2026

生成式人工智能教育产品发展蓝皮书

从对话式到智能体、多模态与端侧：
五场景 · 产品图谱 · 评测 · 路线图

AI-SLI

2026 · AI 辅助研制 · 研究版

编写说明：本报告的人工智能辅助研制过程

AI-SLI · 人工智能辅助研制

本蓝皮书是 AI-SLI 在“人工智能辅助知识生产”方向上的一次系统性探索与尝试。报告的研制全程引入生成式人工智能作为研究助手，在研究团队的选题、设计与把关之下，承担了四类劳动密集型工作。其一，循证调研与数据核验——对厂商公开资料、政策原文、学术文献与权威评测展开多源检索，对全书引用的市场规模、评测分数、产品规格与政策文号等关键数据逐项溯源、以不少于两个独立来源交叉印证并予以勘误，并对前沿产品与文献严格区分“已核实”与“待核实”。其二，产业测算与文献综合——梳理“赋能教学、支持学习、支持教研、智能评价、治理与安全”五场景下的产品图谱与技术脉络，对市场规模与竞争格局进行多源测算与交叉验证，并据此构建面向五场景的评测框架。其三，文本撰写、图表绘制与引文规范——按统一的论证结构与制图规范完成各章撰写、产品图谱与数据图绘制，并以分级可信度标注维护逐条可溯源的引用体系。其四，中英双语并行——在统一术语表的约束下产出语义对齐的中英文两个版本。

需要郑重说明的是：人工智能在本报告中承担的是检索、测算、起草、制图与引文管理等密集性工作，而研究选题、价值判断、产业研判与最终结论，均由研究团队主导并负责把关；报告所涉数据与引文均要求真实来源、可供复核，市场数字均要求不少于两个独立来源交叉印证，无法核实之处一律标注 [待补] 而不臆造。我们谨以此报告作为一种面向未来的研制工作流参考，供教育产业同仁批评指正——它是一次对知识生产新范式的真诚尝试，而非对专家研判与尽职调查的替代。

目录

第 1 章　引言：从对话式到智能体、多模态与端侧——生成式 AI 教育产品的范式迁移

1.1 从"报告"到"蓝皮书"：一次研究口径的升级

本蓝皮书是 AI-SLI 智慧教育皮书系列在生成式人工智能教育产品方向上的年度旗舰研究。它承接既有的《生成式人工智能产品发展报告》所奠定的"教学—学习—教研"三场景分析传统，但在研究对象、方法工具与结论口径上做了系统性的重构。作出这一重构的判断依据，是过去两年间产品形态的实质位移：当研究口径写定之时所观察的对象，与今日校园中正在部署的对象，已不再是同一类东西。

本报告面向三类读者，并试图同时满足其不同需求：对**教育决策者与学校管理者**，本报告提供一套判断"某类产品是否值得引入、引入时需守住哪些底线"的框架；对**教育研究者与一线教师**，本报告梳理三条技术主线的机理与边界，帮助其理解产品宣称背后的真实能力与局限；对**产业界**，本报告以循证的横向比较，指出范式迁移的真实前沿所在与尚待补齐的短板。三类需求的公约数，是一套**克制、可核验、既讲能力也讲边界**的分析语言——这正是本报告不同于一般产品导购或技术乐观叙事的自我定位。

上一版报告成文之时，生成式 AI 教育产品的主流形态还高度集中于"对话式大模型"这一单一范式：以自然语言对话为交互入口，以单轮或多轮问答为核心能力，以云端通用大模型为底座，通过提示工程与轻量微调适配教育场景。这一范式的开端，可以 2023 年 3 月可汗学院随 GPT-4 同日发布 Khanmigo 为标志——一个以对话形式提供答疑与辅导的 AI 导师，代表了当时产品想象力的上限。这一范式在概念普及、内容生成、答疑辅导等任务上展现了突破性价值，但也在教育这一强场景、强约束、强责任的领域暴露出结构性局限——缺乏可控的任务

编排（问答止于一轮，无法自主完成多步工作流）、缺乏对多模态学习材料的原生理解（看不懂板书、手写与实验演示）、缺乏对师生真实物理情境的感知（不在课堂现场、不知道学生此刻在做什么）、缺乏可审计的过程与责任链条（生成即结束，难以追溯依据与责任）。北京师范大学卢宇、汤筱玮在《电化教育研究》2025 年第 6 期的研究中即指出，现阶段生成式人工智能在课堂教学中的应用"大多停留在浅层次的工具性使用层面……局限于资料检索、教学资源生成等基础性任务，尚未充分挖掘其深层赋能价值"——这一判断，恰是本蓝皮书选择以"范式迁移"而非"产品评述"为主线的学理起点。三条主线的展开，正可视为对上述四项局限的逐一回应：智能体化回应"缺乏任务编排"，多模态化回应"缺乏多模态理解"与"缺乏情境感知"（尤其在与端侧硬件结合时），端侧化回应"缺乏可信、低时延、合规的部署形态"，而三线共同指向"可审计的过程与责任链条"这一治理诉求。

进入 2026 年，产品形态已发生实质迁移。本蓝皮书据此将研究口径由"对话式产品评述"升级为"范式迁移的产业与政策研究"，并做出两项结构性调整：其一，在原有三场景之外**新增"智能评价"与"治理与安全"两个场景**，使分析框架从"如何用 AI 教与学"扩展到"如何用 AI 判断学习成效"与"如何让 AI 在教育中可信可控"；其二，将产品图谱的观察焦点由对话界面转向**智能体（Agent）、多模态与端侧硬件**这三条并行推进的技术主线。这两项调整并非编者的主观取舍，而是对 2024—2026 年间三组可核验事实的回应：教育部基础教育教学指导委员会于 2025 年 5 月 12 日同时发布《中小学人工智能通识教育指南（2025 年版）》与《中小学生成式人工智能使用指南（2025 年版）》，将"辅助教师教学、促进学生成长、推动教育管理智能化"与"保障学生隐私与数据安全"并列写入规范文本；主流教育企业自 2025 年初起密集完成智能体化与端侧化产品迭代；而多模态与端侧小模型的工程条件在同期趋于成熟。三者叠加，构成了"从对话式到智能体、多模态与端侧"这一总判断的经验基础。

1.1.1 为何以"范式迁移"而非"产品迭代"立论

需要澄清的是，本蓝皮书刻意选用"范式迁移"（paradigm shift）而非"产品迭代""版本升级"一类表述，是因为过去两年发生的并非同一类产品在同一能力维度上的连续改良，而是产品的**能力构成、交互形态与责任结构**同时改变。一个仅在对话框内回答问题的系统，与一个能够拆解目标、调用工具、跨会话记忆学情、并在本地设备上离线感知手写与语音的系统，二者之间不是"更好"与"更差"的量的差别，而是"能做什么"的质的差别。用库恩式的语言说，评价前者的标尺（单轮回答的准确性、流畅性）已不足以评价后者；后者需要一整套关于"任务完成度、过程可控性、责任可追溯性"的新标尺。这正是本蓝皮书同时重构"研究对象"（产品图谱）与"评价工具"（评测体系）的根本原因：范式变了，度量范式的尺子也必须一起变。

同时，我们对"迁移"一词保持克制：迁移意味着方向，而非完成。就产品在校园中的真实渗透而言，本蓝皮书所观察到的更接近一幅"新旧并存、深浅不一"的图景——多数在售产品仍以对话式答疑为主体功能，智能体化、多模态化、端侧化的能力常以"卖点"形式叠加于其上，其可靠性与可用性尚待独立验证。因此本报告在肯定方向的同时，始终把"厂商宣称"与"可复现能力"分列两栏，避免以趋势叙事替代事实陈述。

1.1.2 研究对象、时间断面与姊妹卷分工

本蓝皮书的研究对象，界定为"面向教育场景、以生成式人工智能为核心能力的产品与系统"，涵盖软件形态（App、平台、智能体服务）与软硬一体形态（学习机、AI 答疑笔、AI 眼镜、教育机器人等），覆盖 K-12 与高等教育、兼及部分职业与终身学习场景。研究的时间断面主要落在 2024—2026 年，重点观察这一区间内产品形态的结构性位移；对更早的技术脉络仅作必要回溯，对更远的趋势仅作有据前瞻。

在皮书系列内部，本蓝皮书与两部姊妹卷形成明确分工、互不重复：本卷聚焦"生成式 AI 教育产品"这一软硬件谱系的整体图景与五场景分析；《AI 智能眼镜教育产业蓝皮书 2026》就

端侧多模态硬件这一切面做纵深展开，是本章"端侧化"主线在眼镜品类上的专卷；《全球教育机器人发展白皮书 2026》则从具身智能切面展开，是"智能体化+多模态+端侧"三线在机器人形态上的会合。读者可将三卷参照阅读：本卷提供框架与全景，两部姊妹卷提供品类纵深。

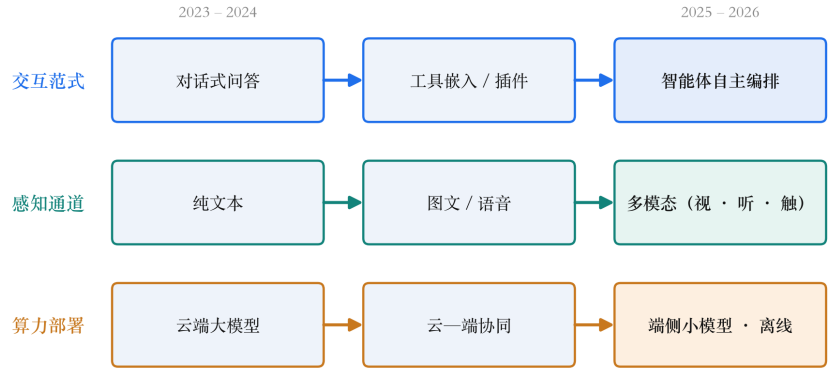
> 本蓝皮书的姊妹卷《AI 智能眼镜教育产业蓝皮书 2026》与《全球教育机器人发展白皮书 2026》分别从端侧多模态硬件与具身智能两个切面，对本章所述范式迁移做了纵深展开，可相互参照，详见本院上述两部报告。

1.2 三条主线：智能体化、多模态化、端侧化

我们将 2024 至 2026 年生成式 AI 教育产品的演进，归纳为三条相互交织、彼此赋能的技术主线。三者并非线性替代关系，而是叠加与融合关系：智能体提供"会做事"的能力骨架，多模态提供"能感知"的输入输出通道，端侧提供"在身边、可信任、低时延"的部署形态。理解这三条主线的关键，在于把握它们的互补性——一个只会对话、看不懂板书、又必须联网上云的系统，与一个能规划任务、原生理解图文声像、且可在本地设备上即时响应的系统，二者在教育场景中的可用性与可信度不可同日而语。

需要说明的是，这三条主线并非在 2024 年凭空出现。教育技术对"类智能体"的追求由来已久：从早期的教学机器、计算机辅助教学（CAI）、智能导师系统（ITS），到自适应学习平台，教育界始终在探索"能感知学生、能自主施教"的技术形态。顾小清、郝祥军将这一脉络概括为"赋能教育的类智能体百年发展历程"。生成式 AI 的真正贡献，不在于发明了"智能体"这一理念，而在于第一次让通用语言与多模态理解能力足够强，使得"规划、对话、感知、生成"能够在同一个模型底座上统一实现，从而把过去需要大量规则工程与领域定制的智能导师，变成了可以相对通用地构建、快速迭代的产品。三条主线，正是这一质变在"能力（智能体）、感知（多模态）、部署（端侧）"三个维度上的投影。

图 1　生成式 AI 教育产品的三条演进主线



来源：本报告分析框架（详见第 1 章）。

1.2.1 智能体化：从"应答"到"办事"

对话式范式的产品逻辑是"用户问、模型答"，交互闭环止于一次生成。智能体范式的产品逻辑则是"用户提出目标、系统自主规划并调用工具完成任务"，交互闭环延伸到真实的教育工作流之中。其技术要件通常包括：任务规划与分解（planning）、工具/插件调用（tool use）、检索增强生成（RAG）、长期记忆与状态维持（memory），以及多智能体协作编排（multi-agent orchestration）。

从机理上看，这五个要件各自解决了对话式范式的一个结构性短板。**规划**把一个笼统目标（"帮我备一节二次函数的课"）分解为可执行的子步骤序列，并在执行中根据中间结果动态调整——这使系统从"一次性生成"转向"多步推进"。**工具调用**让模型能够跳出自身参数、去调用外部能力：查题库、算数值、生成 PPT、写入学习管理系统（LMS），从而把"说"变成"做"。**RAG**（检索增强生成）在生成前先从外部知识库检索相关材料并注入上下文，使模型的回答被约束在教师提供的课程材料与权威知识库之内，而非仅凭参数记忆自由发挥；这一机制对教育尤为关键，因为它把"回答的依据"从模型不可控的内部记忆迁移到教师可控的课程语料——研究者称这一优势为"教师得以掌控大模型所依据的底层文献"。**长期记忆**使系统跨会话保持对同一学生学情的连续认知，把碎片化的一问一答织成一条可追踪的学习轨迹。**多智能体编排**则把不同职能拆分给分工协作的多个智能体，由一个协调者统筹。近两年学界对"智能

体化 RAG"（Agentic RAG）与"AI 智能体时代的记忆"已形成较系统的综述性梳理，为教育智能体的工程实现提供了可复用的方法论骨架。

在教育语境下，智能体化意味着产品可以承接更完整的职能单元——例如自动完成"备课—生成学案—布置作业—批改—生成反馈"的教学链条，而非仅在其中某一环节提供片段式协助。华东师范大学顾小清、郝祥军在《华东师范大学学报（教育科学版）》2025 年第 5 期以"悟空的毫毛"为喻，追溯了"赋能教育的类智能体百年发展历程"，剖析从教学机器到智能学伴的数智化轨迹，并系统梳理了"从独立智能体走向多智能体协同"对学习技术系统的重塑，指出教育智能体正从单点工具演化为可协作的"智能主体"。卢宇、汤筱玓提出的四层框架进一步为这一演化给出了刻度：其"人机协同与创新激活"中级形态，被明确界定为"生成式人工智能成为具有较高自主性的教学智能体，与教师、学生形成多向互动关系"，其"认知融合与思维塑造"高级形态则指向"多元智能体共生的融合赋能"，与本蓝皮书所称"智能体化"在内涵上高度一致。

学术界近两年已涌现出一批以"教师智能体—学生智能体—评价智能体"分工协作为骨架的多智能体教育框架，为这一形态提供了工程参照。例如有研究提出统一的多智能体架构，同时整合"学生层个性化—教育者层自动化—机构层智能"三个层次；也有框架以"教授—学生"关系为原型，设置一个协调者智能体统领"验证者—执行者"团队，分工完成讲义内容的规划、检索、设计与交付；针对作文反馈的框架，则让助教智能体先行评阅、再由教师智能体仲裁，并允许学生对评分提出申诉。这些设计的共性，是把"一个模型独立作答"改写为"多个有分工、可互评、可协商的智能体协同完成一项教育任务"——这正是"从应答到办事"的技术实质。

产品侧最具规模化证据的案例之一是可汗学院（Khan Academy）的 AI 导师 Khanmigo。据其官方披露，Khanmigo 在合作学区中的学生与教师用户，从 2023—24 学年的约 6.8 万人增长到 2024—25 学年的逾 70 万人，合作学区从 45 个扩展到逾 380 个，并于 2025—26 学年目标突破百万；面向教师端，Khanmigo 已完全免费，用于差异化教学、生成教案、出题、学生分

组、评分量规等"事务性"工作。值得注意的是，Khanmigo 的产品设计刻意区别于通用聊天机器人——它"不直接给答案，而是以充分的耐心引导学习者自行求解"，这一设计取向本身即是对"智能体化不等于自动化替代学习"这一教育原则的产品化回应。可汗学院并于 2025 年 10 月至 2026 年 4 月间开展了一系列严格的产品实验，以检验哪些改动能真正提升导师的有效性——这种"以受控实验检验智能体教学效果"的做法，恰是本蓝皮书所倡导的循证态度。

为使"从应答到办事"的差异更为具体，可设想一个典型工作流。在对话式范式下，教师要备一节课，需分别向模型发问："帮我写一份二次函数的教学目标""再给我三道例题""把这段讲解改得更口语化"——每一次都是一轮独立问答，衔接与整合仍由教师手工完成。在智能体范式下，教师只需提出一个目标："帮我准备一节面向初二、时长 40 分钟的二次函数入门课，配 5 道分层练习和一份课堂小测"。系统随即规划出子任务序列，检索教师上传的教材与课标（RAG），据此生成教案、例题与小测，调用工具排版成可下载的课件，并记忆这位教师以往的风格偏好以保持一致。教师的角色，从"逐条指挥"变为"审阅与修订"。这一工作流的价值不在于某一步做得更好，而在于把分散的片段整合成一件可交付的成果——这正是"办事"与"应答"的分野。

产业侧的宏观判断同样清晰。Gartner 于 2025 年预测，到 2028 年将有 33% 的企业级软件应用内置智能体能力（agentic AI），而这一比例在 2024 年尚不足 1%；同时该机构亦发出审慎警示——受成本攀升、业务价值不清与风控不足所累，到 2027 年底可能有超过 40% 的智能体项目被取消。这一"高预期与高失败率并存"的判断，对教育这类高责任场景尤具警示意义：智能体化不是自动的进步，而是一项需要严格评测与治理约束的系统工程。它对产品评测提出了新要求：评价对象从"单轮回答质量"转向"多步任务的完成度、可靠性与可控性"——一个九步流程中每步 95% 可靠的系统，整体成功率仅约 63%，这种"可靠性随步数衰减"的特性，使多步任务的稳健性成为智能体产品的生死线。同时它也对治理提出新问题：当一项教育任务由多个自主智能体协作完成，一旦出错，责任该如何在"教师—产品—模型"乃至"智能体—

智能体"之间归因，成为不可回避的制度设计难题。本蓝皮书把"责任链可追溯"列入治理场景的核心指标，正是对这一难题的正面回应。

1.2.2 多模态化：从"文本对话"到"图文声像的原生理解"

教育天然是多模态的：板书、教材插图、实验演示、语音讲解、手写作业、课堂视频，都是学习信息的载体。对话式范式以纯文本为主，难以直接处理这些材料——它要么把图像交给一个独立的识别模块转成文字再送入语言模型，要么干脆无法处理。以 OpenAI 的 GPT-4o 与 Google 的 Gemini 系列为代表的"全模态/原生多模态"模型，能够在同一架构内对文本、图像、音频乃至视频进行统一的理解与推理，而不再依赖为每一种模态单独拼接的适配器。这一"原生多模态"与"拼接式多模态"的区别并非术语游戏：原生架构能在模态之间保留更完整的语义关联（例如同时听到"这里"并看到手指所指的图上位置），从而支持更接近真实教学的交互。这一工程跃迁使 AI 得以"看懂"一张几何图、"听懂"一段口语表达、"读懂"一份手写答卷。卢宇、汤筱玓亦将"多模态信息理解、知识推理和内容生成"列为生成式人工智能赋能教育的三项核心能力之首。

多模态化不仅拓宽了输入输出通道，更重塑了可被 AI 支持的教育任务边界——口语测评、作业拍照批改、实验过程识别、无障碍学习支持等任务由此具备了产品化基础。以口语测评为例，纯文本模型无法评价发音与语调，而具备语音理解能力的模型可以直接对学生朗读进行音准、流利度与语调的评分；以作业批改为例，多模态模型能够识别手写答卷、定位错误步骤并生成针对性反馈，而不再需要学生手工誊录题目。

产品侧的落地已相当密集，且高度集中于 2025 年。触发点之一是 2025 年初国产开源推理模型 DeepSeek-R1 的发布：网易有道于 2025 年 2 月 6 日宣布接入 DeepSeek-R1，学而思于 2 月 8 日宣布其学习机与智能硬件接入 DeepSeek 并内置"深度思考模式"，作业帮、猿辅导等亦相继跟进，形成一轮行业性的模型底座升级。此后多模态教育硬件密集问世：网易有道于 2025

年发布基于自研"子曰"教育大模型的 AI 答疑笔等硬件，将拍照识别、语音交互与推理答疑整合于一支笔的形态之内；猿辅导集团于 2025 年 4 月 15 日发布首个教育领域 AI 范式"小猿 AI"及配套小猿 AI 学习机，其模型层由自研"猿力大模型"与 DeepSeek-R1 组成矩阵，以"平板为眼、耳，智能基座为手"的伴学机器人形态，承接每日学习计划、学情诊断、知识图谱与全流程辅导。这些产品的共同特征，是把多模态感知（看作业、听朗读）与智能体化流程（诊断—规划—辅导）耦合在同一款终端之中——这也印证了本章开篇的判断：三条主线在真实产品中并非各自独立，而是叠加融合。

多模态化的另一半价值在"输出"侧，常被产品叙事忽略。当模型不仅能理解、还能生成图像、语音与视频，教育内容的生产方式随之改变：一段抽象的物理原理可被即时生成为可视化动画，一份外语课文可被合成为标准发音的听力材料，一道几何题可被配上分步绘制的辅助线示意图。这种"按需生成多模态学习材料"的能力，对个性化与无障碍教育尤具意义——为视障学生把图表转述为结构化语音描述、为听障学生把讲解实时转为字幕、为不同水平的学生生成难度各异的图文讲解，都从依赖人工制作转向可规模化生成。

与此同时，学术界也在为多模态模型的教育适用性建立更审慎的评测标尺——已有研究以 K-12 课堂视频为基准，检验多模态大模型能否真正"看懂"科学教学过程并进行教学法推理，其结论提醒业界：模型"能看图"与"能读懂课堂"之间，仍横亘着可观的能力鸿沟。这一鸿沟对教育产品评测具有直接含义：厂商演示中"识别一张作业照片"的成功，不能等同于"在真实课堂连续视频中理解教学意图"的成功，后者需要更贴近真实场景的独立基准来检验。同样地，多模态生成能力也需警惕——生成一段听起来流利的讲解，不等于生成一段内容正确的讲解，多模态的表现力有时会掩盖内容的错误，反而增加识别幻觉的难度。

1.2.3 端侧化：从"云端集中"到"端云协同"

第三条主线是算力与部署形态的下沉。随着端侧模型压缩、专用推理芯片与小型化多模态模型的发展，一部分推理开始从云端迁移到学习机、平板、AI 眼镜、教育机器人等终端设备上。支撑这一下沉的技术底座，是近两年迅速成熟的小语言模型（Small Language Model, SLM）与量化压缩技术。以微软 Phi 系列、Google Gemma 系列、阿里 Qwen 系列为代表的 SLM，参数规模多在数亿到数十亿之间，配合 4-bit 等低比特量化后仅需数 GB 内存即可在手机、平板与轻量边缘硬件上运行，且在若干文本与多模态任务上逼近甚至局部超越更大规模的云端模型；其中部分小模型已能在单一架构内同时处理语音、视觉与文本，为教育终端的本地化多模态推理提供了现成的模型底座。这一进展的意义在于：它使"离线可用的 AI 辅导"从工程构想变为可量产的产品前提。

端侧化的驱动力在教育场景中尤为突出，可归纳为三条。一是**数据合规与隐私**：未成年人学习数据高度敏感，本地化处理使数据不必上传云端，从源头降低外泄风险——这一诉求已被《中小学生成式人工智能使用指南（2025 年版）》以"严禁师生输入考试试题、个人身份信息等敏感数据""建立健全工具白名单制度""切实保障学生隐私与数据安全"等条款正面回应，构成端侧化在中国教育场景的政策合理性基础。二是**时延与可用性**：课堂随堂反馈、作业实时批改等场景要求毫秒级到秒级的即时响应，且不能因网络波动而中断；端侧推理天然满足这一约束。三是**成本**：教育场景的调用高频且重复，若每一次都走云端 API，长期推理成本可观，而端侧一次性硬件投入可将高频调用的边际成本大幅摊薄。

端侧化并不意味着云端退场，而是走向"端云协同"：轻量、隐私敏感、实时的任务在端侧完成，复杂推理、知识更新与大参数模型能力仍由云端承担，二者通过任务路由动态分工。其分工逻辑大致遵循三条判据：**隐私敏感度**（涉及未成年人身份、学情等敏感数据的处理尽量留在端侧）、**时延要求**（需即时反馈的交互走端侧，可容忍延迟的深度推理走云端）、**任务复杂度**（简单意图识别与常见问答由端侧小模型直接处理，跨学科复杂推理与长文生成上送

云端大模型）。这一分工在真实产品中已成主流架构——前述学习机产品普遍采用"端侧小模型负责即时交互与隐私敏感处理、云端大模型负责深度推理"的混合形态。值得注意的是，端云协同也带来新的评测维度：一款产品在断网状态下还能做什么、哪些功能会退化，成为衡量其"端侧成色"的实测指标，而不能仅凭厂商是否宣称"支持本地大模型"来判断。端侧化的产业热度亦有市场数据佐证：据洛图科技（RUNTO）口径，2024 年中国学习平板全渠道销量约 592.3 万台、同比增长 25.5%，销售额约 190.6 亿元；另据 IDC 口径，2025 年第二季度中国学习平板出货量约 154 万台、同比增长约 44.6%，DeepSeek 接入潮与"国补"政策被普遍视为 2025 年这一轮增长的重要推手。（不同调研机构对同一市场的口径与统计范围存在差异，本报告在产业章节统一交叉核验后引用，此处数据仅用于说明端侧化的产业热度。）

这一形态在硬件产品线上还有更前沿的投影。作为本蓝皮书关联机构的网龙网络（NetDragon，港交所代码 HK:0777），自 2023 年起战略投资 AI 眼镜与 AR 交互企业 Rokid（灵伴科技），围绕教育与交互场景布局端侧多模态硬件；据公开报道，网龙对 Rokid 的战略投资额自 2000 万美元起，后续有增资安排。Rokid 的一体式 AI+AR 眼镜 Rokid Glasses 在 CES 2025 上获颁 Best of CES 2025 奖项，并于 2025 年内推进量产上市。此类"端侧感知—端云协同推理"的硬件形态，正是本蓝皮书姊妹卷《AI 智能眼镜教育产业蓝皮书 2026》的纵深研究对象，本章不再展开，仅在此点明其在三条主线中的坐标：端侧化为智能体与多模态提供了"贴身、可信、低时延"的落地载体，把 AI 从"云端的对话框"带回"学习者身边的设备"。（网龙相关投资额、Rokid 产品上市节奏等具体数据以产业章节交叉核验后的口径为准。）

