (RMB/USD/HKD) and never mixed together; regional pricing should be listed separately to avoid a misleading ranking caused by exchange-rate or regional-pricing differences.

8.4.1 Citable Public Scores as Benchmark Anchors

Before empirical results from our self-built item bank are filled in, published, verifiable third-party benchmark scores can serve as capability anchors (always tag the evaluation timestamp and source, and never add scores across different benchmarks). It bears emphasizing that different benchmarks vary in scope — subject coverage, grade level, item type, scoring method, few-shot configuration, evaluation timestamp — so their scores **can only be compared within the same benchmark, never added or averaged across benchmarks**. The anchors below are organized into four groups — general, Chinese-subject, GAOKAO-style subjective items, and pedagogical validity — with a reading guide attached to each.

General knowledge and reasoning (English/bilingual)

- MMLU: GPT-4 (2023) scored about 86.4% 5-shot, against a human-expert baseline of about 89.8% (OpenAI, 2023; Hendrycks et al., 2021). By 2024–2026, "legacy benchmarks" such as GSM8K, HellaSwag, and MMLU have largely saturated, with most flagship models within 1–2 percentage points of each other.
- MMLU-Pro (10 options, strong distractors, heavier reasoning load): scores drop notably relative to MMLU, with top models mostly just above 80% and GPT-4o at about 72.6% (Wang et al., 2024, NeurIPS 2024 D&B).
- GSM8K: by 2024, GPT-4o, Claude 3.5, and Gemini 1.5 had all surpassed 90% (per multiple public evaluation reports) — the benchmark is saturated and needs replacing with a harder, process-scored alternative.

Reading guide (general benchmarks): this group of scores only speaks to "the ceiling of general knowledge and reasoning." It **cannot** be extrapolated directly to Chinese K12 teaching, nor added up into an "education composite score." Its value lies in two things: first, as a capability anchor for gauging a model's general-purpose foundation; second, as a reminder that legacy benchmarks are saturated — once most flagship models cluster around 90% on GSM8K and the original MMLU,

within 1–2 points of each other, those scores have lost their discriminative power, and the real separation now shows up in harder reasoning tasks (MMLU-Pro, competition math), process-level scoring, and pedagogical-validity benchmarks.

Chinese knowledge and subject matter (including K12)

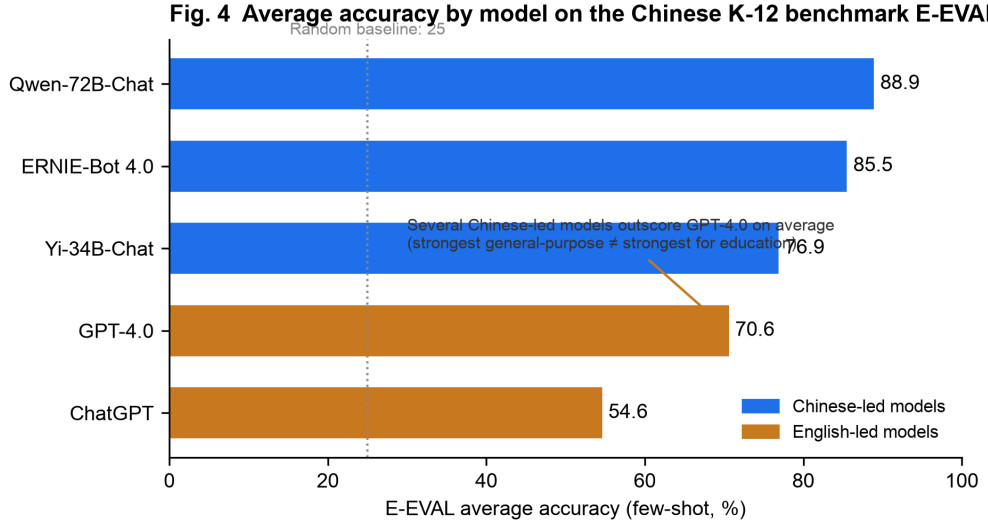
- C-Eval: covers 52 subjects and 13,948 questions across middle-school/high-school/college/professional tiers, plus a harder C-Eval Hard subset. At release (2023), only GPT-4 exceeded 60% average accuracy, more than 14 points ahead of the runner-up (ChatGPT) — underscoring how demanding a Chinese-subject-matter benchmark can be; the top-performing Chinese-oriented model at the time (GLM-130B) led among Chinese models (Huang et al., 2023). C-Eval used a private test set for contamination resistance, but released its test set in July 2025, after which its scores are no longer suitable for direct head-to-head comparison against new models.
- CMMLU: a Chinese-language multi-subject knowledge benchmark spanning multiple levels including K12, built primarily on four-option factual items; commonly used alongside C-Eval, AGIEval, GAOKAO-Bench, and Xiezhi to assess Chinese foundational knowledge (specific model scores should follow each model's official or third-party reports [TBD: CMMLU model-score methodology]).
- E-EVAL (Chinese K12, 4,351 items) — average accuracy by model (Hou et al., 2024, few-shot; per its Table 4. To keep the figures verified, only the reviewed average-accuracy column is listed here; grade-level and humanities/STEM breakdowns should be taken from the original paper's Table 4 at arXiv:2401.15927 and are omitted from this table):

Model	**E-EVAL average accuracy (few-shot, %)**
Qwen-72B-Chat	88.9
ERNIE-Bot 4.0 (文心一言 4.0)	85.5
Yi-34B-Chat	76.9
GPT-4.0	70.6
ChatGPT	54.6
Random baseline	25.0

Reading guide: (1) On Chinese K12 content, Chinese-dominant models outperform English-dominant models overall, with several Chinese models beating GPT-4.0's average score — the reverse of the ranking on English-language leaderboards, and a direct demonstration of scenario validity. (2) Every model scores higher on humanities than on STEM subjects (language comprehension is a strength, logical reasoning a weakness). (3) The counterintuitive "elementary ≤ middle school" pattern is widespread; the paper's showcase example is a bare-bones elementary math item that the top three models (Qwen-72B, ERNIE-Bot 4.0, Yi-34B) all answered incorrectly, making a commonsense error along the lines of treating 92 seconds

as less than 75 seconds. All of the above reflects that paper's own measured methodology (few-shot); reproducing it requires locking down the same prompts, same versions, and same time point.

Fig. 4 Average accuracy by model on the Chinese K-12 benchmark E-EVAL



Source: Hou et al., 2024 (E-EVAL, arXiv:2401.15927, few-shot, Table 4 average column).

GAOKAO-style subjective items and process capability

- GAOKAO-Bench: contains actual college-entrance-exam questions — about 1,781 objective items plus 1,030 subjective items — spanning multiple subjects and item types including Chinese, math, and physics (Zhang et al., 2023). Because real exam questions circulate widely, this benchmark's contamination resistance is weaker than E-EVAL's, and head-to-head comparisons should be especially wary of memorization-driven high scores.
- GAOKAO-Eval (54 teachers manually scoring GPT-4o, Qwen2-72B, GLM-4, Yi-34B, Mixtral-8x22B, and WQX): the central finding is that a high score does not necessarily reflect human-aligned capability. Models exhibited semi difficulty-invariant scoring (near-identical scores across items of very different difficulty, unlike the human pattern of declining scores as difficulty rises) and high variance on items of the same difficulty. The 54 teachers' inconsistent-scoring rate on models' subjective-item responses exceeded 32%, with humanities subjects (politics reaching 41.18%) markedly higher than STEM — reflecting that models are less stable

on open-ended items requiring contextual and value judgment (Lei et al., 2024). The takeaway is that head-to-head comparisons on subjective items must report scoring consistency, not just a single average score; a model with a high average but high variance and difficulty-invariant scoring has an unreliable high score.

Pedagogical ability (process validity) — this is the group of benchmarks with the most incremental value for education-vertical evaluation, and the one that most directly exposes the gap between "answering well" and "teaching well":

- OpenLearnLM (Lee et al., 2026): proposes a three-axis Knowledge/Skill/Attitude framework. The knowledge axis has 2,304 items (918 content-knowledge plus 1,386 pedagogical-knowledge); the skill axis is organized in a hierarchy of 6 hubs / 11 roles / 46 scenarios / 81 sub-scenarios with over 120,000 items difficulty-graded by Bloom's taxonomy; the attitude axis includes 14 situational items and introduces "alignment faking" detection (comparing whether a model behaves consistently when it believes it is being monitored versus not). Testing seven frontier models — Claude-Opus-4.5, GPT-5.2, Gemini-3-Pro, Grok-4.1-fast, Kimi-K2-thinking, GLM-4.7, and DeepSeek-v3.2 — the key finding is a stark gap between pedagogical process correctness (about 56.6%) and answer accuracy (about 97.3%) (per the original paper, this particular 56.6%/97.3% pairing is carried over from prior work it cites, while the cross-axis correlations and individual model scores are OpenLearnLM's own measurements); the three axes correlate weakly or negatively with each other ($r \approx -0.51$ to -0.63); Claude-Opus-4.5 scored lowest on content knowledge (about 66.3%) yet highest on the skill axis — confirming the need for multi-axis evaluation.
- MRBench (Maurya et al., 2025, NAACL): 192 tutoring dialogues comprising 1,596 responses from seven tutoring systems (including human tutors), with gold annotations across an eight-dimension pedagogical rubric. The authors also found empirically that LLM judges (such as Prometheus2) correlate poorly with human labels on fine-grained pedagogical dimensions — i.e., LLM-as-judge is unreliable here.
- MathTutorBench (Macina et al., 2025, EMNLP Oral): uses a "scaffolding reward model" (trained by contrasting expert and novice tutor dialogue) to score tutoring utterances for scaffolding quality, achieving high precision in distinguishing expert-style from novice-style dialogue. It covers general LLMs, LLM tutors, and dedicated math reasoners, concluding that problem-solving ability does not automatically translate into pedagogical ability, and that the two trade off against each other as specialization deepens.
- Answer-leakage robustness (Zhao et al., 2026, ACL): stress-tests tutors' answer-leakage rate using an "adversarial student agent" that repeatedly presses for answers, and proposes a fine-tuned adversarial-student agent as the core of a standardized stress test along with several anti-leakage strategies — directly applicable to D6 safety evaluation.

A synthesized reading of all four anchor groups. Taken together, these four groups support three judgments that form the empirical backbone of this chapter's core argument. First, **rankings flip depending on the benchmark** — the GPT series leads on the English-language MMLU yet is overtaken by several Chinese models on the Chinese K12 benchmark E-EVAL (Qwen-72B-Chat averaging 88.9 versus GPT-4.0's 70.6), proving that "strongest in general use" does not mean "best for education," and that language and grade level are hard constraints. Second, **the frontier of difficulty has moved from knowledge items to process and pedagogy items** — once GSM8K/MMLU saturated, the real discriminative power emerged in MMLU-Pro, process-level scoring on GAOKAO subjective items, and pedagogical-validity benchmarks. Third, **pedagogical validity is the acknowledged weak point today** — multiple 2025–2026 benchmarks consistently show a gap between getting the answer right and getting the process/pedagogy right (56.6% versus 97.3%), which is simultaneously a risk and an opportunity window for education-vertical models. On this basis, this institute's self-built item-bank comparison should focus on process-level scoring and five-scenario pedagogical validity, rather than retreading already-saturated knowledge items.

8.4.2 Head-to-Head Results Table (to Be Filled In from Self-Built Item-Bank Testing)

Head-to-head results should be presented as a structured table (empirical data from this institute's self-built five-scenario item bank is pending; the anchor scores above serve only as capability reference points and should not be added across benchmarks to the scores in the table below):

Model/Product Tested	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**Teaching-scenario weighted score**	**Learning-scenario weighted score**	**API cost (currency labeled)**	**On-device feasibility**
[TBD: Model/Product A]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]
[TBD: Model/Product B]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]
[TBD: Model/Product C]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]	[TBD]

8.5 Capability Radar Charts: An Evidence-Based Visual Language

The seven capability dimensions lend themselves naturally to a **radar chart**, which renders a single model's "capability shape" and supports overlaying multiple models for comparison. Radar charts suit education evaluation particularly well because the core question in model selection is never "who scores highest" but "whose capability shape fits my scenario." Two models with near-identical composite scores can have completely different capability shapes: Model A might excel at knowledge and reasoning but lag on pedagogy, while Model B excels at pedagogy and grade-level fit but trails slightly on knowledge — a difference a composite score conceals but a radar chart reveals at a glance. The real value of a radar chart is how plainly it exposes a lopsided profile: a model strong on knowledge and reasoning but weak on pedagogy and safety will show a silhouette that leans visibly toward D1/D2 and caves in around D3/D6 — signaling that it behaves more like a problem-solver than a teaching tool. This is the visual expression of exactly the "answering well ≠ teaching well" phenomenon that OpenLearnLM (56.6% pedagogical process correctness versus 97.3% answer accuracy) and MathTutorBench (problem-solving ability ≠ pedagogical ability) both independently confirm.

- **Single-model radar.** Displays one model's seven-dimension score profile, making strengths and weaknesses immediately identifiable; the "dents" in the silhouette are exactly where a buyer should be most cautious.
- **Multi-model overlay.** Overlays 2–3 models on the same chart to support selection decisions [TBD: models included and their measured scores]; the overlay makes complementary strengths visually obvious, providing the evidentiary basis for choosing different models for different scenarios.
- **Scenario-weighted radar.** Redraws the same model's chart under different weightings — teaching, learning, research/PD — to show plainly that changing the scenario changes the ranking. For example, the same model's silhouette might contract under "learning" weighting (which elevates D3 pedagogy and D4 grade-level fit) if those two dimensions happen to be weak, and expand under "research/PD" weighting (which elevates D1 knowledge and D5 anti-hallucination) if those dimensions happen to be strong — a single chart demonstrating that the same model's fit changes with the scenario, and that model selection cannot be divorced from scenario context.

A few charting conventions matter here: (1) scores on each dimension must first be normalized to a common scale (e.g., 0–100), since raw scores in different units — accuracy percentages versus a 0–5 rubric score — cannot be plotted on the same chart otherwise; (2) the order in which dimensions are arranged should be fixed and disclosed, to avoid misleading area impressions caused by reordering; (3) the evaluation timestamp and tested version should always be labeled.

Figure 8-1 Seven-dimension capability radar for education-vertical LLMs (illustrative; data [TBD: measured scores]). Charting methodology, models tested, and evaluation timestamp are given in Section 8.3.3.

A radar chart should always be paired with its corresponding head-to-head results table: the radar shows the "shape," the table shows the "numbers," and the two corroborate each other so readers do not draw conclusions from visual impression alone. One caution: never treat radar-chart "area" as a stand-in for a composite score — area scales with dimension ordering and unit choice and is not comparable across charts, and using it that way simply reproduces the single-score trap in a different visual form.

8.6 Evolution Timeline and Trend Assessment

Laying successive rounds of evaluation results along a timeline makes it possible to observe the pace and inflection points of education-vertical models' capability evolution. Based on the public facts surveyed in this chapter, two threads — **benchmark evolution** and **capability evolution** — support the following **qualitative** judgments (containing no unverified specific figures).

On the **benchmark evolution** thread: around 2023, knowledge-oriented multiple-choice leaderboards dominated (MMLU, C-Eval, GAOKAO-Bench). In 2024, two distinct strands emerged — "difficulty-extended" knowledge benchmarks (MMLU-Pro reopening discriminative power on a saturated MMLU via 10-option strong distractors) and purpose-built Chinese K12 benchmarks (E-EVAL), alongside meta-evaluation work questioning whether "high score" equals "high capability" (GAOKAO-Eval, using 54 teachers' manual scoring to expose semi difficulty-invariant scoring). Starting in 2025, pedagogical-ability (process-validity) benchmarks arrived in a wave (MRBench/NAACL, MathTutorBench/EMNLP, the BEA 2025 shared task). By 2026, the frontier extended further toward "trustworthy floors" (OpenLearnLM's three axes and alignment-faking detection, answer-leakage robustness/ACL 2026). The direction of benchmark evolution is clear: from testing knowledge, to testing reasoning process, to testing pedagogical process, and finally to testing the floor for safety and integrity.

On the **capability evolution** thread, three trends stand out:

- **From dialogue to agents.** Single-turn Q&A is giving way to multi-step planning, tool use, and persistent memory, making the pedagogy dimension (D3) the key differentiator. Domestic vertical products are rapidly adopting stronger reasoning and agentic capability: iFLYTEK released its "Spark" deep-reasoning model X1 in December 2024, built for deep reasoning on fully domestic compute platforms. After DeepSeek R1 was open-sourced in early 2025, a large number of education teams quickly integrated it into their internal production pipelines — NetDragon (HK:0777)'s education business integrated R1 immediately after its release and, building on its already extensive deployment of frontier AI tools, constructed an internal

capability platform called AI Hub and a complete AI production line, with plans to deploy an AI production line in Hong Kong to expand educational content resource production. On the evaluation side, this requires expanding from single-turn item accuracy to multi-turn pedagogical validity and correctness of tool use, echoing the agentic orchestration paradigm covered in Chapter 9 of this Blue Book.

- **From cloud to on-device.** The shift on-device is pushing up the weight of D7, requiring evaluation to incorporate latency, power consumption, and offline availability, connecting with the AI-hardware evaluation (glasses/learning tablets) covered in Chapter 7. Learning-tablet head-to-head comparisons have already become a major battleground for on-device education LLMs: the Yinhe/Spark/Jiuzhang/ERNIE/Kanyun models run respectively on hardware from Zuoyebang, iFLYTEK, TAL, Baidu's Xiaodu, and Yuanfudao, and their cloud-side capability may well rank differently from their on-device performance (a model can be strong in the cloud yet underdeliver on-device once compute and latency constraints bite; specific devices and rankings should follow third-party testing [TBD: on-device head-to-head methodology and measured scores]). This requires evaluation to report "real-device end-to-end usability" and "cloud API accuracy" as two separate figures rather than conflating them.
- **From "smarter" to "more trustworthy."** Hallucination resistance (D5) and safety alignment (D6) are shifting from bonus categories to admission requirements, especially in scenarios facing minors. New dimensions — answer-leakage robustness (guarding against students extracting answers, Zhao et al., 2026) and alignment-faking detection (whether a model behaves consistently when monitored versus unmonitored, Lee et al., 2026) — are entering the academic evaluation field of view, signaling that the center of gravity in education evaluation is shifting from capability ceilings to trustworthiness floors. The direct policy implication is that K12-facing products should treat these floor dimensions as mandatory gates for market entry, rather than patching them after something goes wrong.