1.2.4 三线交汇：融合而非替代

必须再次强调，智能体化、多模态化、端侧化并非三条各自独立的技术曲线，而是在真实产品中彼此咬合、相互赋能的整体。缺少任何一条，另外两条的价值都会打折：一个能规划任务的智能体，若看不懂学生的手写作业（缺多模态），其"办事"能力就止于文字层面；一个原生多模态模型，若必须把每一张作业照片上传云端处理（缺端侧），就同时背上了时延与

未成年人数据合规的双重负担；一个跑在本地设备上的小模型，若只能一问一答、不会调用工具与记忆学情（缺智能体化），就退回到了对话式范式的老路。

把三者叠加起来看，一个理想形态是：**端侧多模态智能体**——它在学习者身边的设备上（端侧）原生地看、听、读学习材料（多模态），并自主规划、调用工具、连续记忆地完成一项完整的教育任务（智能体）。这一形态在 2026 年尚未成为主流，但学习机、AI 答疑笔、AI 眼镜等产品线已在朝此方向收敛。本蓝皮书据此把"三条主线的交汇程度"作为观察产品成熟度的一把标尺：越是把三者深度融合、而非把某一条作为营销标签简单叠加的产品，越接近范式迁移的真实前沿。

1.3 五场景框架：本蓝皮书的分析主轴

基于上述三条主线，本蓝皮书将生成式 AI 教育产品置于**五个场景**之下加以考察。前三个场景承接既有传统并予以更新，后两个场景为 2026 版新增。这一框架并非凭空设定：教育部基础教育教学指导委员会 2025 年发布的《中小学生成式人工智能使用指南（2025 年版）》所归纳的"辅助教师教学、促进学生成长、推动教育管理智能化"三类应用场景，与本框架的"支持教学、支持学习、支持教研/治理"存在显著呼应；而卢宇、汤筱玮"教、学、评、辅"多元场景赋能的表述，则为"智能评价"独立成场提供了学理依据。

为何是"五个"而非沿用"教—学—评—辅"或"教—学—研"等既有划分？本报告的考量有三。其一，**把"评"从"教"中独立出来**：在对话式范式下，评价往往被视为教学的附属环节（出题、批改），可并入"支持教学"；但在生成式范式下，AI 开始独立承担评分、诊断与反馈，其可靠性与公平性构成一组自成体系的技术与伦理问题，必须单列考察。其二，**新增"治理与安全"作为横切场景**：治理并非与教、学、评、研并列的又一类"应用"，而是贯穿所有应用的约束层；把它显式列为一个场景，是为了强制每一类产品分析都回答"它安全吗、合规吗、责任可追溯吗"，而非把治理当作可选的附录。其三，**保留"支持教研"以承接机构记忆**：教研是中国

基础教育的独特建制，也是教师专业发展的主渠道，生成式 AI 对教研的赋能不宜被"教"或"学"吸收，故独立保留。五场景由此构成一个"三纵（教、学、研）+ 两横（评、治）"的分析网格，既承接传统、又回应新变。

场景	核心问题	相较旧版的更新重点	代表产品形态
支持教学	AI 如何帮助教师教	从备课/出题助手扩展到教学智能体与课堂多模态感知	教学智能体、备课/出题助手、课堂分析工具（详见后续各章产品图谱）[待补：入选产品名单与筛选口径]
支持学习	AI 如何帮助学生学	从对话答疑扩展到端侧学伴、多模态辅导与个性化路径	AI 学习机、AI 答疑笔、AI 学伴 App（如小猿 AI、有道相关产品；详见后续各章）[待补：完整名单]
支持教研	AI 如何支持教师专业发展与教研	从资源生成扩展到教研智能体与循证分析	教研智能体、循证教研平台（详见后续各章）[待补：产品名单]
智能评价（新增）	AI 如何评判学习与教学成效	过程性评价、多模态测评、生成式题目与自动批改的可靠性	口语测评、作业批改、自动组卷与评分系统（详见后续各章）[待补：产品名单]
治理与安全（新增）	AI 在教育中如何可信可控	内容安全、未成年人保护、数据合规、可解释与责任链	工具白名单、内容安全过滤、合规审计机制（如《使用指南》所要求）[待补：制度/产品清单]

前三个承接性场景在本版中并非原样保留，而是随三条主线一并升级。**支持教学**从早期的"备课/出题助手"扩展到能够感知课堂、编排教学流程的教学智能体，其评价重心也从"生成内容是否可用"转向"是否真正减轻教师负担并改善教学"——如前述以多智能体降低教师工作量的研究所示，减负本身已成为可量化的产品目标。**支持学习**从对话式答疑扩展到端侧学伴与个性化学习路径，Khanmigo"引导而非直接给答案"的设计、以及国产学习机"学情诊断—知识图谱—每日一练"的闭环，都是这一升级的产品化表达；与此同时，学界关于"过度自动化可能削弱学生元认知参与与自我调节学习"的警示，提醒这一场景必须警惕"帮太多反而学更少"的悖论。**支持教研**从零散的资源生成扩展到教研智能体与循证分析，使 AI 从"给老师做素材"走向"帮老师做研究"。

新增"智能评价"场景，回应的是生成式 AI 从"辅助生成"走向"辅助判断"的能力跃迁——当 AI 开始参与评分、诊断与反馈，其可靠性、公平性与偏差控制便从技术问题上升为教育问题。这一跃迁的风险不容低估：已有系统综述指出，"技术可靠性与幻觉"是生成式 AI 教育应用中被讨论最多的挑战，其后依次为过度依赖、评价演进与公平性、隐私与安全；针对 LLM 自动批改的安全性研究更揭示，批改智能体在鲁棒性与可信度上仍存在可被利用的脆弱面——例如答卷中的特定诱导性文本可能操纵评分结果。当一次评分可能影响学生的升学与自我认知，"评分模型会不会被操纵、会不会系统性偏向某类学生"就不再是技术细节，而是教育公平问题。

正因如此，智能评价的产品形态不宜设计为"AI 独立裁决"，而应走向"人机协同评价"：AI 承担初评、标注可疑之处、给出证据与置信度，最终裁量权保留在教师手中。前述让教师智能体对助教智能体的评分进行仲裁、并允许学生申诉的多智能体作文反馈框架，正是这一取向的一种技术实现。这也意味着，智能评价场景的评测重点，不仅在于 AI 评分与人工评分的一致性（如相关系数、评分差），更在于其错误的**可发现性与可纠正性**——一个偶尔出错但错误易被教师识别与推翻的系统，可能比一个准确率略高却难以质疑的"黑箱评分器"更适合进

入课堂。这些证据表明，把"评价"单列为一个需要独立评测与约束的场景，并非概念上的求全，而是回应真实风险的必要之举。

新增"治理与安全"场景，回应的则是产品大规模进入真实校园后不可回避的责任议题：面向未成年用户的内容安全、学习数据的合规使用、模型幻觉与错误信息的防护，以及人机责任边界的界定。这一场景的紧迫性在于其对象的特殊性——面向未成年人的 AI 系统面临成人产品所没有的额外风险，包括儿童特有的内容安全脆弱性、以及"AI 以权威口吻呈现的错误信息更易被学生当作真理"的认知风险。可喜的是，治理正从抽象倡议进入可测量、可工程化的阶段：针对未成年人内容风险的专用基准（如 MinorBench）、面向教育 LLM 统一"安全—有用—教学性"三者的评估框架相继出现，使"安全"从口号变为可打分、可比较的指标。在政策层面，《中小学生成式人工智能使用指南（2025 年版）》确立的五大应用原则（育人导向、教育公平、价值引领、需求驱动、底线思维）与分学段使用规范（小学阶段在教师家长帮助下使用开放式内容生成、初中阶段指导交叉验证生成内容的合理性、高中阶段结合技术原理自主评估生成内容的社会影响），则为这一场景提供了本土化的制度锚点。这套"越低龄、越受约束"的分学段设计，本身即是一种可被产品化的治理机制——理想的教育 AI 产品应当能够识别用户学段并据此调整其自主度与内容开放度，而非对所有用户一视同仁。

综合来看，五场景构成一条从"造能力"到"验成效"再到"控风险"的完整链路：支持教学、支持学习、支持教研解决"AI 能为教育做什么"，智能评价解决"AI 做得到底好不好、判得准不准"，治理与安全解决"AI 做这些事时是否安全、合规、可追责"。三者缺一不可：只有能力而无评价，产品的宣称无从检验；只有能力与评价而无治理，风险无从约束。这两个新增场景使本蓝皮书的框架从"能力图谱"延伸为"能力—评价—治理"的完整链路——只讲能力而不讲如何评判其成效、如何约束其风险的产品叙事，是不完整、也不负责任的。

1.4 研究方法与循证原则

作为一部定位于产业与政策研究的旗舰蓝皮书，本报告在方法上坚持"研究先行、循证优先、保守准确"。这一方法论姿态，直接源于本报告所观察对象的两个特点：其一，生成式 AI 教育产品迭代极快，任何静态的"能力清单"都可能在成书之时已经过期，因此本报告更重视提供一套可持续观察的**分析框架**，而非一份易朽的产品排名；其二，教育是强责任领域，任何被高估的能力宣称，代价可能由学生承担，因此本报告对事实的态度是"宁缺毋滥、宁留白毋臆造"。具体体现为四个方面。

- **产品图谱化**：以场景为轴、以能力为维，对市场在售的生成式 AI 教育产品进行结构化梳理与横向比较，力求呈现范式迁移的真实分布，而非个案罗列。图谱化的价值在于，它能显示"三条主线的能力在不同产品间如何分布"——哪些产品只做了对话式答疑，哪些叠加了多模态，哪些真正实现了端侧智能体化，从而把笼统的"范式迁移"落成一张可核对的结构表。入选口径以"面向 K-12 或高等教育、具备生成式能力、2024—2026 年间在售或公开发布"为基本门槛。[待补：最终入选产品数量与逐项筛选口径]
- **能力评测化**：引入面向教育垂类的大模型评测、AI 硬件评测（如学习机、AI 眼镜）与产品横评，以能力雷达、时间线等可视化方式呈现，避免以厂商宣称替代独立观察。我们尤为重视区分"厂商宣称能力"与"第三方可复现能力"——如前文所引 K-12 课堂视频基准所示，二者可能存在系统性落差；同时也重视区分"通用基准得分"与"教育场景可用性"，一个在通用问答基准上得分高的模型，未必能在真实课堂中稳定、安全、合乎教学法地工作。[待补：本报告采用的评测维度、样本量与数据截止时点]
- **新范式解剖**：对智能体编排、RAG、长期记忆等支撑范式做机理层面的拆解，说明其在教育场景中的适配条件与失效边界。例如 RAG 的教育价值在于"把回答约束在教师可控的课程材料之内"，其失效边界则在于检索质量差或知识库过期时反而放大错误——一个

检索到错误资料的 RAG 系统，会以更自信的口吻给出更难被察觉的错误；长期记忆提升了个性化连续性，却也同步放大了未成年人数据的合规风险，被记住的每一条学情，都是一条需要被保护的隐私。多智能体编排提升了复杂任务的完成度，却使"出错时如何归因"变得更加困难。机理与边界并陈，是本报告避免技术乐观主义的方法自觉：本报告讲每一项能力，都尽量同时讲清它在什么条件下会失效。

- **来源可核验**：本报告严格区分"已核事实""趋势判断"与"分析推断"三类陈述，并在行文与页脚予以区隔。凡涉及市场规模、出货量、份额、用户数、具体政策文号与文献引用等事实性陈述，均以真实公开来源为据并在页脚标注；证据不足者宁留占位、不作臆断。就市场规模而言，第三方研究机构对"AI+教育"全球及中国市场未来规模的测算区间较大（不同口径、不同边界下差异显著），本报告在正文中不采用单一未经交叉核验的数字，相关口径统一于产业章节交叉比对后处理；涉及币种的数据严格区分人民币、美元与港元，绝不混算。[待补：本报告采用的市场数据来源、口径、币种与截止时点]

在数据边界上，本报告还坚持一条"关联企业审慎"原则：对作为本机构关联企业的网龙网络（HK:0777）及其相关产品与投资，本报告采用与其他企业完全一致的循证标准，凡具体数据均以公开可核验来源为准、不因关联关系而放宽或收紧，以维护研究的独立性与可信度。

需要特别说明的是，本蓝皮书由 AI-SLI 研制，编写过程本身应用了人工智能辅助工具。这既是本报告研究对象的一种"以身试法"式印证，也带来了额外的自我约束义务：为保持研究的严谨与可信，我们对 AI 生成的每一处事实性内容均执行了来源回溯与人工复核——凡无法回溯至可信一手来源的具体数据、产品名、公司名、基准分数与政策文号，一律不写入正文或以占位标注。这一做法本身，正是我们在"治理与安全"场景中所倡导的"人机责任边界界定"原则的自我实践，相关说明见本报告"编写说明"页。

1.5 本蓝皮书的结构安排

全书围绕"范式—场景—评测—治理—前瞻"的逻辑展开，形成一条从总判断到具体分析、再回到政策建议的闭合链路。本章（第 1 章）确立范式迁移的总判断与五场景分析框架，是全书的坐标系。其后各章大致依三个层次推进：**第一层是产业与产品的整体图谱**，对生成式 AI 教育产品做结构化分类与横向比较，回答"市场上究竟有哪些产品、它们在三条主线上分布如何"；**第二层是五场景的分场景纵深**，依次就支持教学、支持学习、支持教研、智能评价、治理与安全五个场景，拆解其机理、盘点其代表产品、指出其适配条件与失效边界；**第三层是评测与前瞻**，引入教育垂类大模型评测与 AI 硬件评测，以独立观察校验厂商宣称，并在此基础上给出面向政策制定者、教育机构与产业界的发展建议与趋势前瞻。三个层次层层递进，共同支撑本报告"能力—评价—治理"一体的分析主张。[待补：各章章名、章序与页码的最终对应]

我们希望，本蓝皮书不仅是一份产品清单式的观察，更是一套帮助教育者、决策者与产业界理解"生成式 AI 教育产品正在成为什么"的分析框架。范式迁移仍在进行之中——正如 Gartner 关于智能体项目高失败率的警示、以及学界关于幻觉与评价公平性的持续关切所提示的，这一进程既非线性、也不自动向善：智能体化可能带来"多步任务的可靠性坍塌"，多模态化可能带来"看得到却读不懂"的能力错觉，端侧化可能在便利与合规之间制造新的张力。承认这些张力，恰恰是负责任地推进这场迁移的前提。本报告所呈现的，是这一进程在 2026 年这一时间断面上的结构性图景，以及一套用以持续观察它的循证工具；我们期待它成为一个可被检验、可被修正、可被逐年更新的观察起点，而非一锤定音的结论。

本章参考来源

1. 卢宇、汤筱玮．《生成式人工智能赋能课堂教学的形态层级与进阶路径》[J]．电化教育研究，2025，46(6)：75-82+106．北京师范大学教育技术学

院 . DOI:10.13811/j.cnki.eer.2025.06.010 . （原文 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 ；亦见 https://aic-fe.bnu.edu.cn/docs/2025-07/a4d9baa90d9e49d9920ee2c46dd283b7.pdf ）

2. 顾小清、郝祥军. 《悟空的毫毛：正在重塑学习技术系统的多智能体》[J]. 华东师范大学学报（教育科学版），2025 年第 5 期. 华东师范大学教育学部. URL: https://ercdee.ecnu.edu.cn/b8/9d/c16571a702621/page.htm
3. 教育部基础教育教学指导委员会. 《中小学生成式人工智能使用指南（2025 年版）》与《中小学人工智能通识教育指南（2025 年版）》. 2025-05-12 发布. （全文见北京日报客户端 https://xinwen.bjd.com.cn/content/s68217324e4b0ec1c3d96f32d.html ；解读见中国教育和科研计算机网 CERNET https://www.edu.cn/xxh/focus/zc/202505/t20250513_2667992.shtml ）
4. 猿辅导集团. "小猿 AI 暨智能硬件战略发布会"（发布小猿 AI、小猿 AI 学习机，底座为猿力大模型 +DeepSeek-R1 ），2025-04-15 . 中国日报网报道：https://tech.chinadaily.com.cn/a/202504/17/WS680071c2a310e29a7c4a9b3f.html （另见新浪科技 https://finance.sina.com.cn/tech/roll/2025-04-15/doc-inetfxpu4284697.shtml ）
5. 多知网. 《网易有道推出新款教育智能硬件（有道 AI 答疑笔 Space X 等，基于"子曰"教育大模型）》，2025. URL: http://www.duozhi.com/industry/insight/2025022317027.shtml
6. 量子位 / 北京日报. 在线教育企业接入 DeepSeek-R1 时间线（网易有道 2025-02-06、学而思 2025-02-08 等 ） . URL: https://www.qbitai.com/2025/02/254011.html ；https://xinwen.bjd.com.cn/content/s67b5d6bae4b068c68f1001dd.html
7. 动点科技. 《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

获 Best of CES 2025、2025 年内量产上市，见证券时报 https://www.stcn.com/article/detail/1579878.html）

8. Gartner．"Agentic AI to feature in 33% of enterprise software applications by 2028"（及"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 导师用户规模与形态（2024-25 学年合作学区学生/教师用户增至逾 70 万，教师端免费）．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（教育 RAG 导师应用见 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. 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

第 2 章　产品图谱总览（2025–2026）：赛道、形态与竞品坐标

2.1　为什么需要一张"图谱"：从对话式产品到智能体化生态

2024 年前后，生成式人工智能教育产品的主流形态是"对话式大模型 + 教育场景包装"：一个通用或微调后的语言模型，套上学科提示词、教材知识库与前端对话界面，即构成一款"AI 学伴""AI 助教"或"智能批改"产品。这一代产品的能力边界，基本等同于底层模型的对话能力边界，产品之间的差异更多体现在内容运营与界面体验，而非底层范式。以 OpenAI 在 2024 年 5 月推出、面向高校的 ChatGPT Edu 为例，其定位仍是"把企业版 ChatGPT 以合规、可管控的方式交付给校园"，底座为 GPT-4o，主打跨文本与视觉推理、支持 50 余种语言、可基于院校自有数据构建自定义版本，并承诺对话数据不用于训练——这是典型的"通用模型 + 教育合规封装"路线，牛津大学、宾大沃顿商学院、得州大学奥斯汀分校、亚利桑那州立大学、哥伦比亚大学等院校在企业版阶段的先行实践直接催生了该产品。

进入 2025—2026 年，三条技术主线的叠加，使产品形态发生了结构性迁移，也使"用一张静态清单罗列产品"的旧方法失效：

- **智能体化（Agentic）**：产品从"一问一答"转向"可规划、可调用工具、可多步执行"的智能体。这一转折的标志性事件是 Google 于 2024 年 12 月发布 Gemini 2.0，官方明确将其定位为"面向智能体时代（the agentic era）"的模型，原生支持工具调用（Google 搜索、代码执行、第三方自定义函数）与多模态输入输出。教学、学习、教研任务被拆解为可编排的工作流，产品价值从"回答得准"转向"任务能闭环"。

- **多模态（Multimodal）**：文本之外，语音、图像、板书、屏幕、实物、课堂视频成为一等输入。GPT-4o 的原生"音频到音频"交互（可捕捉语气中的犹疑、急促、反讽），以及 Gemini Live API、GPT Realtime 类接口把首个音频块延迟压到约 300–600 毫秒区间，使产品开始"看得见课堂、听得懂讨论"；评价与反馈从文本作业扩展到过程性、情境化的多模态证据。
- **端侧化（On-device）**：算力从云端下沉到学习机、平板、AI 眼镜等终端，带来低时延、可离线、数据本地化等新特性，也重塑了"硬件—模型—服务"的价值分配。2024—2025 年中国 AI 学习机市场的爆发（详见 2.5 节）即是端侧化在教育消费市场的直接投影。

这三条主线并非彼此独立，而是相互强化：智能体化提出了"多步任务需要跨模态感知与工具"的需求，多模态提供了"看懂课堂、听懂讨论"的输入通道，端侧化则回答了"感知发生在真实物理场景、数据不宜全量上云"的部署约束。三者叠加的净效应，是把产品的价值评价标准从"单轮回答的正确率"整体上移到"一段真实教育任务能否被稳定、可控、合规地完成"。这也解释了为何 2024 年那种"一张 Excel 罗列几十款对话式产品"的静态清单会迅速失效——当产品在"范式—模态—部署"三个维度同时移动时，任何一维的清单都无法刻画产品之间真正的竞争关系。

因此，本章不再以"产品名录"为骨架，而是构建一张二维以上的**竞品坐标（competitive landscape）**：横轴刻画产品的**能力范式**（对话式 → 智能体编排式），纵轴刻画产品的**形态与部署方式**（纯云软件 / 平台中台 / 端侧硬件），并叠加**教育场景归属**（教学 / 学习 / 教研 / 评价 / 治理）这一维度。这一坐标既用于定位单个产品，也用于观察整条赛道的迁移方向。需要强调的是，坐标是"分析工具"而非"排行榜"：同一象限内的产品并不构成简单的优劣序列，其真实竞争力取决于场景适配度、合规成熟度与数据闭环质量，这些将在第 4—6 章以评测数据逐一展开。

关于端侧硬件形态的深度拆解（尤其 AI 眼镜与学习机的产业链、市场测算与竞品格局），详见本院《2026 AI 智能眼镜教育产业蓝皮书》与《全球教育机器人发展白皮书 2026》；本章仅在总览层面纳入其坐标位置，不重复展开。

2.2 赛道划分：五大场景 × 三类形态

本蓝皮书在旧版"教学—学习—教研"三场景基础上，新增"智能评价""治理与安全"两场景，形成**五大应用场景**；与之正交的是产品的**三类形态**。二者构成本章的基本分析网格。这一"五场景"划分并非本院独创的分类臆想，而与产业与政策口径高度吻合：2025 年 9 月华为与科大讯飞联合发布的"星火教育、医疗大模型场景一体机解决方案"，即把教育侧能力归纳为**助教、助学、助研、助管、助交流**五类场景；2026 年 4 月教育部等五部门印发的《"人工智能+教育"行动计划》，在"应用融合"一节明确提出"赋能学生学习、赋能教师教学、赋能学校治理、赋能科学研究"。本蓝皮书的五场景与之基本对齐，并把"评价"从教学中独立出来单列，以回应生成式 AI 承担评价职能这一核心争议。

2.2.1 五大应用场景

场景	核心诉求	2025–2026 典型能力迁移	代表能力关键词
教学（Teaching）	减负备课、生成资源、课堂辅助	从"生成教案/课件"到"课堂协同智能体"	备课智能体、板书理解、AI 助教
学习（Learning）	个性化辅导、答疑、练习	从"对话答疑"到"苏格拉底式引导 + 长期记忆学伴"	引导式学习、错题归因、记忆化学伴
教研（Research/PD）	资源沉淀、教师发展、数据洞察	从"素材检索"到"教研 RAG 知识库 + 听评课分析"	教研 RAG、课例分析、教师发展
智能评价（Assessment）※新增	过程性评价、多模态证据、能力画像	从"客观题批改"到"多模态过程性评价"	主观题评分、过程画像、Rubric 对齐

治理与安全（Governance）※新增	合规、内容安全、数据与伦理	从"内容过滤"到"教育垂类安全对齐与审计"	内容安全、数据合规、分学段使用规范

后两个场景是 2026 版图谱相对旧版《生成式人工智能产品发展报告》的关键增补：**智能评价**回应"生成式 AI 能否可信地承担评价职能"这一核心争议（研究证据见 2.4 节与第 5 章）；**治理与安全**回应产品规模化落地后凸显的合规、内容安全与伦理风险——中国《中小学生成式人工智能使用指南（2025 年版）》以"应用为王、治理为基"为总纲、按学段划定使用红线，UNESCO《生成式 AI 教育与研究指南》（2023）建议将平台独立使用年龄下限设为 13 岁，均说明"治理"已从产品的外部约束变为产品设计的一等约束。二者的评测方法与详细产品分析分别见本蓝皮书第 5 章、第 6 章。

2.2.2 三类产品形态

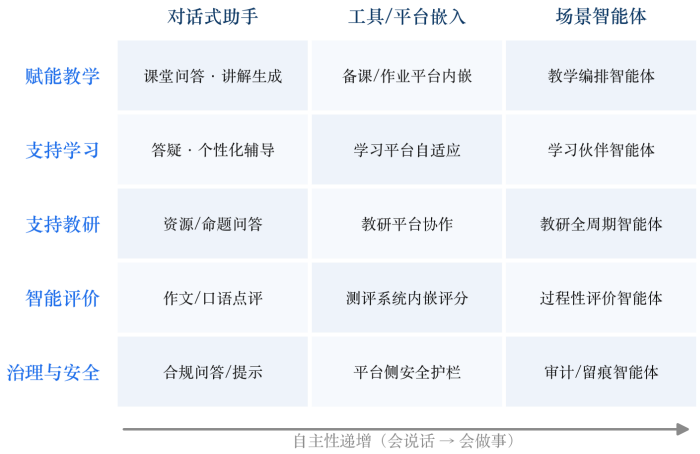
- **软件应用层（App）**：面向师生的终端应用与网页/小程序，直接承载对话、答疑、批改等交互。此层产品数量最多、迭代最快，但同质化程度也最高。代表如 OpenAI ChatGPT Edu、Google Gemini for Education、Anthropic Claude for Education、字节"豆包爱学"等。
- **平台/中台层（Platform / Agent Infra）**：为区域、学校或产品团队提供模型接入、智能体编排、RAG、记忆、评测与安全治理的底层能力。2025—2026 年，该层从"模型 API 转售"演化为"智能体运行时 + 教育知识底座"，是本轮产品竞争的价值高地。代表如科大讯飞星火教育大模型、华为—讯飞"场景一体机"、阿里通义千问教育垂类模型（Qwen3-Learning）等。
- **端侧硬件层（Device）**：学习机、平板、AI 眼镜、课堂交互终端等承载端侧模型的硬件。此层将模型能力"实体化"进入真实教学场景，是多模态与端侧化趋势的直接载体。代表如科大讯飞 AI 学习机、作业帮/步步高学习平板、讯飞 AI 黑板、Rokid Glasses（网龙战略投资）等。

三类形态之间存在清晰的价值传导与锁定关系：平台/中台层向下为 App 层与硬件层提供模型、编排、记忆、评测与安全能力，一旦某中台成为区域或学校的"教育知识底座"，就对上层应用形成较强的迁移成本与生态锁定；App 层贴近师生、迭代最快，但因底座可替换而议价能力有限、同质化严重；硬件层则把能力"实体化"进物理课堂，其壁垒来自供应链、渠道（如线下体验与国补政策叠加带来的销量弹性）与端云协同的工程能力。理解这一传导关系，是判断"哪一层在攫取价值"的前提——2.5 节将用"技术能力价值"增速快于"产品服务总盘"的数据（中国口径 45% vs 37%）从商业结构上印证：价值正在向中台层集中。

将五大场景与三类形态交叉，即得本章的**产品图谱主网格**。下表所填代表产品，均来自本章检索到的公开来源，用于示意"格子"的填充方式；完整的产品横评见第 6 章。

形态＼场景	教学	学习	教研	智能评价	治理与安全
软件应用层	ChatGPT Edu；Gemini in Classroom	Claude for Education（Learning Mode）；豆包爱学	Khanmigo 教师端（教案/分组/进度）	LLM 主观题评分工具（研究阶段为主）	各学段"使用指南"合规模块 [待补：成熟商用产品]
平台/中台层	讯飞星火"助教"；华为—讯飞场景一体机	通义千问 Qwen3-Learning	教研 RAG 知识底座 [待补：具名平台]	校本评价引擎 [待补：具名平台]	教育垂类安全对齐/审计中台 [待补：具名平台]
端侧硬件层	讯飞 AI 黑板 2.0	讯飞/作业帮/步步高 AI 学习机	[待补：教研专用终端]	[待补：过程性评价终端]	Rokid Glasses 等可穿戴（合规与隐私为核心变量）

图 2 五场景 × 三形态：生成式 AI 教育产品图谱框架



来源：本报告分析框架：单元格示例产品经检索核实（详见第 2－7 章）。

2.3 形态迁移的四条驱动线索

在给出竞品坐标之前，有必要说明驱动产品形态迁移的四条底层线索。它们既解释"为什么产品在向右上方（智能体化 + 端侧化）迁移"，也构成后续章节评测指标的设计依据。

1. **从"模型能力"到"任务闭环"**：竞争焦点从单点问答质量，转向能否稳定完成一个完整教育任务（如"从课标到分层教案再到课堂实施建议"）。这要求产品具备规划（把任务拆成有序子步骤）、工具调用（检索教材、生成图表、写入学情库）与错误恢复（发现前一步偏差后回退重做）三类能力，而这恰是"对话式"范式所缺失的。Khanmigo 教师端一次性生成教案、评分量规、分组策略与学情摘要，即是"任务闭环"而非"单点问答"的典型；Gemini 2.0"面向智能体时代"的官方定位、以及其原生支持 Google 搜索与代码执行的工具链，本质上是把这种闭环能力从"应用层的编排技巧"下沉为"模型层的原生能力"。对产品方而言，这意味着竞争的护城河从"提示词工程"上移到"工作流设计与任务成功率的稳定性"，后者需要长期的真实场景数据打磨，门槛更高。

2. **从"通用大模型"到"教育垂类底座"**：通用模型难以同时满足学科知识准确性、教学法适配、未成年人安全三项刚性约束，教育垂类大模型与领域知识库（RAG）因此成为差异化关键。

Google 的 LearnLM 系列基于 Gemini、依据学习科学微调，围绕"主动练习与反馈、认知负荷管理、个性化、激发好奇、元认知"五条学习科学原则设计，2025 年 I/O 已将其能力并入 Gemini 2.5，官方并援引对照实验称：接受 LearnLM 短时辅导的学生，在后续新题上的解题概率较仅由人类导师辅导的对照组高出约 5.5 个百分点。阿里于 2025 年 12 月 4 日发布 Qwen3-Learning 教育垂直大模型（该发布信息据艾瑞 2026 报告转述），是中国侧同一逻辑的落子。垂类化的价值不仅在"答得更准"，更在于把教学法（如苏格拉底式引导而非直接给答案）与安全边界（如分学段的内容红线）固化进模型行为，从而降低下游产品的对齐成本。其评测方法见本蓝皮书第 4 章。

3. **从"无状态对话"到"长期记忆"**：学习是长周期过程，产品从单次会话转向跨会话、跨学期的学习者记忆与画像，带来个性化红利，也带来数据合规与伦理挑战。艾瑞《2026 年中国 GenAI+教育行业发展报告》调查显示，家长辅导中"37.1% 固定使用一两款、35.6% 多款混合使用"，成年学习者"40.9% 混合使用多款 AI 工具"——这种碎片化、跨产品的使用现状，恰恰凸显"跨会话、跨产品的统一学习者记忆"仍是未被满足的刚需；但记忆能力同时把"未成年人画像的采集、存储与再利用是否合规"这一治理问题从边缘推到中心，这也是本版把"治理与安全"独立成场景的现实动因之一。

4. **从"云端纯软件"到"端云协同硬件"**：端侧化把部分推理与感知放到终端，换取时延、隐私与情境感知三重优势——低时延满足课堂实时互动，本地化处理缓解未成年人数据外流的合规压力，第一视角/近场感知则提供云端拿不到的情境信号。硬件因此重新成为教育 AI 竞争的关键变量。这一线索在中国消费市场表现得尤为明显：2024 年中国 AI 学习机全渠道销量 592.3 万台、同比增长 25.5%，销售额 190.6 亿元、同比增长 37.6%（详见 2.5 节）。值得注意的是，端侧化并不意味着"完全离线"，主流形态是"端云协同"——端侧负责低时延感知与隐私敏感处理，云端负责重推理与知识更新，二者的算力与数据分工，正是硬件层差异化的核心，详细产业拆解见本院 AI 眼镜与学习机相关评测。

这四条线索共同指向同一个方向：产品在坐标图上向"右上方"（智能体化 + 端侧/多模态）迁移。它们也构成本蓝皮书后续评测指标的设计依据——第 4 章的教育垂类模型评测对应线索 1、2，第 5 章的评价与治理评测对应线索 3 中的合规议题，第 6 章的硬件与产品横评对应线索 4。

2.4 竞品坐标：两轴定位与聚类

综合上述框架，本章以**能力范式（对话式 ↔ 智能体编排式）为横轴、部署形态（纯云软件 ↔ 端侧硬件）为纵轴**建立竞品坐标，将主要产品聚为若干簇。选择这两轴而非"厂商规模""价格"等常见维度，是因为它们恰好对应 2.3 节四条驱动线索中最具结构性的两条（任务闭环、端云协同），因而能最灵敏地反映赛道迁移方向。坐标中每一个具体产品的归位与坐标值均需真实数据支撑，此处给出聚类结构与已核实的代表玩家，精确坐标值（气泡大小对应的规模口径）见后续评测章节。

需要说明产品在坐标中的"移动性"：同一款产品可能同时具有多个形态版本（如 Khanmigo 既有面向学习者的对话学伴、又有面向教师的智能体化教研工具），因此其在坐标上并非一个点而是一段轨迹。以 Khanmigo 为例，其规模化速度是 A→B 迁移中"学习+教研双轮"的一个可核实注脚：据 Khan Academy 公开信息，2024–25 学年其 K-12 学生用户显著放量，并预期在 2025–26 学年迈向百万量级、进入更多美国学区并延伸至印度、巴西、菲律宾等国的课堂；课堂访问须经由学校或学区统一部署开通。这类"经由机构统一部署"的分发方式，本身就把治理（象限 E）内生进了产品的落地路径。

> 关于 Khanmigo 用户规模，坊间流传多种口径（如"从数万到百万级""795 个学区"等），不同来源的基准与时点不一。本蓝皮书仅采用可回溯到 Khan Academy 官方渠道、口径与时点明确的表述，其余高增长数字标注为 [待补：需核实官方口径] 而不直接引用，以免以讹传讹。

聚类（象限）	范式定位	部署定位	典型玩家（已核实公	竞争要点

			开来源）	
A. 云端对话式学伴/助教	对话为主	纯云软件	ChatGPT Edu；Gemini for Education；豆包爱学	内容质量、学科覆盖、合规封装
B. 智能体化教研/教学平台	智能体编排	云 + 中台	讯飞星火教育大模型；华为—讯飞场景一体机；Khanmigo 教师端	编排能力、RAG/记忆、安全治理
C. 端侧学习硬件（学习机等）	对话→轻智能体	端云协同	讯飞/作业帮/步步高 AI 学习机；讯飞 AI 黑板	端侧模型、离线能力、亲子/护眼合规
D. 多模态可穿戴（AI 眼镜等）	多模态智能体	端侧为主	Rokid Glasses（网龙 HK:0777 战略投资）	第一视角感知、时延、隐私与伦理
E. 垂类评价/治理引擎	智能体 + 规则	云/私有化	LLM 自动评分与"使用指南"合规模块（多处于研究/试点阶段）	评价效度、合规审计、可解释性

坐标解读（定性）：

· 产品的整体迁移方向是从象限 A（左下：云端对话式）向象限 B/D（右上：智能体化 + 多模态/端侧）移动，反映"任务闭环 + 情境感知"的价值升级。Claude for Education 与 Gemini 的"引导式学习/Learning Mode"（以苏格拉底式提问替代直接给答案）代表 A 向 B 迁移中"学习范式升级"的一支；讯飞、华为等"场景一体机 + 助教/助学/助研/助管/助交流"代表 B 象限的中台化落子。

- 象限 B（智能体化平台/中台）是**价值高地与竞争最激烈处**，因为它同时承接教学、教研与治理三类需求，且对下游 App 与硬件具有底座锁定效应。艾瑞报告测算的"GenAI 技术能力价值"以 45% 的年复合增速快于产品服务总盘子的 37%（见 2.5 节），从商业结构上印证了中台层的价值锐度。
- 象限 D（多模态可穿戴）是**最高不确定性区**：技术前景显著，但成熟度、成本、伦理与教育场景适配仍待验证。网龙于 2023 年 C 轮以领投方战略投资 Rokid（该轮总额 1.12 亿美元、投后估值 10 亿美元），并签署五年战略合作，是本院关联产业在该象限的直接布局；其产业细节见本院 AI 眼镜蓝皮书。
- 象限 E（评价/治理引擎）是**新生但战略性**的聚类，直接对应本版新增的两大场景。当前证据表明该象限"效度尚不稳定、须人机协同"：一方面，多项 2025 年同行评议研究显示 LLM 自动评分已具备相当潜力——微调后的 ChatGPT 在 EFL 作文评分上可达较高一致性（Yavuz et al., 2025：组内相关系数 ICC 约 0.972、默认版约 0.947）；但同类研究结论并不一致——Flodén（2025）发现 ChatGPT 评分倾向回避极端分档、与人工评分完全一致的比例仅约三成、学科可靠性不足；这表明其效度并不均衡：模型对语法、用词、句式等表层维度把握较好，而对连贯性、论证清晰度与说服力等高阶、构念性维度仍不稳定，且在不同题目与不同学生群体间的可靠性存在漂移，因而研究者普遍强调"人工复核（human moderation）与显式效度论证"不可省略。这一"可用但不可全信"的证据状态，正是判断产品能否规模化进入正规教育体系评价环节的关键，也是第 5 章评测把"评价效度 + 人机协同边界"列为核心指标的原因。就产业成熟度而言，象限 E 目前以研究与试点为主、缺乏具名的成熟商用产品，故本章相关格子多标注 [待补]，不以研究原型冒充商用产品。

图 2-1（拟）：以"对话式↔智能体编排式"为横轴、"纯云软件↔端侧硬件"为纵轴的竞品坐标气泡图，五聚类 A–E，气泡面积表示 [待补：市场规模/装机量口径，数据见第 4—6 章评测]。图注中的精确坐标值与气泡口径待评测章节回填，占位不臆造。

2.5 市场规模与结构

产品图谱的商业底盘需要真实市场数据支撑。本节所引数据均标注机构、年份与口径，并严格遵守**币种防火墙**：人民币（RMB）与美元（USD）分列，不做汇率折算或跨币种加总；不同机构口径不同，横向比较仅作趋势参考，不作等价替换。

2.5.1 全球市场（USD 口径）

按 Research and Markets《AI in Education Market Report 2026》（2026 年 4 月发布）：全球"AI 教育"市场规模由 2025 年约 **75.2 亿美元** 增至 2026 年约 **106 亿美元**（2025→2026 CAGR 约 40.9%），并预计到 2030 年达约 **424.8 亿美元**（2026→2030 CAGR 约 41.5%）；区域上北美 2025 年领先，亚太增速最快。需要提示的是，不同机构对"AI 教育"边界界定差异极大：Precedence Research 给出 2025 年约 70.5 亿美元、2035 年约 1367.9 亿美元的长周期口径；MarketsandMarkets、Mordor Intelligence、Grand View Research、Technavio 等各家的基准年数值与 CAGR 亦不一致。故本蓝皮书**只引用口径与年份明确的单一机构数值**，不做跨机构拼接。

维度	口径与机构	数值
全球 AI 教育市场规模	Research and Markets，2026 发布，USD	2025≈75.2 亿美元；2026≈106 亿美元
全球 AI 教育市场规模预测	Research and Markets，2026 发布，USD	2030≈424.8 亿美元（2026→2030 CAGR≈41.5%）
长周期口径（供对照，勿与上行加总）	Precedence Research，2025 发布，USD	2025≈70.5 亿美元；2035≈1367.9 亿美元

2.5.2 中国市场（RMB 口径）

按艾瑞咨询《2026 年中国 GenAI+教育行业发展报告》（2026 年 3 月 3 日发布）：

维度	口径说明（RMB）	数值
GenAI+教育产品服务总规模	全年，2025 预测	2025≈3442 亿元；2028≈8910 亿元（CAGR≈37%）
其中"GenAI 技术能力价值"	纯技术能力口径	2025≈664 亿元；2028≈2023 亿元（CAGR≈45%）
技术能力价值占比	平均值	2025 年约 20%（即每 5 元产品服务中约 1 元多用于购买技术能力）
C 端教育市场总规模（背景盘）	消费端教育支出	2025≈1.3 万亿元；2028≈1.5 万亿元
B 端教育信息化数字化经费（背景盘）	学校端投入	2025≈5515 亿元；2028≈6802 亿元
用户侧渗透	2025 年 12 月口径	家长辅导使用 GenAI 比例 57.3%；成年学习者备考/学习使用 25%

学校端采购结构上，艾瑞报告给出：普通高校中 GenAI 相关项目占比约 27%（单项预算多在 100 万—400 万元），职业院校约 35%（150 万—400 万元），中小学约 46%（100 万—500 万元）。这组数据说明 B 端已进入"从硬件薄改转向数字基座 + 大模型深水区"的分梯队推进阶段（一线城市已入深水区，二三线仍以软硬件协同或硬件改造为主）。用户侧还有两个值得记录的基础事实：截至 2025 年 12 月，中国生成式 AI 用户规模约 6.02 亿人、普及率约 42.8%；孩子首次接触 GenAI 中约 74.8% 是通过通用（而非教育垂类）AI——这意味着"通用入口先行、垂类产品后置"是当前用户触达的现实路径，也为治理（内容安全、分学段规范）提出了先于产品成熟度的紧迫要求。

政策是中国市场不可忽视的需求侧变量。2025 年 5 月，教育部基础教育教学指导委员会发布《中小学生成式人工智能使用指南（2025 年版）》与《中小学人工智能通识教育指南（2025 年版）》，前者以"应用为王、治理为基"为总纲，按学段划定使用边界：小学阶段禁止学生独自使用开放式内容生成功能、教师可在课内适当使用，初中阶段可适度探索生成内容的逻辑性分析，高中阶段允许结合技术原理开展探究性学习。2026 年 4 月，教育部、国家发展改革委、工业和信息化部、科技部、国家数据局五部门联合印发《"人工智能+教育"行动计划》（教科信〔2026〕1 号），提出到 2030 年"人工智能与教育深度融合格局基本形成"的总目标，部署人才培养与素养提升、教育深度融合、基础环境、开放生态四大重点任务，并在应用融合部分明确"赋能学生学习、赋能教师教学、赋能学校治理、赋能科学研究"。这一顶层文件的意义在于：它把 GenAI 从"学校自选动作"变为"全学段系统工程"，直接决定了 B 端未来数年的预算流向与场景优先级，是本图谱"五场景"划分的政策依据，也是象限 B（中台）与象限 E（评价/治理）需求扩容的制度性推力。国际侧，UNESCO《生成式 AI 教育与研究指南》（2023）给出的"独立使用最低 13 岁"年龄门槛，与中国分学段红线一道，构成了产品设计必须内嵌的合规基线。

2.5.3 端侧硬件：AI 学习机（RMB 口径）

学习机是端侧化在中国消费市场最成熟的载体，多家机构口径可相互印证：

维度	口径与机构	数值
全渠道销量	2024 全年	592.3 万台，同比 +25.5%
全渠道销售额	2024 全年	190.6 亿元，同比 +37.6%
全渠道销量	2025Q1	126.5 万台，同比 +29.4%；销售额 40.2 亿元，同比 +15.8%
出货量	2025Q2（IDC 口径，学习平板）	154 万台，同比 +44.6%
竞争格局	2025Q2 销售额第一（IDC）	科大讯飞 AI 学习机（其入行以来销

		售额首度登顶）
竞争格局	2025Q1 销量/销售额双第一	作业帮（销量份额 38.8%、销售额份额 31.1%；2001–4000 元主流价位段份额 46.7%）

注：学习机口径由多家机构（IDC、艾媒、行业半年报等）分别测算，销量/销售额/出货量三种口径不可混用；上表按来源逐行标注。AI 学习机与 AI 眼镜的完整产业链、装机量测算与竞品格局详见本院《2026 AI 智能眼镜教育产业蓝皮书》。

2.5.4 结构性判断

综合三组数据可得三点结构性判断：其一，**全球与中国均处于 30%–45% 区间的高速增长**，但绝对盘子与口径差异极大，横向比较须锁定同一机构、同一口径；其二，**"技术能力价值"增速快于产品服务总盘**（中国 45% vs 37%），意味着价值正在向 2.4 节象限 B 的中台层集中；其三，**端侧硬件是当前商业化最确定的落点**（学习机连续多季度双位数增长），而可穿戴（象限 D）与评价/治理引擎（象限 E）仍处早期，规模数据尚不足以支撑精确气泡口径，故图 2-1 的气泡面积口径保留 [待补]。

2.6 时间线：2025—2026 关键节点

为呈现赛道演进节奏，本章设置一条产品与政策交织的时间线。下表所有条目均来自本章检索到的公开来源；未能核实到确切日期者标注为区间或 [待补]。

时间	事件类型	事件（已核实来源）	来源要点
2024-05	产品发布（国际·App）	OpenAI 发布面向高校的 ChatGPT Edu（底座 GPT-4o）	openai.com 官方公告
2024-12	范式（国际·智能体）	Google 发布 Gemini 2.0，	blog.google

		定位"面向智能体时代"	
2025-04-02 前后	产品发布（国际·App）	Anthropic 发布 Claude for Education，含 Learning Mode（苏格拉底式引导）	anthropic.com
2025-05	范式（国际·学习科学）	Google I/O 2025：LearnLM 并入 Gemini 2.5，Guided Learning 上线	blog.google
2025-05-12 前后	政策/标准（中国·治理）	教育部基教指委发布《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）》	edu.cn/CERNET
2025-06-24	硬件（中国·端侧）	科大讯飞 AI 学习机暑期发布会：AI 一对一精准学/答疑辅学/互动课等 16 项升级	qbitai.com
2025-07-15	产品发布（国际·渠道）	Claude for Education 上架 AWS Marketplace	aws.amazon.com
2025-09	平台/中台（中国·B 端）	华为全联接大会：华为 × 科大讯飞发布"星火教育、医疗大模型场景一体机"（助教/助学/助研/助管/助交流五场景）	e.huawei.com
2025-10	硬件（中国·端侧）	科大讯飞发布新一代 AI 黑板 2.0	edu.iflytek.com
2025-12-04	产品发布（中国·垂类模型）	阿里发布 Qwen3-Learning 教育垂直大模型	艾瑞报告转述

2026-03-03	报告（中国·市场）	艾瑞《2026 年中国 GenAI+教育行业发展报告》发布	iResearch
2026-04-02 印发/04-10 公开	政策（中国·顶层）	教育部等五部门印发《"人工智能+教育"行动计划》（教科信〔2026〕1 号），提出到 2030 年目标与四大任务	moe.gov.cn
2026-04-07	报告（国际·市场）	Research and Markets《AI in Education Market Report 2026》发布	Research and Markets

图 2-2（拟）：2025—2026 生成式 AI 教育产品演进时间线，按"国际产品 / 中国产品与硬件 / 政策与标准"三泳道排布，节点与来源同上表；跨币种规模数据不进入本图，避免混算。

2.7 小结

本章确立了 2026 版产品图谱的分析框架：**以五大场景（教学、学习、教研、智能评价、治理与安全）× 三类形态（App、平台/中台、端侧硬件）**为主网格，**以能力范式 × 部署形态**两轴竞品坐标定位主要玩家（A 云端对话式、B 智能体化中台、C 端侧学习硬件、D 多模态可穿戴、E 评价/治理引擎），并识别出产品整体向"智能体化 + 多模态/端侧化"迁移的主线。

相较旧版以对话式产品为主的清单式呈现，本图谱强调**范式迁移、场景扩容与循证定位**三点变化，并以三组可核实数据锚定商业底盘：全球 AI 教育市场（Research and Markets，USD 口径）2025→2026 由约 75.2 亿增至约 106 亿美元；中国 GenAI+教育（艾瑞，RMB 口径）2025 约 3442 亿元、CAGR 约 37%，其中技术能力价值以 45% 更快增速向中台层集中；端侧学习机 2024 全年销量 592.3 万台、2025 上半年延续双位数增长。三组口径分币种、分机构陈列，不混算。

需要坦承的边界是：象限 D（可穿戴）与象限 E（评价/治理）尚处早期，缺乏可支撑图 2-1 精确气泡口径的公开规模数据，相关格子与数值在本章保留为 [待补]，由第 4—6 章的教育垂类大模型评测、AI 硬件评测与产品横评以真实数据逐一填充，确保"图谱有骨架、数据有出处"。

本章参考来源

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. 艾瑞咨询. 2026 年中国 GenAI+教育行业发展报告. iResearch, 2026-03-03（界面新闻转载全文）. 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（Yahoo Finance 转载）. 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. 华为. 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". 华为企业业务, 2025-09. https://e.huawei.com/cn/news/2025/industries/education/spark-education-medical-large-model

11. 量子位. 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. 量子位. 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. 中华网. 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. 中国教育和科研计算机网（CERNET）. 《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. 中华人民共和国教育部. 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. 智通财经. 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. 腾讯新闻. 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

第 3 章　赋能教学：机理、产品形态与发展建议

3.1 引言：从"对话式辅助"到"智能体协同"的教学范式迁移

在生成式人工智能进入课堂的初期，"赋能教学"通常被界定为大语言模型对教师备课、授课、课堂互动等环节的对话式支持。这一界定在 2023 年具有解释力，但已难以覆盖 2026 年的技术现实：教学侧的生成式人工智能正从"被教师调用的问答工具"演进为"可被编排、具备记忆、能跨模态感知并主动介入教学流程的智能体（agent）"。本章在保留"机理—产品—建议"叙事骨架的基础上，将分析对象从单一对话式大模型扩展至**教育垂类大模型、教学智能体、多模态与端侧硬件**三类载体，并引入"能力—落地成熟度"双轴视角，避免将实验室演示与规模化课堂应用混为一谈。

这一范式迁移背后有三条相互交织的技术驱动力。其一是**模型能力从"通用问答"向"教育对齐"演进**：以教育语料做继续预训练与微调、以学习科学原则做对齐（如 Google 的 LearnLM 明确以主动学习、认知负荷管理、个性化、激发好奇、元认知五项学习科学原则微调），使模型输出更贴近教学法而非泛泛而谈。其二是**从"单轮生成"向"多工具编排"演进**：智能体框架让模型能够拆解任务、调用工具、串联流程，把"生成一段文字"升级为"完成一件教学工作"。其三是**从"云端"向"端侧"下沉**：批改一体机、交互白板、教师视角硬件把推理放到课堂现场，以应对实时性与数据合规的双重约束。三条驱动力共同把"对话式辅助"推向"智能体协同"，也正是本章从机理到产品重新组织叙事的原因。

本章需要在开篇明确一条贯穿全章的判断纪律：教学侧最具确定性的规模化落地，仍集中于**内容生成与备课辅助、习题与讲解生成、课堂语言支持、作业批改减负**等"教师增效"型场景。已有多项独立调查为这一判断提供了量化支撑：Walton 家族基金会与盖洛普（Gallup）2025 年 3—4 月对 2232 名美国公立中小学教师的调查显示，至少每周使用 AI 的教师平均每周节约

5.9 小时，折合一学年约六周时间；这些教师报告 AI 最常见的用途正是备课（lesson planning，约 37%）、制作学习单/活动（约 33%）、按学生需要改编材料（约 28%）等内容生成任务，而评分（约 16%）、一对一教学（约 14%）、学生数据分析（约 12%）等更接近决策层的用途占比明显更低。RAND 公司对 2024—2025 学年美国教师的抽样亦显示，全体 K-12 教师中约 53% 在该学年为教学目的使用过生成式 AI，较 2023—2024 学年的约 25% 翻倍（其中英语、科学学科教师的使用率明显高于数学与小学教师），采纳曲线明显陡峭上升；同期美国约 48% 的学区（2024 年秋）报告已开展教师 AI 培训，较上一年增长约 25 个百分点，说明制度性支持正在跟进但仍滞后于一线使用。相较之下，"自主授课智能体""实时多模态学情感知""AI 情绪识别"等场景多处于试点、演示或受法规约束的阶段。本章对每一类产品形态均标注其成熟度区间，凡无法用真实来源核实的具体项一律留白，不作臆断。