Figure 8-2 Capability evolution timeline for education-vertical LLMs (illustrative; specific events and dates [TBD: complete evaluation batch and time-window data]; known anchor points: C-Eval/GAOKAO-Bench 2023; E-EVAL/MMLU-Pro/GAOKAO-Eval 2024; MRBench 2025/NAACL and MathTutorBench 2025/EMNLP; OpenLearnLM and answer-leakage robustness 2026/ACL).

8.7 Limitations and Usage Recommendations

8.7.1 Inherent Limitations of Evaluation

Every evaluation is "measurement under assumptions," and its boundaries deserve honest disclosure:

- **Snapshot effect.** Conclusions are valid only for the evaluation timestamp and the specific version tested; models need retesting after every iteration. With models updating monthly or

even weekly, a head-to-head comparison from three months ago may already be stale. Legacy benchmarks (GSM8K, original MMLU) have saturated, with most flagship models packed into the high-score range within measurement noise of each other, and need replacing with process-level scoring and harder benchmarks (MMLU-Pro, competition math, pedagogical-validity benchmarks).

- **Item-bank bias and contamination.** The coverage and representativeness of the item bank directly bound how far its conclusions can generalize; training-set leakage systematically inflates scores (decontamination can knock GSM8K accuracy down by as much as roughly 13 percentage points for some models), which is why private test sets, contamination probes, and periodic rotation are all necessary. An item bank covering only a handful of subjects or a single grade band cannot be generalized into a claim about a model's overall capability — that would be mistaking a local measurement for a global judgment.
- **High score ≠ strong capability.** GAOKAO-Eval demonstrates that a model can maintain a high score while exhibiting patterns inconsistent with human behavior, such as semi difficulty-invariant scoring and high variance on same-difficulty items; the 54 teachers' scoring-inconsistency rate exceeded 32% (41.18% for politics). This means an attractive average score can sit on top of a highly unstable, distinctly non-human-like response distribution. Subjective items must therefore report scoring consistency and variance, guarding carefully against the illusion of a high score.
- **Scoring bias.** LLM-as-judge exhibits self-preference bias (overrating its own or same-family outputs), position bias (swapping candidate order alone can shift pairwise judgment accuracy by more than 10 points), and verbosity bias (favoring longer answers), and correlates poorly with human labels on fine-grained pedagogical dimensions. Without bidirectional blind review and human spot-audit cross-checking, scaling up automated scoring simply writes these systematic biases into the conclusions.
- **Single-score trap.** Capability axes correlate weakly or even negatively with each other (OpenLearnLM reports r as low as roughly −0.6); substituting a single composite score for multidimensional evidence conceals a lopsided profile — a model that aces knowledge but fails pedagogy and safety might still post a decent "average," while being entirely unsuitable for direct student-facing use. Model selection for education should always come back to your specific scenario, grade level, and constraints.
- **Cultural and contextual bias.** Most pedagogical-validity benchmarks (MRBench, MathTutorBench, OpenLearnLM, and others) are built primarily around English-language math tutoring; transplanting their rubrics and conclusions to Chinese, to the humanities, or to the Chinese curriculum-standard context requires local recalibration rather than a direct transfer of scores.

8.7.2 A Decision Checklist for Model Selection (Recommendations for Practitioners)

(1) **Fix the scenario before you look at the score.** First establish who the user is (teacher or student), the grade level, the subject, the setting (classroom or after-school), and the deployment mode (cloud or on-device) — then consult the corresponding scenario-weighted score, not a general composite score or a headline-leaderboard rank.

(2) **Weigh effectiveness alongside cost and on-device feasibility.** Decide using effectiveness score, API cost, latency, and on-device feasibility together; costs must always be currency-labeled (RMB/USD/HKD) and never mixed across currencies. Large-scale deployments especially need to account for total cost of ownership, not just a single-point effectiveness figure.

(3) **Treat floor dimensions as pass/fail gates.** Set safety, hallucination resistance, and protection against answer leakage/ghostwriting/exam leaks as admission requirements, not features to be traded off against effectiveness gains. For any product facing minors, clear the safety bar before discussing how smart it is.

(4) **Look at both process and consistency.** For subjective items, check scoring consistency and variance; for objective items, check contamination-resistance design — neither can be skipped. Be wary of a model with a high average but low stability.

(5) **Establish a recurring retesting cycle.** Treat the "evaluation timestamp" as the operative reference and turn evaluation into a standing practice — item-bank rotation, contamination probes, periodic retesting — rather than a one-time procurement input.

(6) **Multiple axes, not a single score.** Use the radar chart to see the "shape" and the head-to-head table to see the "numbers," to identify a lopsided profile; do not put stock in "area" or a "composite score."

Specific evaluation data, item banks, and lists of models tested will be filled in with empirical results in this Blue Book's data supplement and subsequent updates; every [TBD] gap will be labeled with a genuine source once filled, never padded with a guessed figure.

Chapter References

1. Hendrycks D., Burns C., Basart S., et al. "Measuring Massive Multitask Language Understanding (MMLU)." ICLR 2021. See MMLU leaderboard and documentation. (57 subjects, approx. 15,900 items; GPT-4 5-shot approx. 86.4%, human expert baseline approx. 89.8%)

2. OpenAI. "GPT-4 Technical Report." 2023. https://cdn.openai.com/papers/gpt-4.pdf (MMLU 86.4%, etc.)

3. Wang Y., Ma X., Zhang G., et al. "MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark." NeurIPS 2024 Datasets & Benchmarks. https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-

Paper-Datasets_and_Benchmarks_Track.pdf (10 options; approx. 16–33% drop relative to MMLU; GPT-4o approx. 72.6%)

4. Huang Y., Bai Y., Zhu Z., et al. "C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models." NeurIPS 2023. https://cevalbenchmark.com/ ; paper https://arxiv.org/abs/2305.08322 (52 subjects, 13,948 items; at release only GPT-4 exceeded 60% average; test set released starting 2025-07)

5. Hou J., Ao C., Wu H., et al. "E-EVAL: A Comprehensive Chinese K-12 Education Evaluation Benchmark for Large Language Models." (Shenzhen Institute of Advanced Technology, CAS, et al.) arXiv:2401.15927, 2024. https://arxiv.org/abs/2401.15927 ; data https://eevalbenchmark.com (4,351 items, 9 subjects, elementary/middle/high; Table 4 gives per-model scores; documents the counterintuitive "elementary ≤ middle school" phenomenon and the "92<75" error case)

6. Zhang X., Li C., Zong Y., et al. "Evaluating the Performance of Large Language Models on GAOKAO Benchmark." arXiv:2305.12474, 2023. https://arxiv.org/abs/2305.12474 (approx. 1,781 objective items + 1,030 subjective items)

7. Lei F., Liu P., Zhang Z., et al. "GAOKAO-Eval: Does High Scores Truly Reflect Strong Capabilities in LLMs?" arXiv:2412.10056, 2024. https://arxiv.org/abs/2412.10056 (54 teachers manually scoring; inconsistent-scoring rate >32%, politics 41.18%; "semi difficulty-invariant scoring"; tested GPT-4o, Qwen2-72B, GLM-4, Yi-34B, Mixtral-8x22B, WQX)

8. Maurya K. K., Srivatsa K. V. A., Petukhova K., Kochmar E. "Unifying AI Tutor Evaluation: An Evaluation Taxonomy for Pedagogical Ability Assessment of LLM-Powered AI Tutors." NAACL 2025. arXiv:2412.09416. https://arxiv.org/abs/2412.09416 ; code https://github.com/kaushal0494/UnifyingAITutorEvaluation (MRBench: 192 dialogues, 1,596 responses, 8-dimension rubric; finds LLM-as-judge unreliable on pedagogical dimensions)

9. Kochmar E., Maurya K. K., Petukhova K., Srivatsa K. V. A., Tack A., Vasselli J. "Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors." BEA 2025. arXiv:2507.10579. https://arxiv.org/pdf/2507.10579 (four axes: mistake identification / mistake location / pedagogical guidance / actionability)

10. Macina J., Daheim N., et al. "MathTutorBench: A Benchmark for Measuring Open-ended Pedagogical Capabilities of LLM Tutors." EMNLP 2025 (Oral). arXiv:2502.18940. https://aclanthology.org/2025.emnlp-main.11/ ; code https://github.com/eth-lre/mathtutorbench (scaffolding reward model; problem-solving ability ≠ pedagogical ability trade-off)

11. Lee et al. "OpenLearnLM Benchmark: A Unified Framework for Evaluating Knowledge, Skill, and Attitude in Educational Large Language Models." arXiv:2601.13882, 2026. https://arxiv.org/html/2601.13882 (Knowledge/Skill/Attitude three-axis framework; pedagogical process correctness ≈56.6% vs. answer accuracy ≈97.3%; cross-axis correlation r≈−0.51 to −0.63;

tested Claude-Opus-4.5, GPT-5.2, Gemini-3-Pro, Grok-4.1-fast, Kimi-K2-thinking, GLM-4.7, DeepSeek-v3.2)

12. Zhao J., Knežević M., Käser T. (EPFL ML4ED). "Evaluating Answer Leakage Robustness of LLM Tutors against Adversarial Student Attacks." ACL 2026. arXiv:2604.18660. https://arxiv.org/abs/2604.18660 ; code https://github.com/epfl-ml4ed/tutor-robustness-eval (adversarial student agent inducing tutor "answer leakage"; anti-leakage stress testing and defense strategies)
13. Wataoka K., et al. "Self-Preference Bias in LLM-as-a-Judge." arXiv:2410.21819, 2024. https://arxiv.org/abs/2410.21819 (judge self-preference bias); related: Ye J., et al. "Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge." arXiv:2410.02736, 2024 (position/verbosity and other biases; swapping order in pairwise judgment shifts accuracy by more than 10 points)
14. iFLYTEK — "Spark" deep-reasoning model X1 launch (2024-12, deep reasoning on fully domestic compute platforms); reported by stcn.com and others. https://www.stcn.com/article/detail/1098277.html (capability claims per official announcement [TBD: X1 specific evaluation methodology])
15. Overview of education-vertical products including TAL's "Jiuzhang" (MathGPT, formerly the Xueersi Math Model), NetEase Youdao's "Ziyue," Yuanfudao's "Kanyun," and Zuoyebang's "Yinhe"; industry coverage and learning-tablet comparisons per Zhihu, 36Kr, and others (specific versions and scores per vendor announcements and third-party testing [TBD: evaluation methodology for each vertical model]). https://36kr.com/p/3381670432831494
16. NetDragon (HK:0777) education business: coverage scale of 101 Education PPT, integration of DeepSeek R1 following its release and construction of the internal AI Hub capability platform, planned deployment of an AI production line in Hong Kong, etc.; per NetDragon's official site and coverage by China News Service and Securities Times. https://www.nd.com.cn/products/education.shtml ; https://www.chinanews.com.cn/cj/2024/01-25/10152751.shtml

Chapter 9 Evaluating AI Education Hardware: AI Glasses, Learning Tablets, and On-Device Agents

9.1 Why Hardware Evaluation Needs Its Own Chapter in 2026

In its first phase, generative AI entered education mostly as a software interface layered on top of conversational large models, and evaluation naturally focused on language ability, knowledge coverage, and conversational quality. By 2026, a structural shift is under way: generative capability is moving **from the cloud and the browser down onto on-device hardware** — away from "opening a chat box" and toward embodied interaction that is worn, situated, and multimodal. The market signals behind this shift are hard to miss. According to IDC, global smart-glasses shipments grew 64.2% year-over-year in the first half of 2025 to reach 4.065 million units; for the full year, global shipments reached roughly 14.773 million units, up 44.2% year-over-year, with China alone shipping about 2.46 million units — an 87.1% increase that far outpaced the global average (IDC, 2025–2026; 21st Century Business Herald, 2026). Running in parallel is the smartification of home learning terminals: in the first quarter of 2025, all-channel sales of AI learning tablets in China reached 1.265 million units, up 29.4% year-over-year (RUNTO, 2025). AI glasses, next-generation learning tablets, and on-device agents running on local chips together make up the tangible, wearable, touchable face of generative AI in education.

Hardware evaluation differs fundamentally from software evaluation. Software can be benchmarked head-to-head at a uniform API layer, but hardware evaluation has to answer three questions at once: **first, the capability of the model it carries** (multimodal understanding, generation, reasoning); **second, the engineering constraints of the device itself** (compute, power draw, latency, battery life, wearability/interaction form); **third, fit for the education context** (grade-level alignment, content compliance, parent/teacher controllability, data and privacy). Evaluating any single layer in isolation distorts the picture — a pair of glasses with a very capable cloud model may still be unusable for a full class period if on-device latency is too high or battery life too short; a learning tablet with striking screen specs may be unsuitable for younger learners if it lacks content governance and simply hands over answers. Hardware pins the "abstract model score" to "the concrete moment of use," which is exactly why it demands an evaluation methodology distinct from pure software benchmarking.

Viewed as an industry trajectory, the three categories of education AI hardware did not emerge in isolation — they mark the convergence of three technology curves between 2024 and 2026. The first is the maturing of wearable computing and optical displays, which has carried glasses from "geek toy" to a consumer product light enough (40–70 g) for daily wear. The second is the arrival of

vertical education large models and agent orchestration, which has turned learning tablets from "problem-bank lookup tools" into "study companions that ask follow-up questions." The third is the breakthrough in on-device chip compute and small-model quantization, which has moved "running a good-enough model locally" out of research papers and into shipping devices. The common driver beneath all three curves is generative AI's migration from centralized cloud services toward distributed, embodied form factors. For education, this migration matters more than it might elsewhere — learning has always happened in specific times, places, and situations (a classroom, an experiment, a trip, a wrong answer on a worksheet), and only by pulling AI out of the browser tab and embedding it in those real situations can the field move from "retrieving knowledge" to "accompanying learning."

This chapter treats "education AI hardware" as a standalone object of evaluation, builds a cross-form-factor framework for comparing it, and works through evaluation dimensions, methods, and observations separately for AI glasses, learning tablets, and on-device agents. A note on evaluation discipline is warranted up front: every product specification, price, and market figure cited in this chapter comes from a public, primary source — manufacturer websites, launch-event materials, authoritative institutional reports, mainstream media — each listed at the end of the chapter; any cross-sectional capability score not yet measured by our institute's evaluation lab is marked `[TBD: lab measurement]` rather than filled in with an estimate. This is not a formality. Hardware evaluation is uniquely vulnerable to being hijacked by "marketing specs" — manufacturers are happy to disclose screen resolution, camera megapixels, and battery capacity (the numbers that look good on paper), but rarely publish end-to-end latency, weak-network usability, or subject-matter hallucination rates (the numbers that reveal what actually happens in use). It is precisely the latter category that determines whether a product succeeds or fails in a real classroom. This chapter therefore deliberately separates "publicly verifiable specifications" — cited, sourced, checkable today — from "capability scores that require lab testing" — left blank pending measurement, never fabricated. For the systematic classroom application of AI glasses, see our institute's *AI Glasses in Education: 2026 Development Report*; for embodied classroom agents, see our *Global Education Robotics Blue Book 2026*; for on-device agent orchestration and memory paradigms, see Chapter 7 of this Blue Book; for vertical education large-model benchmarks, see Chapter 8.

9.2 A Framework for Evaluating Education AI Hardware

9.2.1 Three Categories of Evaluation Target

- **AI glasses (first-person wearables).** Built around first-person camera capture, voice interaction, and either near-eye display or pure audio feedback, AI glasses are organized around the idea of "see it, ask about it" on-the-go assistance. By display capability, the market splits into three tiers: audio-only/no-display (e.g., the standard Xiaomi AI Glasses, Rokid AI Glasses Style, Huawei

smart glasses 2), monocular light-display (e.g., Meta Ray-Ban Display), and binocular display (e.g., Rokid Glasses). Typical education uses include real-time guidance during labs and hands-on training, situated language learning and translation, accessibility support, and identification/explanation during field trips.

- **AI learning tablets (home learning terminals).** These are dedicated terminals for K12 home self-study, built around subject-content systems, wrong-answer/learning-diagnostic engines, eye-care design, and parental controls. The key upgrade in 2026 is the embedding of, or connection to, vertical education large models and an agent-driven "AI study companion." Leading manufacturers include iFLYTEK, Bbkeducation (Step by Step), Xueersi, Zuoyebang, Squirrel AI, Youdao, and Xiaodu.
- **On-device agents (local-inference form factors).** This category spans NPU-equipped learning terminals, offline voice assistants, desktop study-companion robots, and phones/tablets running local foundation models — with the defining trait being the ability to complete core inference and interaction under **weak- or no-network conditions**, balancing privacy against latency. The three categories are not mutually exclusive: a learning tablet can double as an on-device agent, and a pair of glasses may embed a small offline model (Rokid Glasses, for instance, claims offline translation for 6 languages).

9.2.2 A Six-Dimension Evaluation Framework

This chapter proposes a six-dimension framework tailored to education, with observable indicators nested under each dimension:

Dimension	**Core Question**	**Representative Indicators (examples)**
Model capability	Can multimodal understanding and generation handle learning tasks?	Subject Q&A accuracy, multimodal recognition accuracy, reasoning-chain completeness, refusal/correction rate `[TBD: lab measurement]`
On-device engineering	Can the hardware sustain interaction reliably?	End-to-end latency, offline usability rate, battery life, heat/power draw, weight `[TBD: lab measurement]`
Interaction experience	Is it comfortable and sustainable to use?	Speech-recognition character error rate, wake/response success rate, wearing/operation comfort, near-eye display clarity `[TBD: lab measurement]`
Education fit	Does content match grade level and curriculum standards?	Grade-level coverage, alignment with curriculum standards/textbook editions, completeness of content

		system `[TBD: lab measurement]`
Governance and safety	Is it safe, compliant, and controllable?	Harmful-content interception rate, degree of data localization, granularity of parent/teacher controls, recording-notice compliance `[TBD: lab measurement]`
Sustainability	Can it be used and iterated on long-term?	OTA update frequency, ecosystem openness, per-unit learning cost, content subscription model `[TBD: data]`

The "model capability" row shares the same benchmark suite as our institute's evaluation of vertical education large models (Chapter 8), which lets us run a "software baseline—hardware measurement" comparison end to end: the same subject-matter question is first scored for baseline accuracy at the cloud API layer, then re-tested on the actual device (on-device inference, weak network, worn/handheld interaction). The gap between the two is what we call "hardware loss" — a unique output of this framework that a pure software evaluation cannot produce.