3.2 赋能教学的作用机理

3.2.1 三层机理框架

生成式人工智能赋能教学的作用路径，可归纳为**内容层、交互层、决策层**三层递进机理：

- **内容层（生成与重组）**：模型基于教材、课标与教师意图，生成或重组教学内容——教案、课件、讲解文本、分层习题、情境案例、多语言与无障碍材料等。其本质是把教师从"从零创作"转向"审阅与再加工"，机理上依赖大模型的语言生成与知识重组能力，是当前最成熟、最可靠的赋能层。盖洛普调查中教师报告的"备课（37%）、制作学习单/活动（33%）、按学生需要改编材料（28%）"等高频用途，均落在此层。
- **交互层（对话与多模态感知）**：模型在课堂或备课过程中与教师、学生进行多轮对话，并逐步引入语音、图像、板书、屏幕等多模态输入，实现"讲—问—答—纠"的交互闭环。此层的成熟度取决于多模态识别精度与实时性，端侧硬件（详见本院《2026 AI 智能眼镜教育产业蓝皮书》对教师第一视角课堂系统的产业分析）是其重要载体。

· **决策层（编排与介入）**：智能体依据教学目标、课堂状态与学情信号，自主决定何时提示、何时补充、何时转交教师，即所谓"教学编排（orchestration）"。此层引入 RAG（检索增强生成）以约束内容于校本知识库、引入记忆机制以跨课时保持连续性，是 2026 年的前沿方向，但规模化课堂验证仍不充分。

三层机理并非线性替代，而是叠加增强：内容层保证"生成得对"，交互层保证"交流得顺"，决策层保证"介入得当"。早期报告主要停留在内容层与交互层的对话式实现，本章将决策层（智能体编排）作为 2026 版的核心增量，同时强调：越往决策层走，人机责任边界越须清晰，教学决策权始终应保留于教师。

值得强调的是，三层机理的成熟度与教师的接受度呈现明显正相关：内容层因价值直观、风险可控而被广泛接纳，决策层则因触及教学判断而更易引发"是否越界"的疑虑。这一接受度梯度，与前述盖洛普调查中"备课类用途占比高、评分与学情分析类用途占比低"的数据高度吻合，说明教师的实际选择本身就是对三层机理成熟度的一种投票。

从认知负荷的角度看，三层机理对应教师工作负担的三种再分配。内容层削减的是"外在认知负荷"（extraneous load）——把排版、检索、初稿撰写等低创造性的机械劳动转移给模型，让教师把注意力集中在"生成相关认知负荷"（germane load）即教学法判断上；这解释了为何备课与批改的减负效应最容易被量化验证。交互层削减的是课堂即时应答的"工作记忆压力"——当教师面对突发提问、板书识别或多语言需求时，模型提供的实时补充相当于一个可即时调用的外部记忆。决策层则试图承担部分"教学监控"负荷，即判断"此刻该做什么"，这恰恰是教学专业性最集中、也最难被自动化的部分，因而其成熟度最低、责任约束应最严。这一负荷分层与本章"从内容到决策、确定性递减、约束递增"的总判断是一致的。

3.2.2 备课—授课—作业—反馈四环节的作用机理

若把三层机理落到教师的具体工作流，可进一步拆为备课、授课、作业、反馈四个环节，每个环节都有其独特的机理与瓶颈：

- **备课（内容层为主）**：机理是"意图对齐 + 知识重组 + 课标校验"。教师给出学段、教材版本、课时目标与学情约束，模型据此生成教案框架、教学设计、分层课件与情境案例。瓶颈在于课标对齐与学科准确性——通用模型缺乏对具体教材版本与课标条目的精确记忆，容易产出"看似合理、实则错位"的目标与活动，这正是校本 RAG 成为教育垂类产品标配的根本原因。盖洛普调查显示备课（37%）是教师使用 AI 的第一大用途，与该环节机理最成熟的判断吻合。
- **授课（交互层为主）**：机理是"实时感知 + 即时生成 + 低延迟反馈"。模型在课堂现场承担讲解补充、追问生成、板书/屏幕识别、多语言即时转写等任务。瓶颈是时延与可靠性——课堂是强实时、低容错场景，云端往返的秒级延迟与偶发幻觉都会打断教学节奏，这既推动了端侧化部署，也决定了"自主授课"在可预见的将来仍不成立。
- **作业（内容层 + 决策层）**：机理是"命题生成 + 自动批改 + 学情归因"。前端是分层命题与题目变式生成，后端是主客观题批改与错因分析。客观题批改已高度成熟（可视作规则+识别问题），主观题批改则进入决策层——需要模型对开放作答做评分与理由说明，其一致性与可解释性是核心瓶颈，必须保留人工复核与申诉机制。讯飞星火智能批阅机把"批改—学情分析—个性化作业"整合为一条端侧流水线，是该环节产品化的代表。
- **反馈（决策层为主）**：机理是"学情建模 + 差异化建议 + 教学再设计"。模型基于作答与课堂数据，为教师生成"下一步教什么、对谁补什么"的建议，并回流到下一轮备课，形成闭环。瓶颈在于因果归因的可靠性与数据合规——学情信号往往相关而非因果，且高度依赖对未成年人数据的采集，稍有不慎即触碰隐私与情绪识别红线（见 3.4）。这也是四个环节中产品边界最模糊、与"智能评价""教研支持"耦合最深的一环。

四环节的机理差异，直接决定了产品的落地节奏：备课与作业批改因瓶颈可控而率先规模化，授课因实时性约束而依赖硬件演进，反馈因合规与因果难题而最为审慎。后文的产品剖析与成熟度时间线，均可回置到这一四环节框架来解读。

3.2.3 关键技术范式的教学映射

下表将 2026 年的技术范式映射到具体教学环节，说明其作用机理与当前成熟度定位。成熟度定位综合了本章检索到的公开案例与部署口径，个别缺乏公开评测的项目仍以占位保留：

技术范式	教学环节映射	作用机理	落地成熟度（本章判断）
教育垂类大模型	备课、命题、讲解、批改	以教育语料对齐课标与学段，降低通用模型的知识错配	推广（如好未来 MathGPT、讯飞星火教育模型；评测口径见 3.5）
教学智能体编排	课堂流程调度、多工具协同	按教学目标拆解任务、调用工具、决定介入时机	试点（如 MagicSchool "Raina"、Khanmigo 工具集合入口）
RAG（检索增强）	校本知识库讲解、答疑	将生成约束在可信教材/校本资源，抑制幻觉	试点—推广（校本语料接入已成教育垂类产品标配）
记忆机制	跨课时连续教学	保持对班级/单元的长期上下文	演示—试点 [待补：跨课时记忆的规模化课堂验证]
多模态感知	板书识别、学情观察	融合语音/图像/屏幕实现课堂状态感知	演示—试点（受情绪识别法规约束，见 3.4）
端侧化部署	无网/低延迟课堂、批改一体机	本地推理降低时延、保护数据、适配弱网环境	试点（如讯飞星火智能批阅机；硬件口径见蓝皮书姊妹册）

3.2.4 通用大模型与教育垂类模型的教学能力分野

教师侧产品能力的可信度，很大程度上取决于其底层是"通用大模型"还是"经教育对齐的垂类模型"，二者在教学任务上的分野可从三个方面理解：

其一是**知识对齐**。通用模型的知识以互联网语料为主，缺乏对特定教材版本、课标条目、学段认知规律的精确记忆，容易在"知识点归属哪一册、难度是否匹配学段、术语是否符合课标"等细节上出错；教育垂类模型则以教材、课标、题库等教育语料做继续训练与对齐，可显著降低这类"知识错配"。中国信息通信研究院等机构已开展面向教育大模型的分级评估（好未来 MathGPT 获 4+ 级评级即出自此类评估），为区分二者提供了初步的第三方锚点，但面向具体教学任务的细粒度评测仍有待补充。

其二是**教学法对齐**。通用模型倾向于给出"正确但平铺直叙"的答案，而好的教学需要循序、追问、留白、纠错等教学法结构。Google 的 LearnLM 以五项学习科学原则做微调、讯飞星火教师助手融合"优秀教师育人经验"，本质都是在补足这一层——让模型不只是"答得对"，而是"教得像老师"。

其三是**安全与价值对齐**。面向未成年人的教育场景对内容安全、价值导向、隐私保护有远高于通用场景的要求。教育垂类产品普遍内置更严格的内容护栏、面向学生的独立安全环境（如 MagicStudent），并需符合教育专门法规（中国《生成式人工智能服务管理暂行办法》与算法备案、美国 FERPA/COPPA、欧盟 GDPR 及 AI 法案）。这也是通用对话产品被教师"借用"时最需警惕的风险来源。

需要澄清的是，垂类并不等于"更强"，而是"更对"：在开放推理与跨学科综合上，前沿通用模型往往能力上限更高。理想的教师侧产品，是以强通用模型为底座、以教育语料与 RAG 做对齐、以任务化工具与安全护栏做封装，兼取"能力上限"与"教育可信"两端之长。

3.3 产品形态图谱

3.3.1 从对话式工具到智能体的四类形态

2026 年教学侧生成式人工智能产品可归为四类形态，构成一条从"轻"到"重"的连续谱：

1. **对话式教学助手**：以自然语言对话为主入口，服务备课问答、内容生成、讲解润色。此为覆盖面最广、门槛最低的一类。其形态又可细分为两支：一支是通用对话产品（如 ChatGPT、文心一言、通义、豆包等）被教师"借用"于教学任务，优点是能力上限高、成本低，缺点是缺乏课标对齐与教育安全约束，需教师自行承担提示工程与内容审核；另一支是从通用产品分化出的教育专用对话界面（如 Khanmigo、星火教师助手的对话入口），以教育语料与安全护栏做了二次约束。二者的分野正是本章反复强调的"通用可用"与"教育可信"之别。
2. **嵌入式教学插件**：内嵌于备课平台、办公套件、交互白板与教学一体机的生成式功能模块，教师在既有工作流中就近调用，迁移成本低。国内典型如网龙旗下 101 教育 PPT 在备授课一体化软件中集成 AI 生成能力，国际典型如 Google 将 30 余项 AI 工具嵌入 Google Classroom 的"Gemini in Classroom"、微软将 Khanmigo 与 Copilot 能力接入教育版办公套件。此类形态的价值主张是"不改变教师既有工作习惯"，因而在规模化采纳上往往优于独立 App。
3. **教学智能体与智能体平台**：可配置角色、绑定工具、接入校本知识库并具备一定自主编排能力的智能体，面向"一键生成整单元教学包""对话式选工具""跨工具串联任务"等复合任务。典型如 MagicSchool 的教师侧对话智能体"Raina"，用户可用一句话描述需求、由智能体判断该调用哪个工具或直接生成并在对话中迭代；Khanmigo 则把 25 余项工具组织为可按 Plan/Create/Differentiate/Support/Learn 分类调度的集合入口。此类形态是"决策层编排"

的产品化早期，多处于试点—推广早期，其自主性目前仍限于"工具选择与串联"，而非"教学决策"。

4. **多模态与端侧教学硬件**：以教学一体机、交互白板、智能批阅机、教师第一视角智能眼镜等为载体，将生成式能力下沉至端侧设备。此形态是"内容/交互能力"与"物理课堂"的接口，既解决实时性与弱网问题，也是数据合规的重要抓手（数据不出端）。硬件形态、出货与评测请参见本院《2026 AI 智能眼镜教育产业蓝皮书》与《全球教育机器人发展蓝皮书 2026》的产业链与硬件评测部分，本章不重复市场测算，仅从教师赋能机理角度纳入分析。

四类形态并非互斥，而是同一能力在不同"重量级"下的封装：越轻越易普及但越难保证教育可信，越重越能保障合规与实时但部署成本越高。教师侧产品的竞争，实质是在"可用性—可信性—部署成本"三角中寻找与具体教学场景匹配的平衡点。

3.3.2 教师教学全流程的产品落点

按教师教学的时间线，可将产品能力落点映射为"课前—课中—课后"三段：

- **课前（备课与命题）**：教案生成、课件生成、分层习题与命题、跨学科情境设计、多语言与无障碍材料生成。此段落地最成熟，价值最确定，也是当前各产品功能密度最高的区间。中国教育部《"人工智能+教育"行动计划》明确提出"赋能教师教学，推动构建覆盖课前、课中、课后全环节的智能应用"，课前是其中最先被产品化的环节。
- **课中（授课与互动）**：实时讲解补充、课堂提问生成、板书与屏幕内容识别、即时多语言支持、课堂互动设计。此段依赖多模态与端侧硬件，成熟度差异大，须审慎区分"可用"与"可演示"；涉及生物特征与情绪推断的功能受法规红线约束（见 3.4）。
- **课后（反馈与再设计）**：作业批改与学情分析、课堂实录的结构化复盘、教学行为分析、下一课时的自适应再设计。此段既有已规模落地的成熟功能（如智能批阅），又与"智能评价"及教研支持高度耦合，产品边界正在模糊化。

三段之间存在一条被产品化不断打通的数据回路：课后的学情分析结果，理应回流到课前的下一轮备课与命题，形成"教—学—评—备"闭环。当前多数产品仍以单点功能为主，闭环打通尚不充分；能否把课后反馈稳定地转化为课前再设计的输入，是判断一个教师侧平台"是否真正进入决策层"的关键标志。教育部《"人工智能+教育"行动计划》所提"覆盖课前、课中、课后全环节的智能应用"，其深层要求正是这一闭环，而非孤立的功能堆叠。

3.3.3 内容生成与编排的实现范式

在剖析具体产品前，有必要说明教师侧产品共用的三种实现范式，它们决定了产品的能力天花板与可信度下限：

- **任务化封装（prompt-to-tool）**：把常见教学任务固化为参数化工具（如"输入学段+主题→输出量规"），教师无需掌握提示工程即可调用。MagicSchool 的 60—80 余项工具、Khanmigo 的五类工具集、Google Classroom 的 30 余项工具均属此类。其优点是可靠、可复用、可评测；代价是灵活性受限于工具设计者的预设。
- **检索增强（RAG）+ 校本适配**：在生成前先从教材库、课标库、区校本资源中检索相关材料，再据此约束生成，从而抑制幻觉、提升课标对齐度。讯飞星火教师助手基于"自建+区校本+UGC"资源做检索整合、Google 强调按学习目标与教材样例生成，均是此范式的实例。它是教育垂类产品区别于通用产品的技术分水岭。
- **对话式编排（agentic orchestration）**：由智能体解析教师意图，自主决定调用哪个工具、以何顺序串联、何时回问澄清、何时把结果交还教师。MagicSchool 的"Raina"是当前较成熟的教师侧实例。此范式提升了复合任务的一次性完成度，但也把"判断该做什么"的部分权重交给了系统，因而必须以清晰的责任边界与人工终审为前提。

三种范式在真实产品中往往叠加使用：以 RAG 保证"生成得对"，以任务化封装保证"用得顺手"，以对话式编排保证"串得起来"。理解这三条范式，有助于在 3.5 节的横评中把"看起来相似"的产品按其真实技术深度区分开。

3.3.4 代表性教师侧产品剖析

以下按"国际—国内"两组，剖析本章实际检索到的、以教师为直接服务对象的代表产品，重点是其功能构成、真实规模口径与证据强度。所有数据均标注来源年份，未获公开核实者留白。

（一）Khan Academy Khanmigo for Teachers（美国）

Khanmigo 是可汗学院（Khan Academy）推出的 AI 教学助手与辅导系统，其教师侧（Khanmigo for Teachers）自 2024 年 5 月起由微软（Microsoft）合作提供、并基于 Azure OpenAI 服务运行。据微软教育官方博客（2024 年 8 月）介绍，教师版免费提供 25 项以上教师专用工具，按 Plan（规划）、Create（创作）、Differentiate（差异化）、Support（支持）、Learn（学习）五类组织，代表工具包括课堂导入设计（Lesson Hook）、命题生成（Question Generator）、评语与成绩单批注（Report Card Comments）、班级通讯（Class Newsletter）等；该版本以英语在 49 个国家免费开放。其教学价值在于将"需要专业提示词技巧的通用模型"封装为"教师无需学习提示工程即可调用的任务化工具"。

Khanmigo 的课堂证据主要来自区域试点：美国新泽西州纽瓦克公立学区（Newark Public Schools）2023—2024 学年试点后扩大部署，据 Chalkbeat 报道，该学区当年参加标准化测试的学生数学通过率从 2023 年约 15% 升至 17.7%，但该学区同时说明无法量化归因于 Khanmigo 辅导本身；截至相关报道，合作已覆盖该学区数十所学校、约 2.9 万名学生。需要强调：Khanmigo 兼具学生辅导与教师助手双重定位，其中学生辅导侧的因果证据仍在积累，教师侧的价值目前更多体现为备课与批改效率，而非直接的学业提升。

（二）MagicSchool AI（美国）

MagicSchool 是增长最快的教师侧生成式 AI 平台之一。据其官方公告，公司于 2025 年 2 月 11 日完成 4500 万美元 B 轮融资，由 Valor Equity Partners 领投，Bain Capital Ventures、Adobe Ventures 等参投；公告称平台注册教育者已超 600 万，合作学校逾 1 万所，覆盖约 160 个国家。至 2026 年初，多方口径显示注册教育者数进一步增长，覆盖学校逾 1.3 万所。其产品以 60—80 余项任务化工具著称，涵盖分层教案、评分量规（Rubric Generator）、试题、家校沟通、IEP（个别化教育计划）辅助等；第三方评述普遍援引"教师每周节省 7—10 小时"的自报口径。

以其评分量规生成器（Rubric Generator）为例，可管窥其任务化封装的机理：教师只需选择学段、填入作业标题、描述与目标、设定分值档位，工具即产出表格化、标注了评价维度与各档描述的量规。这类工具把"教师需要花半小时琢磨的评价框架"压缩为分钟级操作，正是其"每周节省数小时"口径的微观来源。MagicSchool 同时提供面向学生的独立安全环境 MagicStudent，教师可创建受控的"学生房间"，与教师侧工具形成"教师生成—学生使用"的联动。

MagicSchool 在 2025 年引入了教师侧对话智能体"Raina"，其定位不是通用聊天机器人，而是"经教育教学法训练、能在对话中理解需求并引导至相应工具或直接生成内容"的编排入口，并在对话中保持上下文以便迭代修改——这正是本章 3.2.1"决策层编排"在教师侧的一个可用化实例。合规方面，MagicSchool 宣称符合 FERPA、COPPA、GDPR 与 SOC 2 Type 2，并援引 Common Sense Media 给出的较高隐私评级。需要注意的是，此类"节省时长""好评率"多为自报或厂商口径，缺乏独立随机对照验证；且据公开信息，MagicSchool 于 2026 年初对其学生侧产品做了改版并调整了 AI 助手的人格化设计，反映出教育产品在"拟人化"与"工具化"之间仍在探索边界。选型时应与 3.5 节的循证评测建议对照使用。

（三）Google Gemini for Education / LearnLM（美国）

Google 面向教育的路线是"通用模型 + 学习科学微调 + 工作流嵌入"。其 Gemini for Education 由 Gemini 2.5 Pro 驱动并融入 LearnLM——一组以学习科学原则（主动学习与反馈、认知负荷管理、个性化、激发好奇、元认知）微调的模型族。教师侧最具规模的落点是"Gemini in Classroom"：Google 将 30 余项 AI 工具嵌入 Google Classroom，支持按学习目标生成教案、从零或依样例生成测验与量规、按学生水平重置文本难度、针对常见误解设计引导提示等，且对 18 岁以上的 Workspace for Education 用户免费。Google 2025 年年终回顾称，其教育 AI 能力已进入逾千所美国高校、触达千万量级学生（该口径以学生侧为主，教师侧规模 [待补：教师用户口径]）。LearnLM 的一项常被引用的内部结果是"接受 LearnLM 短时辅导的学生，在后续新题上的解题正确概率较仅由人类辅导者辅导的学生高约 5.5 个百分点"——但这属于学生辅导侧证据，且为厂商自评，宜谨慎对待。Google 路线的独特之处在于其"生态嵌入"策略：绝大多数中小学教师已在使用 Google Classroom 与 Workspace，因而 AI 工具的采纳几乎不产生额外迁移成本，这也是其能快速触达大规模用户的结构性原因；反过来，这也意味着其教育可信度高度依赖底层通用模型的对齐质量与 LearnLM 微调的有效性。

（四）好未来"九章大模型"（MathGPT）（中国）

好未来（TAL）自研的九章大模型（MathGPT）是国内较早发布的数学领域千亿级大模型，核心为解题与讲题算法，能力覆盖小学、初中、高中数学的计算题、应用题、代数题等题型，并延伸出作文批改等功能。据公开信息，九章大模型在中国信息通信研究院的教育大模型评估中获得 4+ 级评级（属该评估当时公布的较高等级）。在教师侧，好未来智慧教育提供覆盖"备课+上课"的智慧备课方案，支持查题、组卷、制作讲义与互动课件、录制剪辑微课等。需注意九章的强项在于数学学科的解题与讲题一致性，跨学科通用性与课标对齐的独立评测 [待补：分学段/分学科第三方评测结果]。

（五）科大讯飞"星火教师助手"（中国）

科大讯飞基于讯飞星火认知大模型推出"星火教师助手"，定位"懂教学、更专业的教师 AI 伙伴"，通过对话式交互为教师生成大单元教学规划、教学设计、贴合情境的课件，并覆盖教学反思、课题灵感、学生评语、班会设计、家访沟通提纲等场景；其课件生成可基于讯飞自建、区校本与 UGC 资源做检索、分析与整合，属典型的 RAG 校本适配路线。据讯飞智慧教育披露的一线教师使用数据：教学设计效率提升约 56.52%、资源检索便捷度提升约 56.22%、课件制作效率提升约 64.18%、教师好评率约 93%；同时称产品已覆盖全国 4000 余所学校、20 万余名教师，教师周均对话生成约 5.3 次。除助手外，讯飞 2024 年 6 月发布的星火智能批阅机是典型的端侧化教师减负硬件，官方称可将作业批改时间由约 90 分钟降至约 5 分钟、学情分析由约 60 分钟降至约 1 分钟。上述均为厂商自报口径，独立第三方评测 [待补]。

（六）网龙 101 教育 PPT（中国，本机构关联企业）

网龙（NetDragon，港股代码 HK:0777）旗下 101 教育 PPT 是一款面向教师的备授课一体化工具软件，覆盖义务教育、高中及中职等不同版本教材，提供免费教案、课件、微课等教学资源与学科/课堂互动工具。据公开披露口径，101 教育 PPT 已为全国约 970 万教师提供服务、累计装机量超过约 3576 万，用户覆盖全国 32 个省级行政区。网龙是国内较早将 AI 引入教育的企业之一（2018 年即发布"AI 助教"），并在其教育产品线中持续融合 AI 生成能力以辅助备课与课件制作。需要说明：本蓝皮书对关联企业口径亦坚持来源核验，凡涉及最新装机、教师数与 AI 具体功能的动态数据，均以公开披露为准，未公开者留白 [待补：101 教育 PPT 生成式 AI 模块的功能清单与教师采纳率]；关于网龙在 AI 眼镜领域的投资（含与 Rokid 相关的产业布局），本章不展开，详见本院《2026 AI 智能眼镜教育产业蓝皮书》。

综观六个案例，可归纳出教师侧产品的三条共性演进线：其一，从"通用对话"走向"任务化工具集"，降低教师的提示工程门槛；其二，从"单点工具"走向"对话式编排入口"（如 Raina），提升复合任务的一次性完成度；其三，从"云端生成"走向"校本 RAG + 端侧减负硬件"，同时

应对内容可信与数据合规两重约束。各产品覆盖学科、部署规模与教师采纳率的横向可比数据仍不充分，本章建议以 3.5 节的循证工具补齐。

3.3.5 国际与国内两条落地路径的比较

将上述案例并置，可辨识出教师侧生成式 AI 在国际与国内呈现出两条既相通又有别的落地路径：

- **国际路径以"平台工具集 + 生态嵌入"为主**。MagicSchool、Khanmigo、Google 均走"把大量教学任务封装为工具、再嵌入教师既有生态（办公套件、Classroom、LMS）"的路线，规模扩张极快（MagicSchool 数百万至千万量级教育者、Khanmigo 通过微软合作覆盖数十国）。其优势是普及速度与工具丰富度，短板是校本教材/课标的本地化适配相对薄弱，且高度依赖底层通用大模型（多为 OpenAI/Google 系）。合规上以 FERPA/COPPA/GDPR 等既有框架为约束。
- **国内路径以"垂类模型 + 备授课一体化 + 端侧硬件"为主**。好未来 MathGPT、讯飞星火教师助手与批阅机、网龙 101 教育 PPT 更强调对国内教材版本、课标与考试的深度适配，并常与交互白板、批阅机、学习机等硬件绑定，形成"软件+硬件+区校本资源"的一体化方案。其优势是课标对齐与场景闭环，短板是跨学科通用性与国际可比评测数据相对稀缺。合规上以《生成式人工智能服务管理暂行办法》、算法备案、数据分类分级及新近的《教师生成式人工智能应用指引》等为约束。

两条路径的共同挑战是一致的：如何把"效率提升的自报口径"转化为"教学质量提升的独立证据"，以及如何在多模态与学情数据上守住未成年人隐私与情绪识别的红线。这也正是本章 3.4、3.5 两节所要回应的核心问题。