The six dimensions are not equally weighted; their relative importance shifts with grade level and use case. For younger learners (pre-K through early elementary), governance/safety and education fit should carry substantially more weight than model capability — a product that occasionally can't answer is far safer than one that hands out an incorrect solution step or inappropriate content. For older, more self-directed learners (middle and high school), model capability and interaction experience matter more. For special education and accessibility use cases, interaction experience — particularly speech-recognition error rate and response stability — is close to a hard veto. The radar charts in this chapter will therefore report both a "general weighting" and a "grade-adjusted weighting" once lab testing is complete, so a single composite score doesn't paper over context-specific differences. There are also genuine engineering trade-offs among the six dimensions: on-device engineering's "offline usability" often comes at the cost of some "model capability" (local small models trail cloud-hosted large ones); interaction experience's "lightweight wearability" often comes at the cost of battery life (a smaller battery); and governance/safety's "data localization" often trades against the pace of cloud-side iteration under "sustainability." The point of evaluation isn't to find a product that maxes out all six dimensions — no such product exists — but to show where each product sits on that trade-off curve, so buyers can weigh the dimensions according to their own priorities.

9.2.3 Methodological Principles

- **Scenario-based testing comes first.** The unit of evaluation should be a real learning task (a problem, an experiment, a translation, a class period), not an isolated benchmark run.

Translation latency on glasses should be tested at the pace of real conversation; explanation quality on a learning tablet should be tested across a complete wrong-answer-correction workflow.

- **End-to-end timing.** Latency should be measured across the full chain from "user initiates" to "usable response is delivered," with on-device inference and cloud round-trip measured separately — don't let "time to first token" mask "time to a complete, readable answer."
- **Weak-network/no-network stress testing.** Dedicated tiers for no connection and weak connection (e.g., throttled to 1 Mbps) test on-device fallback capability — this is the normal condition in real education settings (outdoor field trips, rural schools, dorm-network peak hours), not an edge case.
- **Double-blind checks on education fit.** Subject-matter teachers should double-blind-score curriculum alignment, explanation correctness, and guided-instruction design, so "marketing selling points" don't substitute for "teaching effectiveness."
- **Reproducibility and evidentiary traceability.** Test-set composition, scoring criteria, and device configuration should be disclosed publicly; every publicly reported score should be traceable `[TBD: methodology documentation source]`.

On definitions, this framework deliberately separates several metrics that are easily conflated, precisely because that conflation is where "numbers that look good" diverge from "numbers that are comparable." **Latency** must distinguish "time to first token" (the user's sense that a response has started) from "time to a complete, readable answer" (the moment it's actually usable) — for translation applications, "steady-state latency across continuous conversation" must also be measured, since a fast cold start doesn't guarantee the device can keep pace with natural speech. **Accuracy** must distinguish "closed-book accuracy" (the model's intrinsic knowledge) from "open-book/RAG accuracy" (drawing on local documents); education use cases place more weight on the traceability of the latter. **Offline usability** should be measured by "task completion quality," not "whether the device powers on" — an assistant that launches without a connection but answers off-topic should be scored as having low offline usability. **Battery life** must distinguish "standby/light-use rated figures" from "measured figures under heavy load" (continuous translation, continuous capture, extended inference), which can differ by a factor of several. **Character/word error rate (CER/WER)** must be measured under real ambient noise (classroom background chatter, outdoor wind noise), not a quiet lab. These distinctions may look like fine print, but they are the primary source of the gap between "what the manufacturer claims" and "what the user actually experiences" — and they are the core value this framework adds beyond marketing copy.

9.3 AI Glasses: Evaluating the First-Person Learning Assistant

9.3.1 Product Landscape and Capability Generations

By 2026, AI glasses in education are making the leap from "audio assistant" to "multimodal visual assistant": first-person camera capture plus voice interaction plus large-model understanding is turning "pull out your phone to photograph the problem" into "look at it and ask." This shift in interaction paradigm is worth unpacking. Photographing a problem with a phone requires five steps — pull out the phone, open the app, aim, shoot, wait — and photo-solver apps have long been criticized as glorified answer-copying tools. Glasses compress that interaction to nearly zero steps, and because what the glasses see is "the real scene the learner is actually looking at" (a page in a book, an experiment, a conversation), the form factor naturally leans toward "aiding understanding" rather than "substituting for the answer." But there's a flip side to this coin: a sharp rise in privacy risk, since a first-person camera means continuously recording other people and the surrounding environment — this is the fundamental governance challenge that sets glasses apart from phones. In terms of form factor, the market splits roughly into audio-only/no-display, monocular light-display, and binocular display tiers, with weight, display capability, battery life, and price all scaling accordingly — as does education fit: no-display models win on lightness and low intrusiveness, suited to voice Q&A and translation playback; display models win on visualizing information (captions, teleprompting, step overlays), but add weight, power draw, and motion-sickness risk. The table below summarizes specifications for several representative public models, illustrating generational differences (specifications drawn from manufacturer websites and launch materials; cross-sectional capability scores await lab testing).

Model	Display Form	Weight	Chip/Camera	Battery Life (rated)	Reference Price	Key Capabilities
Rokid Glasses	Binocular monochrome Micro-LED + diffractive waveguide	49 g	Qualcomm Snapdragon AR1 + 12MP camera	~4 hours (charging case enables repeated top-ups)	RMB 2,499	Real-time translation (89 languages online, 6 claimed offline), teleprompting, AI Q&A, photo capture
Meta Ray-Ban Display	Monocular monochrome display (right lens, 600×600, 20° FOV, 30–5000 nit)	69 g	Paired with Neural Band EMG wristband for interaction	~6 hours (charging case tops up to ~24 hours); wristband ~18 hours	US$799 (including wristband)	Display overlay, gesture/EMG interaction, translation, capture
Xiaomi AI	No display	~40 g	12MP, 105°	`[TBD:	RMB 1,999	Photo/video

Glasses	(audio + camera)	(titanium frame)	wide angle, 0.8s capture	official rated battery life]`	and up (electrochromic version also available)	capture, Q&A, translation, calls, payment
Huawei smart glasses 2	No display (audio)	`[TBD: official weight]`	Audio-focused	`[TBD: official rated battery life]`	RMB 1,699 and up	Translation, playback, health monitoring, phone pairing
Rokid AI Glasses Style	No display (audio + camera)	38.5 g	12MP camera	`[TBD: official rated battery life]`	`[TBD: official price]`	ChatGPT/Gemini access, real-time translation, hands-free calls

In 2025, China's domestic AI glasses market entered what's been called a "hundred glasses war," with Huawei, Xiaomi, OPPO, Alibaba (Quark), Baidu (Xiaodu), Li Auto, and others piling in; product lines rapidly expanded from "audio-only glasses" to "camera-equipped" and then to "display-equipped" (Tencent News/Sina Finance, 2025). Price bands have widened accordingly — from basic Bluetooth-audio models in the low hundreds of RMB (Xiaomi AI Glasses' entry model at roughly RMB 499), to thousand-RMB photo-plus-translation models (Xiaomi at RMB 1,999 and up, Huawei smart glasses 2 at roughly RMB 1,699 and up, titanium/square-frame models at RMB 2,299), up to display models above RMB 2,000 (Rokid Glasses at RMB 2,499, Meta Ray-Ban Display at US$799). For education buyers, the relationship between price band and capability band deserves attention: more expensive doesn't automatically mean better suited to education. A RMB 2,499 display model with insufficient information density or battery life that can't cover a full class period may deliver less educational value than a cheaper audio-only model focused on translation with longer battery life — what matters is matching scenario to capability, not matching price to tier.

Rokid's product naming needs clarifying to avoid a common mix-up: **Rokid Glasses** (launched November 18, 2024 at Rokid Jungle in Hangzhou, initial price RMB 2,499, roughly 49 g, binocular monochrome Micro-LED diffractive-waveguide display, Snapdragon AR1 platform, 12MP camera, roughly 4-hour battery life paired with a rechargeable charging case) is a display-equipped AI+AR consumer product that won recognition at CES 2025 and went on sale in Q2 2025. **Rokid AI Glasses Style** (roughly 38.5 g), by contrast, is a no-display audio model that connects to ChatGPT/Gemini and offers real-time translation. The two differ in positioning, form factor, and price, and should not be conflated. At the overall market level, multiple institutions project that global AI smart-glasses shipments will grow from roughly 6 million units in 2025 to roughly 20 million units in 2026, with market size expanding from roughly US$1.2 billion to roughly US$5.6 billion, and China and the US together accounting for nearly 80% of the share; IDC projects global smart-glasses shipments could exceed roughly 23.687 million units in 2026 (Wall Street CN/Huxiu, IDC, 2025–2026). A caveat is

in order: the figures above are "forecasts" from various institutions, not "actual" results, denominated in US dollars — they should not be mixed with RMB-denominated figures. Independent statistics on education-specific penetration remain scarce `[TBD: education-sector shipment/share data]`.

The relationship between NetDragon (HK:0777) and AI glasses needs to be stated precisely. In November 2023, NetDragon made a US$20 million strategic investment in Rokid and signed a five-year strategic cooperation agreement; in early 2025, when Rokid's ecosystem drew capital-market attention, NetDragon's share price swung markedly at times. This has led NetDragon, alongside its core online-gaming and edtech businesses, into the AI glasses space and adjacent software/content tracks (JRJ.com/Zhitong Finance, 2025). As of this chapter's writing, public information indicates that NetDragon's collaboration with Rokid consists primarily of strategic investment and ecosystem cooperation; no verifiable, independently sourced production model of education-specific AI glasses sold under the NetDragon brand has been found in our search `[TBD: if a NetDragon-branded glasses model and its specifications exist, they must be verified against official announcements]`. NetDragon's own education product lines center on software and interactive hardware — 101 Education PPT (cumulative installations exceeding 35.76 million, with a built-in "AI Teaching Assistant" feature), Promethean interactive displays, Future Labs, and others (NetDragon Websoft Education website, 2024). Any specifications, pricing, or launch dates concerning NetDragon glasses must be verified against official announcements; anything uncertain is marked `[TBD]`.

9.3.2 Typical Education Use Cases

- **Lab and hands-on training guidance.** First-person recognition of procedural steps enables real-time prompts on technique and safety risk, merging "watch the demonstration" and "do the operation" into "be guided while doing." The pain point in traditional lab instruction is that a teacher can't watch every workstation at once; the first-person view from glasses turns "what each student is doing right now" into an information stream that AI can monitor and prompt on — whether a burette reading is correct, whether an alcohol lamp was extinguished properly, whether a circuit connection has shorted, all of it can be flagged the instant it happens. This applies to physics/chemistry/biology labs, vocational training (CNC machining, culinary arts, cosmetology/hairdressing), and medical skills training (CPR technique, sterile procedure). This use case demands fine-grained recognition of hand movements, and the evaluation focus is the stability of step recognition — and the false-positive rate — under shake, occlusion, and fast motion; a false positive (flagging a correct action as wrong) does more damage to learning trust than a missed detection.
- **Situated language learning and translation.** This is currently the most mature and scalable education use case for AI glasses. Language-acquisition research has long emphasized the

importance of "comprehensible input" and authentic context, and translation glasses deliver exactly that immersive condition — seeing an object and instantly getting a target-language label — at low cost. Rokid Glasses, for example, claims support for translation in 89 languages online and 6 languages offline (Chinese, English, Japanese, German, French, Spanish; offline translation is powered by its in-house LLM), with the display able to overlay translated captions onto the field of view in real time (the online translation engine is provided by Microsoft), and support for instant translation by looking at text (menus, street signs, documents); no-display models (Xiaomi, Huawei, Rokid Style) deliver translated speech via audio playback, and Rokid Style connects directly to ChatGPT/Gemini. This has direct value for study-abroad preparation, foreign-language immersion, bilingual classrooms, and instant translation of menus, street signs, or original-language texts. The evaluation focus is continuous translation latency at the pace of real conversation, accuracy on proper nouns and subject-specific terminology (biological species names, chemical formulas, historical place names), and the quality gap between offline and online modes — whether the 6 offline languages can support real communication is the key test of whether "offline" is more than a marketing checkbox.

- **Accessibility and special education.** Providing environmental description, text read-aloud, and obstacle alerts for learners with visual impairments, and real-time speech-to-caption for learners with hearing impairments, the first-person form factor naturally suits accessibility needs for hands-free, always-on assistance. Compared with a phone, which must be raised, aimed, and tapped, glasses that "work by looking" are friendlier to learners with limited physical coordination. This use case demands extremely high accuracy in speech recognition and environmental description (misjudging a step, for instance, is a safety issue), and since it serves the population most in need of protection, governance and safety carry the highest weight here.
- **Field trips and outdoor learning.** Recognizing plants, animals, cultural artifacts, geological features, and astronomical phenomena to trigger generative explanations turns "I was there" into "I learned there," making museums, nature reserves, and historical sites into "living textbooks." This is precisely the scenario where connectivity is least reliable (mountains, underground galleries, international travel without roaming), which is exactly why offline capability matters most here for evaluation purposes — a field-trip pair of glasses that goes "blind and mute" the moment signal drops renders its advertised naturalist features useless.

9.3.3 Evaluation Focus and Observations

- **Multimodal recognition accuracy** is the core threshold for the glasses category. First-person imagery is significantly affected by lighting, shake, and occlusion, and recognition accuracy shows a systematic gap versus phone photography — this must be measured under real motion/lighting conditions, not static staged shots `[TBD: lab measurement]`.

- **End-to-end latency** determines whether the device is usable in class and in conversation. Translation applications in particular need to be tested for "keeping pace with speech," not single-utterance cold-start latency; display models also need evaluation of near-eye display clarity (resolution, FOV, brightness) and motion-sickness/fatigue risk — Meta Ray-Ban Display's display specs are 600×600, 20° FOV, 30–5000 nit, which puts it at an "information prompt" level rather than "immersive reading" level, and the readability of education content on it needs to be verified empirically `[TBD: lab measurement]`.
- **Battery life and wearing comfort** are the real-world constraints on whether a device can "get through a full class period." Current mainstream display glasses mostly rate battery life in the 4–6-hour range (Rokid Glasses at roughly 4 hours, Meta Ray-Ban Display at roughly 6 hours), which shrinks noticeably under heavy load (continuous translation, continuous capture); weight sits in the 40–70 g range, and nose-bridge and ear pressure during extended wear need to be factored into the comfort score `[TBD: lab measurement]`.
- **Voice interaction quality** is the primary interaction channel for no-display glasses, and an important supplement for display models. Evaluation should measure wake-word success rate (including false-wake rate), speech-recognition character error rate under real ambient noise, and context retention across multi-turn conversation. Background noise in classrooms, subways, and outdoor settings meaningfully raises error rates, and younger learners' less-standard pronunciation and unsteady pacing place extra demands on recognition robustness — a pair of glasses that consistently "can't understand what the kid is saying" will never build a usage habit, no matter how many features it has.
- **Privacy and compliance** shift from "optional" to "a gating requirement" in the glasses form factor. First-person camera capture involves other people's likenesses and classroom recording, and its sensitivity far exceeds phone photography — taking a photo with a phone is a visible, deliberate act that bystanders can notice, whereas glasses-based recording is a quiet, ongoing default that the person being recorded may never notice. That asymmetry is the crux of the privacy controversy. Evaluation should check: whether there is a clear recording indicator (an externally visible light or audible tone that the user cannot silently disable), whether image and audio are processed locally, whether data-retention and upload policies are transparent and verifiable, and whether the device meets the requirements of China's Personal Information Protection Law, Data Security Law, Regulation on the Protection of Minors in Cyberspace (effective January 1, 2024), Measures for the Security Administration of Facial Recognition Technology Applications (effective June 1, 2025), and Measures for Compliance Audits of Personal Information Protection (effective May 1, 2025). Campus deployment requires particular care: classroom recording captures the likenesses and voiceprints of other students, and questions of informed consent, data minimization, retention limits, and deletion mechanisms cannot be waved away with "it's needed for teaching." For a systematic governance framework

for glasses in classroom settings, see our institute's *AI Glasses in Education: 2026 Development Report*.

9.4 AI Learning Tablets: Evaluating the Agentification of the Home Learning Terminal

9.4.1 From "Problem-Bank Lookup" to "AI Study Companion"

The core competitive edge of learning tablets is shifting from "the scale of content and problem banks" to "how agentic the AI study companion is." By "agentic," we mean the product upgrading from passive response (student asks one question, machine answers one question) to active orchestration: the AI companion can take a learning goal and independently decompose it into sub-tasks — explaining, questioning, diagnosing, recommending, reviewing — and dynamically adjust the path based on the student's real-time responses. This tracks the agent paradigm discussed in Chapter 7, with one key difference: the agent on a learning tablet serves the user group that most needs patience and accuracy, leaving far less room for error. The key change in 2026 is that learning tablets are broadly connecting to vertical education large models and using **agent orchestration** to string "explain—question—diagnose—recommend" into a closed loop built around a persistent learning-profile memory. In an ideal state, a truly agentic learning tablet, when a student gets a problem wrong, wouldn't just supply the answer — it would diagnose the underlying knowledge gap, ask a follow-up question to pin down the type of misunderstanding, push a targeted micro-lesson and variant problems, and then proactively re-test that knowledge point days later to confirm it has stuck — weaving discrete "answering questions" into continuous "cultivation." This shift was directly catalyzed by the surge of interest in DeepSeek's open-source models in early 2025 — Xueersi, NetEase Youdao, Seewo, Xiaoyuan, Gaotu, Zuoyebang, and other leading players all announced integration with DeepSeek in short order (The Beijing News, 2025). Xueersi, for one, has taken an "open-source base model plus industry-data post-training" approach: its proprietary "Jiuzhang" large model uses DeepSeek-V3 as one of its foundations, layered with proprietary education-data fine-tuning; its fourth-generation learning tablet runs a "Jiuzhang plus DeepSeek dual-engine" architecture, and the company claims all-channel sales surpassed 150,000 units in November 2025, up 130% year-over-year. Squirrel AI, alongside its own proprietary adaptive-learning model, has also integrated DeepSeek (accounting for roughly 10% of its stack) and claims cumulative shipments surpassing 200,000 units (Zhihu Q&A/Securities Times, 2025 — both manufacturer-reported figures, not independently audited).

The essence of this upgrade cycle is the learning tablet's transformation from a "storage medium" into a "service medium." The old moat was content — whoever bought exclusive rights to more star-teacher courses, whoever had the fuller problem bank, won. Today's moat is service — whoever has

the more insightful AI study companion, the more accurate learning-diagnostic engine, the faster multimodal grading, wins. This also explains why online-education giants (Zuoyebang, Xueersi, Youdao, Xiaoyuan) were able to leverage their model and data advantages to move quickly into hardware in 2025 and outpace traditional hardware manufacturers on sales rankings: once the competitive center of gravity shifted from "stacking content" to "refining models," the players with research data and algorithm teams claimed the new high ground.