3.4 典型场景与风险边界

3.4.1 高价值、高确定性场景

- **备课减负**：把教师从重复性内容生产中释放，价值最易验证。盖洛普—Walton 调查中"每周节省 5.9 小时、约六周/学年"与讯飞"教学设计效率提升逾 56%"等口径，指向同一方向。风险在于内容准确性与课标契合度，需教师终审。
- **作业批改与学情反馈**：智能批改一体机、批改工具已在部分区域规模落地，是课后段确定性最高的减负点；风险在于主观题评分的一致性与可解释性，须保留人工复核与申诉通道。
- **分层与个性化命题**：面向差异化学情快速生成分层练习与量规；风险在于难度标定与知识点覆盖的可控性，宜以校本题库与课标做 RAG 约束。
- **课堂语言与无障碍支持**：多语言课堂、听障/视障场景的实时文字/语音支持；此场景与端侧硬件强相关，其产业与硬件评测口径见本院 AI 眼镜蓝皮书。
- **家校沟通与事务性文书**：家长通讯、学生评语、家访提纲、通知与总结等事务性写作，是教师隐性工时的重要消耗点，也是各产品（Khanmigo 的 Class Newsletter、讯飞的家访沟通提纲、MagicSchool 的家校沟通工具）高频覆盖的场景。其价值确定、风险较低（内容多为程式化文书），但仍须教师核对个别学生信息的准确与得体，避免机械化措辞伤害家校关系。

上述场景的共同特征是：任务边界清晰、错误代价可控、人工终审成本低。它们之所以能率先规模化，正是因为在"生成得对"这一最成熟的内容层就能创造确定价值，而不必依赖尚不成熟的多模态感知与自主编排。

3.4.2 需审慎对待的场景与红线

- **"自主授课智能体"宣称**：当前多为演示或高度受控的试点，严禁将其表述为可无人监督地替代教师授课。中国教育部 2025 年发布的《教师生成式人工智能应用指引（第一版）》明确要求教师"始终发挥育人主导作用，将生成式人工智能仅作辅助工具使用"，并规定在价值观引导、道德教育、情感培养、心理支持等关键育人环节"必须由教师主导完成，不得交由技术替代"。这为产品定位划出了不可逾越的责任边界。
- **课堂多模态感知与情绪识别**：涉及未成年人生物特征与情绪数据，须严守法规红线。欧盟《人工智能法案》第 5 条（Article 5(1)(f)）自 2025 年 2 月起禁止在工作场所与教育机构中基于生物特征数据推断自然人情绪（医疗或安全目的除外），违者最高可处全球年营业额 7% 的罚款；其立法理由（Recital 44）直指此类系统科学基础不足、可靠性与泛化性有限。这意味着"摄像头专注度/情绪监测"类课堂功能在欧盟辖区原则上被禁止，在其他辖区亦应以最小采集、明确告知与可退出为前提审慎设计。
- **内容质量、偏见与幻觉**：多项 2024—2025 年研究表明，AI 生成教案在缺乏角色约束、范例与量规等脚手架时，易出现目标笼统、认知层次偏低、差异化流于表面等问题，甚至在教材信息不足时"编造事实"（hallucination）；在盲评中，人工撰写的教案在质量维度上仍普遍高于 AI 生成教案。据此，生成内容必须经教师专业审阅方可进入课堂，产品应内置量规约束与来源可溯机制。
- **数据出域与端侧合规**：课堂音视频、学生作答等敏感数据的采集与传输，端侧化部署在降低时延之外亦是一条合规路径，但不能替代制度性的数据治理安排。中国《教师生成式人工智能应用指引》亦要求相关企业依法落实数据分类分级、安全评估与算法备案。
- **教师专业能力的"去技能化"隐忧**：过度依赖生成式工具，可能弱化青年教师独立设计教学、诊断学情的核心专业能力，形成"会用工具、不会教书"的空心化风险。这一风险难

以量化，却关系教师队伍的长期专业性，应在教师培训与产品设计中以"辅助而非替代思考"为原则加以对冲——例如让工具输出"可编辑的初稿+设计理由"而非"不可质疑的成品"。

3.4.3 证据强度分级：区分自报、试点与因果

本章检索到的"证据"在可信度上差异极大，必须分级看待，避免以低强度证据支撑高强度结论：

- **厂商自报口径（最弱）**：如"节省 7—10 小时/周""效率提升 56%/64%""好评率 93%"。这类数据来自厂商或其合作方，样本、对照与测量方法通常不公开，只能作为"方向性参考"，不能作为选型的决定性依据。
- **第三方调查口径（中等）**：如盖洛普—Walton 对 2232 名教师、RAND 对全国教师样本的抽样调查。这类数据有明确样本与方法，可信度较高，但反映的是"教师自报的使用与感受"，仍非对学习结果的因果度量。
- **准实验/区域试点口径（中等偏上）**：如纽瓦克学区 Khanmigo 试点后数学通过率变化。这类数据接近真实课堂，但常缺乏对照组、难以排除混杂因素，学区自身也往往声明"无法量化归因"。
- **随机对照/受控实验口径（最强）**：目前针对"教师侧生成式 AI 提升教学质量"的严格随机对照证据仍稀缺；已有的部分受控实验多集中于"学生辅导"侧（如 LearnLM 的解题正确率、AI 辅导对比研究），不能直接迁移到"教师赋能"侧的结论。

这一分级提示了本领域的核心证据缺口：备课与批改的"减负"已有中等强度证据支撑，但"减负是否转化为更好的教学质量与学生学习结果"仍缺乏高强度因果证据。选型与政策制定应据此保持审慎乐观，并优先投资于可产出高强度证据的评测设计。

3.5 循证评测建议：让"赋能教学"可比较、可核验

厂商披露的"节省时长""效率提升百分比""好评率"多为自报或受控条件下的口径，缺乏统一方法与独立复核。为避免产品能力停留在宣传口径，本章建议在蓝皮书中引入三类循证工具：

- **教育垂类模型的教学任务评测**：面向教案生成、命题质量、讲解准确性、课标对齐度、批改一致性等教学专项设计评测集，分学段、分学科报告能力表现，并公开评分标准与样本口径 [待补：统一评测方法与结果]。已有的行业级评估（如中国信通院教育大模型评估的分级）可作为参考锚点，但需补充面向具体教学任务的细粒度证据。
- **产品横评与能力雷达**：对同一形态产品在"内容生成、交互体验、编排能力、校本适配、合规与安全、端侧性能"六个维度做统一评分并以雷达图呈现，明确区分厂商自报口径与独立复核口径 [待补：横评样本与评分]。
- **落地成熟度时间线**：以时间线标注各类形态从"演示—试点—推广—规模化"的迁移节点，区分"能力可达"与"课堂已落地"。例如：备课与批改类工具已进入"推广—规模化"，对话式编排入口处于"试点—推广早期"，跨课时记忆与多模态学情感知处于"演示—试点"，而受法规约束的情绪识别在多数辖区不进入产品化。

在评测方法上，本章进一步建议三条可操作原则，以把上述工具落到实处：

- **任务锚定而非能力泛谈**：评测应绑定具体教学任务（如"人教版八年级物理某单元教案生成""初中数学开放题批改一致性"），由学科教师依据统一量规盲评，而非笼统评价"模型强不强"。教案质量可从课标对齐度、目标可测性、认知层次分布、差异化设计、活动可行性等维度打分；批改质量可用与教师评分的一致性系数（如二次加权 Kappa）与错因解释的准确率衡量。

- **人机对照与增量度量**：不仅比较不同产品，更要比较"教师独立完成"与"教师+AI 完成"的质量与耗时差异，从而度量 AI 的真实增量；理想情况下引入随机分组，以逼近因果证据，填补 3.4.3 指出的证据缺口。
- **口径透明与利益披露**：任何被引用的效率或效果数据，均须标注样本量、对照设置、测量方法与数据来源方（厂商/第三方/学术），并在图表中以不同标识区分厂商自报与独立复核，杜绝把营销口径混入循证结论。

上述评测须坚持来源可核验原则：任何评分、排名与规模数据均应在脚注标注真实来源与口径，未获实证前以占位保留；尤应警惕将"学生辅导侧证据"错置为"教师赋能侧证据"（如 LearnLM、Khanmigo 的部分学业结果实为学生辅导路径的证据）。

3.6 发展建议

面向教学侧的生成式人工智能产品发展，本章提出六点建议，分别面向产品方、学校与教育行政方，以及评测与治理方：

1. **以"教师增效"而非"教师替代"为产品定位主线**。优先做实课前备课、课后批改与课中语言支持等高确定性、已有量化证据的场景；把决策层的智能体编排作为渐进增量而非营销噱头，成熟度未达即不宣称，并在产品与合同中明确"关键育人环节由教师主导"的红线，与教育部《教师生成式人工智能应用指引》保持一致。产品的输出宜以"可编辑初稿 + 设计理由"的形态呈现，而非"不可质疑的成品"，以保护而非侵蚀教师的专业判断。
2. **推动 RAG 与校本知识库成为教学智能体的默认底座**。以校本教材与课标约束生成，抑制幻觉与知识错配，是教育垂类产品区别于通用产品的关键（讯飞星火教师助手的校本资源检索、Google 的课标对齐生成均是此路线的实例），也是可信落地的前提。生成内容应可溯源、可被教师量规审阅；对超出知识库覆盖范围的请求，产品应"知之为知之"地提示不确定，而非编造。

3. **将合规与安全前置为产品设计约束**。针对课堂多模态数据、未成年人数据与情绪识别红线，在产品架构层面落实最小采集、端侧优先与可审计；面向欧盟等辖区须默认关闭基于生物特征的情绪推断功能（欧盟《人工智能法案》第 5 条已将其列为禁止实践），面向中国须落实数据分类分级、安全评估与算法备案。具体治理框架详见本蓝皮书治理与安全相关章节。

4. **建立可核验的教学评测与横评机制**。以教育专项评测、能力雷达与成熟度时间线支撑选型决策，杜绝以演示替代实证、以自报替代独立复核，并严格区分教师赋能侧与学生辅导侧的证据来源；相关硬件形态的产业与评测口径可交叉参考本院《2026 AI 智能眼镜教育产业蓝皮书》《全球教育机器人发展蓝皮书 2026》。

5. **把教师数字素养培训与产品部署同步推进**。RAND 数据显示美国学区教师 AI 培训虽快速增长仍显滞后于一线使用，国内教育部《"人工智能+教育"行动计划》亦将"全面提升教师数字素养"列为重点。学校在引入教师侧产品时，应同步开展"如何审阅 AI 输出、如何守住数据与育人边界"的培训，避免"工具先行、素养滞后"造成的误用与去技能化。

6. **以"教—学—评—备"闭环而非孤立功能作为平台选型标准**。真正的价值不在于工具数量，而在于课后学情能否稳定回流到课前再设计。学校与行政方在采购时，应重点考察平台能否打通全环节数据回路、能否与既有教学生态与校本资源对接，而非被工具清单的长度所迷惑。

综上，2026 年赋能教学的核心命题，已不再是"生成式人工智能能否辅助教学"——多项独立调查已证实其在备课与批改等环节的减负价值——而是"以何种形态、在何种成熟度、受何种治理约束地进入课堂"。本章以三层机理框架与四环节拆解、四类产品形态与三种实现范式、六个真实产品剖析、证据分级与循证评测建议给出结构化回答。需要清醒看到的是：减负的证据已相对充分，而"减负转化为更好教学质量与学习结果"的高强度因果证据仍稀缺；守住

教师育人主导权与未成年人数据红线，是这一切进入课堂的前提。仍存疑或缺乏公开核实的具体数据以［待补］保留，待后续循证补充。

本章参考来源

1. Walton Family Foundation & Gallup. "Teaching for Tomorrow: Unlocking Six Weeks a Year With AI"（AI Dividend / 五点九小时·六周口径，2232 名美国公立中小学教师，调查期 2025-03-18 至 2025-04-11）. 2025. https://www.waltonfamilyfoundation.org/the-ai-dividend-new-survey-shows-ai-is-helping-teachers-reclaim-valuable-time ；Gallup 版本："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+ 工具、五类分组、Azure OpenAI、49 国免费）. 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. 关于纽瓦克公立学区 Khanmigo 试点与数学通过率（15%→17.7%）及扩大部署的报道. 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"（4500 万美元 B 轮、2025-02-11、Valor Equity Partners 领投、600 万+ 教育者、1 万+ 学校、160 国）. 2025. https://www.magicschool.ai/blog-posts/series-b-fundraise-for-teacher-ai

7. MagicSchool. "Rubric Generator" 与 "AI Tools for Teachers / Raina"（教师侧对话智能体、FERPA/COPPA/GDPR/SOC2 合规、Common Sense Media 隐私评级）. 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+ 工具、Gemini 2.5 Pro、LearnLM 五项学习科学原则）. 2025. https://blog.google/products-and-platforms/products/education/gemini-iste-2025/ ； 年终回顾 https://blog.google/outreach-initiatives/education/google-for-education-year-in-review-2025/

9. 科大讯飞智慧教育. "星火教师助手"（教学设计效率 +56.52%、课件制作 +64.18%、好评率 93% 、覆盖 4000+ 校 20 万 + 教师）. 2023—2025. https://edu.iflytek.com/solution/school/teachers-assistant ； 公司新闻 https://edu.iflytek.com/about-us/news/company-news/795.html

10. 科大讯飞智慧教育. "讯飞星火 V4.0 / 星火智能批阅机"（批改 90→5 分钟、学情分析 60→1 分钟）. 2024-06-27. https://edu.iflytek.com/about-us/news/company-news/1095.html

11. 好未来九章大模型（MathGPT）官方与介绍（数学千亿级模型、解题/讲题、信通院教育大模型评估 4+ 级）. https://www.mathgpt.com/ ； https://www.100tal.com/

12. 101 教育 PPT 官网（备授课一体化工具、覆盖各版本教材、免费资源）；网龙教育 AI 布局公开报道（970 万教师、装机量 3576 万口径）. https://ppt.101.com/

13. 欧盟《人工智能法案》第 5 条（禁止在工作场所与教育机构基于生物特征推断情绪，2025-02 起适用，最高罚款 7% 全球年营业额）与 Recital 44. https://artificialintelligenceact.eu/article/5/ ； Future of Privacy Forum 解读 https://fpf.org/blog/red-lines-under-eu-ai-act-unpacking-the-prohibition-of-emotion-recognition-in-the-workplace-and-education-institutions/

14. 中华人民共和国教育部. 《教师生成式人工智能应用指引（第一版）》（首份面向教师的生成式 AI 应用规范；关键育人环节由教师主导、企业依法数据分类分级/安全评估/算法备

案）. 2025-11. 专家解读见 https://www.edu.cn/xxh/focus/li_lun_yj/202512/t20251201_2704690.shtml

15. 教育部等五部门. 《"人工智能+教育"行动计划》（赋能教师教学、覆盖课前课中课后全环节；四大重点任务）. 2026. http://www.moe.gov.cn/jyb_xwfb/s271/202604/t20260410_1433232.html
16. 教育部. 《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. 关于 AI 生成教案质量、偏见与幻觉的实证研究综述（arXiv 2510.19866 高中物理教案评测；socialinnovationsjournal 教案生成器的教学法偏见；MIT Sloan 关于 AI 幻觉与偏见的教学指引；盲评中人工教案质量更高的结论）. 2024—2025. https://arxiv.org/pdf/2510.19866 ；https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias/

第 4 章　支持学习：机理、产品形态与发展建议

4.1 引言：从"答疑工具"到"学习伙伴"的范式跃迁

在 2024 版《生成式人工智能产品发展报告》中，"支持学习"场景主要围绕对话式大模型的即时答疑、知识讲解与作业辅助展开，产品形态以聊天窗口为核心交互界面。进入 2026 年，随着智能体（Agent）编排、检索增强生成（RAG）、长期记忆机制与端侧多模态能力的成熟，学习支持的技术底座与产品逻辑均发生了结构性变化。生成式人工智能不再只是被动响应提问的"答疑工具"，而正在演化为具备目标管理、过程陪伴、个性化适配与跨会话记忆能力的"学习伙伴"。

这一转型的理论坐标，可回溯到教育心理学家本杰明·布卢姆（Benjamin Bloom）1984 年提出的"两个标准差问题"（2 Sigma Problem）：接受一对一掌握式辅导的普通学生，其成绩可超越传统班级授课环境下约 98% 的同伴，效应量高达约两个标准差；然而一对一辅导的师资成本使其无法规模化普及。四十年来，教育技术界始终把"以可负担的成本把一对一辅导的效果规模化"视为核心命题。生成式人工智能之所以在 2024—2026 年引发教育界的高度关注，正是因为它第一次让"每个学习者都拥有一位随叫随到、可个性化的辅导者"从经济学上变得可能——尽管，正如本章后文将反复强调的，"技术上可行"距离"教育上有效且安全"仍有相当距离。

本章沿用"机理—产品—证据—建议"的分析结构，但在若干维度上作了重制升级：其一，机理层面从单轮问答的"知识调取"扩展到多轮、跨模态、有记忆的"学习过程建模"；其二，产品层面从对话式应用扩展到智能体化学习应用、端侧学习硬件（学习机、学练机、AI 眼镜等）与多模态学习环境，并以真实在售产品为例；其三，本版新增"辅导效果证据"专节，系统梳理 2024—2026 年围绕生成式 AI 辅导效果的随机对照试验（RCT）与元分析，区分"厂商宣称

"与"可核验证据"；其四，建议层面从"如何用好工具"上升到"如何在人机协同中保障学习者的主体性与认知发展"。需要说明的是，本章聚焦学习者视角（learning，学生侧）的支持形态，与第 3 章"赋能教学"（教师视角）、第 6 章"智能评价"存在天然的交叠与互补，读者可参照阅读。

4.2 支持学习的作用机理

4.2.1 认知支架：可调节的最近发展区

生成式人工智能对学习的支持，其核心机理可置于维果茨基"最近发展区"（Zone of Proximal Development）与"认知支架"（scaffolding）的理论框架下理解。传统学习支持受限于师资与资源，难以为每一位学习者提供恰好匹配其当前水平的引导。生成式模型通过对学习者输入的实时理解，能够动态生成难度可调、粒度可变的解释、示例、追问与提示，从而在学习者"已知"与"未知"之间搭建可伸缩的支架。

与传统自适应学习系统（如智能导师系统 ITS）相比，生成式模型搭建支架的方式发生了质变。ITS 依赖专家预先编写的知识组件、规则库与固定反馈模板，其支架是"离散、预置"的；而生成式模型可在推理时即时生成自然语言解释、类比与追问，其支架是"连续、生成"的——理论上可覆盖近乎无限的题目变体与提问方式。这一"连续支架"能力，正是它相较既有教育技术的核心增量：它把过去只能对少数高频题目预置讲解的"点状覆盖"，扩展为对任意输入即时应答的"面状覆盖"。

这一机理的关键在于**支架的动态撤除（fading）**：理想的学习支持不是持续给出答案，而是随着学习者能力提升逐步减少提示强度，最终实现独立完成。值得注意的是，2024—2026 年效果较好的辅导系统，几乎无一例外地把"不直接给答案、以启发式追问引导"这一教育学原则显式编码进了交互策略。Khan Academy 的 Khanmigo 明确设计为"不直接给出答案，而以苏格拉底式提问引导学习者逐步逼近解答"；哈佛大学团队自建的物理辅导智能体 PS2 Pal 亦将"

主动学习式的分步引导而非直接作答"作为首要设计原则（详见 4.5 节效果证据）；谷歌 DeepMind 面向教学微调的 LearnLM 也被监督教师评价为"擅长起草引发深度反思的苏格拉底式问题"。这标志着支持学习产品的机理设计，正从 2023 年的"有问必答"向"有教育学约束的引导"演进。

必须指出，"连续支架"的实现有其前提约束。生成式模型的支架质量高度依赖于提示工程（prompt engineering）与教学策略的显式编码：同一底层模型，在"直接作答"与"苏格拉底式引导"两种系统提示下，其教育价值可能截然相反。这解释了一个反直觉但被证据反复印证的现象——**模型能力的强弱，并不直接决定辅导效果的高低；把教育学原则编码进交互策略的工程质量，才是关键变量。** 这一判断贯穿本章始终。

4.2.2 个性化路径：从内容适配到过程适配

个性化是生成式人工智能支持学习的第二重机理。其演进呈现两个层次：

- **内容适配**：根据学习者的知识状态、兴趣与语言水平，调整讲解内容的表达方式、示例情境与呈现模态。例如对同一道数学题，为基础薄弱的学生给出更细的分步拆解，为学有余力者给出更凝练的思路提示。
- **过程适配**：在多轮、跨会话的学习过程中，依托长期记忆机制记录学习者的易错点、偏好与进度，据此调整后续的学习节奏与内容序列。

从内容适配到过程适配的跨越，依赖于 2025—2026 年逐步落地的**记忆（Memory）与用户建模能力**。在学生侧产品中，这一能力已有具体落地：猿辅导"小猿 AI"公开宣称其研发了"记忆模式"，对用户的年级、练习内容与练习结果形成记忆，"同一道题不同学生练习或提问，系统会根据过往使用数据匹配不同的讲解方式"。这类设计使个性化从"单次会话内的语气与难度调节"迈向"跨会话的学习画像驱动"。

过程适配的另一支撑，是把学情结构化为可计算的"知识画像"。国内学习机产品普遍以知识图谱为骨架，将学科拆解为细粒度的知识点，并通过学习者在诊断测评与日常练习中的表现，估计其对每一知识点的掌握概率，据此驱动"薄弱点定位—精准推题—讲解—再测"的闭环。生成式大模型的接入，使这一闭环中的"讲解"环节从"调取预置视频"升级为"按学习者具体错误即时生成的分步讲题"，从而把结构化的诊断能力与生成式的讲解能力结合起来。这也是2024—2026 年学习机从"题库机"向"AI 辅导机"叙事跃迁的技术内核。

但过程适配也带来新的机理性风险：记忆的准确性、时效性与隐私边界成为影响学习支持质量的关键变量。一旦用户模型对学习者水平的判断出现偏差，个性化反而可能把学习者"锁死"在错误的难度区间，形成"越练越窄"的负反馈；而对未成年学习者练习数据、情绪状态的长期留存，更直接触及数据最小化与隐私保护的合规底线（详见第 5 章"治理与安全"）。因此，个性化能力越强的产品，越需要配套的可解释性与可纠偏机制，让学习者与家长能看见并修正系统对学习者的判断。

4.2.3 知识锚定：以检索增强抑制幻觉

第三重机理关乎学习支持的可靠性底线：**知识锚定**。生成式模型的固有短板是幻觉（hallucination）——以流畅、自信的语气生成事实性错误。在闲聊场景中幻觉尚可容忍，在学习场景中却可能直接传播错误知识，其危害因学习者难以辨识而被放大。

检索增强生成（RAG）是当前缓解这一问题的主流工程路径。其机理是：在生成回答前，先从一个受控的知识库（教材、课程讲义、权威题解等）中检索相关片段，再要求模型基于检索到的材料作答，从而把回答"锚定"在可核验的外部知识上，减少对模型内部参数化知识的依赖。研究表明，将回答锚定于检索文本，可显著降低相较纯生成模型的幻觉率；在教育场景中，RAG 还带来两项附加价值：一是可注入最新的、领域特定的、经教师核定的知识，二

是可要求模型给出引用出处，使回答可溯源、可质疑。这与学习场景对"理论与数学正确性"的高要求高度契合。

对学生侧产品而言，RAG 的意义不仅在于降错，更在于**把知识边界与责任边界显性化**：挂载权威教材的辅导系统，其回答范围与依据变得可界定、可审计，这为面向未成年人的内容合规提供了工程抓手。国内教育垂类大模型普遍以自建学科题库与解题过程数据为知识源，正是这一机理在中国市场的具体形态。当然，RAG 并非万灵药——检索不到相关材料时模型仍可能"自由发挥"，且检索质量、知识库时效与冲突消解本身构成新的工程挑战。

4.2.4 多模态具身：贴近真实学习情境

第三重机理来自多模态与端侧化。学习并非只发生在文字对话中，而嵌入于阅读、书写、观察、动手操作、口语表达等具身情境。多模态大模型使系统能够"看见"学习者的书写、拍照上传的题目与实验操作、"听见"其朗读与口语表达，从而在更贴近真实学习活动的情境中提供支持。

这一机理在两类学生侧产品中体现得尤为直接。其一是**拍照解题类**：以 Photomath（2023 年被 Google 收购）为代表，通过光学字符识别与计算代数系统识别拍照上传的数学题，给出分步骤解题过程与动画讲解；国内学习机普遍内置类似的"拍照搜题—分步讲解"能力。其二是**口语陪练类**：Duolingo Max 的 Roleplay（角色扮演）与 Video Call（视频通话）功能，让学习者以语音与 AI 角色进行开放式对话，把"听说"这一高度具身的语言技能纳入 AI 支持范围。

口语陪练之所以是多模态机理的高价值应用，在于它命中了传统学习的结构性痛点：语言的"输出性"技能（说、写）远比"输入性"技能（读、听）更依赖高频、即时、无评判压力的练习机会，而这类机会在传统课堂中因师生比与开口焦虑而严重稀缺。生成式 AI 的语音对话把开口练习的边际成本降至近零，且 AI 不知疲倦、不作评判，恰好补上了这块最难规模化的短板——这也是语言学习成为学生侧生成式 AI 最早、最成熟落地场景之一的深层原因。

端侧化则降低了时延、保护了隐私、突破了联网限制，使学习支持得以延伸到课堂、书桌乃至户外等更广的物理场景。学习机类产品普遍采用"本地小模型+云端大模型"的混合部署以兼顾离线可用与算力；AI 眼镜等穿戴设备则把多模态感知推向"所见即所学"的免手交互形态。关于端侧多模态硬件的形态演进与安全评测，详见本院《AI 眼镜教育应用发展蓝皮书 2026》。

4.2.5 机理的两面性：赋能与依赖的张力

必须审慎指出，上述机理具有内在两面性。同一套即时、流畅、拟人的支持能力，既可能促进学习，也可能诱发**认知卸载（cognitive offloading）**与**思维惰性**——学习者过度依赖模型代为思考，反而削弱了自身的知识建构与元认知能力。

从学习科学看，这一张力有其深刻根源：有效学习往往需要"必要难度"（desirable difficulties）——提取练习、间隔重复、必要的困惑与试错，恰是长期记忆与迁移能力形成的关键。而生成式 AI 的核心卖点是"降低难度、消除摩擦"：它把原本需要学习者自己经历的检索、组织、纠错过程，替换为一次流畅的问答。当被消除的"摩擦"恰是学习赖以发生的"必要难度"时，即时便捷便与深层学习构成直接冲突。这正是 4.5.3 节将呈现的"更快完成、更少学习"现象的理论解释。

这一张力并非纯理论推演，而已有实证支撑。微软研究院与卡内基梅隆大学 2025 年一项针对 319 名知识工作者、936 个真实 AI 使用案例的调查发现：**对 AI 的信任度越高，使用者投入的批判性思维越少；而对自身能力的信心越高，批判性思维投入越多**——即 AI 有可能把认知努力从"解决问题"置换为"核验 AI 输出"，若核验也被省略，批判性思维便被整体外包。在学习场景中，这一风险因学习者尚处认知发展关键期而更为突出：有研究以"元认知懒惰"（metacognitive laziness）描述学习者把元认知负荷卸载给 AI、从而减少自身反思与调节的现象。

值得注意的是，风险的分布并不均匀。前述调查与相关研究提示，年龄越小、对自身能力信心越低者，越易陷入过度依赖；而受教育程度、AI 素养对认知卸载具有缓冲作用。这一发现有两层政策含义：其一，面向低龄学习者的产品尤须内建主体性保护，因为其自我调节能力尚未成熟；其二，AI 素养教育本身即是一种保护性干预，而非仅是"会用工具"的技能训练。因此，支持学习产品的机理设计不能仅以"响应质量"为目标，还须内建对学习者主体性的保护机制——这正是"支架撤除"原则之所以重要的深层原因。理想的学习伙伴，应当在"帮学习者渡过难关"与"保留必要难度以促成长"之间动态权衡，而非一味追求"最省力的答案"。这一"赋能与依赖"的张力，构成本章效果证据分析与发展建议的核心关切。

4.3 产品形态图谱（2026）

相较 2024 版以对话式应用为主的图谱，2026 年支持学习类产品呈现"应用智能体化、载体多元化、能力多模态化"的分化趋势。下表给出概览性分类框架，代表产品仅列本章已核实者。

产品类别	核心形态	典型能力	代表产品（已核实）
对话式/苏格拉底式辅导应用	App/网页对话界面	启发式答疑、分步引导、讲解	Khanmigo（Khan Academy）
语言学习与口语陪练	App+语音/视频对话	角色扮演、口语陪练、答案解释	Duolingo Max（Roleplay/Video Call）
拍照解题与分步讲解	拍照识别+计算引擎	数学分步解题、动画讲解	Photomath（Google）、各家学习机搜题
端侧学习硬件（学习机/学练机）	专用终端+本地/云混合模型	拍照搜题、分步讲题、学情诊断、家长管控	学而思 xPad、作业帮学习机、小猿学练机、科大讯飞 AI 学习机
教育垂类大模型	底层模型能力	学科解题、作文批改、口语对话	学而思"九章"、作业帮"银河"、猿辅导"猿力/看云"

AI 眼镜等穿戴设备	端侧多模态穿戴	情境化提示、实时翻译、免手操作	详见《AI 眼镜教育应用发展蓝皮书 2026》

4.3.1 从"对话"到"智能体"：学习应用的编排化

2026 年最显著的产品变化，是学习应用从单轮对话向智能体化编排演进。智能体化学习应用不再被动等待提问，而能围绕一个学习目标（如"两周内掌握某单元"）自主拆解任务、调用工具（检索、计算、代码执行）、跨多个会话追踪进度，并在过程中主动发起追问与复盘。RAG 使其能挂载权威、可控的知识源以降低幻觉；记忆机制使其能跨会话延续学习脉络。

Khanmigo 是这一路线中被研究与讨论最多的学生侧代表。它构建于 OpenAI 的 GPT-4 系列之上，但经过面向教育的定制约束——最核心的设计是"不直接给答案"，而以苏格拉底式提问引导学生自主推理，兼具答疑、写作陪练、口语练习等功能。2024 年 5 月，微软宣布通过捐赠 Azure OpenAI 基础设施，使"Khanmigo for Teachers"对全美教师免费开放（此前教师版收费约每月 4 美元）；迁移至 Azure OpenAI 服务后，Khanmigo 可调用 GPT-4o、GPT-4、Whisper、DALL·E 3 等多种模型。在采用规模上，截至 2023 年底约 45 个学区、4 万名学生参与试点；至 2024—2025 学年结束，美国已有约 795 个学区、77 万名学生使用。面向学生的账号仍向学区收费，据报道单价约每生每年 35 美元，Khan Academy 表示正努力将其降至 10—20 美元区间。需要强调：**采用规模并不等于学习效果**，Khanmigo 自身的因果效果证据仍在积累中（详见 4.5 节）。

从技术栈看，智能体化学习应用大致由四层构成：底层是大模型（通用或教育垂类）；其上是 RAG 知识层，把回答锚定于受控教材与题库；再上是记忆层，跨会话维护学习者画像；最上是编排与策略层，负责任务拆解、工具调用（检索、计算、代码执行、绘图）与教学策略（何时提示、何时追问、何时撤除支架）。四层中，真正区分"智能体化学习应用"与"套壳聊天机器人"的，恰是记忆层与策略层——前者决定它能否跨会话延续学习脉络，后者决定它是

"引导者"还是"答案机"。这也解释了为何本章反复强调"教育学设计"而非"模型能力"是效果关键：模型是四层中最易获得、最同质化的一层，差异化空间主要在其上三层。

在国内，学生侧的智能体化更多以"学练机/学习机内的 AI 辅导系统"和"教育垂类大模型"的形态出现，而非独立的通用对话智能体，这与国内对未成年人使用通用大模型的审慎监管取向有关。这一取向的合理性在于：未成年学习者的自我调节能力尚未成熟（见 4.2.5 节风险分布），把 AI 支持约束在内容合规、可管控的专用载体内，比放任其使用开放式通用助手更利于风险防控；其代价则是灵活性与能力上限的部分让渡。

4.3.2 端侧化与硬件化：学习支持走出屏幕

第二条主线是载体的多元化与端侧化，这在中国市场表现得尤为突出。学习机/学练机类专用终端凭借离线可用、家长管控、护眼与内容合规等特性，在家庭学习场景中形成独特定位；生成式大模型的接入，使其从"题库检索+视频点播"升级为"AI 分步讲题+学情诊断+个性化推题"。

市场规模上，据洛图科技（RUNTO）数据，2024 年中国学习平板（含 AI 学习机）全渠道销量约 592.3 万台，同比增长约 25.5%；销售额约 190.6 亿元，同比增长约 37.6%。2025 年增长延续：第一季度全渠道销量约 126.5 万台、同比增长约 29.4%；第二季度学习平板出货量约 154 万台、同比增长约 44.6%。据 IDC 数据，2025 年第二季度科大讯飞首次登顶 AI 学习机市场销售额第一，此前其已连续多年稳居高端市场首位。就整体市场份额，有咨询机构统计 2025 年初作业帮约占 31.8%、学而思约 20.9%、小猿约 11.7%（不同口径统计差异较大，具体以各机构原始报告为准）。国家以旧换新与教育硬件补贴（"国补"）被普遍认为对 2024—2025 年的需求释放起到刺激作用。

代表产品与其大模型能力：

- **学而思 xPad 系列**：搭载学而思自研的"九章大模型"（英文名 MathGPT）。九章由数学切入、以解题与讲题算法为核心，是国内首批通过备案的教育垂类大模型之一，现已扩展至全学段全学科，支持单题批改、作文辅助与批改、口语对话练习等；xPad 上线基于九章的"数学随时问"，官方称对中小学数学题"约 80% 可即问即答"，暂不能作答者最快 1 小时内上传真人讲解、20 分钟内生成 AI 视频解析。
- **作业帮学习机**：搭载自研"银河大模型"，覆盖多学科解题、多语言对话等场景。据报道，其智能学习机内置 8.5 亿以上题库，覆盖 K12 全学段，含同步课与 AI 专题互动课等资源。
- **小猿学练机（猿辅导）**：全系产品接入推理大模型 DeepSeek，并与自研的"猿力大模型"深度融合；猿辅导自研大模型（看云大模型）已通过备案。小猿 AI 主打"记忆模式"个性化，并首次引入"心理健康守护"功能——通过大模型识别用户情绪，以"共情—安慰—行动"路径提供情感支持。据公开报道，小猿学练机销量在约 16 个月内突破百万台。
- **科大讯飞 AI 学习机**：以讯飞星火大模型为底座，长期占据高端市场，2025 年 Q2 首度登顶行业销售额第一。

相关硬件的能力横评、雷达对比与安全评测，详见本院《AI 眼镜教育应用发展蓝皮书 2026》及本蓝皮书第 7 章。需提醒：上述"约 80% 即问即答""8.5 亿题库"等为厂商公开口径，宜与第三方评测结果对照理解，不应等同于独立验证的效果证据。

4.3.3 学科垂类与教育大模型：从通用到专用

第三条主线是垂类化。通用大模型在学科强逻辑任务（如数学分步推理、编程调试）上的可靠性仍有局限，且存在幻觉与内容合规风险，催生了面向具体学科的垂类工具与教育垂类大模型。国内三家头部教育企业均已推出并备案自研教育大模型（学而思"九章"、作业帮"银河"、猿辅导"猿力/看云"），并普遍叠加 DeepSeek 等通用推理模型以增强能力。其共同思路是：以

领域数据、过程性标注（不仅标注答案、更标注解题步骤）与专用评测，追求更高的学科正确率与"讲题"这一教学动作的适切性，而非仅仅"算对"。

在语言学科，垂类化则体现为**内容生产的规模化**。Duolingo 借助生成式 AI，于 2025 年一次性推出 148 门新语言课程，使其课程数量翻倍以上——据其官方说明，"前 100 门课程耗时约 12 年，而借助生成式 AI、共享内容系统与内部工具，约一年内即完成近 150 门课程的开发上线"。这一案例说明：生成式 AI 对学习支持的影响，不止于面向学习者的交互端，也深刻改变了教育内容的供给侧生产方式。

值得记录的是，这一供给侧变革伴随着显著的组织与舆论震荡。2025 年 4 月，Duolingo CEO 提出"AI-first"战略，计划逐步以 AI 替代可自动化的外包任务、并要求团队在证明工作已充分自动化前限制新增招聘，随即引发用户对"质量下降"与"以 AI 取代人"的强烈反弹，社交平台出现退订声浪。CEO 其后澄清不裁减全职员工、仅涉及少量从事重复性任务的小时工，招聘节奏不变。尽管争议一度轻微影响用户增长，公司当季营收仍超预期。这一插曲对教育产品研发具有超出个案的警示：**学习产品承载着用户对"人的关怀"的期待，"AI-first"的效率叙事若压过"以学习者为本"的价值叙事，可能招致信任成本**。

从教育公平的角度，生成式 AI 对学习支持的普惠潜力与新的鸿沟并存。一方面，Khanmigo 教师版经微软资助后对全美教师免费、Duolingo 免费层依然存在，都在降低优质学习支持的获取门槛，呼应了"以低成本规模化一对一辅导"的初心；另一方面，最强的能力（如 Duolingo Max 的视频通话、学区为 Khanmigo 学生账号支付的每生每年费用、高端学习机的价格）仍以付费为门槛，可能形成"能力可及性"上的新分层。这提示，评估一款学生侧产品的社会价值，不能只看其技术上限，还须看其把有效能力普惠给资源较弱学习者的能力。

相关教育垂类大模型的评测方法与结果，详见本蓝皮书第 7 章"教育垂类大模型评测"。

4.4 拍照解题、语言陪练与通用助手：三类学生侧交互形态

在智能体与硬件两条主线之外，从"交互形态"横切来看，学生侧生成式 AI 支持可归为三类高频形态，各有其机理侧重与风险画像。

其一，拍照解题与分步讲解。 以 Photomath 为原型。Photomath 采用增强型光学字符识别与计算代数系统，用户以手机相机扫描题目后，屏幕即呈现分步骤解题过程；其付费版另提供教材解答与动画讲解。2022 年 Google 宣布收购，经欧盟审查后于 2023 年完成交割，并于 2024 年初将其纳入 Google Play 发行体系，其识别与解题能力有望反哺 Google Lens、Search 等产品。这一形态贴近"作业辅助"的真实刚需，但也最容易滑向"抄答案"——机理上的价值取决于产品是引导理解还是直供结果，这也是国内学习机把"分步讲题""AI 视频解析"作为核心卖点、而非仅给最终答案的原因所在。

其二，语言学习与口语陪练。 以 Duolingo Max 为代表。其 Roleplay 让学习者与 AI 角色进行"在巴黎咖啡馆点单""讨论旅行计划"等情境化开放对话；Video Call（与虚拟角色 Lily 的视频通话）进一步把练习推向真实语速的语音交流；"Explain My Answer"则对学习者的作答给出针对性的准确性与复杂度反馈。这类形态的独特价值在于把"开口说"这一最难在传统 App 中练习的技能低成本化。

其三，通用对话助手用于学习。 大量学习者直接使用 ChatGPT、Gemini、DeepSeek 等通用助手完成查资料、改作文、写代码等任务。据 Common Sense Media 2024 年调查，约 70% 的美国青少年使用过至少一种 AI 工具，其中约四成用于学校作业类任务；Pew 研究中心 2023—2024 年调查亦显示约 26% 的美国中学生使用过 ChatGPT，且多数受访青少年认为"用它查资料"可接受、"用它代写作文"不可接受——这一"接受度分化"本身是宝贵的教育契机：学习者对"辅助"与"代做"的边界已有朴素直觉，学校的任务是把这种直觉转化为明确、可操作的使用规范。

值得注意的一个结构性张力是：**专用辅导应用与通用助手在设计取向上截然相反**。 Khanmigo、PS2 Pal 等专用产品刻意"不给答案、以引导为纲"；而通用助手的产品目标恰是"高效满足用户诉求"，学习者一句"帮我把这道题做出来"往往立即得到完整答案。因此，学习支持效果的高低，很大程度上取决于学习者实际使用的是哪一类工具、以何种方式使用。通用助手最灵活、能力上限最高，但也最缺乏教育学约束与内容护栏，与学业诚信、认知卸载的张力最为尖锐，构成学生侧治理的重点对象（详见第 5 章）。这也解释了国内在面向未成年人时，为何更倾向以"内置教育约束的学习机/学练机"而非"开放的通用对话助手"作为主要载体。

4.5 个性化与辅导效果：证据的谱系与边界

这是本版新增的核心一节。面对厂商普遍的效果宣称，本节主张严格区分不同强度的证据：**（a）平台使用与成绩的相关性证据；（b）严格设计的随机对照试验（RCT）证据；（c）汇总多项研究的元分析证据；（d）揭示潜在风险的反向证据**。 四类证据强度、适用边界各不相同，共同构成 2024—2026 年生成式 AI 辅导效果的真实图景。

4.5.1 相关性证据：使用越多、进步越多，但非因果

以 Khan Academy 为例。其 2022—2023 学年对约 35 万名 3—8 年级学生的研究显示：按推荐方式每周使用平台 30 分钟以上（全年约 18 小时以上）的学生，其学业进步比预期约高 20%，效应量约为 0.36（以 MAP Growth 测评衡量）；相关方法后经一项发表于《美国国家科学院院刊》（PNAS）的研究，通过控制教师与学生特征等未观测因素予以强化。**但须清醒指出两点边界**：第一，该效应属于"使用越多、进步越多"的相关性（辅以准实验控制），并非对某单一功能的干净因果归因；第二，也是最关键的，这些收益主要来自 Khan Academy 已积累十余年的**练习题与内容体系**，而非生成式 AI 组件 Khanmigo——把平台层面的效果直接归功于 Khanmigo，是一种常见但不成立的推断。截至本报告写作时，尚无针对 Khanmigo 本身、

经同行评议的因果效果研究公开发表；一项由 J-PAL 与多伦多大学在加拿大 6—8 年级课堂开展的 Khanmigo RCT，据报道结果预计在 2026 年中前后公布，值得持续关注。

4.5.2 因果证据：设计良好的 AI 辅导可显著优于优质课堂

真正提供因果证据的是若干小规模但设计严谨的 RCT。

哈佛大学物理辅导 RCT（Kestin 等，2025 年 6 月发表于《Scientific Reports》） 是迄今被引用最广的一项。研究采用交叉随机对照设计，194 名哈佛物理本科生分别接受"AI 辅导"与"优质主动学习课堂"两种处理。自建 AI 辅导系统 PS2 Pal 严格植入研究支持的教学原则：引导式分步作答而非直供答案、认知负荷分段管理、成长型思维语言、针对具体误解的个性化反馈，以及为抑制幻觉而设计的富提示。结果显示，AI 组学习收益翻倍以上（后测中位数 4.5 对 3.5），报告效应量约 0.73—1.3 个标准差；且 AI 组耗时更短（中位 49 分钟对 60 分钟）、参与度（4.1/5.0 对 3.6/5.0）与动机（3.4/5.0 对 3.1/5.0）更高。这一结果被视为"AI 辅导可在更短时间内达到甚至超过优质课堂"的有力例证。

但对该研究的解读须极为克制。 其自身与评论者列出的局限包括：仅两周、单次的短时干预，无法排除新颖效应（novelty effect）随时间衰减，也未测量长期保持；样本为哈佛高选拔性本科生，向社区学院、低龄学生或技术可及性较弱群体的外推存疑；聚焦布卢姆分类中"理解—分析"的中阶认知技能，未涉及高阶思维、协作与社会情感能力；且其对照是"优质主动学习课堂"而非平庸教学。换言之，这项研究证明的是"当 AI 辅导被严格按教育学原则精心设计时，短期内可以很有效"，而非"任意 AI 辅导都优于课堂"——**教育学设计的质量，而非 AI 本身，才是效果的决定变量。**

英国中学数学 RCT（2025 年 12 月，arXiv 2512.23633） 提供了更贴近真实课堂的补充证据。该探索性 RCT 覆盖 5 所英国中学、165 名学生，把面向教学微调的生成式模型 LearnLM 接入 Eedi 数学平台的聊天式辅导。研究采用"专家教师监督 AI"的稳妥设计：教师有权修改 AI 拟

发的每条消息直至满意，结果 AI 拟稿的约 76.4% 被零修改或极小修改地采纳；获得 LearnLM 支持的学生在后续新题上的解题正确率（66.2%）比仅由人类教师辅导者（60.7%）高约 5.5 个百分点。教师反馈特别称许其擅长生成引发深度反思的苏格拉底式问题，部分教师甚至表示从中学到了新的教学法。

综合上述，因果证据支持一个有条件的乐观结论：**经教育学微调、以引导而非直供为原则的 AI 辅导，能在受控条件下带来真实且往往显著的学习收益。** 两项 RCT 的一个共同启示是：它们真正拉开效果的，都不是"更强的模型"，而是"更好的教学设计"——哈佛的 PS2 Pal 靠的是植入的教学原则，英国的 LearnLM 靠的是教学微调与教师监督。这再次印证了 4.2.1 节的判断：教育学设计的质量，是效果的决定变量。

4.5.3 元分析证据：整体效应显著，但存在系统性偏倚

超越单项研究，元分析提供了更稳健但也需审慎解读的整体图景。Ma 等（2025，《Journal of Computer Assisted Learning》）对 34 项（准）实验研究的元分析报告，生成式 AI 对总体学习结果的合并效应量约为 0.68，其中认知维度（g≈0.80）与能力维度（g≈0.71）效应更强、情感维度（g≈0.51）中等；学科类型是显著调节变量，而干预时长、知识类型、先验知识水平等未见显著调节作用。其他元分析结论方向一致：多项综述报告生成式 AI 对学业成就、学习动机的中到大效应，动机维度亦有正向提升。

在"高阶思维"这一教育界最关切的问题上，证据更为微妙。有基于 29 项实验/准实验的元分析显示，生成式 AI 对学习者高阶思维（批判性思维、创造性思维、问题解决等）总体呈正向作用，但同时警示"对 AI 生成内容的过度依赖可能削弱自主学习与自我调节能力"——即正向与负向效应可能在不同使用方式下同时存在。尤其值得决策者关注的是**干预时长与效果之间的"倒 U 形"关系**：有研究发现约 8—16 周的中等时长干预效果最佳，过短不足以形成能力、过

长则可能积累过度依赖的负面后果。这一发现对产品与课程设计有直接启示：AI 辅导并非"用得越久越好"，而存在一个需要经验设计的最优剂量区间。

对元分析证据须持三点保留。其一，**发表偏倚**：阳性结果更易发表，可能整体高估效应。其二，**干预质量偏倚**：进入元分析的多是精心设计、短期、由研究者亲自实施的干预，与学习者在真实生活中散点式、无监督的日常使用相去甚远；把实验室效应量（如 0.68）直接外推到日常使用，是常见误读。其三，**测量口径**：多数研究以短期后测的知识/技能得分为结果变量，较少测量长期保持、迁移与对学习习惯的长远影响。因此，元分析可支持"在良好设计下生成式 AI 对学习有正向作用"的结论，但不足以支持"任意使用都有益"或"效应可长期维持"的强主张。

4.5.4 反向证据：便捷可能以牺牲学习深度为代价

与因果证据同等重要的，是揭示风险的反向证据。前述微软/CMU 关于"AI 信任度越高、批判性思维投入越少"的调查（4.2.5 节）已给出机理层面的警示；在学习任务上，亦有实验直接观察到"便捷换深度"的权衡。一项针对数学题练习的实验研究（arXiv 2605.21629，标题即"更快完成，更少学习"）发现：可用生成式 AI 时，学生完成数学题的用时显著缩短，但据此建立的知识却更少——即 AI 提升了"完成效率"，却未必提升、甚至可能损害"学习效果"。这与 4.2.5 节"必要难度"的理论解释互相印证：当 AI 抹平了学习者本应自己经历的检索与试错，效率的提升恰以学习的削弱为代价。

这一反向证据在真实世界中有其对应的行为图景。据 Common Sense Media 2024 年调查，约 70% 的美国青少年使用过至少一种 AI 工具，其中约四成用于学校作业类任务；相关统计亦显示，在承认使用 AI 的大学生中，用于作业者占比极高。当"用 AI 完成作业"从"辅助理解"滑向"替代思考"，其后果恰是反向证据所刻画的"更快完成、更少学习"。这提示：产品形态（引

导 vs 直供）与使用方式（过程性 vs 结果性）共同决定了同一项 AI 能力究竟促进还是侵蚀学习——技术本身在此是中性的。

语言学习方向的证据则较为正面但也需谨慎：一项针对 Duolingo Max 的两项前后测调查（合计 385 名学习者，学习法语或西班牙语）显示，使用 Roleplay 与 Explain My Answer 一个月后，学习者的自我效能感显著提升——研究一中原先不认同"我已准备好在真实情境中使用外语"的学习者，后测中转为认同者显著增多；研究二在七项自我效能指标中的六项上录得显著提升，86%—99% 的学习者认为 AI 功能有效支持了学习，63%—73% 表示把所学应用到了 App 之外的真实情境。但研究者明确承认：无对照组、依赖自我报告、存在参与者流失、且未直接测量实际语言水平。因此其只能作为"信心与主观投入提升"的证据，而非"客观语言能力提升"的证据——自我效能与学习结果虽相关，但不能等同。这一案例典型地说明了 4.5 节开篇所强调的证据分级之必要：满意度与自我效能调查，是三类证据中最弱的一档。

方法论提示：一份效果证据的核验清单。 评价一款学生侧 AI 辅导产品的效果宣称，可依次追问六个问题：（1）**证据类型**——是相关性、RCT，还是仅为满意度/自我效能调查？三者强度递减。（2）**证据来源**——来自厂商自评，还是独立第三方或同行评议？（3）**对照设置**——对照的是优质教学、常规教学，还是"无干预"？对照越强，结论越可信但效应通常越小。（4）**测量口径**——测的是即时后测得分，还是数周乃至学期后的长期保持与迁移？（5）**干预时长与生态效度**——是研究者亲自实施的短期精心干预，还是学习者日常真实使用？（6）**归因边界**——效果是否被恰当地归因到"被评估的那个 AI 功能"，还是被笼统归给整个平台？以本章案例映射：Khan Academy 的 0.36 属"相关性且归因于平台而非 Khanmigo"，哈佛与英国 RCT 属"因果但短期、样本受限"，Duolingo Max 的自我效能提升属"最弱的满意度类且未测客观能力"。厂商宣传中大量"提分""高效"的表述，若不满足上述条件，宜谨慎对待。

4.6 发展建议

4.6.1 面向产品研发者：把"教育学"编进系统

- **将支架撤除机制产品化**：以促进独立学习为目标设计交互策略，避免"有问必答"式的答案直供。Khanmigo 的苏格拉底式引导、PS2 Pal 的分步作答原则、LearnLM 的引导式提问，均已被证据证明是效果的关键来源，宜作为默认模式而非可选项；应提供分步提示、启发式追问与"先尝试后揭示"等模式，并把"是否让学习者先行尝试"作为可衡量的产品指标而非隐性策略。
- **以知识锚定守住可靠性底线**：面向未成年学习者的产品，宜以 RAG 等方式把回答锚定于经审核的教材与题解知识库，降低幻觉、支持出处可溯，并明确界定"知识库覆盖范围之外不作答或明确提示不确定性"的行为边界，避免模型在无据时自由发挥。
- **审慎设计记忆与用户建模**：在提升过程适配能力（如小猿"记忆模式"）的同时，明确记忆的可见、可编辑、可删除边界，保护学习者隐私，并防范用户模型误判把学习者锁死在错误难度区间（详见第 5 章）。个性化能力越强，越应配套可解释、可纠偏的界面，让学习者与家长能看见系统"如何判断我"。
- **以"最优剂量"而非"最长时长"为设计目标**：元分析提示干预时长与效果可能呈倒 U 形，产品不应把"日均使用时长"作为唯一增长指标，而应关注学习成效与依赖风险的平衡，必要时内建使用节律与"脱手"提醒。
- **以真实、分级的证据驱动迭代**：学科正确率、幻觉率、教学适切性等应通过第三方或公开可核的评测衡量；对外披露效果时，应明确区分相关性、RCT 与满意度证据，标注干预时长、对照组与测量口径，避免以平台整体效应或短期实验效应误导为产品因果效果。有条件者宜投入或支持独立 RCT，以真正的因果证据建立差异化信任。相关方法参见本蓝皮书第 7 章。