On market structure, in Q1 2025 the top six brands by sales — Zuoyebang, Xueersi, iFLYTEK, Bbkeducation (Step by Step), Xiaoyuan, and Xiaodu — together held roughly 74.4% share (in some monthly rankings, Zuoyebang led with roughly 31.8%, followed by Xueersi, Xiaoyuan, and iFLYTEK). Bbkeducation, with 28 years in education, maintains more than 18,000 offline points of sale — its offline channel and after-sales network remain a moat not to be dismissed. On market size, China's smart-tablet learning-machine market was worth roughly RMB 27.072 billion in 2024, up roughly 48.27% year-over-year, with forecasts putting it above RMB 50 billion by 2027 (iiMedia Research/RUNTO, 2024–2025). A caveat: the sales, share, and market-size figures above use inconsistent bases (unit sales vs. revenue, quarterly vs. annual, all-channel vs. online-only, institutional estimates vs. manufacturer self-reporting) — cite them with the basis and source clearly labeled, and don't add figures across bases; manufacturer-reported cumulative shipments and growth rates in particular should be flagged as "manufacturer-reported, not independently audited."

9.4.2 What Makes Evaluation Distinctive Here

- **Subject-content system and curriculum alignment.** Unlike a general-purpose tablet, a learning tablet's value depends heavily on its alignment with curriculum standards and textbook editions. Evaluation should check coverage of multiple textbook editions (PEP, Beijing Normal University Press, Jiangsu Education Press, etc.) across all grade levels from pre-K through high school, and whether content updates track curriculum-standard revisions `[TBD: curriculum-alignment lab measurement]`. iFLYTEK's T30 Ultra, for instance, claims coverage of "all subjects, pre-K through high school," and comes with 188 Oxford University Press-licensed graded English readers (15 difficulty levels) built in (Kuaikeji, 2024).
- **Wrong-answer and learning-diagnostic engine.** Evaluation should test whether the memory mechanism genuinely accumulates an individual student's learning profile and whether recommendations converge on weak points, rather than generalizing broadly. This is the dividing line between genuine "agentification" and mere "problem-bank lookup" — a true study companion should remember, across sessions, that "this child keeps making mistakes on the discriminant of quadratic equations," and orchestrate the subsequent path accordingly `[TBD: lab measurement]`.

- **Eye care and usage controls.** Screen eye-care parameters and the granularity of parental controls are hard constraints for a home terminal. iFLYTEK's T30 Ultra, for example, claims a "natural-light-like plus micro-nano paper-like" eye-care screen, 3K resolution, 120Hz refresh rate, 247 PPI, hardware-level low blue light, a 14.7-inch display, 12GB+1T storage, a 12,000 mAh battery, four cameras, and a "NearLink" AI stylus (with over 10,000 levels of pressure sensitivity), launching at RMB 11,699 (Kuaikeji, 2024); Bbkeducation, Xueersi, and others each have their own eye-care approaches. The controls dimension should evaluate granularity of usage-time limits, app whitelisting, content filtering, and parental remote viewing `[TBD: checklist of control items]`.
- **AI study-companion quality (teaching effectiveness).** This is the hardest dimension to evaluate and the most consequential. Evaluation should look at explanation correctness, guided (rather than answer-first) instruction, and the ability to decline to answer or self-correct. Xueersi's Jiuzhang model, for instance, claims a "Socratic" approach to explanation — rather than giving the answer directly, it first analyzes the knowledge point and problem type, then guides the student to derive the solution step by step through successive follow-up questions (Xueersi public materials, 2025). Whether this kind of "guided" design genuinely promotes thinking should be assessed with process indicators (steps the student completes independently, improvement in independent-completion accuracy) rather than "answer-correct rate" alone `[TBD: lab measurement]`.

9.4.3 Notable Risks

The agentification of learning tablets introduces three prominent risks that evaluation must weight heavily.

The first is **giving away answers undermining the learning process.** If a product markets itself on "instant answers" or "one snap and you're done," it risks turning the learning tablet into a de facto cheating device — the learner bypasses thinking altogether and is left only with the act of copying. This is the deepest paradox of "AI-assisted learning": the more convenient the technology, the more it risks stripping away the "productive struggle" that learning actually requires. Evaluation should reward interaction design that guides rather than substitutes, treating "does it promote the thinking process" as the core yardstick — concretely testable as: does the explanation ask a follow-up question before revealing anything; does it offer scaffolded hints rather than the direct answer when a student gets stuck; does it ask the student to restate the reasoning after getting an answer right. Xueersi's claimed Socratic-style explanation in its Jiuzhang model, and the "guided mode/answer mode" toggle some products offer, are examples of design pointed in the right direction, but their real-world effectiveness still needs to be verified with process indicators through lab testing.

The second is **hallucination creeping into subject-matter problem solving.** In quantitative reasoning especially, an incorrect intermediate step is more misleading than an incorrect final answer, because the learner imitates the reasoning path and internalizes the flawed method. A problem where the final answer happens to be right but a step of logic along the way is wrong can do more harm to learning than getting the whole thing wrong — it teaches the student a "looks right" but flawed technique. The emerging consensus among universities and manufacturers is "large model plus RAG plus knowledge graph" to suppress hallucination and stay aligned with the curriculum: retrieval anchors answers to the textbook, and the knowledge graph constrains which reasoning paths are legitimate (53AI application case studies, 2025). Evaluation should treat subject-matter accuracy, traceability of reasoning steps (can each step be traced back to a textbook passage or formula), and self-correction rate as hard metrics, stratified by subject — the hazard from step-level hallucination is greatest in math, physics, and chemistry, and those subjects warrant separate, dedicated re-testing.

The third is **over-reliance among minors and data risk.** Emotionally engaging, anthropomorphic interaction increases stickiness and can ease loneliness, but it can also induce excessive emotional dependency in younger users and erode real-world socializing and independent-learning motivation; policy proposals have already suggested folding highly interactive, high-stickiness interaction design into anti-addiction mechanisms (various policy proposals, 2025). At the same time, learning tablets accumulate extremely sensitive data on minors — learning profiles, behavior, even facial/voiceprint data — and their data governance must be benchmarked against the Regulation on the Protection of Minors in Cyberspace and the Personal Information Protection Law, treating parental informed consent, data minimization, and retention/deletion mechanisms as gating requirements. Evaluation should check: whether emotionally engaging interaction has boundaries and anti-addiction design, whether sensitive data is localized or encrypted, and whether parents can view and delete their child's data profile. Wherever younger learners are involved, all three of these risk categories should be weighted above "how smart is the model" `[TBD: lab measurement]`.

9.5 On-Device Agents: Evaluating Local Inference, Privacy, and Offline Capability

9.5.1 Why On-Device Matters So Much in Education

The value of on-device processing in education goes beyond latency — it's fundamentally about **privacy and usability**. Minors' data is highly sensitive, and local processing inherently shrinks the surface for data leakage, which aligns with the minimization principle in the Personal Information Protection Law and the broader "privacy by design" philosophy. Network conditions vary widely across schools and homes, and offline usability directly determines whether a device works "at the moment it's actually needed" — outdoor field trips, rural schools, and dorm-network peak hours are

all real battlegrounds where on-device fallback either holds up or doesn't. By 2026, as on-device chip compute improves and small language models (SLMs) along with quantization/distillation techniques mature, "running a good-enough agent locally" is moving from concept to deployment.

On the hardware side, Qualcomm's Snapdragon 8 Gen 3 claims the ability to run models with tens of billions of parameters, generating roughly 20 tokens per second for a 7-billion-parameter LLM; MediaTek's Dimensity 9300 supports running 1-billion-, 7-billion-, and 13-billion-parameter models on-device (manufacturer public materials/tech media, 2024–2025). On the model side, Apple's third-generation on-device foundation model, introduced at WWDC 2025, has roughly 3 billion parameters, is optimized specifically for Apple silicon, and ships with the Foundation Models framework, which lets developers call the on-device model directly, free of charge, and offline (Apple ML Research, 2025). Progress on the open-source side has been even more rapid — Qwen3-4B scores roughly 83.7 on MMLU-Redux (surpassing some models more than twice its size); Phi-4-mini (3.8B) scores roughly 67% on the full MMLU (comparable to Llama 3.1 8B while using roughly half the memory); Qwen2.5-7B scores roughly 74.2 (various teams' technical reports, 2025). Quantization pushes the deployment bar even lower — Gemma 3 4B, through quantization-aware training, can shrink from roughly 8GB (BF16) to roughly 2.6GB (int4), and Gemma 3 1B from roughly 2GB to roughly 0.5GB, with quality loss held to a few percentage points (Google, 2025); 4-bit quantization (Q4_K_M) can reduce a certain 13B model's memory footprint from roughly 13.2GB to roughly 4.8GB and cut on-device inference latency from roughly 4200ms to roughly 980ms (tech-media lab testing, 2025). Taken together, these numbers point to one conclusion: a model capable of subject-matter Q&A and explanation at the middle- and high-school level can now run locally on flagship phones and next-generation learning terminals, without needing to reach the cloud for every task. For education, this is not just "faster" — it's "more controllable." The model, the data, and the logs all stay on the device, giving parents and schools a physical-layer grip on "what is my child talking to, and what did they say."

9.5.2 Technical Paradigms for On-Device Agents

- **Small model plus Retrieval-Augmented Generation (RAG).** A local knowledge base fills in the gaps in a small model's knowledge, turning textbooks, wrong answers, notes, and lecture handouts into a searchable memory. This addresses the small model's "insufficient knowledge" problem while also easing hallucination — answers can be traced back to a local document. In education, RAG lets a local 3–7B model effectively "memorize this semester's textbook."
- **Local memory and personalization.** A long-term learning profile accumulates on the device itself, reducing reliance on cloud-side profiles and keeping "this child's weak points" local. This is architecturally the same idea as the learning-diagnostic engine discussed in Section 9.4, but with the emphasis on "data never leaves the device."

- **Device-cloud collaboration and graceful degradation.** Routine tasks are handled on-device, complex tasks go to the cloud, and the system degrades gracefully to local capability when the connection drops. Underneath this is a routing logic: the device first judges a task's difficulty and privacy sensitivity — simple or sensitive tasks (vocabulary drills, mental-math grading, questions involving personal information) stay local, while complex or non-sensitive tasks (long-form writing coaching, open-ended inquiry) go to the cloud. Apple's combination of a roughly 3-billion-parameter on-device model with Private Cloud Compute is a paradigmatic engineering expression of this approach — its cloud compute also commits to not retaining or training on user data, an attempt to combine cloud-scale capability with device-level privacy guarantees (Apple, 2025). For more on emerging paradigms in agent orchestration, RAG, and memory, see Chapter 7.

Device-cloud collaboration also reshapes the cost structure of education hardware, a facet of the "sustainability" dimension that shouldn't be overlooked. A pure-cloud approach has marginal costs that rise linearly with usage — every Q&A session burns tokens — which for high-frequency learning use translates into an ongoing inference bill, ultimately passed on as subscription fees. An on-device approach front-loads the compute cost into the hardware purchase (paying once for a stronger chip), after which local inference runs at close to zero marginal cost. For large-scale school deployment and long-term use by lower-income families, the on-device model's "pay once, use free indefinitely, and works offline" carries distinct value for access — it decouples "being able to afford AI" from ongoing network and subscription spending. Evaluating "per-unit learning cost" should combine hardware depreciation, subscription fees, and data charges, rather than looking at the bare device price alone.

9.5.3 Evaluation Focus

- **Offline usability rate.** How completely core functions (subject Q&A, translation, explanation, grading) perform with no connection is the single most differentiating metric for on-device agents. Glasses offer a useful reference point in Rokid Glasses' tiered claim of "89 languages online, 6 offline" — the breadth and quality of offline capability must be labeled separately and never allowed to hide behind online-capability numbers `[TBD: lab measurement]`.
- **On-device inference latency and power draw.** The speed of local inference and its thermal/power cost. Prefill (reading the prompt) and decode (generating the response) should be measured separately, along with device temperature rise and power consumption under sustained inference, since heat buildup triggers throttling that in turn slows latency `[TBD: lab measurement]`.
- **The gap between local capability and the cloud (hardware loss).** The quality difference for the same task across pure-on-device, device-cloud-collaborative, and pure-cloud tiers,

quantifying exactly how much is lost by going offline. This is where the "software baseline—hardware measurement" comparison described in Section 9.2.2 lands in practice `[TBD: lab measurement]`.

- **Privacy verification.** Manufacturer claims alone aren't sufficient — packet capture and traffic analysis are needed to verify whether data is genuinely kept local, whether there is hidden uploading, what gets uploaded, and where it goes. The gap between a manufacturer's claim of "local processing" and the reality of "logs/profiles still being uploaded" can only be exposed through actual testing. Evaluation methods can include: checking whether core functions still work with the connection off (if going offline breaks the function, "local" is questionable), analyzing outbound traffic and destinations while connected, and comparing the privacy-policy text against actual behavior. This is the key step in turning a "privacy promise" into "provable privacy," and it echoes the shift the 2025 Measures for Compliance Audits of Personal Information Protection is pushing — from "having done it" to "being able to prove it" (law-firm compliance review, 2025) `[TBD: methodology/source]`.
- **Ongoing update of model capability and content.** One cost of going on-device is "model calcification" — unlike the cloud, a local model can't be hot-updated at will, so its knowledge may lag and its capability may fall behind the latest cloud models. Evaluation should examine whether the OTA update mechanism is sound, whether the local knowledge base (RAG corpus) can be synchronized with textbook revisions, and what long-term maintenance commitment the manufacturer makes for the on-device model. An on-device agent that stops receiving updates will gradually see both its knowledge and its safety guardrails go stale — a hidden risk under the sustainability dimension `[TBD: OTA update-frequency data]`.

9.6 Cross-Category Comparison and Capability Radar (Evidence-Based Visualization)

To make the three categories comparable on a common footing, this chapter uses the six-dimension framework to produce a **capability radar chart** and a **product comparison table**, along with a **capability-evolution timeline** showing generational change. All underlying cross-sectional scores in these visualizations come from our institute's evaluation lab; anything not yet tested is left blank rather than filled with an estimate. The specifications in the table below are drawn from public sources and can be verified immediately; capability scores await lab testing.

Form Factor	Model Capability	On-Device Engineering	Interaction Experience	Education Fit	Governance & Safety	Sustainability
AI glasses	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`
AI learning	`[TBD: lab	`[TBD: lab	`[TBD: lab	`[TBD: lab	`[TBD: lab	`[TBD: lab

tablet	test]`	test]`	test]`	test]`	test]`	test]`
On-device agent	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`	`[TBD: lab test]`

Based on public specifications and use cases, a **qualitative portrait** of each form factor can already be sketched (non-quantitative; quantification awaits lab testing). AI glasses are irreplaceable for "on-the-go first-person context and situated translation," but constrained by battery life (roughly 4–6 hours), weight (40–70 g), and display information density (prompt-level rather than reading-level), they aren't yet suited to carrying sustained, systematic learning — their proper role is "situated assistant" rather than "primary learning terminal." Learning tablets are the most mature on "subject-content systems, eye care, parental controls, and wrong-answer diagnostics," making them the primary vehicle for systematic home learning — a 14.7-inch screen and a configuration priced in the five-figure RMB range (as with iFLYTEK's T30 Ultra) are positioned to rival one-on-one tutoring, though giving away answers and subject-matter hallucination remain the key governance concerns. On-device agents are not a standalone category so much as a "capability trait" that can attach to glasses, learning tablets, or phones; they hold a structural advantage on "offline, privacy, low marginal cost," serving as the fallback for weak-network and data-sensitive scenarios, though local small models still trail the cloud in capability and need RAG and device-cloud collaboration to close the gap.

The three form factors have a clear **complementary, not substitutive, relationship**, which means the right approach to procurement isn't "pick one of three" but "combine by scenario." An ideal home-learning ecosystem might look like this: the learning tablet handles systematic subject learning and wrong-answer correction (primary role), AI glasses handle language immersion, field-trip recognition, and lab guidance (situated supplement), and on-device capability runs through both, guaranteeing offline usability and keeping data on the device (the foundation). All three should ideally share a single learning profile and a single governance policy — the spoken language a student practices on the glasses and the wrong answers corrected on the learning tablet ought to feed into one unified growth record, and the more that record is localized and parent-controllable, the better it aligns with the direction of minors' data protection. The current market reality is that the three product categories operate in silos, with data fragmented across them; unifying learning profiles and governance across devices remains an open gap — both a pain point and a potential entry point for manufacturers like NetDragon that span software, content, and hardware ecosystems `[TBD: case studies of cross-device learning-profile integration in production]`.

Figure 9-1 Six-dimension capability radar for three categories of education AI hardware (cross-sectional scores from: our institute's evaluation lab, `[TBD: test batch/date]`; specifications from: manufacturer public materials)

Figure 9-2 Capability-evolution timeline for education AI hardware (2024 Rokid Glasses launch → 2025 DeepSeek catalyzes learning-tablet agentification, AI glasses' "hundred glasses war" → 2025 Meta Ray-Ban Display and Apple's on-device model → 2026 deepening on-device adoption; sources at chapter end)

9.7 Findings, Conclusions, and Procurement Guidance

Key findings (qualitative judgments based on the framework; quantitative conclusions await lab testing):

1. The bottleneck in education AI hardware is shifting from "is the model smart enough" to "can on-device engineering hold up reliably." The common shortfall across all three categories concentrates in latency, battery life, and offline capability: glasses are constrained by battery life and display information density, learning tablets by hallucination and answer-giveaway design, and on-device agents by the gap between local and cloud capability `[TBD: lab evidence]`.
2. Agentification is the shared direction across all three categories, with DeepSeek's 2025 open-source release serving as a major catalyst; but the design split between "guided learning" and "handing over answers" predicts educational value better than parameter count does — a learning tablet that uses Socratic follow-up questioning may deliver more educational value than a larger-parameter product that just spits out answers instantly.
3. Governance and safety shift from "optional" to "a gating requirement" in hardware form factors, especially around compliance for first-person camera recording of likenesses, localization of minors' data, and the risk that subject-matter hallucination poses to misleading younger learners. The Regulation on the Protection of Minors in Cyberspace, the Measures for the Security Administration of Facial Recognition Technology Applications, and the Measures for Compliance Audits of Personal Information Protection, all issued in quick succession in 2024–2025, have collectively raised the compliance bar.
4. Product naming and corporate relationships must be verified precisely: Rokid Glasses (the display model, RMB 2,499) and Rokid AI Glasses Style (the audio model) should not be confused with each other; NetDragon (HK:0777)'s relationship to Rokid consists of a US$20 million strategic investment in November 2023 plus a five-year cooperation agreement, and NetDragon's own education hardware centers on interactive displays and software (101 Education PPT) — any claim about a production glasses model must be verified against official announcements.

Procurement guidance (for schools and families; directional, without endorsing specific brands):

- **Choose the form factor by scenario.** Prioritize glasses for on-the-go situated assistance, language immersion, and accessibility; prioritize learning tablets for systematic home learning, wrong-answer correction, and subject-matter improvement; prioritize on-device agents (or

devices with equivalent offline capability) for privacy-sensitive, weak-/no-network scenarios where data must stay on the device.

- **Treat "governance and safety" as an absolute veto criterion.** Content interception, data localization, parent/teacher controls, and recording-notice compliance are all non-negotiable; for glasses, additionally verify compliance around first-person camera capture of people's likenesses.
- **Favor guided design over answer-substitution design.** Use "does it promote the thinking process" rather than "does it produce answers quickly" as the procurement yardstick, and be wary of "instant answer" marketing language.
- **Assess offline capability by tier, not as a blanket claim.** Require suppliers to clearly mark the boundary between "online capability" and "offline capability" (e.g., which languages, which problem types can be solved offline), and don't accept online capability being used to paper over offline shortfalls.
- **Demand verifiable evaluation evidence.** Ask suppliers for public, reproducible testing methods and data (end-to-end latency methodology, weak-network tier definitions, subject-matter accuracy test sets, privacy traffic reports) rather than isolated benchmark numbers and marketing specs; for market-size/share figures, check the basis and source, and watch for cross-basis addition, manufacturer self-reports being treated as lab measurements, and currency mixing (US-dollar forecasts should not be conflated with RMB-denominated market size) `[TBD: source for evaluation checklist]`.
- **Compute total cost, not just the bare device price.** Combine hardware purchase, cloud subscription, data charges, and content-update fees into a "three-year total cost of ownership" before comparing, paying particular attention to the access value that on-device solutions' "pay once, usable offline, low marginal cost" model offers for large-scale and lower-income scenarios.

Outlook. Looking ahead to 2026–2028, education AI hardware appears to be on three fairly well-defined trajectories. First, on-device adoption will keep deepening — as stronger on-device chips and smaller-yet-capable models reach production, more and more core capability will move onto the device itself, with privacy and offline usability shifting from selling points to defaults, while the cloud increasingly handles complex long-running tasks and cross-device synchronization. Second, form factors will move toward convergence and coordination — the data silos among glasses, learning tablets, and phones may be bridged by a "unified learning profile plus unified governance policy," and manufacturers with full-stack software, content, and hardware capability (such as NetDragon and its ecosystem partner Rokid) are structurally positioned to benefit from this convergence, provided that cross-device data integration is built on a foundation of localization, parental control, and provable compliance. Third, governance will move from "after-the-fact compliance" to "built into the design" — as regulations such as the Regulation on the Protection of Minors in Cyberspace, the Measures for the Security Administration of Facial Recognition

Technology Applications, and the Measures for Compliance Audits of Personal Information Protection take hold, privacy-by-design, anti-addiction mechanisms, content guardrails, and recording compliance will harden from "nice-to-have" into "the price of admission," forcing products to solve these problems at the architecture level rather than the marketing level. For evaluation institutions, this means the center of gravity will keep shifting from "specs and benchmark scores" to "usage truth and governance evidence" — which is precisely the direction this chapter's framework has tried to establish. All cross-sectional capability scores in this chapter will be completed and published, along with full methodology, once our institute's evaluation lab finishes lab testing — at which point the placeholders in this text will be replaced with traceable, reproducible, evidence-based data.

Chapter References

1. "49g Ultra-Light AR glasses with AI — Rokid Glasses" · Rokid official website · 2025 · https://global.rokid.com/products/rokid-glasses
2. "Real-Time Translation Glasses | AI-Powered Smart Glasses" · Rokid official website · 2025 · https://global.rokid.com/pages/rokid-glasses
3. "Rokid AI Glasses Style (Non-Display) 38.5g Ultra-Light" · Rokid official website · 2025 · https://global.rokid.com/pages/rokid-ai-glasses-style
4. "RMB 2,499! Rokid Glasses Launches, AR Glasses Enter the Consumer Era at a Run" ["2499 元！Rokid Glasses 发布，AR 眼镜跑步进入消费时代"] · QbitAI · 2024 · https://www.qbitai.com/2024/11/225625.html
5. "I traveled 5,000 miles with Rokid Glasses — this Meta Ray-Ban Display rival impressed me" · Tom's Guide · 2025 · https://www.tomsguide.com/computing/smart-glasses/rokid-glasses-review
6. "Strategic Investment in Rokid: 'Pure AI Play' NetDragon (00777) Opens Up New Imaginative Space" ["战略投资 ROKID，"纯正 AI 股"网龙(00777)打开全新想象空间"] · JRJ.com/Zhitong Finance · 2025 · https://m.jrj.com.cn/madapter/finance/2025/01/14153147428755.shtml
7. "NetDragon Invests in Rokid, Stakes a Claim in the New AI Glasses Wave" ["网龙投资 ROKID，掘金 AI 眼镜新风潮"] · Sohu/financial media · 2025 · https://www.sohu.com/a/848866851_122004016

8. "NetDragon Integrates Digital Technology into Education Products" ["网龙将数字技术融入教育产品之中"] · Digital China Summit official website · 2024 · https://www.szzg.gov.cn/2024/szzg/szfj/202405/t20240514_4823981.htm

9. "NetDragon: Positioning in the New AI Track, Exploring a New Digital Journey" ["网龙：布局 AI 新赛道，探索数字新征程"] · Securities Times Net · 2024 · https://stcn.com/article/detail/1098277.html

10. "101 Education PPT Official Website" ["101 教育 PPT 官网"] · NetDragon Websoft Education · 2024 · https://ppt.101.com/

11. "New Meta Ray-Ban AI-Powered Display Glasses and Neural Band" · Meta official website · 2025 · https://www.meta.com/ai-glasses/meta-ray-ban-display/

12. "Meta Ray-Ban Display: AI Glasses With an EMG Wristband" · Meta Newsroom · 2025 · https://about.fb.com/news/2025/09/meta-ray-ban-display-ai-glasses-emg-wristband/

13. "Meta Ray-Ban Display: Full Specification" · VR-Compare · 2025 · https://vr-compare.com/headset/metaray-bandisplay

14. "Xiaomi AI Glasses (Product Page)" ["小米 AI 眼镜（产品页）"] · Xiaomi official website · 2025 · https://www.mi.com/prod/xiaomi-ai-glasses

15. "Huawei AI Glasses Specifications" ["华为 AI 眼镜规格参数"] · Huawei official website · 2025 · https://consumer.huawei.com/cn/audio/ai-glasses/specs/

16. "Is the AI Feature Just a Gimmick? Reviewing 10 AI Glasses Including Xiaomi and Rokid" ["AI 功能是噱头吗？测评小米、Rokid 等 10 款 AI 眼镜"] · Southern Metropolis Daily · 2025 · https://m.mp.oeeee.com/a/BAAFRD0000202507191104492.html

17. "Global AI Smart Glasses Market to Reach $5.6 Billion by 2026, China and US to Account for 80% of Share" ["2026 年全球 AI 智能眼镜市场将达 56 亿美元，中美占 80% 份额"] · Huxiu (citing IDC/institutional forecasts) · 2026 · https://www.huxiu.com/article/4857057.html

18. "Giants Accelerate Market Entry: AI Glasses Set to Fight a New Round of Positioning Battles in 2026" ["巨头加速入局，AI 眼镜 2026 年打响新一轮排位赛"] · 21st Century Business Herald · 2026 · https://www.21jingji.com/article/20260114/herald/29460ed43ae258bde7c914301cb29cf6.html

19. "Launching at RMB 11,699: iFLYTEK AI Learning Tablet T30 Ultra Goes on Sale, Featuring the Industry's First NearLink AI Stylus" ["首发 11699 元 科大讯飞 AI 学习机 T30 Ultra 开售：配备行业首款星闪 AI 手写笔"] · Kuaikeji · 2024 · https://news.mydrivers.com/1/992/992671.htm
20. "iFLYTEK AI Learning Tablet T20 Series Launched: Powered by the Spark Cognitive Model, Starting at RMB 7,299" ["讯飞 AI 学习机 T20 系列发布：搭载星火认知大模型，首发 7299 元起"] · DoNews · 2024 · https://www.donews.com/news/detail/4/3489459.html
21. "The Wildly Popular AI Learning Tablet: Why Has It Become One of the Three Major Hardware Tracks?" ["卖疯了的 AI 学习机，为何成为硬件三大赛道之一？"] · 21st Century Business Herald · 2025 · https://www.21jingji.com/article/20250729/herald/59577065a816e441b1b2c0e136eddbe3.html
22. "iiMedia Gold List | Top 10 Brands in China's Smart Tablet Learning Machine Market, 2025" ["艾媒金榜 | 2025 年中国智能平板学习机十大品牌"] · iiMedia Research · 2025 · https://www.iimedia.cn/c880/106929.html
23. "DeepSeek's Arrival: The Contest and Reflection Behind Education Companies' Race to Get In" ["DeepSeek"来袭"，教育企业抢滩背后的博弈与思考"] · The Beijing News · 2025 · https://m.bjnews.com.cn/detail/1739588622168866.html
24. "Xueersi Launches Fourth-Generation Learning Tablet, Powered by the Jiuzhang Model and DeepSeek Dual-Core Architecture" ["学而思发布第四代学习机，搭载九章大模型与 DeepSeek 双核架构"] · Zhihu (Q&A/manufacturer-reported) · 2025 · https://www.zhihu.com/question/1904505543529313225
25. "Charting the Future, Empowering New Growth in Learning: 2025 China AI Learning Tablet Market Insight White Paper" ["智启未来 学赋新生 2025 年中国 AI 学习平板市场洞察白皮书"] · RUNTO · 2025 · http://runtotech.com/uploadfiles/2026/01/【洛图】2025 年中国 AI 学习平板市场洞察白皮书.pdf
26. "Introducing the Third Generation of Apple's Foundation Models" · Apple Machine Learning Research · 2025 · https://machinelearning.apple.com/research/introducing-third-generation-of-apple-foundation-models

27. "Apple Intelligence gets even more powerful with new capabilities across Apple devices" · Apple Newsroom · 2025 · https://www.apple.com/newsroom/2025/06/apple-intelligence-gets-even-more-powerful-with-new-capabilities-across-apple-devices/

28. "Qwen2.5 Technical Report" · Qwen Team / arXiv:2412.15115 · 2025 · https://arxiv.org/pdf/2412.15115

29. "Small Language Models Can Still Pack a Punch: A Survey" · arXiv:2501.05465 · 2025 · https://arxiv.org/pdf/2501.05465

30. "The Best Open-Source Small Language Models (SLMs) in 2026" · BentoML · 2026 · https://www.bentoml.com/blog/the-best-open-source-small-language-models

31. "2025 On-Device Large Model Technology Analysis: A Deep Dive from Technical Principles to Deployment Practice" ["2025 端侧大模型技术分析：从技术原理到落地实操的深度拆解"] · Longteng Yatai (AI Technology and Consulting) · 2025 · https://www.longtengyatai.com/info/264

32. "A Review of 2024 Data Compliance Legislation and Regulation, and a Look Ahead to 2025" ["2024 数据合规立法监管的回顾与 2025 年展望"] · AllBright Law Offices, Shanghai · 2025 · https://www.allbrightlaw.com/CN/10475/63efeb45f0af5dbe.aspx

33. "2025 Annual Review of China's Cybersecurity and Data Protection, and a Look Ahead to 2026" ["2025 中国网络安全与数据保护年度回顾与 2026 年展望"] · Secrss · 2025 · https://www.secrss.com/articles/86785

34. "Application Cases of Large Models in the Education Industry (RAG + Knowledge Graph to Overcome Hallucination)" ["大模型在教育行业的应用案例（RAG+知识图谱克服幻觉）"] · 53AI · 2025 · https://www.53ai.com/news/neirongchuangzuo/2025050196740.html

35. "Second Wave of Frenzied Competition: Xiaomi, Baidu, China Telecom and Nearly Ten AI Glasses Products Set to Launch in a Cluster" ["大厂疯抢第二季！小米、百度、中国电信等近十款 AI 眼镜产品或扎堆发布"] · Tencent News · 2025 · https://news.qq.com/rain/a/20250319A0774T00

36. "AI Glasses Concept Stocks Hit Limit-Up: Watching the Rokid Ecosystem Chain Opportunity" ["AI 眼镜概念涨停潮：关注 Rokid 生态链概念机会"] · Sina Finance · 2025 · https://finance.sina.com.cn/stock/jsy/2025-02-20/doc-inemcpfh2390679.shtml

37. "Hong Kong Stock Movers | NetDragon (00777) Up Over 9% in Morning Trade as Rokid Glasses Expected to Launch in Q2; Company Made 2023 Strategic Investment in Rokid" ["港股异动 | 网龙(00777)早盘涨超 9% Rokid Glasses 有望二季度开售 公司 23 年战略投资 Rokid"] · Zhitong Finance · 2025 · https://cn.investing.com/news/stock-market-news/article-2678278

Chapter 10 New Paradigms and Recommendations: Agent Orchestration, RAG, and Memory; Policy, Teacher Literacy, and Equity; An Implementation Roadmap

The preceding nine chapters examined the mechanisms, product landscape, and representative forms of generative-AI education products across five scenarios — Empowering Teaching, Supporting Learning, Supporting Research (PD), Intelligent Assessment, and Governance and Safety — each grounded in evidence. This chapter closes the book with two final tasks. First, at the technical-architecture level, it distills the **three new paradigms** underpinning how these scenarios are evolving: agent orchestration, Retrieval-Augmented Generation (RAG), and long-term memory. Together these form the common foundation on which 2026 products are shifting from single-model conversational tools toward agentic, multimodal, on-device systems. Second, at the industry and policy level, it offers **recommendations for a diverse set of stakeholders** and a **staged, verifiable implementation roadmap**.

One methodological point needs stating up front. Every product and data point cited in this chapter comes from a primary source that our editorial team located and verified through open search — vendor announcements, official documentation, academic conference papers, and original policy texts — each listed individually in the "Chapter References" at the end. Wherever a specific figure, share, benchmark score, or date could not be traced to a credible source, we mark it "[TBD: ...]" rather than substitute an estimate or an unverifiable label. This is the evidentiary discipline this book has followed from the outset, and it is also, fittingly, a self-demonstration of the very RAG paradigm discussed in this chapter's first section: holding our own writing to the same standard we apply to products — sources must be identifiable, and conclusions must be traceable back to them.

10.1 Three New Technical Paradigms: From Talking to Doing — Traceable, and With Memory

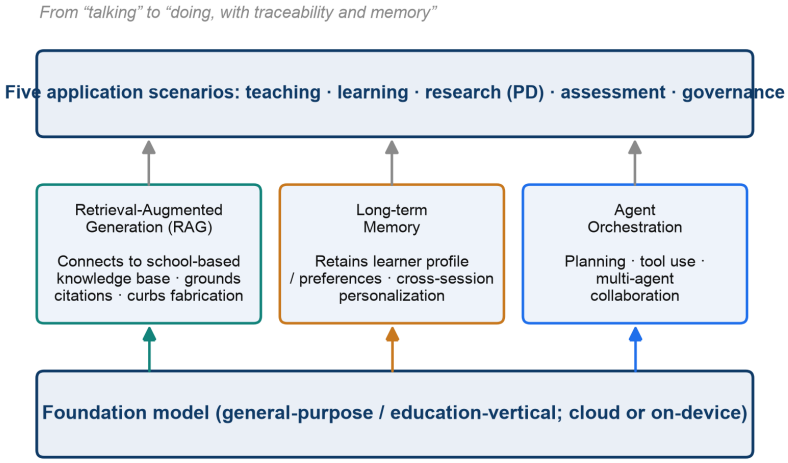
The 2024 edition of this report took conversational large language models as its throughline, with capabilities largely confined to single- or multi-turn natural-language generation. The model functioned as a responder: input a question, output text, and forget the context the moment the session ended — unable to call on outside resources, and unable to cite a source for its own claims. That form factor was serviceable for casual Q&A, but the moment it met a real educational task, three weaknesses surfaced immediately. It **couldn't act** — unable to break a compound task like "produce a tiered lesson worksheet aligned to an assessment rubric" into steps and carry them out. It

couldn't cite sources — unable to say which textbook or curriculum standard a given conclusion came from, leaving teachers with no way to verify it. And it **had no memory** — unable to recall where a given learner went wrong last week or how far along they'd progressed, so "personalization" amounted to a one-off performance.

The 2026 education-product landscape has expanded precisely by closing these three gaps: models are now organized into systems that **can plan, call tools, cite sources, and accumulate state over time**. We describe this shift through three interlocking technical paradigms: orchestration gives the system the ability to act, RAG gives it the ability to cite sources, and memory gives it the ability to remember. These are not three separate selling points but three facets of a single system, and understanding their mechanics and limits is the key yardstick for telling a genuine agent apart from a chatbot in a fancy wrapper.

The urgency behind this shift is borne out by how fast adoption is accelerating. A RAND survey found that the share of American teachers using AI roughly doubled over two school years, from 25% to 53%. A Gallup/Walton Family Foundation survey found that 60% of K-12 teachers used some form of AI tool in their work during the 2024–25 school year. Pew Research recorded that 26% of American teenagers used ChatGPT for schoolwork in 2024, up from 13% in 2023 (RAND, Gallup, and Pew 2024–2025, as compiled by tutorbase). Market-sizing estimates vary across firms — which is exactly why this book treats market-size figures with caution and declines to rely on any single source — but they converge on rapid growth: Precedence Research puts the 2025 global "AI + education" market at roughly US$7.05 billion, while Mordor Intelligence estimates roughly US$6.9 billion. The orders of magnitude are similar even where the specific figures diverge, so this book draws only the robust conclusion that the market sits in the billions of dollars and is growing at a high compound rate; a unified figure remains [TBD: standardized market-size estimate]. The wider the adoption and the larger the investment, the sharper the question of whether these products actually hold up — and the answer turns squarely on the engineering quality of the three paradigms.