4.6.2 面向学校与教师：在协同中守护主体性

- **确立人机协同的边界与规范**：明确哪些学习环节适合 AI 支持、哪些必须由学习者独立完成，防范认知卸载对深层学习的侵蚀。英国 LearnLM RCT 中"教师监督 AI 输出"的模式，为课堂内负责任地引入 AI 辅导提供了一种可借鉴的稳妥范式。
- **培养学习者的 AI 素养**：将"如何提问、如何质疑、如何核验"纳入日常教学。研究提示，对自身能力有信心者更倾向保持批判性思维、受教育程度对认知卸载有缓冲作用——这意味着 AI 素养教育本身即是抵御过度依赖的保护性因素，应使学习者成为工具的驾驭者而非依赖者。

4.6.3 面向学习者与家长：从"要答案"到"要成长"

- 鼓励以"过程性使用"替代"结果性使用"，善用 AI 做检验、做陪练、做复盘，而非直接索取答案——尤其警惕拍照解题、通用助手代写作业等"便捷换深度"的高风险用法。
- 关注端侧硬件的用眼健康、使用时长与内容合规。面对国内学习机市场的功能同质化与高价争议，家长选购时宜穿透"AI 一对一""亿级题库"等营销话术，关注其是否真正提供"引导式讲题"而非"答案直供"，并结合第三方评测判断；家庭场景宜就使用时段与用途建立必要约定。

4.6.4 面向政策与治理：为学习支持划定安全底线

面向未成年学习者的生成式产品，宜在内容安全、数据最小化、算法透明与防沉迷等方面接受更审慎的治理约束。国内对教育大模型实行的备案制度（学而思"九章"、作业帮"银河"、猿辅导相关模型均已完成备案）为准入设立了基线，但备案侧重合规准入，尚不等同于对"教学有效性"与"儿童认知安全"的持续评估；后者需要独立、长期的效果与安全评测机制予以补充。

具体的治理框架、标准与合规要点，详见第 5 章"治理与安全"，以及本院相关标准化研究成果。

4.7 小结

2026 年，"支持学习"正经历从对话式答疑到智能体化、多模态、端侧化学习伙伴的深刻转型。其作用机理由单轮知识调取扩展为有记忆、可适配、贴近具身情境的学习过程建模，产品形态由聊天窗口分化为苏格拉底式辅导应用、语言陪练、拍照解题、学习机/学练机与教育垂类大模型等多元载体。

在效果层面，本章的核心判断是：**证据支持"设计良好的 AI 辅导可显著促进学习"，但决定性变量是教育学设计的质量，而非 AI 技术本身。** 哈佛与英国的 RCT 证明了引导式、经教育学微调的辅导之价值；而"更快完成、更少学习"的反向证据与认知卸载研究，则提醒技术能力的跃升始终伴随主体性弱化的张力。Khan Academy 的案例更警示：平台层面的相关性效应不能简单归因于其 AI 组件，采用规模亦不等于因果效果。

在中国市场，学生侧的落地更多以"学习机/学练机+教育垂类大模型"的合规化路径展开：2024 年学习平板全渠道销量约 592.3 万台、销售额约 190.6 亿元，2025 年增长延续；学而思、作业帮、猿辅导、科大讯飞等以自研并备案的教育大模型驱动"拍照搜题—分步讲题—学情诊断"的闭环。这一路径以内容合规与家长管控为特色，但也面临功能同质化、高价争议与"效果证据不足"的共同挑战——其真正的差异化，终将回到"是否以引导促成长"这一教育学本质上来。

本章据此主张：支持学习产品的价值，最终不取决于它能替学习者做多少，而取决于它能帮助学习者成长为怎样的独立学习者。把"支架撤除"编进系统、把"知识锚定"守住底线、把"证据分级"用于评价、把"主体性保护"作为红线，是这一场转型行稳致远的前提。这一判断，也为后续"智能评价"与"治理与安全"两章埋下伏笔。

本章参考来源

1. Kestin, G., et al. "AI tutoring outperforms in-class active learning: a randomized controlled trial."（哈佛大学物理辅导 RCT）· Scientific Reports · 2025 · https://www.nature.com/articles/s41598-025-97652-6 ；研究综述见 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 英国中学数学 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"（约 35 万学生、效应量 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 免费与采用规模）· 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"（学区收费口径、795 学区/77 万学生）· 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"（生成式 AI 规模化生产课程）· 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 自我效能前后测，385 人）· 2025 · https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2025.1499497/full

12. Wikipedia. "Photomath"（Google 收购与功能）· 2024 · https://en.wikipedia.org/wiki/Photomath；另见 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 信任度与批判性思维，319 名知识工作者）· 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. 量子位.《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. 智研咨询.《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》（洛图 RUNTO 数据，2024 年 592.3 万台 /190.6 亿元）· 2025 · https://www.chyxx.com/industry/1224843.html

17. 新浪财经.《学习机拼 AI，谁是赢家？》（作业帮/学而思/小猿市场份额）· 2025-07 · https://finance.sina.com.cn/tech/roll/2025-07-01/doc-infcyazn3425896.shtml

18. 极客公园.《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》（九章/MathGPT、xPad）· https://www.geekpark.net/news/337563 ；另见腾讯新闻《2024 服贸会｜大模型加持学习机》· 2024-09 · https://news.qq.com/rain/a/20240913A00E1O00

19. 中国日报网.《全面拥抱 AI 时代 小猿教育产品深度融合 DeepSeek 和猿力大模型》（小猿记忆模式、心理健康守护、备案）· 2025-02 · https://bj.chinadaily.com.cn/a/202502/19/WS67b5d47ba310510f19ee7ea5.html

20. 南方都市报.《垂类教育 AI 比拼什么？猿辅导王向东：瞄准三重 AI 布局》与《好未来 CTO 田密：通用模型技术狂飙，垂直模型必须更懂行业》（教育垂类大模型思路）· 2025 · https://m.mp.oeeee.com/a/BAAFRD0000202504161069982.html

21. Common Sense Media / Pew Research（青少年生成式 AI 使用与学业诚信态度调查，2024）· 转引自 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 "AI-first" 战略与舆论反弹）· 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 抑制幻觉、知识锚定于教师核定材料）· 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"（高阶思维正向作用、过度依赖警示、干预时长倒 U 形）· 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"（动机与投入元分

析）· 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 依赖、批判性思维与 AI 素养缓冲）· ScienceDirect · 2025 · https://www.sciencedirect.com/science/article/pii/S0001691825010388

第 5 章　支持教研：教研全周期的智能体化重构、产品形态与发展建议

5.1　教研场景的再定义：从"备课辅助"到"教研智能体"

教育研究（教研）是连接课程标准、教学实施与教师专业发展的中枢环节。在我国的教育治理结构中，它既包括校本层面的集体备课、听评课、课例研究、命题与作业设计、校本课程建设，也包括区县与省市层面的区域教研、学科教研员的循证指导与教研数据治理。相较于"支持教学"（第 3 章）直接面向课堂交付、"支持学习"（第 4 章）直接面向学生个体，教研场景的核心对象是**教师群体的专业实践**，以及**教学知识的沉淀、迁移与再生产**。它决定着一线课堂能否持续改进，也决定着优质教学经验能否跨越"个体—学校—区域"的边界扩散。

生成式人工智能进入教研，并非从 2026 年才开始，但其形态在近两年发生了实质跃迁。在早期（对话式大模型阶段），"支持教研"主要指以通用大模型辅助教师完成教案生成、素材检索、命题参考等离散任务——本质上是"更强的搜索与写作工具"。教育部工程研究中心（智能技术与教育应用教育部工程研究中心）与北京市数字教育中心 2025 年 7 月发布的《人工智能赋能基础教育应用蓝皮书（2025 年）》，把这一阶段的教研支持归纳为"以'智'助研"，指出其价值在于"整合教学与科研数据，支持精准分析与科学决策，减轻教师在资料整理、效果评估等事务中的负担"，并判断教师角色正由"经验推动者"向"数据驱动者"转变。这一判断抓住了方向，但仍以"任务辅助"为主基调。

2026 年，随着智能体（Agent）、多模态理解与端侧算力三条主线的成熟，教研支持的形态正从"单点问答工具"演进为"贯穿教研全周期的教研智能体（Research-oriented Teaching Agent）"。北京师范大学卢宇、汤筱玓在《电化教育研究》2025 年第 6 期发表的《生成式人工智能赋能

课堂教学的形态层级与进阶路径》中，构建了"四个逐级递进、相互关联的层级"框架：以"劳动替代与任务辅助"为基础形态、"能力增强与边界拓展"为初级形态、"人机协同与创新激活"为中级形态、"认知融通与思维塑造"为高级形态。虽然该框架直接面向课堂教学，但其"从工具赋能到智能体协同"的演进逻辑同样适用于教研——教研智能体的判别特征在于：

- **由被动应答转向主动编排**：智能体可依据教研目标自主拆解任务、调用工具（检索、批改、可视化、课堂数据分析），并对中间产物进行自检与迭代；这与卢宇等所述"具有较高自主性的教学智能体"在能力构成上同源。
- **由单模态文本转向多模态证据**：课堂录像、板书、学生作品、语音互动等成为可被理解与结构化的教研素材。这是 2026 年相较早期报告最显著的能力增量，也是"听评课从主观印象走向循证诊断"的技术前提。
- **由通用知识转向可溯源的领域记忆**：以检索增强生成（RAG）与长期记忆机制，将校本教学资源、课程标准与教研历史沉淀为可复用、可引用的知识库；RAG 的价值正在于"将生成过程锚定在检索到的文档上，使模型基于真实证据作答，并增强结果可追溯性"。

值得强调的是，这一演进不仅是技术叙事，也已进入国家政策的顶层部署。2026 年 4 月，教育部等五部门（教育部、国家发展改革委、工业和信息化部、科技部、国家数据局）印发《"人工智能+教育"行动计划》（教科信〔2026〕1 号），明确提出"利用智能技术分析课堂教学行为，开展人工智能循证教研实践，构建适应智能时代的教师研修模式"，并在教育治理侧部署"推动智能命题、智能组卷、智能监考、智能评卷等应用"，同时要求"研发教育评价智能化工具，探索开展学生学习全过程纵向评价、德智体美劳全要素横向评价"。"循证教研"由此从学术概念上升为政策术语，为本章讨论的产品形态提供了制度坐标。

需要辨明的是，本章所谓"教研智能体"并非要以一套自动化系统取代教研组的集体智慧，而是要在教研的每一个环节嵌入可被教师驾驭的智能能力。教研之所以是生成式 AI 落地的高价值场景，源于其三重特征：其一，**任务密度高而重复性强**——教师在备课、找资源、出题、

写观课记录、整理教研简报等事务上耗费大量时间，这些正是生成式 AI 擅长"起草"的部分；其二，**证据本可结构化却长期停留在主观印象**——听评课的"这节课互动不错""提问偏浅"等判断，长期缺乏可量化、可比较的证据支撑，而多模态理解恰好把课堂现场转化为可统计的证据流；其三，**知识本可沉淀却随人流失**——优秀教师的隐性经验难以显性化、难以跨校迁移，而 RAG 与长期记忆为"校本教研资产化"提供了新的技术载体。正是这三重特征，使教研成为四层能力（生成、多模态、检索记忆、智能体编排）能够同时发力、且能形成闭环的少数场景之一。

说明：本章聚焦生成式 AI 在教研环节的产品化机理与形态，硬件侧的第一视角课堂采集能力（如智能眼镜在听评课中的应用）详见本院《2026 AI 智能眼镜教育产业蓝皮书》。

5.2 赋能机理：四层能力如何作用于教研全周期

5.2.1 教研任务的结构化拆解

教研工作可沿"输入—加工—产出—反馈"的链条拆解为若干可被 AI 介入的子任务。下表给出典型教研环节与 AI 能力的对应关系。成熟度一列为定性判断，其依据是本章检索到的产品落地证据与权威报告表述（如《人工智能赋能基础教育应用蓝皮书（2025 年）》对智能备课、智能出题、智能组卷成熟度的分述），具体量化评测口径见 5.4 节。

教研环节	核心任务	主要 AI 能力	2026 典型范式	成熟度（定性）
集体备课	教学目标对齐、教案生成、分层设计	生成、结构化、知识对齐	RAG + 课标知识库	较成熟（资源检索/内容生成已规模化落地）
听课观课	课堂过程采集、行为标注	多模态理解、语音转写	端侧采集 + 事件抽取	快速成熟（硬件+大模型一体化已建示范区）

评课议课	证据归纳、维度化诊断	归纳、可视化、量规映射	课堂分析智能体	应用中（多指标报告已生成，标准待统一）
课例研究	迭代改进、经验沉淀	长期记忆、对比分析	教研记忆库	早期（跨周期记忆机制尚不成熟）
命题作业	试题生成、难度与信度参考	生成、约束满足	领域垂类模型	应用中（生成效率高但质量需教师核验）
区域教研	资源汇聚、教研数据治理	检索、聚合、脱敏	教研中台 + 智能体编排	早期（治理与合规要求高，标准化不足）

5.2.2 支撑教研的四层技术能力

· **生成与重构层**：以教育垂类或通用大模型完成教案、量规、命题、教研报告的初稿生成与风格重构。其机理是将教师的隐性经验以自然语言指令转译为结构化产物，降低"从零起草"的认知负荷。《人工智能赋能基础教育应用蓝皮书（2025 年）》指出，此类"教学内容生成"可"根据教师输入的主题、年级、学科等基础信息，自动生成匹配教学要素的教学内容，如试题、教学方案、课件等"，是"减轻教师备课负担"的关键。以网易有道"子曰"教育大模型为例，其 2025 年 8 月开源的子曰 3 数学模型宣称"覆盖全学科高频需求，实现备课、出题、批改、答疑的全流程、多角色赋能"，把生成能力贯穿到教研上游各环节。

· **多模态理解层**：对课堂录像、板书图像、学生作品与师生对话进行识别、转写与语义标注，把非结构化的课堂现场转化为可检索、可统计的教研证据。这是 2026 年相较早期报告最显著的能力增量。希沃（Seewo）课堂智能反馈系统即通过"智能设备实时采集课堂教学

行为、师生互动等多维数据"，生成学情画像与可视化诊断报告，把听评课从主观印象推进为可讨论的证据。

· **检索增强与记忆层**：通过 RAG 将校本资源、课程标准与教研历史接入生成过程，使输出"有据可依"；通过长期记忆机制跨教研周期保留教师偏好、班情学情与既往改进项。2025 年多份 RAG 综述指出，该技术"通过从外部数据库检索相关知识，增强生成的准确性与可信度"，并可"仅更新知识库即纳入新知识、实现快速迭代、增强结果可追溯性"——这三点恰对应教研对"可溯源、可更新、可复核"的刚性要求。对教研而言，RAG 的意义有三层：一是**贴地**，把通用模型接入本地课标与校本资源，解决"通用而不贴地"的问题；二是**可溯**，每一条生成内容都能回指到具体文档，使教研结论可被复核；三是**可更新**，教材改版、课标调整时只需替换知识库文档而无需重训模型。长期记忆机制则进一步把"单次会话"升级为"跨周期资产"——教师的分层偏好、某班的学情基线、上一轮课例研究的改进项，都可被保留并在下一周期复用。卢宇等亦指出，当前教学智能体已"初步实现多模态感知、检索增强生成、推理与规划等关键能力"，但对"教学资源的属性、关联与语义信息的精准解析能力有待提升"——这提示 RAG 在教研中的效果，很大程度上取决于校本知识库本身的结构化质量与标注粒度。

· **智能体编排层**：以规划—执行—反思的循环，将上述能力组织为可自动推进的教研工作流，并在关键节点交由教师审核（human-in-the-loop）。业界普遍将教育智能体描述为"以大语言模型为核心推理引擎、结合 RAG 处理教育知识库、支持文本/语音/图像多模态交互"的系统，并强调其"能够追踪学习轨迹、理解个体差异、基于历史数据持续优化策略"。以一次完整的课例研究为例，编排层可将"确定研究主题→采集与转写课堂录像→按量规抽取诊断维度→与历史课例对比→生成改进建议→回写教研记忆库"组织为一条可自动推进但节点可控的流水线：智能体负责"规划任务、调用工具、汇总证据、生成初稿"，教师负责"确认选题、复核诊断、定稿建议"。需要清醒认识的是，卢宇等明确判断"当前基

于大模型的教学智能体构建尚处于起步阶段"，现有智能体"对教学资源的属性、关联与语义信息的精准解析能力有待提升""对教学对象的行为模式、语言特征与潜在意图的识别精度尚需提高""对教学过程中各类互动行为、活动设计与目标达成的深度理解仍需加强"。这三处"待加强"恰恰是编排层在教研中最容易"越权下结论"的地方，因此其落地应遵循"高价值、低风险环节先行、关键节点必留人工审核"的稳健路径。

5.2.3 形态成熟度：四层能力沿教研全周期的差异化分布

若把卢宇等的四级形态框架映射到教研全周期，可以观察到明显的成熟度梯度。在**备课与资源检索**环节，能力已推进到"能力增强与边界拓展"的初级形态——生成与检索能力规模化落地，《人工智能赋能基础教育应用蓝皮书（2025 年）》称其可破解一线教师"备课之难"，通过对"教师的备课需求、教学目标、教学场景和学生学情等信息"建模，"动态排序形成匹配教师备课需求的资源推荐列表"。在**听评课与课堂分析**环节，能力正处于"人机协同"的门槛上——多模态采集与诊断报告已在示范区规模化生成，但"证据到结论"的推理仍高度依赖预设量规与教师解读。在**课例研究与区域教研治理**环节，能力仍停留在"任务辅助"与"能力增强"之间——跨周期的长期记忆、多校数据的脱敏聚合、共性问题的自主识别，都尚未形成稳定可靠的产品闭环。这一梯度提示：教研智能体的价值兑现将是**分环节、分阶段**的，试图一步到位构建"全自动教研大脑"既不现实、也不稳健。

5.2.4 机理的边界与前提

需要审慎指出，上述机理成立的前提是**教师主导、证据可溯、隐私合规**。生成式 AI 在教研中的作用是"放大专业判断"而非"替代专业判断"。这一定位与卢宇等强调的"建立起'人机协同'而非'机器替代'的应用理念与行为准则，充分发挥自身的主导作用"完全一致，也与《"人工智能+教育"行动计划》"有效防范利用人工智能伪造诈骗、学术造假、应试内卷、泄露隐私等问题"的底线要求相呼应。

从边界看，至少有三条前提不可动摇。**第一，教师主导不可让渡。** 生成的诊断结论、命题难度与改进建议均需教师复核——《人工智能赋能基础教育应用蓝皮书（2025 年）》直言智能出题"生成的题目、解析和答案等可能存在错误，需教师核验后方可用于实际教学"。教研的专业性恰在于对"为什么这样教""这节课问题出在哪"的价值判断，这一层不可外包给模型。**第二，证据链不可断裂。** 教研的说服力来自"结论可回溯到证据"。一个仅给出"本节课以封闭式提问为主"结论、却无法定位到具体时间戳与对话片段的系统，其结论既难被质疑也难被信服；RAG 与证据锚定的意义正在于此。**第三，隐私合规不可事后补救。** 其调用的课堂多模态数据涉及师生（尤其是未成年学生）隐私，须在采集端即遵循数据最小化与脱敏原则，而非在数据汇聚后再考虑合规（相关治理要求详见本蓝皮书第 7 章"治理与安全"）。这三条前提共同构成了教研智能体"可用、可信、可持续"的底座。

5.3 产品形态图谱：从对话工具到教研智能体

2026 年支持教研的产品可归纳为四类形态，呈现从"通用对话"向"领域智能体"迁移的总体趋势。需说明的是，下列产品均以厂商公开信息与权威媒体报道为据，本节仅陈述已披露的能力与规模，不对未经核实的市场份额或评测排名作断言。

5.3.1 对话式教研助手（通用底座）

以通用大模型为底座、面向教师的对话式助手，覆盖教案生成、素材检索、答疑与文案润色等高频离散任务。其优势在于零门槛与广覆盖，局限在于缺乏校本语境与课堂证据支撑。

国际上，可汗学院（Khan Academy）推出的 Khanmigo 面向教师端提供免费的"标准对齐的教案设计、学习目标与量规（rubric）、出场券（exit ticket）、分层教学与测验题"生成，官方称"原本需数小时的分层设计、教案、测验题、学生分组、导入活动等可在数分钟内完成"，其"教案设计器（Lesson Planner）"由"可汗学院教育者创建的、基于研究的结构"引导，"创建

导入（Create a Hook）"功能则提供故事、活动、诗歌、示例等多种情境化选项，并可为多语家庭邮件与班级简报生成初稿。Khanmigo 之所以值得作为标杆，不在其技术领先，而在其把"教师从事务性工作中解放出来、聚焦教学设计"的产品哲学落到了免费、开放、标准对齐的细节上。

国内以通用大模型（如讯飞星火、豆包、文心一言、通义千问、DeepSeek 等）为底座衍生的教师助手亦属此类，特点是通用能力强、迭代快、成本低，教师可通过自然语言直接获得教案框架、试题草稿与教研简报初稿。其共同短板是：**对国家课程标准与校本资源的对齐度，高度依赖提示词工程与外挂知识库**——通用模型并不"天然"知道某地某版教材某单元的课时安排与学情基线，若不接入本地资源，其产出容易"通用而不贴地"。这也解释了为什么产品形态会自然地向下一类"垂类模型 + 教研中台"演进。

5.3.2 教育垂类大模型与教研中台

面向 K12 与教师专业发展训练或微调的教育垂类大模型，强调课标对齐、学科知识准确性与教研合规性，通常以"教研中台/平台"形态部署于区域或学校，对接资源库与教研管理系统。代表性产品包括：

- **科大讯飞星火教育大模型**：2025 年迭代至讯飞星火 X1.5 等版本，以"懂教育"为定位，其"星火教师助手"面向教师"提供建议性的课程大纲和教学方案，帮助教师减少查找资料、整理信息和制作素材的时间"，并在讯飞 AI 黑板等硬件上实现"备课、授课、评测等全链路智能化"；2025 年 9 月，华为与科大讯飞发布"星火教育、医疗大模型场景一体机解决方案"，将昇腾算力与星火大模型融合，指向本地化、可控化部署。
- **网易有道"子曰"教育大模型**：2025 年 8 月 20 日发布多款 AI 新品，并提出教育 AI 应用能力 L1–L5 分级——据新华网报道，业界当前正"从 L3 主动学习辅导加速迈向 L4 虚拟老师（已具备接近人类教师的思考能力）"。这一 L1–L5 分级与本章借用的四级形态框架异

曲同工，都试图为"教育 AI 到底做到了哪一层"提供可对话的坐标。在教研上游能力上，其 2025 年开源的子曰 3 数学模型宣称"覆盖全学科高频需求，实现备课、出题、批改、答疑的全流程、多角色赋能"；子曰 3.0 小语种模型支持 38 种语言实时互译。合规维度上，子曰教育大模型通过中国信息通信研究院（信通院）可信 AI 教育大模型评估，获当时最高评级 5 级——这是少数进入第三方合规评价视野的教育垂类模型。

- **好未来"九章大模型（MathGPT）"**：定位为面向全球数学爱好者与科研机构、以解题与讲题算法为核心的数学垂类大模型，是国内首批通过备案的教育大模型之一。官方称其为教师提供"设计课程、布置作业和进行学生评估"的工具支撑，依托学而思多年积累的教研教学数据与用户数据训练。数学学科对推导严谨性、解析正确性的高要求，使 MathGPT 成为观察"垂类深耕能否换来命题与解析质量"的典型样本——而这恰是通用模型在理科复杂题上的公认短板。

垂类模型的合规底座亦在完善：截至 2024 年底，已有约 302 款生成式人工智能服务在国家网信办完成备案（其中当年新增约 238 款）；到 2025 年，备案数量持续增长（据中央网信办公告，累计已备案的生成式人工智能服务规模进一步扩大），教育大模型是其中活跃的一类。合规备案是产品进入 K12 校园的前置门槛，也是本章在陈述任何教育大模型时首先核对的事项。需要提醒的是：各产品的参数规模、训练语料构成与第三方评测口径，应以厂商与评测机构公开信息为准，凡未公开者本章不作推断，更不将厂商自报的"准确率""覆盖率"等同于第三方结论。

5.3.3 多模态课堂分析与听评课产品

以课堂录像与师生互动为输入，输出结构化的课堂行为分析、师生对话比例、提问认知层级、教学环节时序等诊断维度，服务于听评课与课例研究。此类产品在 2026 年因端侧算力与多模态模型进步而快速演进，是本章判断中"落地最扎实"的一类教研支持产品。

以希沃（Seewo）课堂智能反馈系统为代表，其技术构成是"软硬件结合"：以 AI 摄像头（据报道采用四镜头方案）无感采集课堂音视频，以教学大模型进行数据深度分析，从而生成"精准学情画像"与可视化诊断报告。在教师侧，系统对"提问、追问、引导"作精准统计，并"基于教师个体差异提供个性化提问建议"；在学生侧，系统统计抬头率、举手率与参与度，"客观呈现课堂互动与投入水平"。这一诊断维度体系，本质上是把弗兰德斯互动分析系统（FIAS）等经典课堂观察编码的人工工作，交由多模态模型自动完成——从而把过去需要教研员逐帧标注、耗时数小时的课堂分析，压缩为课后自动生成的一份报告。