Fig. 6 The new-paradigm technology stack for generative-AI education products



Source: this report's analytical framework (see Chapter 10).

10.1.1 Agent Orchestration: From Single-Model Answers to Multi-Role Collaboration

Mechanism: from answering directly to a perceive-plan-act-reflect loop. Agent orchestration uses a large language model as the reasoning core — often described in the industry as the agent's "brain" — organized around a continuously running loop: perceive the environment and input, plan the next step, invoke tools to execute it, then reflect on and correct the outcome. The two most influential frameworks in the literature are ReAct and Reflexion. ReAct (Reasoning and Acting) interleaves step-by-step reasoning (chain-of-thought) with grounded action (calling tools, querying databases, spawning sub-agents), so the model no longer produces an answer in one shot but instead thinks a step, acts a step, and observes a step. Reflexion adds self-critique and memory on top of that, letting an agent revise its subsequent behavior based on the previous round's failures (EmergentMind, Medium 2024–2025). New 2025 work goes further in decoupling planning from execution: ReflAct, for instance, has the agent continuously check its own "belief state" against the task goal rather than fixating only on the next action; hierarchical planning-execution architectures — breaking a high-level goal into sub-problems and running a standard ReAct loop on each — are used to curb "concept drift" across long multi-step chains and improve stability (arXiv:2505.15182 and related work, 2025).

Understanding this loop is what makes clear the three advantages orchestration holds over a single model answering directly — and which educational needs each one serves:

- **Tasks become decomposable.** "Produce a tiered lesson worksheet" breaks into sub-steps — retrieve the curriculum standard, diagnose the learner's current level, generate tiered tasks, align to an assessment rubric, format the final document — each handed to whichever agent or tool is best suited to it. This speaks directly to the reality that instructional-research work has many stages, each demanding its own expertise.

- **The process becomes interruptible.** The intermediate output at each step — a retrieved curriculum-standard entry, a draft task — can be checked, edited, or vetoed by a teacher or another agent, rather than a black-box result being dumped on the user all at once. This gives "teacher in the loop" an actual technical foothold.
- **Results become verifiable.** Introducing an independent "reviewer" role that scores interaction quality makes what was previously an implicit judgment about teaching quality explicit and scorable, giving the automated pipeline a built-in capacity for self-inspection.

None of this comes free, though. A 2025 survey makes the trade-off explicit: adding explicit reflection, separating planning from execution, and forcing goal restatement do systematically mitigate common failure modes — context drift, error accumulation, hallucination, inefficient tool calls — but they also raise compute cost, lengthen prompts, and add complexity to multi-agent orchestration (arXiv 2025). The lesson is that orchestration isn't a matter of "more agents is always better" — it's an engineering trade-off between reliability and cost.

The engineering foundation has matured. By 2025, mainstream orchestration frameworks are reasonably complete. LangChain provides an abstraction layer for models, prompts, memory, and workflows that supports multi-step and multi-agent orchestration. Microsoft released the **Microsoft Agent Framework** in public preview on October 1, 2025, merging its earlier AutoGen (multi-agent orchestration patterns) with Semantic Kernel (enterprise-grade AI governance) — a signal that orchestration capability has moved from research prototype to operable middleware (Kubiya, TrueFoundry 2025). These general-purpose frameworks give education-vertical products a reusable scheduling backbone, sparing smaller teams from having to build an orchestration layer from scratch.

Three recurring orchestration patterns have emerged in education, drawn from the last two years of academic and product work:

- **Tutor-student-reviewer triads (self-play content generation and quality control).** One agent role-plays a learner to surface common misconceptions, another plays a tutor offering scaffolded guidance, and a third serves as reviewer scoring the interaction's quality. The three spar repeatedly, which both mass-produces instructional content grounded in realistic misconceptions and builds a quality-control loop directly into the generation process. Its value isn't in seeming "more human" — it's in converting a judgment that used to require manual spot-checks ("was this a good tutoring exchange?") into something automatable and cumulative.
- **Multi-agent systems split along instructional function.** Splitting a tutoring task across dedicated agents for lesson planning, dialogic instruction, and assessment/diagnosis is now the mainstream approach in 2025 intelligent tutoring system (ITS) research. The **GenMentor** framework, presented at ACM Web Conference 2025, first uses a fine-tuned LLM to map a learner's goals onto the skills required, then schedules a learning path based on a dynamic, multi-dimensional learner profile — forming a "goal-skill-path" orchestration chain. Related

work includes multi-agent collaboration built on a "preview-analyze-reason" pipeline for conversational learning diagnostics (Wang et al., WWW 2025 Companion; survey at arXiv:2503.11733, 2025).

- **Teacher-facing lesson-prep orchestration (human in the loop).** Teacher-centered multi-agent systems break "generate a tiered worksheet / differentiated lesson packet" into sub-tasks — parsing requirements, generating content, tiering difficulty, formatting — assigned to different agents, with the teacher reviewing and deciding at key checkpoints. The **FACET** framework on arXiv (2025) supports exactly this: teachers implementing differentiated instruction at scale, with its role explicitly framed as "assisting the teacher" rather than "replacing the teacher" — closely aligned with this book's position on where human and machine responsibility should divide.

One commercial example already at scale: Khanmigo's "constrained tutor" orchestration. Khan Academy's Khanmigo is among the most widely deployed commercial products applying the orchestration paradigm to K-12 tutoring today, with users growing from roughly 68,000 in the 2023–24 pilot year to more than 700,000 in 2024–25 (Khan Academy / reruption 2025). Architecturally, it is not simply "GPT-4 exposed directly to students" — it's a constrained orchestration layer. A system prompt first intercepts the student's intent and identifies which step of the problem-solving logic they're currently on, then constrains the response to giving hints, asking follow-up questions, and pointing out specific grammatical or computational errors — deliberately **not** rewriting or supplying the complete solution. The industry describes this as "a pedagogical agent whose top-level instruction is to guide rather than to answer" (reruption, mlopsaudits 2025). It also routes tasks to different models based on cost and capability — GPT-4 handles tutoring, lighter models handle auxiliary tasks, and a separate model handles writing — and includes built-in content moderation to keep the conversation focused on education and to decline inappropriate questions; for teachers, it offers tools for grading, generating handouts, and drafting discussion questions (Khan Academy 2025). What this case demonstrates is that orchestration's real engineering weight isn't "a stronger model" — it's "tighter constraints and clearer division of labor." Encoding Socratic guidance, ethical boundaries, and model routing directly into the orchestration layer is what sets it apart from an ordinary chatbot. Notably, its accuracy on multi-step math computation only improved through continuous updates by 2025 — further evidence that reliability across long chains of steps takes sustained work to get right.

A conservative read on maturity. Today's at-scale deployments concentrate on instructional-research and grading tasks with **relatively fixed workflows and clearly bounded tools** — every step in the orchestration has a well-defined input artifact and verification criterion, so failures can be caught promptly. Open-ended, long-horizon "fully autonomous teaching agents" that require cross-session independent decision-making remain largely at the demo and small-pilot stage; their reliability isn't yet sufficient for unsupervised classroom deployment. Errors at any point in a multi-step chain compound as they propagate (error accumulation), and the more autonomy an agent has to

decide its own next step, the harder the challenges of explainability and control become. This book's position, accordingly, is that educational agent orchestration should be prioritized for workflows **with a teacher in the loop and clear verification checkpoints**, treating autonomy as a controlled increment rather than a default goal. For related performance comparisons and failure modes, see the assessment discussion in Chapter [TBD: assessment chapter number].

10.1.2 Retrieval-Augmented Generation (RAG): Making Answers Traceable

Mechanism: anchoring answers to identifiable sources. RAG retrieves relevant passages from a controlled knowledge base before generation and injects them into the prompt as context, anchoring the model's output to **identifiable sources** rather than pure parametric memory. The literature calls this behavior — showing where a result came from so the user knows the origin of the information — grounding: constraining the answer to retrieved documents substantially reduces a large model's tendency to generate content that "sounds plausible but is wrong" (Springer, *Business & Information Systems Engineering*, 2025). Appreciating RAG's value requires distinguishing two kinds of hallucination: **intrinsic hallucination** (output inconsistent with the supplied reference context) and **extrinsic hallucination** (output that cannot be verified against the context at all). What RAG chiefly addresses is the tendency to present unsupported claims as established fact, making it possible to classify the relationship between an answer and its evidence as "supported / not mentioned / contradicted / supplementary" (arXiv:2505.04847 and related work, 2025).

For educational settings, this paradigm answers three hard requirements:

1. **Factuality and traceability.** Answers can be traced back to a textbook, curriculum standard, or school-authored material; teachers can verify them, and students can ask where they came from. For minors, learning to ask "where did this come from" is itself a transferable media-literacy skill whose educational value is not less than the answer itself.
2. **Governable knowledge.** A knowledge base curated by an institution can be added to, edited, and audited, preventing the model from treating outdated or incorrect information as settled fact. Teachers can upload their own school's course materials so explanations are grounded in that specific curriculum — Google's Learn About already offers exactly this: the ability to "upload your own source documents so explanations are grounded in them," constraining in-lesson explanations to teacher-supplied material (Google I/O 2025).
3. **Vertical adaptation at low cost.** Subject-matter corpora and school-specific knowledge can be connected without retraining the model, transferring a general LLM's capabilities cheaply to a specific grade level and textbook edition — substantially lowering the barrier to vertical deployment. This is the technical premise that lets "one general-purpose LLM plus one school-specific knowledge base" rapidly serve different regions and different textbook editions.

Where it's heading: from naive concatenation to an engineered combination of techniques. Education-sector RAG in 2026 is moving away from the naive approach of retrieving a passage and pasting it straight into the prompt, toward a modular, benchmarkable engineering stack. The consensus view in the industry is that RAG has already made the leap from "usable" to "genuinely good," and the key lies in modular decomposition, graph structuring, and agentic orchestration (Synthimind, Data Nucleus 2025):

- **Hybrid retrieval (dense + sparse).** Vector retrieval captures semantic similarity; keyword retrieval catches terminology and proper nouns (specific formula names, historical events) that vector search alone tends to miss on exact matches — the two complement each other.
- **Reranking.** After the initial retrieval pass, a high-precision cross-encoder scores each candidate passage individually against the query, narrowing dozens of initial candidates down to the true top 5 or 10 before feeding them to the model — trading a modest amount of compute for a substantial gain in retrieval precision.
- **Graph retrieval / GraphRAG.** Combining a knowledge graph's symbolic reasoning with dense semantic retrieval supports explainable, context-aware multi-hop question answering. Education already has a curriculum-focused example in the **KA-RAG** framework (*Applied Sciences*, 2025), which pairs a knowledge graph with agentic retrieval in a dual-retrieval strategy to answer curriculum-related questions, balancing explainability with semantic coverage — a particularly good fit for subjects like math and physics where concepts have clear prerequisite relationships.
- **Agentic RAG.** Introducing "planning" and "reflection" lets an agent decide, based on how complex the question is, whether to retrieve in multiple steps, whether to call external tools (databases, APIs), and whether to self-correct before generating a final answer — upgrading RAG from a one-shot retriever into an agent that "thinks it through before searching, and tries another path if the first one comes up empty." This is orchestration bleeding into RAG, further evidence that the three paradigms share common roots.

Stringing these pieces together, a "good enough" 2026 education RAG pipeline runs roughly as follows: a student asks "why do you extract the coefficient of the quadratic term first when completing the square?" An agent recognizes this as a concept question requiring textbook backing and triggers hybrid retrieval, pulling both vector-similar passages and passages containing the keywords "completing the square" and "quadratic coefficient" from the school's knowledge base. A cross-encoder reranks the candidates and selects the handful of textbook passages most relevant to the question. The model generates its explanation constrained to those passages, tagging each step of the explanation with the textbook section number it draws on. A verification step checks whether every generated sentence is actually supported by the retrieved passages, flagging any "not mentioned / contradicted" sentence for regeneration. What the student sees isn't just an answer — it's

a traceable citation ("this explanation comes from Section X of this textbook") they can question further, and the teacher can verify it with one click. What actually determines the experience in this pipeline is the granularity of how the knowledge base is segmented, the precision of the reranking, and the strictness of the verification step — all of which belong to "curation and engineering," not "model parameters."

Trustworthy only if measurable. RAG can only deliver on its promise of "evidence-based" answers if its own faithfulness is itself measurable. Evaluation methods for RAG hallucination and faithfulness matured quickly in 2025: benchmarks emerged using sentence-level labels ("supported / contradicted / not mentioned / supplementary," with the latter two classified as "unfaithful") to annotate the relationship between an answer and its evidence, alongside frameworks using a high-capability model as an "LLM-as-Judge" to assess retrieved-passage relevance, citation faithfulness, and answer completeness — some methods now match human annotation on faithfulness judgments at a fairly high level of agreement (arXiv:2505.04847, ACL/EMNLP 2025 Industry track, and related work). This gives education buyers something concrete to ask for: third-party, reproducible faithfulness evaluations from vendors.

RAG also brings a capability that's underappreciated in education: **the ability to say "I don't know" with justification**. When retrieval turns up no reliable support, a responsible educational RAG system should decline to answer or state plainly that "the textbook doesn't cover this" rather than fabricate a plausible-sounding answer. Work specifically evaluating RAG hallucination under an "abstention policy" already exists as of 2025 (ResearchGate 2025). For minors, a system willing to say "I'm not sure — let's check the textbook together" is more trustworthy than one that always answers confidently. Building "knowing what you don't know" into product behavior is itself a form of honest educational modeling.

This engineering path shares its roots with this Blue Book's own evidentiary discipline. Every conclusion needs a source; we'd rather leave a placeholder than fabricate one — RAG, in a sense, is simply this editorial fact-checking discipline internalized into a technical product. Domestic vendors are pursuing hallucination governance along the same lines: iFLYTEK states publicly that its Spark model uses "a hallucination-governance technique based on multi-path sampled verification and factuality-constrained reinforcement learning," yielding a meaningful improvement in response reliability on RAG and related tasks (QbitAI, 53AI 2025).

Real-world evidence from education deployments. RAG's application to higher-education tutoring already has peer-reviewed pilot data: a 2025 *ScienceDirect* pilot study on AI tutoring specifically evaluated RAG models applied to tutoring; Springer has also published pedagogical work using RAG to improve digital literacy; and, on classroom Q&A, researchers are systematically comparing vector retrieval against graph retrieval to establish best practice (arXiv, *Aligning LLMs for the Classroom with Knowledge-Based Retrieval*, 2025). Together these studies point to a conclusion with real

guidance value for industry: what differentiates educational RAG products comes chiefly from **knowledge-base curation quality** and **retrieval-and-citation engineering**, not the underlying model's parameter count. Investing in curating a high-quality, well-structured, continuously updated school knowledge base tends to pay off better than chasing a bigger model.

This conclusion carries particular weight for the domestic industry. China's many textbook editions, wide regional variation, and curriculum standards that spiral upward by grade level are exactly the conditions where "one general-purpose LLM plus regionally and grade-specific school knowledge bases" delivers the most value: rather than retraining a separate model for every textbook edition, it makes more sense to curate a structured knowledge base tagged to curriculum standards and kept continuously updated, letting the same underlying model serve different regions through different knowledge bases. In other words, RAG converts "the locality of education" from a burden on model training into an asset in knowledge curation — whoever has the more authoritative knowledge base, updated more often and structured for better retrieval, has the more trustworthy product. This is also why Section 10.2.2 argues that engineering should take priority over the parameter race: in the education vertical, the moat lies more in data and engineering than anywhere else.

10.1.3 Long-Term Memory: Making the System Remember This Learner

Mechanism: turning a stateless responder into a stateful learning partner. Memory gives a product the ability to retain a learner's profile, error patterns, preferences, and progress toward goals across sessions and over time — the technical precondition for truly personalized, "a thousand faces for a thousand learners" learning. A product without memory treats every conversation as a first meeting; "personalization" then depends entirely on the user re-explaining their context each time, which is both inefficient and impossible to build on. Memory can be roughly divided into three layers:

- **Short-term / working memory**: the context window within a single session, serving immediate responses and topical coherence.
- **Long-term memory**: a structured learner profile accumulated across sessions — mastery of specific concepts, recurring error patterns, learning pace, progress toward goals — the core of "remembering this particular learner."
- **Episodic and semantic memory**: records of specific learning events (episodic — e.g., "got stuck on completing the square three times on March 5") and the stable knowledge state abstracted from them (semantic — e.g., "this student has not yet mastered completing the square"). The former supports traceable, process-oriented assessment; the latter supports a stable judgment of current level.

A concrete scenario shows how the three layers work together: a student makes the same mistake on checking solutions to rational equations for two weeks straight. **Episodic memory** logs the specific events — "missed the extraneous-root check on November 3, 7, and 12." The system abstracts a

semantic memory from this — "this student hasn't internalized why checking solutions is necessary; this is a procedural gap, not a conceptual misunderstanding" — as a stable judgment. At the next tutoring session, **working memory** loads that semantic conclusion into the current context, so the tutor can add scaffolding at the verification step directly, without needing to ask again. A product without memory can't do this: it starts from zero every time, teaching the same general method over and over, wasting the student's time and never accumulating a picture of "this learner's growth trajectory." Memory, then, isn't a nice-to-have on top of personalization — it's the precondition that makes the word "precise" mean anything at all.

On the engineering side, the MemGPT / Letta line of work treats the LLM as analogous to an "operating system": the system automatically summarizes conversations as they grow, moves infrequently used information into a retrievable database, and stores and edits details like names, dates, and preferences — turning a "stateless pattern processor" into "a stateful agent that keeps learning and adapting across long stretches of time" (Letta / MemGPT 2025). The value of this "operating system" metaphor is that it turns "what to remember, how long to keep it, when to retrieve it, when to forget it" into an explicitly designable, auditable policy, rather than an uncontrollable side effect of the model — which is exactly what education demands from memory governance. On the education side, 2025 surveys and new work identify "reusable, parameterized learner profiles" as the breakthrough: the old single-prompt, single-model approach struggles to maintain a consistent learner profile at scale, whereas a multi-agent system can maintain a profile that stays consistent across interactions while capturing both psychological and cognitive dimensions (survey at arXiv:2503.11733; EMNLP 2025 Findings). Worth repeating: good personalization shouldn't focus only on cognitive dimensions — motivation, emotion, and self-concept shape learning outcomes just as profoundly. A system that remembers which problems a student got wrong but can't read their frustration is only personalizing half the picture.

Product status: memory is now a default capability. Memory has moved from a research concept into the default configuration of commercial products. ChatGPT's **Study Mode**, launched by OpenAI on July 29, 2025, tailors lesson difficulty "based on questions that assess skill level **and memory of past conversations**," and implements Socratic guidance — asking follow-up questions, giving hints, prompting self-reflection rather than handing over answers directly — through custom system instructions (OpenAI 2025). Google has likewise built LearnLM's profiling and personalization capability into custom quizzes within the Gemini app and other settings (Google 2025). Domestic learning-hardware makers treat cross-session learning-record persistence as a core selling point too: at its 2025 summer product launch, iFLYTEK's AI learning tablet named "AI 1-on-1 precision learning / Q&A tutoring / interactive lessons" as its three core functions, with a teacher-assistant feature that can "automatically generate learning reports and quickly diagnose learning gaps" — personalization here depends on continuously recording learner state (QbitAI 2025). TAL Education Group similarly embeds its Jiuzhang LLM into learning-tablet features like on-demand Q&A,

precision learning, and essay grading, all of which likewise depend on learning-record persistence across sessions (TAL 2025).

A risk flag (tightly bound to the governance chapter). The stronger memory gets, the more directly it touches the **collection boundaries, data minimization, retention limits, and right to be forgotten** that apply to minors' data. UNESCO's 2025 *Guidance for Generative AI in Education and Research* recommends mandatory data-privacy protections and age limits on the use of generative AI tools (UNESCO 2025); China's own *Guidelines for the Use of Generative AI by Primary and Secondary School Students (2025 Edition)* likewise sets explicit usage boundaries by grade level (Ministry of Education 2025; see Section 10.2.1 for detail). This book accordingly maintains that **memory should not be maximized for its own sake — it should be sufficient, controllable, and deletable.** Any accumulation of a learner profile must be tightly bound to the red lines set out in the "Governance and Safety" chapter [TBD: governance chapter number], treating data minimization, purpose limitation, auditability, and deletability as baseline requirements rather than optional add-ons. A responsible educational memory system should be able both to remember the learner and to forget them completely on request — the right to be forgotten needs an actual operable channel, not just language in a privacy policy.

10.1.4 How the Three Paradigms Work Together: A Unified Product Foundation

The three paradigms are not isolated features — they should be understood as **three facets of one system**: orchestration provides the skeleton for acting, RAG provides the factual bedrock for citing sources, and memory provides the continuity of having a personality. Stacked together, they form an education agent that can act, cite its sources, and remember the learner; missing any one of the three, a product falls back into some degraded form — orchestration without RAG means "doing a lot, but getting a lot wrong too"; RAG without memory means "starting from scratch getting to know you every time"; memory without orchestration means "remembering, but unable to carry out a compound task."

The past two years' flagship products are exactly this convergence in practice. At I/O 2025, Google folded LearnLM directly into Gemini 2.5 and benchmarked it against competitors including Claude 3.7, GPT-4o, and OpenAI o3 across "five pedagogical principles" — Gemini 2.5 Pro came out preferred in every category. Education and pedagogy experts favored it across a range of learning scenarios, judged not just on answer accuracy but on pedagogical factors like "does it offer appropriate scaffolding" and "does it correct the student's mistakes" (Google 2025). Its capabilities have been built into NotebookLM, Learn About, and custom quizzes in the Gemini app, and it supports uploading source documents for grounded explanations — exactly the convergence of orchestration (breaking learning into guided steps), RAG (grounded in source documents), and memory/profiling (tailoring instruction to the learner's level). OpenAI's Study Mode echoes this pattern: moving fast with system-instruction-based behavior first to gather real student feedback,

with the intention of eventually training that behavior into the flagship model itself (OpenAI 2025). Leading vendors have clearly made the three paradigms their product foundation rather than an optional feature — a signal to domestic products that the next phase of competition won't be decided by any single model capability, but by the engineering quality of how the three paradigms are integrated together.

Layering in **multimodal input** and **on-device inference** completes the picture, giving rise to the 2026 product form factor: agentic, multimodal, on-device. It's worth spelling out how these last two dimensions relate to the three paradigms. Multimodality extends orchestration's "perception" step beyond plain text to voice, images, handwriting, and first-person video — a student can photograph a handwritten mistake or narrate a classroom explanation aloud, and only then can the system truly "see" the learning moment as it happens. On-device inference, meanwhile, answers to privacy and network constraints specific to education: keeping a lightweight model and part of memory on the local device, going to the cloud only when necessary, cuts latency and bandwidth dependence while also keeping the most sensitive data about minors local — a natural fit with the "minimized, controllable, deletable" approach to memory governance described in Section 10.1.3. "On-device" in an education context, then, isn't just a performance choice — it's an architectural choice about privacy and equity: it lets network-constrained regions still get memory-backed personalized tutoring without having to upload all their learning data. For how this plays out in embodied hardware, see this institute's *AI-SLI 2026 Blue Book on the AI Smart Glasses Education Industry* and *AI-SLI Global Education Robotics Development White Paper 2026* — the former demonstrates multimodal perception and first-person interaction (our affiliated enterprise NetDragon (HK:0777) made a strategic investment of US$20 million in Rokid in November 2023 and signed a five-year strategic cooperation agreement; Rokid Glasses went on sale in Q2 2025, and per Omdia data, Rokid ranked first globally in the display-equipped AI-glasses segment in 2025 — Zhitong Finance, Stockstar 2025); the latter demonstrates the boundaries of embodied agents operating in physical space. Software's three paradigms and hardware's embodiment cross-reference each other, forming a complete loop across this institute's Blue Book series: software handles thinking, searching, and remembering; hardware handles seeing, hearing, and being present. The two together approach the next generation of educational interaction — ever-present, seamless, and evidence-based.

10.2 Recommendations for a Diverse Set of Stakeholders

Building on the paradigm assessment above, we offer structured recommendations to four groups of stakeholders: policy and standards bodies; industry and product teams; schools and teachers; and researchers working on equity. Wherever a specific metric, checklist, or document number has not yet been verified, we mark it as a placeholder pending confirmation. All four sets of

recommendations share a common thread: turning "new paradigm" from a technical selling point into **deliverable reliability** and **governable equity**.

10.2.1 Policy and Standards: Organized Around Usability, Trustworthiness, and Governability

China built out a policy foundation for AI in education over the course of 2025. On May 12, 2025, the Ministry of Education released the *Guidelines for General AI Literacy Education in Primary and Secondary Schools (2025 Edition)* and the *Guidelines for the Use of Generative AI by Primary and Secondary School Students (2025 Edition)*, forming a two-track system of "general literacy plus usage norms." The former centers on literacy development through a spiraling curriculum — from hands-on curiosity in primary school, to understanding technical principles in middle school, to systems thinking and innovative application in high school. The latter sets usage boundaries graded by school level: primary students are prohibited from using open-ended content-generation features unsupervised, though teachers may use them appropriately in class; middle-school students may explore analyzing the logical structure of generated content to a moderate degree; high-school students are permitted to conduct inquiry-based learning that engages with the underlying technical principles (Ministry of Education; China Education Online, CERNET, DoNews 2025). The concurrently released *China Smart Education White Paper (May 2025)* disclosed that 23 provincial-level education administrations have already deployed AI education initiatives in primary and secondary schools, with Beijing, Guangzhou, and other localities having issued specific implementation plans (Ministry of Education 2025). Building on this foundation, we recommend:

- **Establishing admission and evaluation standards for educational AI products.** Push for capability and safety evaluation standards tailored to the education vertical, covering four dimensions: subject-matter accuracy, values safety, protection of minors, and traceability. Existing third-party evaluation practice offers a template: the China Academy of Information and Communications Technology (CAICT) has already conducted evaluations of educational LLMs with tiered ratings, and TAL's Xiangzhang (Jiuzhang) model was among the first to pass and to receive the current top tier (Level 4+) (TAL 2025). Evaluations should include the "reproducible faithfulness measurement" described in Section 10.1.2 as a mandatory test item, to avoid mistaking "appears to cite sources" for "citations are actually reliable." The definitive standard document number and the body responsible for setting evaluation criteria remain [TBD: standard document number/source].
- **Clarifying compliance boundaries for data and memory.** Provide operable rules on the collection, retention, cross-border transfer, and deletion ("right to be forgotten") of learner data, implementing the principles of data minimization and purpose limitation in alignment with the memory-governance red lines described in Section 10.1.3. "Memory-enabled products" in

particular should face stricter retention-period limits and deletion-response requirements than ordinary applications.

- **Aligning with existing regulation and international consensus.** Domestically, this means connecting with the two Guidelines above and the Smart Education White Paper; internationally, it means drawing on UNESCO's 2025 *Guidance for Generative AI in Education and Research* recommendations on data-privacy protection and age limits on use (UNESCO 2025). Specific provisions and effective dates await the formal text [TBD: policy document number/effective date].

Policy tools should combine hard and soft approaches. Admission thresholds and safety baselines — values alignment, protection of minors, data deletion — warrant mandatory standards; fast-evolving technical forms (specific orchestration architectures, memory strategies) warrant regulatory flexibility, guided by best-practice guidance rather than locked down by rigid rules that can't keep pace with the technology. This mirrors the approach already embedded in China's two Guidelines — grading restrictions by school level, strictest at the youngest ages and progressively relaxed — itself a fine-grained governance model that matches age, capability, and risk. The same logic is worth carrying through to product admission: the same feature should face different data and autonomy boundaries depending on which grade level it targets.

10.2.2 Industry and Products: Turning "New Paradigm" Into Deliverable Reliability

- **Replace claims with evidence.** Product capability claims need third-party, reproducible evaluation behind them, and should avoid overstating the degree of automation. Leading vendors have already modeled a cautious release path — system instructions first, small-scale feedback collection, only later training the behavior into the flagship model (OpenAI's Study Mode, 2025) — rather than declaring "a fully automated teacher" outright. Any vendor claiming "AI one-on-one" or "precision learning" should be able to supply verifiable evidence of outcomes and failure-rate data. For product benchmarking methodology, see the assessment discussion in Chapter [TBD: assessment chapter number].
- **Prioritize engineering RAG and memory over the parameter race.** As argued in Section 10.1.2, differentiation in education comes more from knowledge-curation quality, retrieval-and-citation engineering, and controllable memory than from simply adding parameters. Domestic practice bears this out: iFLYTEK governs RAG hallucination through "factuality-constrained reinforcement learning" (QbitAI 2025); TAL runs a dual-engine approach combining its Jiuzhang LLM with DeepSeek, embedding these capabilities into concrete learning-tablet features like essay grading, on-demand Q&A, and precision learning, with Jiuzhang selected as

one of the 2025 Global Smart Education Excellent Cases (TAL 2025) — the competitive focus has shifted from "whose model is bigger" to "whose deployment is more stable."

- **Give equal weight to on-device and multimodal capability.** For privacy-sensitive, network-constrained school environments, advance on-device inference and lightweight deployment — using lightweight models for auxiliary tasks and keeping sensitive data on the local device — cutting costs while protecting privacy. At the same time, build first-person perception, voice, and image input into the product baseline, in coordination with this institute's two hardware lines (AI glasses and educational robotics), preparing for an ever-present, in-the-moment style of learning interaction.

10.2.3 Schools and Teachers: Literacy First, Human-Machine Collaboration

- **Building teacher AI literacy.** Bring "prompt engineering, output verification, ethics and data awareness, and judgment about human-machine division of labor" into teacher professional development. Teachers should be able to identify and correct an AI's factual and value-based errors rather than accepting them passively — this is exactly the audience that RAG's traceability and orchestration's verifiability are designed to serve: only when the technology lays its intermediate outputs bare can a teacher meaningfully stay in the loop. Literacy-building shouldn't stop at "knowing how to use the tool" — it needs to reach the level of "knowing when the tool can't be trusted." The corresponding literacy framework and required training hours remain [TBD: literacy framework/source].
- **Drawing a clear line on human-machine responsibility.** In assessment and decision-making contexts, hold firmly to "AI assists, the teacher is responsible" — high-stakes judgments (academic standing, admissions, identifying a mental-health crisis) must never be handed entirely to a model. Product design should build "the teacher has final say" into the workflow as a checkpoint that cannot be skipped, as demonstrated by teacher-centered orchestration systems like FACET (arXiv 2025) — turning human-machine division of labor from an aspiration into an architectural constraint. There's an easily overlooked risk here: as AI output becomes more fluent and convincing, the human in the loop can degrade into rubber-stamp review — automation bias, where trust leads teachers to relax their verification. Countering this requires effort on both the product and training sides: products should proactively surface uncertainty and sources (RAG's traceability is a lever here too), while teacher training needs to specifically build the judgment to "stay skeptical of output that's fluent but questionable" — the advanced goal of literacy is precisely knowing when the tool can't be trusted.

10.2.4 Research and Equity: Building the "Digital Divide" Into Design From the Start

- **Closing the twin gaps in access and use.** UNESCO warned in 2025 that "the digital divide is rapidly becoming an AI divide" — as of 2024, nearly a third of the world's population (roughly 2.6 billion people) still lacked internet access, with girls, rural populations, people with disabilities, and marginalized groups disproportionately affected (UNESCO 2025). Without deliberate design for equity, generative AI risks widening existing inequality along two paths: the **access gap** (whether compute, devices, and connectivity are available at all) and the **use gap** (how strong teachers' capability and digital literacy are) — even where devices are available, not knowing how or being afraid to use them creates its own disparity. Research should treat monitoring of regional, urban-rural, and inter-school gaps as routine, to prevent the benefits of the technology from flowing first to those already better off. Relevant local equity data remain [TBD: equity data/source].
- **Designing for accessibility for disadvantaged groups.** Build accessibility, multilingual support, and special-education needs into the product baseline rather than treating them as add-ons — offering learning content and explanations by default in multiple languages, with text-to-speech, magnification, and adjustable difficulty. NotebookLM's Audio Overviews already support more than 80 languages (Google 2025) — a useful reference for treating "usable by everyone" as a design constraint from the start, rather than an "accessibility mode" bolted on after a product matures. The three paradigms can reinforce equity here too: on-device capability lets network-constrained regions use the product offline; multimodality lets students with reading difficulties or visual impairment interact through voice and image; and multilingual knowledge bases built on RAG let students speaking minority languages or non-mainstream native languages get explanations grounded in their local curriculum. Inclusive benefit doesn't happen automatically — equity becomes real only when "usable by the learner in the most disadvantaged position" is written into the design goals and acceptance criteria from the outset.
- **Measuring success by reduced burden, not increased burden.** One study, published in *Nature Scientific Reports*, found that students using AI tutoring learned more in less time than they did through active classroom learning (as compiled by tutorbase, 2025) — but "learning more, faster" should not be allowed to morph into disguised pressure. Research and evaluation should build learner burden, motivation, and mental health into outcome metrics, guarding against converting AI's efficiency dividend one-directionally into more drilling and longer study hours. Being learner-centered means placing "is this person growing with more ease" above "do the metrics look better."

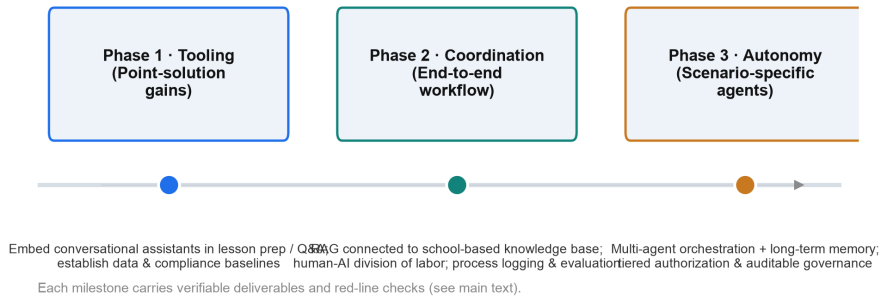
The table below summarizes the core recommendations and priority levers for each of the four stakeholder groups (undetermined items marked as placeholders).

Stakeholder	Core Recommendation	Priority Lever	Key Constraint/Reference
Policy and standards	Establish admission and evaluation standards; clarify data and memory boundaries	Education-vertical evaluation standards; data compliance and deletion rules	The two Guidelines, Smart Education White Paper, CAICT tiered evaluation, UNESCO Guidance; standard document number [TBD]
Industry and products	Replace claims with evidence; engineer RAG/memory; on-device and multimodal	Third-party faithfulness evaluation; knowledge curation; on-device deployment	Cautious release path (Study Mode); hallucination governance (iFLYTEK); evaluation criteria [TBD]
Schools and teachers	Build AI literacy; clarify human-machine responsibility boundaries	Teacher training; embed division of labor as workflow checkpoints	Prompting/verification/ethics literacy; teacher in the loop (FACET); literacy framework [TBD]
Research and equity	Close the access/use gaps; design for accessibility	Gap monitoring; accessibility/multilingual baseline	AI divide (~2.6 billion without access); multilingual accessibility; local equity data [TBD]

10.3 Implementation Roadmap: A Three-Stage, Verifiable Pace of Progress

To make these recommendations actionable and assessable, we propose a three-stage roadmap: **Foundation, Deepening, and Convergence**. Each stage sets qualitative goals and verifiable milestones; specific dates, scale, and evaluation metrics are marked as placeholders for implementing bodies to fill in based on actual conditions. It's worth stressing that the three stages are not strictly sequential — different regions and grade levels can progress in parallel, and more advanced regions can move ahead into stages two and three while still supporting less-developed regions in laying their foundation. The roadmap describes a logical order, not a one-size-fits-all timetable.

Fig. 5 A three-phase roadmap for deploying generative-AI education products



Source: this report's proposed implementation roadmap (see Chapter 10).

10.3.1 Stage One (Foundation): Standards and Infrastructure

- **Goal:** Establish baseline evaluation and data-compliance standards, complete an initial rollout of teacher AI literacy training, and take stock of the current scale of both the access gap and the use gap.
- **Milestones:** Implement school-based usage norms and grade-level red lines based on the two Guidelines; connect to third-party evaluation (such as CAICT's tiered rating of educational LLMs) to establish local procurement admission thresholds [TBD: local admission criteria]; complete an equity baseline assessment of compute, devices, and teacher capability [TBD: local equity data/source]; run a first round of teacher AI-literacy training, focused on "output verification" and "data awareness."
- **Verifiable markers:** Has a school- or district-level usage policy been published? Does procurement require vendors to supply third-party evaluation evidence? Has teacher-training coverage reached a set threshold [TBD]?
- **Why the foundation stage can't be skipped:** The acceleration on the adoption side — teacher usage rates doubling over two school years, teen usage rates doubling, as noted at the opening of Section 10.1 — means the window for "use it first, figure out the rules later" is closing fast. When rules lag behind adoption, hard-to-reverse facts on the ground accumulate: data already collected, habits already formed. The real value of the foundation stage is putting evaluation yardsticks, data red lines, and teacher judgment in place before scale hits, rather than "polluting first and cleaning up later."

10.3.2 Stage Two (Deepening): Paradigms in Practice, Scenarios in Depth

- **Goal:** Stably embed RAG, memory, and agent orchestration across the four scenarios of Empowering Teaching, Supporting Learning, Supporting Research (PD), and Intelligent Assessment, forming a reusable reference architecture, and lock "teacher in the loop" into the product workflow.

- **Milestones:** Establish a standing, cross-scenario evidence-based evaluation mechanism that tracks answer faithfulness, failure rate, and teacher intervention rate as routine metrics [TBD: mechanism/responsible body]; produce reference deployment plans for on-device and multimodal use; scale up first in instructional-research and grading tasks with fixed, verifiable workflows before cautiously extending toward longer-horizon tutoring.
- **Recommended sequencing:** Follow an "easy-to-verify first, hard-to-verify later" progression — first get the three paradigms running reliably on tasks with clear boundaries and verifiable outcomes (lesson prep, objective-question grading, reference-answer Q&A), building up a library of failure cases and teacher-intervention data; only once the evaluation mechanism has matured and reliability can be quantified should deployment cautiously extend to higher-uncertainty, higher-stakes tasks like essay/subjective-question assessment and long-horizon personalized tutoring. This sequencing matches the conservative maturity assessment of orchestration in Section 10.1.1: release autonomy as a controlled increment, not all at once.
- **Verifiable markers:** Is a "teacher in the loop" orchestration workflow in routine use? Do RAG answers link back to the school knowledge base by default? Has a reproducible product-outcomes evaluation report been established, one that discloses both success rates and failure modes?

10.3.3 Stage Three (Convergence): Software-Hardware Coordination and a Governance Loop

- **Goal:** Achieve "software agent, hardware embodiment" coordination, scale deployment within governance red lines, and see meaningful convergence in both the access gap and the use gap.
- **Milestones:** Establish cross-referencing and joint evaluation with this institute's two hardware lines, the *AI-SLI 2026 Blue Book on the AI Smart Glasses Education Industry* and the *AI-SLI Global Education Robotics Development White Paper 2026* [TBD: joint-evaluation criteria]; build a full-lifecycle audit loop for data and memory (collection-use-retention-deletion fully traceable, auditable, and responsive to deletion requests); elevate accessibility from "compliance floor" to "design precondition."
- **Why software-hardware convergence is the endgame:** Software's three paradigms handle thinking, searching, and remembering — but learning happens in real physical and social settings: a hand raised in class, an action at a lab bench, a conversation between classmates. Only when perception and interaction extend from the screen to embodied carriers like first-person glasses and educational robots can AI understand the learning context while actually present in it — genuinely bringing orchestration's perceptual input, RAG's on-the-spot evidence-gathering, and memory's continuous profiling into the real learning moment. This is also the logic behind this institute advancing on three fronts simultaneously — a software Blue Book, an

AI glasses Blue Book, and an educational robotics White Paper: only by cross-referencing and jointly evaluating all three can we offer a complete picture that is ever-present, seamless, evidence-based, and memory-enabled, rather than fragments that each tell only part of the story.

- **Verifiable markers:** Is there an operational channel for exercising the "right to be forgotten" over learner data/memory? Is there a reproducible joint-evaluation report on software-hardware coordination? Are accessibility metrics for disadvantaged groups built into routine monitoring [TBD]?

10.4 Conclusion

From 2024 to 2026, the decisive shift in generative-AI education products has not been models becoming better talkers — it has been systems becoming better at acting, at citing their sources, and at remembering, while gradually moving toward multimodal, on-device, embodied forms. Agent orchestration, RAG, and long-term memory — the ability to act, to cite sources, and to remember — have converged to rebuild the product foundation. This book has restructured the product landscape around "five scenarios plus three new paradigms," and has held to a single research discipline throughout: evidence first, and a placeholder rather than a fabrication whenever evidence runs out.

We are equally clear-eyed about the other side of these paradigms. Orchestration brings complexity and uncontrolled autonomy; RAG's reliability depends entirely on knowledge-base curation and faithfulness evaluation; the stronger memory gets, the higher the data risk. The three paradigms, then, are not three free passes — they are three capabilities that must be evaluated, governed, and constrained. The technology will keep iterating, but the orientation should stay fixed: **learner-centered, teacher-accountable, with equity and safety as the floor, not the ceiling.** As models become more capable of acting, humans' ability to intervene in the process, verify the sources, and control the data should not be weakened — if anything, it needs to be reinforced through institutional and product design. This is exactly where orchestration's verifiability, RAG's traceability, and memory's deletability land at the level of values.

Looking back, each of the three paradigms answers an old question in education. Orchestration answers "how do you scale personalized instruction?" RAG answers "how does knowledge get transmitted in a way that's trustworthy and traceable?" Memory answers "how do you truly come to know and accompany one specific learner?" Generative AI didn't invent these questions — it simply offers the most powerful tool yet for addressing them, and the one that most demands to be handled with care. The more powerful the tool, the more it tests the values of whoever wields it: will it be used to see every child more clearly and support them more steadily, or to push a standardized logic of efficiency to its limit? This book's position is unambiguous: the ultimate measure of what these three paradigms mean is whether every learner — especially those in the most disadvantaged circumstances — grows up more fairly, and with more room to breathe.

We hope this book offers policymakers, industry practitioners, schools, and researchers a shared, verifiable, and lasting point of reference. The endpoint of technology was never a smarter machine — it is every learner being more truly seen, more truly supported, and given a fairer chance to grow.

Chapter References

1. OpenAI. *Introducing study mode*. 2025-07-29. https://openai.com/index/chatgpt-study-mode/
2. OpenAI Help Center. *ChatGPT study mode FAQ*. 2025. https://help.openai.com/en/articles/11780217-chatgpt-study-mode-faq
3. Google (The Keyword / blog.google). *I/O 2025: LearnLM in Gemini 2.5 and more AI updates to help people learn*. 2025-05. https://blog.google/products-and-platforms/products/education/google-gemini-learnlm-update/
4. Google DeepMind. *Evaluating Gemini in an Arena for Learning* (LearnLM technical report, May 2025). 2025-05-19. https://storage.googleapis.com/deepmind-media/LearnLM/learnLM_may25.pdf
5. Wang et al. *LLM-powered Multi-agent Framework for Goal-oriented Learning in Intelligent Tutoring System* (GenMentor). Companion Proceedings of the ACM Web Conference 2025 (arXiv:2501.15749). https://dl.acm.org/doi/10.1145/3701716.3715244
6. *FACET: Teacher-Centred LLM-Based Multi-Agent Systems — Towards Personalized Educational Worksheets*. arXiv:2508.11401, 2025. https://arxiv.org/html/2508.11401v2
7. *LLM Agents for Education: Advances and Applications*. arXiv:2503.11733 / Findings of EMNLP 2025. https://aclanthology.org/2025.findings-emnlp.743.pdf
8. *KA-RAG: Integrating Knowledge Graphs and Agentic Retrieval-Augmented Generation for an Intelligent Educational Question-Answering Model*. Applied Sciences, 15(23):12547, 2025 (MDPI). https://www.mdpi.com/2076-3417/15/23/12547
9. *Exploring the use of retrieval-augmented generation models in higher education: A pilot study on AI-based tutoring*. ScienceDirect, 2025. https://www.sciencedirect.com/science/article/pii/S2590291125004796
10. *Retrieval-Augmented Generation (RAG)*. Business & Information Systems Engineering, Springer, 2025. https://link.springer.com/article/10.1007/s12599-025-00945-3
11. *Aligning LLMs for the Classroom with Knowledge-Based Retrieval: A Comparative RAG Study*. arXiv:2509.07846, 2025. https://arxiv.org/html/2509.07846
12. *Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards*. arXiv:2505.04847, 2025 (includes sentence-level faithfulness labels and LLM-as-Judge evaluation). https://arxiv.org/html/2505.04847v2

12a. *Benchmarking Hallucination Evaluation for RAG Under an Abstention Policy* (RAG hallucination evaluation under an abstention policy; RAGAS/DeepEval/LLM-as-Judge). ResearchGate, 2025. https://www.researchgate.net/publication/399331938

13. Survey and original work on ReAct / Reflexion / ReflAct agent architectures: EmergentMind, *Reason-Act-Reflect (ReAct) Architectures*; *ReflAct: World-Grounded Decision Making in LLM Agents via Goal-State Reflection*, arXiv:2505.15182, 2025. https://arxiv.org/html/2505.15182v2
14. Letta / MemGPT (ksm26). *LLMs as Operating Systems: Agent Memory* (the Letta framework and the MemGPT long-term memory concept). 2025. https://github.com/ksm26/LLMs-as-Operating-Systems-Agent-Memory
15. Kubiya. *Top AI Agent Orchestration Frameworks for Developers 2025* (includes the Microsoft Agent Framework public preview of 2025-10-01, LangChain, and others). 2025. https://www.kubiya.ai/blog/ai-agent-orchestration-frameworks
16. Synthimind. *RAG Optimization Strategies 2025: GraphRAG, Agentic RAG & Hybrid Search Explained*. 2025. https://synthimind.net/blog/rag-optimization-strategies-2025/
17. Ministry of Education / China Education Online. Release of the *Guidelines for the Use of Generative AI by Primary and Secondary School Students (2025 Edition)* [中小学生成式人工智能使用指南（2025 年版）]. 2025-05-12. https://www.eol.cn/zhengce/wenjian/202505/t20250512_2667831.shtml
18. CERNET (China Education and Research Network). Official release of the *Guidelines for General AI Literacy Education in Primary and Secondary Schools (2025 Edition)* [中小学人工智能通识教育指南（2025 年版）]. 2025-05-13. https://www.edu.cn/xxh/focus/zc/202505/t20250513_2667990.shtml
19. Ministry of Education of the People's Republic of China. *China Smart Education White Paper (May 2025)* [中国智慧教育白皮书（2025 年 5 月）] (includes deployment of AI education initiatives by 23 provincial-level departments). 2025-05. https://bm.cugb.edu.cn/jsfzzx/upload/resources/file/2025/05/26/266517.pdf
20. UNESCO. *Guidance for generative AI in education and research*; and *UNESCO convenes global leaders at Digital Learning Week 2025* (the "AI divide," roughly 2.6 billion people without internet access as of 2024, recommendations on data privacy and age limits). 2023–2025. https://www.unesco.org/en/articles/unesco-convenes-global-leaders-digital-learning-week-2025-shape-inclusive-human-centred-futures-ai

21. QbitAI. iFLYTEK accelerates its "AI + education" push again: learning-tablet feature upgrades [科大讯飞"AI+教育"再提速：学习机功能升级] (includes "hallucination-governance technique based on multi-path sampled verification and factuality-constrained reinforcement learning," teacher-assistant learning reports). 2025-06. https://www.qbitai.com/2025/06/300819.html

22a. reruption / mlopsaudits. *Khanmigo: Khan Academy's GPT-4 AI Tutor* (architecture: system-prompt constraints, model routing, content moderation, teacher tools; users grew from roughly 68,000 to over 700,000). 2025. https://reruption.com/en/knowledge/industry-cases/khanmigo-khan-academys-gpt-4-ai-tutor-scaling-education

22b. tutorbase. *EdTech & AI in Education Statistics 2026* (compiles RAND teacher usage 25%→53%, Gallup/Walton 60% of K-12 teachers, Pew teen usage 13%→26%, Nature Scientific Reports learning-outcomes study). 2025–2026. https://tutorbase.com/statistics/edtech-ai

22c. Precedence Research / Mordor Intelligence. *AI in Education Market Size* (global 2025 market estimated at roughly US$6.9–7.05 billion; estimates vary by source). 2025. https://www.precedenceresearch.com/ai-in-education-market

22d. Springer Nature Link. *Enhancing Digital Literacy Through Retrieval-Augmented Generation*. 2025. https://link.springer.com/chapter/10.1007/978-3-032-17604-2_22

22. TAL Education Group (100tal). TAL's Xiangzhang (Jiuzhang) LLM passes CAICT's educational-LLM evaluation (Level 4+, the current top tier); Jiuzhang LLM selected as one of the 2025 Global Smart Education Excellent Cases [学而思九章大模型通过中国信通院教育大模型评估；九章大模型入选 2025 全球智慧教育优秀案例]. 2024–2025. https://www.100tal.com/zh-cn/news/detail/1822

23. Zhitong Finance / Stockstar. Strategic investment in Rokid: NetDragon (00777) invested US$20 million in November 2023 and signed a five-year cooperation agreement; Rokid Glasses went on sale in Q2 2025; Omdia market-segment ranking [战略投资 ROKID：网龙(00777)2023 年 11 月投资 2000 万美元并签五年合作；Rokid Glasses 2025 年二季度开售、Omdia 细分市场排名]. 2025-01/02. https://cn.investing.com/news/stock-market-news/article-2634061

Appendix A Research Methods and Data Conventions

A.1 Research Methodology

This Blue Book was developed by AI-SLI using a "self-built curated knowledge base plus AI-assisted production pipeline" workflow. The team first built a five-scenario analytical framework — empowering teaching, supporting learning, supporting research and instructional design, intelligent assessment, and governance and safety — and then conducted web-based retrieval to gather evidence for each topic. Drafting drew only on verified, primary-source material retrieved this way, with every specific fact tied to an inline citation and every direction and red line reviewed by human researchers. The research team owned topic selection, the analytical framework, judgment calls, and final conclusions; the AI handled the labor-intensive work of retrieval, drafting, tabulation, and citation management. Judgments about the product landscape and product forms are based on vendors' public materials, authoritative media coverage, and institutional reports, not marketing claims.

A.2 Data Sources

Data draws primarily on primary sources: vendors' official product documentation and announcements; the original texts of laws and regulations in each jurisdiction (e.g., the EU AI Act, the U.S. FERPA/COPPA, and China's Interim Measures for the Management of Generative AI Services); reports from international organizations and authoritative institutions; public benchmark results that can be independently verified; and authoritative media coverage. The "Sources for This Chapter" list at the end of each chapter itemizes the chapter's genuine citations (title, institution, year, URL). Wherever products from affiliated enterprises such as NetDragon (HK:0777) are discussed, the official public information from that company is treated as authoritative.

A.3 Principles for Handling Data Conventions

- **Market data follow different conventions.** Different research institutions define "generative AI in education" and "EdTech" differently, and their figures cover different geographies and reference years — market-size figures from different sources should not be compared or added together directly. Citations retain the source institution, its stated convention, and its publication year.
- **Currency firewall.** Chinese yuan (RMB), U.S. dollars, and Hong Kong dollars are never combined in a single calculation. Every monetary figure carries its currency and reference date, with no implicit currency conversion.

- **Forecasts are distinguished from actuals.** Any institutional forecast or estimate is labeled "forecast," together with its publication date, and is presented separately from actual reported figures.
- **Benchmark results must be verifiable.** Model and product benchmark scores are used only when drawn from verifiable public sources; scores of unclear provenance are excluded.
- **Placeholders, not fabrication.** Any specific figure, policy document number, product list, benchmark result, or citation that could not be verified through retrieval is marked "[TBD: …]" — never invented. This document is a research-stage version; all [TBD] items are to be resolved through multi-source verification before the document is finalized.

A.4 Limitations and Next Steps

Generative-AI education products iterate extremely quickly, and this version retains [TBD] placeholders for certain fast-changing specifications, prices, benchmark results, and market-size figures. These will be filled in with verified sources through the fact-checking pipeline to produce the final version. Product examples are used solely to illustrate product forms and underlying mechanisms; they do not constitute a procurement recommendation or investment advice.

Appendix B Glossary

Term	Definition
Generative AI (GenAI)	AI technology capable of generating text, images, speech, code, and other content, typified by large language models.
Agent	An AI system that can understand a goal, decompose it into tasks, invoke tools, and carry out multi-step tasks continuously.
Agent Orchestration	The planning, scheduling, and coordination of multiple agents/tools, moving a system from "answering" to "completing tasks."
Retrieval-Augmented Generation (RAG)	Retrieving external knowledge before generation and grounding output in it, used to improve accuracy, enable citation traceability, and suppress fabrication.
Long-term Memory	A mechanism for retaining learning history, preferences, and context across sessions to support sustained personalization.
Multimodal	The capacity of a unified model to work across multiple information forms — text, image, speech, video, and code.
On-device	Deploying inference capability on a personal device or local terminal, offering low latency, strong privacy, and offline availability.
Education-vertical Model	A specialized large language model optimized for education scenarios through subject-matter corpora, instructional tasks, and safety alignment.
Scenario Validity	The degree to which what a benchmark measures aligns with the actual demands of real educational use cases.
Process-oriented Assessment	An assessment approach that evaluates the learning process — response trajectories, revisions, collaboration, and the like — rather than only the final output.
Guardrail	A mechanism that imposes safety, compliance, and content constraints on a model's inputs and outputs.
Academic Integrity	Norms and judgments concerning originality, attribution, and the reasonable use of AI in human-AI collaborative learning.
Data Minimization	The principle of collecting and processing only the minimum personal data necessary to accomplish a stated purpose.
EU AI Act	The European Union's Artificial Intelligence Act, which

	regulates AI systems under a risk-tiered framework.
FERPA / COPPA	The U.S. Family Educational Rights and Privacy Act / Children's Online Privacy Protection Act.

Appendix C Citation Conventions

This Blue Book uses an "inline attribution plus end-of-chapter sourcing" convention: specific facts in the body text are attributed to their source institution and year either parenthetically or within the sentence, and the "Sources for This Chapter" list at the end of each chapter gives verifiable entries in full (title · institution/author · year · URL). The convention follows these rules:

- Only sources actually retrieved and verified during production are included; unverified sources are excluded, and the corresponding facts are marked [TBD] in the body text.
- Primary sources are given priority — original texts of laws and regulations, vendors' official documentation, institutions' original reports, and official benchmark pages — with media coverage used as a supplement and the outlet named.
- Where multiple sources report the same fact under different conventions, the primary source is retained and the difference in convention is noted in the body text.
- URLs are public links provided to facilitate verification; if a link becomes inactive, the source can be recovered by searching its title and institution name.

(The finalized version will convert all citations to a unified numbered-reference format with a complete bibliography.)

AI-SLI

2026 · AI-Assisted Research Edition