规模层面，据希沃 2026 年 1 月披露，截至 2025 年 12 月 31 日，该系统已建成 19 个重点应用示范区，覆盖超 5600 所学校、超 1.7 万间教室，逾 9.7 万位教师累计生成超 65 万份反馈报告（相较其 2025 年 6 月披露的 3000 余所学校、7000 余间教室、约 36 万份报告，半年内实现了约一倍的增长）。这一增速从侧面印证：多模态课堂分析是当前教研 AI 中需求最刚性、复购最明确的品类。相关实践并入选教育部教育技术与资源发展中心（中央电化教育馆）2025 年教师人工智能应用案例（详见 5.5 节情境二）。

需要说明的是，此类产品的诊断维度虽已相当丰富（浙江部分落地案例宣称"每节课 200+ 项数据指标"），但其价值高度依赖两点：一是**指标与教研目标的对齐**——统计出"提问 62% 为封闭式"只有在指向"增加开放性提问"的可操作建议时才有意义；二是**证据的可回溯**——诊断结论须能定位到具体时间戳与对话片段，方能支撑循证议课。第一视角采集形态可与智能眼镜互补，为教师视角、板书细节、个别辅导等固定机位难以覆盖的场景提供补充证据（详见本院《2026 AI 智能眼镜教育产业蓝皮书》）。

5.3.4 教研智能体与编排平台

以智能体编排、RAG 与长期记忆为核心的新一代产品，能够自主推进"选题—证据采集—诊断—改进—沉淀"的教研闭环，并在节点交由教师审核。这类产品目前多处于早期落地阶段，其可靠性、可解释性与教师接受度仍待系统评测（见 5.4 节）。

业界对教育智能体的典型架构描述为"以 LLM 为推理引擎 + RAG 处理教育知识库 + 多模态交互 + 历史数据驱动的策略优化"，其价值定位被反复强调为"不在替代教师，而在支持教师，将重复性工作转移出去，让教师专注于教学设计与情感关怀"。卢宇等在其框架中进一步描绘了"多元智能体协同"的图景：在课堂教学中，"助教智能体"负责知识答疑与个性指导，"学习同伴智能体"模拟协作学习中的多样化角色，"元认知导师智能体"实时监测项目进展、分析各主体参与度并给出反馈——这套多智能体分工同样可迁移到教研，例如"检索智能体"负责调取校本资源与课标、"分析智能体"负责从课堂证据中抽取诊断维度、"沉淀智能体"负责把定稿的改进方案回写记忆库。标准侧，2025 年世界数字教育大会发布的"教育大模型总体参考框架"联盟标准，为教育智能体从实验室走向规模化提供了权威指引，被视为教育智能体"从实验室走向大规模应用的关键节点"。

国内教育信息化企业也在既有备授课平台上叠加智能体化能力。以网龙网络（港股 HK:0777）旗下 101 教育 PPT 为例，其在备授课一体化平台上叠加了"AI 助教"（可选择或自定义助教形象、添加语料创建个性化内容、在授课中执行课堂指令）与升级版"备课台"（可直接创建课件、获取备课资源与模板）等能力；官方称 101 教育 PPT 已为全国约 970 万教师提供超 77 万个免费教案、课件、微课等教学资源。这一庞大的教师侧存量与资源库，为教研智能体的资源沉淀、记忆库建设与个性化推荐提供了现实基础——对教研智能体而言，"有多少可检索、可复用的优质校本资源"往往比"底座模型有多强"更能决定产品的实用价值。总体看，教研智能体是四类形态中想象空间最大、但成熟度最低的一类：其编排自主性越强，对证据可溯与

人工审核的要求就越高，稳健落地的关键在于"先自动化确定性高的环节、把不确定性留给教师"。

产品能力雷达（拟纳入维度）：课标对齐度、多模态支持、证据可溯性、编排自主性、隐私合规、教师可控性。各产品在上述维度的实测评分需以第三方专项评测数据为据，见 [待补：横评数据/来源]，能力雷达图见图 5-1。

5.4 评测与横评：让"支持教研"可被检验

延续本蓝皮书"循证评测"的定位，本节给出面向教研场景的评测框架，并尽量援引已公开的评测口径与标准。教研类产品的评测不能照搬通用大模型榜单，而应围绕"教研到底有没有变得更好"设计维度，至少覆盖：**课标对齐度**（生成内容是否符合国家课程标准与教材版本）、**学科事实准确性**（尤其命题与解析的正确率）、**证据可溯性**（诊断结论能否回溯到原始证据）、**诊断一致性**（AI 课堂分析与专家人工标注的吻合度）、**编排可靠性**（智能体任务完成率、工具调用正确性、幻觉率与人工干预频次）、**教师可控性与接受度**（教师能否理解、修正与信任其产出）六个方面。

评测的标准化基础正在形成。 2025 年发布的国家标准 GB/T 45288.2—2025《人工智能 大模型 第 2 部分：评测指标与方法》，以及认知智能全国重点实验室（科大讯飞牵头）发布的《通用大模型评测体系 2.0》，为大模型能力评测提供了通用指标与方法框架；面向教育行业的评测则进一步"覆盖 K12 多学科知识能力，对齐我国教育体系，评估模型在智能备课内容生成、个性化学习路径规划等场景的表现"。信通院的可信 AI 教育大模型评估（如前述子曰获评 5 级）则提供了合规与可信维度的第三方参照。

- **教育垂类大模型评测**：在课标对齐、学科事实准确性、命题质量、教研报告可用性等任务上设计题库与评分量规，报告分数、样本量与评测口径。此处尤须警惕"高准确率"表述

的口径差异——不同厂商宣称的课标理解准确率、答疑准确率往往基于各自私有测试集，缺乏横向可比性，具体数据 [待补：第三方评测数据/来源]。

- **多模态课堂分析评测**：以标注课堂录像为基准，评估行为识别、对话归因与环节切分的准确率与一致性。可参照弗兰德斯互动分析系统（FIAS）等经典编码体系作为标注基准，评估 AI 生成的师生对话比例、提问认知层级分布与人工标注的一致性 [待补：评测数据/来源]。
- **教研智能体评测**：评估任务完成率、工具调用正确性、幻觉率与人工干预频次。鉴于卢宇等指出智能体"对教学过程各类互动行为、活动设计与目标达成的深度理解仍需加强"，此类评测应重点考察智能体在缺乏充分证据时是否会"越权下结论"，以及其诊断结论能否回溯到原始证据 [待补：评测数据/来源]。

专题：智能出题/组卷的质量评测尤须审慎。 命题是教研中对"事实正确性"要求最高的环节，也是最容易被"高效率"表象掩盖"质量缺口"的环节。《人工智能赋能基础教育应用蓝皮书（2025 年）》对此有相当冷静的判断：现有智能出题系统虽已支持"选择学段、科目、题型、难度、知识点覆盖范围"等模板化生成，甚至"接入大模型支持文档自动分析、提取核心知识点并生成试题"，但"生成的题目、解析和答案等可能存在错误，需教师核验后方可用于实际教学"，且"尚未真正契合教学实际场景"。该报告进一步指出三处待优化方向：其一，在生成类型上应"聚焦高阶思维、情境化和跨学科试题生成"，而非停留于口算题、填空题等低阶题型；其二，在解析生成上应"着力攻克包含公式推导、特殊解法及几何图形等数学、物理等理科中的复杂难题"，构建"科学严谨、逻辑严密的解析生成规范"；其三，在难度控制上应基于学生认知轨迹"生成涵盖各种难度层次、强化考查个性化高阶思维能力的试题"。智能组卷则可借助"遗传算法、蚁群算法等多目标优化"与知识图谱技术，实现"难度、知识点覆盖率、题型分布"多约束平衡——但"约束满足"意义上的"科学组卷"，并不等于"有育人价值的命题"。因此，面向命题的评测不应只报告"生成速度"与"知识点覆盖率"，更应引入教育测量学的难度、

区分度、信度等指标，并以学科教育专家的人工评审作为质量金标准。相关横评数据 [待补：命题质量第三方评测/来源]。

评测结果时间线：不同产品在关键能力上的迭代节点与版本演进见图 5-2 时间线。所有评测均须标注测试时间、模型版本与数据集，避免以单次结果作绝对结论——这也是本蓝皮书对一切"榜单第一""准确率 99%"类表述保持克制的原因。事实上，本章检索过程中即见到个别渠道宣称某产品"课标理解准确率 98.5%""答疑准确率 99.8%"，但此类数字往往基于厂商私有测试集、缺乏公开口径与横向可比性，本章对其一律仅作"厂商自述"处理，不纳入定量结论。

5.5 典型应用情境（循证叙事）

本节以已公开、可核验的真实案例为主，佐以框架性叙述，力求避免虚构。

- **情境一·校本集体备课与资源沉淀**：教研组以教研助手接入本校历年教案与课程标准，围绕某单元生成分层教案初稿，教师在其上修改并回写，形成可复用的校本教研记忆。一个理想的工作流是：备课智能体先通过 RAG 检索本校该单元的历年教案、优质课件与对应课标要求，据此生成"教学目标—重难点—分层活动—对应习题"的结构化初稿；教师在初稿上按班情学情作删改与本土化调整；系统再把定稿及其修改痕迹回写记忆库，使"下一次备同一单元时更懂本校"。此过程中 AI 承担"起草与对齐"，教师承担"判断与定稿"，二者边界清晰。

《人工智能赋能基础教育应用蓝皮书（2025 年）》将此类能力描述为破解一线教师"备课之难"的资源高效检索与精准推荐，其核心技术是"对教师的备课需求、教学目标、教学场景和学生学情等信息"进行多维建模，并"动态排序形成匹配教师备课需求的资源推荐列表"；该报告特别强调需整合国家中小学智慧教育平台等公共资源与学校校本库，实现公共资源与校本资源的双轮驱动。产品侧，科大讯飞"星火教师助手"以"为教师提供建议性的课程大纲和教学方案、减少查找资料与制作素材的时间"为卖点，其星火教育大模型（2025 年迭代至 X1.5 等版

本）主打"懂教育"；以网龙 101 教育 PPT 为代表的备授课平台，其面向约 970 万教师的海量共享教案与课件，则为这一情境提供了现实的资源素材基础。

· **情境二·多模态听评课与循证改进**：这是 2026 年落地最扎实、也最能体现"循证教研"内涵的情境。浙江省金华永康市花川小学引入希沃 AI 课堂智能反馈系统，每节课生成 200 余项数据指标，系统记录教师的每一次提问、走动轨迹与互动瞬间，生成实时反馈报告并提出教学改进建议，被教师亲切地称为"AI 教研员"。

一组可核验的细节，恰好完整呈现了"循证改进"的因果链：系统捕捉到某教师平均每 3 分钟提问一次、其中 62% 为封闭式问题，据此自动生成优化建议"增加开放性提问比例"；教师据此调整提问策略后，该班学生课堂主动发言率从 35% 提升至 68%。这条"证据→诊断→建议→改进→再测量"的链条，正是把过去"评课凭印象、议课凭经验"的传统教研，转变为"用数据说话"的循证教研。学校由此构建了"证据收集—数据分析—问题诊断—教学改进—迭代优化"的教研全流程闭环，形成"数据驱动—问题导向—分层进阶"的教师发展新样态，推动约 40 位教师从"经验型"向"循证型"转变。这与《人工智能赋能基础教育应用蓝皮书（2025 年）》所述教师由"经验推动者"向"数据驱动者"转变的判断高度吻合，也是《"人工智能+教育"行动计划》"开展人工智能循证教研实践"的一个鲜活注脚。

希沃相关实践并入选教育部教育技术与资源发展中心（中央电化教育馆）2025 年教师人工智能应用案例征集活动，案例名称如《构建智慧教研新范式：利用 AI 推动教师教学行为的循证改进》《GenAI 驱动的循证教研在教师课堂问题诊断与改进中的实践研究》《希沃 AI 课堂反馈系统赋能小学语文"教—研—评"一体化实践案例》。这些案例名称本身即勾勒出多模态课堂分析在教研中的三条落地主线：教师教学行为的循证改进、课堂问题的诊断与改进、以及"教—研—评"一体化。

· **情境三·区域教研数据治理**：区域教研中台在脱敏前提下汇聚多校教研数据，借助智能体完成主题聚类与共性问题识别，为区域教研选题与教研员的循证指导提供依据。例如，

当某区多所学校的课堂分析报告都指向"高阶提问不足""小组合作流于形式"等共性问题时，区域中台可据此确定学期教研主题，并把优质课例定向推送给需要改进的教师，实现"从个案改进到区域提质"的放大。希沃在全国建成的 19 个重点应用示范区，正体现了这种"从单校到区域"的组织形态。《"人工智能+教育"行动计划》提出的"人工智能循证教研实践"与"构建适应智能时代的教师研修模式"，为这一情境提供了政策空间；但区域级数据汇聚对隐私合规与安全防护要求更高，涉及跨校、跨主体的数据流动，须"分类分级确定安全防护标准"，并防范"泄露隐私"（详见第 7 章）。该情境目前多处于示范区试点阶段，具体区域的量化落地成效 [待补：区域案例/来源]，本章不作虚构陈述。

5.6 挑战与风险

- **事实性与幻觉**：垂类模型在学科知识与命题上的错误具有隐蔽性，可能误导教研判断。《人工智能赋能基础教育应用蓝皮书（2025 年）》明确指出智能出题"生成的题目、解析和答案等可能存在错误，需教师核验后方可用于实际教学"，并坦承现有系统生成的试题"尚未真正契合教学实际场景"，在高阶思维、情境化与跨学科命题上仍有差距。事实性风险须以评测与教师复核双重把关。
- **证据可溯性不足**：部分产品的诊断结论缺乏可追溯的原始证据链，削弱教研的说服力与可复核性。RAG 与证据锚定是缓解路径，但并非所有产品都提供"结论—证据"的可回溯映射。
- **命题质量与"应试内卷"隐忧**：智能出题、智能组卷在提升效率的同时，若一味追求题量与覆盖，可能强化机械刷题。《"人工智能+教育"行动计划》专门将"防范应试内卷"列为底线；智能组卷虽可通过遗传算法、蚁群算法等多目标优化实现难度—知识点—题型的多约束平衡，但"科学的组卷"不等于"有育人价值的命题"，须以核心素养为导向。

· **隐私与合规**：课堂多模态数据涉及未成年人与教师隐私，采集、存储与使用须符合数据最小化与合规要求，并"有效防范……泄露隐私等问题"（详见第 7 章）。

· **教师负担的再平衡**：工具引入若缺乏流程再造，可能产生"为使用而使用"的形式化负担，背离减负初衷。真正的减负来自流程闭环（如花川小学的五环节教研闭环），而非工具堆叠。

· **评价标准缺位与口径不可比**：面向教研的 AI 产品尚缺乏公认的能力基准与评测口径；虽有 GB/T 45288.2—2025《人工智能 大模型 第 2 部分：评测指标与方法》等通用标准，以及世界数字教育大会"教育大模型总体参考框架"联盟标准起步，但面向"教研"这一特定场景（课标对齐、命题质量、课堂诊断一致性）的横评基准仍待建立，厂商自报的准确率数据横向可比性弱，这使得学校与区域在选型时缺乏可信的比较依据。

· **模型版本漂移与结论可复现性**：教育大模型迭代频繁（本章检索期内即见到星火、子曰等在数月内多次版本更新），同一提问在不同版本下可能给出不同诊断或命题结果。这对教研的"可复现性"构成挑战——若一份教研结论无法标注其所依赖的模型版本与数据集，其学术价值与可追溯性将大打折扣。因此，评测与教研留痕都应记录模型版本、时间与口径。

· **过度依赖与专业能力退化**：若教师长期将备课、命题、观课诊断"外包"给 AI 而疏于自主判断，可能导致教学设计能力与教研敏感度的退化。卢宇等强调应"发展创新意识和能力，鼓励师生深入探索符合学科特点的人机协同教学模式"，其潜台词正是：AI 应成为教师专业成长的"脚手架"而非"替代物"，教研的最终目的仍是教师本身的专业发展。

5.7 发展建议

1. **以"教师主导、证据可溯"为红线**：明确 AI 在教研中的辅助定位，所有诊断与改进建议须保留可追溯证据链并经教师复核。这与卢宇等"人机协同而非机器替代"及《"人工智能+教育"行动计划》的底线要求一致。
2. **推动教育垂类模型的课标对齐与评测公开**：鼓励厂商公开评测口径与样本，依托 GB/T 45288.2—2025 等国家标准与教育大模型总体参考框架，建立面向教研的公共基准，供第三方（如中央电化教育馆、信通院及相关专业机构）组织评测，破解"各家准确率互不可比"的困局。
3. **优先发展多模态课堂证据能力**：将听评课从主观印象推进为循证诊断（如希沃反馈系统、"AI 教研员"实践所示），并与第一视角采集硬件协同（详见本院《2026 AI 智能眼镜教育产业蓝皮书》）；同时以 FIAS 等经典编码体系为标注基准，保障 AI 诊断的科学性与一致性。
4. **以校本记忆库沉淀教研资产**：通过 RAG 与长期记忆把教案、课例与教研历史转为可复用、可引用的知识资产，避免经验随人流失；充分利用国家中小学智慧教育平台等公共资源与学校校本库，实现"越用越懂本校"。
5. **将治理与安全前置到产品设计**：在数据采集、脱敏、留存与授权环节内建合规能力，落实"分类分级确定安全防护标准"（详见第 7 章"治理与安全"）。
6. **在命题与组卷上坚持核心素养导向**：推动智能出题从"知识点覆盖"走向"高阶思维、情境化、跨学科"的育人价值命题，加强与学科教育专家协作，防范"效率提升异化为应试内卷"。
7. **稳健推进智能体编排落地**：在高价值、低风险的教研环节先行试点智能体自动化，保留关键节点的人工审核，逐步扩展；正视"教学智能体尚处起步阶段"的现实，避免对编排自主性作过度承诺。

8. **把教师专业发展作为最终尺度**：一切教研 AI 的成效，最终都应以"教师是否真正成长、课堂是否真正改进"来衡量，而非以工具使用频次或报告生成数量来衡量。应将 AI 循证教研纳入教师研修体系与智能素养培训（《"人工智能+教育"行动计划》已提出"制定教师智能素养标准""构建情境化测评系统"），使教师既会用工具、又不被工具替代，最终实现"人机协同、教师主导"的教研新样态。

本章参考来源

1. 卢宇、汤筱玮．《生成式人工智能赋能课堂教学的形态层级与进阶路径》．《电化教育研究》2025 年第 6 期（总第 386 期）．北京师范大学教育技术学院 ．DOI:10.13811/j.cnki.eer.2025.06.010 ．https://aic-fe.bnu.edu.cn/docs/2025-07/a4d9baa90d9e49d9920ee2c46dd283b7.pdf
2. 智能技术与教育应用教育部工程研究中心、北京市数字教育中心（北京电化教育馆）．《人工智能赋能基础教育应用蓝皮书（2025 年）》．2025 年 7 月．https://vmcl.bnu.edu.cn/docs/2025-07/3a416f18f01b42f998062babd78b9ca8.pdf
3. 教育部、国家发展改革委、工业和信息化部、科技部、国家数据局．《"人工智能+教育"行动计划》（教科信〔2026〕1 号）．2026 年 4 月．中华人民共和国教育部政府门户网站．http://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html；全文另见中国教育在线 https://www.eol.cn/zhengce/wenjian/202604/t20260410_2727386.shtml
4. 新华网．《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》．2025 年 8 月 21 日．http://www.news.cn/tech/20250821/11dbed39f03b4b3aa0723e0a2ce910d9/c.html
5. 科大讯飞智慧教育．《讯飞星火 X1.5 发布：懂教育、更懂你》（公司新闻）．2025 年．https://edu.iflytek.com/about-us/news/company-news/2535

6. 华为．《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"》．2025 年 9 月． https://e.huawei.com/cn/news/2025/industries/education/spark-education-medical-large-model

7. 希沃（Seewo）．《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》．2026 年 1 月．https://www.seewo.com/article/detail/2889

8. 慧聪教育网．《每节课 200+ 项数据指标？浙江"AI 教研员"让每一节课都有进步》．2025 年 8 月． https://edu.hczyw.com/2025/0801/77901.html （另见希沃 https://www.seewo.com/article/detail/2797）

9. Khan Academy．《Khanmigo for teachers: Free, AI-powered teacher assistant》．2025 年．https://www.khanmigo.ai/teachers ；相关教案能力见 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. 好未来九章大模型（MathGPT）产品资料．2023 年上线、2024–2025 年迭代．（首批通过备案的数学教育大模型）相关介绍见 https://www.aihub.cn/tools/llm/mathgpt/ 及中华网报道 https://m.tech.china.com/digi/digi/20240419/202404191508210.html

11. 网龙网络（HK:0777）101 教育 PPT 官网与产品资料．https://ppt.101.com/ ；平台规模数据引自腾讯软件中心 https://pc.qq.com/detail/8/detail_22548.html 及智通财经报道 https://m.zhitongcaijing.com/contentnew/appcontentdetail.html?content_id=384035

12. 国家市场监督管理总局、国家标准化管理委员会．GB/T 45288.2—2025《人工智能 大模型 第 2 部分：评测指标与方法》．2025 年．https://lib.scu.edu.cn/genai/static/wenjian/GBT45288.2-2025-genaim.pdf

13. 认知智能全国重点实验室．《通用大模型评测体系 2.0》正式发布（认知智能全国重点实验室牵头制定）．https://cogskl.iflytek.com/archives/3296

14. 国家互联网信息办公室（中央网信办）. 《关于发布 2024 年生成式人工智能服务已备案信息的公告》. 2025 年 1 月. https://www.cac.gov.cn/2025-01/08/c_1738034725920930.htm（备案统计另见 C114 通信网报道 https://m.c114.com.cn/w5339-1259288.html）

第 6 章　智能评价（新增场景）：机理、产品形态与发展建议

6.1　场景定位与新增背景

在承接生成式人工智能教育产品原有"赋能教学、支持学习、支持教研"三场景的基础上，本蓝皮书 2026 将"智能评价"单列为独立的新增场景。这一调整回应了两条并行的趋势。

其一，生成式人工智能的能力边界正从"生成内容"向"理解、判断与反馈"延伸。当模型能够阅读一篇作文、比对一份评分量规（rubric）、并生成分数与逐句评语时，它已经在功能上跨过了"内容生产工具"的门槛，具备了介入评价这一判断性任务的技术条件。与前代基于关键词匹配、正则规则或浅层统计特征的自动批改相比，生成式模型对语义、结构与推理路径的把握是质的跃升——这也是本章把评价从其他场景中独立出来的直接原因。

其二，评价长期是教育链条中人力投入最重、反馈周期最长、标准化难度最高的环节。一次省级中考英语听说考试可能涉及数十万份口语作答，一次期中作文批改会占用一线教师整个周末，而反馈往往在学生已经进入下一单元时才姗姗返回。评价环节对"规模化、即时化、个性化"的需求，恰是生成式技术最有可能释放价值之处。以中国的实践为参照，科大讯飞的智能语音评测技术已在全国 29 个省市的高考英语听说测试和 132 个地市的中考英语听说测试中规模化应用，累计服务考生人次超过 4500 万[1]——这类"高利害、大规模、强标准化"的场景，正是评价智能化最先落地、也最需要审慎的地带。

需要强调的是，评价场景与教学、学习、教研场景高度耦合：智能评价既是教学的"出口"，也是学习诊断与教研改进的"入口"。将其单列，并非把它从"评—教—学"闭环中割裂出来，而是凸显评价环节独立的机理特征与产品逻辑。评价一旦被机器介入，其技术可行性与教育适

切性之间会立即出现张力——一个能给分的模型未必是一个能被信任去给分的模型。本章的组织即围绕这一张力展开：先讲清机理（6.2），再盘点产品形态（6.3），随后集中讨论效度与公平风险（6.4），最后给出面向可信构建的发展建议（6.5）。

从政策层面看，把评价单列为独立场景亦有明确的制度背景。2020 年 10 月中共中央、国务院印发的《深化新时代教育评价改革总体方案》，是新中国第一个关于教育评价系统性改革的文件，明确提出破除"唯分数、唯升学、唯文凭、唯论文、唯帽子"的顽瘴痼疾，并要求"充分利用信息技术，提高教育评价的科学性、专业性、客观性"，建立基于大数据和人工智能支持的教育评价机制，开展多维度的过程评价、增值评价与综合评价[19]。2025 年 4 月，教育部等九部门印发《关于加快推进教育数字化的意见》，进一步把"人工智能赋能教育评价"纳入数字化转型的整体部署[19]。这意味着智能评价既有强烈的技术驱动，也有清晰的政策牵引——它不是一个可有可无的"增效工具"，而是被写入国家评价改革路线图的关键环节。也正因如此，本章对其风险的讨论必须格外审慎：当一项技术被制度赋予介入高利害评价的合法性时，它的效度与公平缺陷会被放大为系统性的教育后果。

一个必要的界定是：本章所称"智能评价"，指生成式人工智能（及与之协同的判别式模型、专用评分引擎）介入教育测量与反馈的产品与实践，涵盖诊断性、形成性与终结性三类评价目的，覆盖从命题、作答采集、评分判断到反馈干预的完整链路。它既不等同于狭义的"自动阅卷"，也不局限于考试场景——日常作业批改、写作过程辅导、口语陪练、课堂学情分析乃至综合素质画像，都在其外延之内。正是这种"贯穿评价全链路、横跨高低利害"的特征，使它有别于其他三个场景，值得独立成章。

6.2 机理：生成式人工智能如何介入评价

生成式人工智能介入评价，本质是把"评分"这一判断任务，转化为"理解作答—对照标准—生成判断与反馈"的语言与多模态处理过程。要理解它与前代技术的差别，需先厘清自动评分的两条技术路线。

第一条是**显式特征路线**，以 ETS 的 e-rater 为代表。它对作答文本抽取一组可解释的显式微观特征（microfeatures）——语法错误率、用法、机制（拼写标点大小写）、风格、词汇复杂度、篇章组织与发展等，再以回归或机器学习模型把这些特征聚合为分数[2][5]。其优点是每一分都可追溯到具体特征，可解释、可审计；其代价是特征由人工工程定义，难以捕捉论证质量、思想深度等"语言之下"的构念，容易被形式化写作策略钻空子。

第二条是**生成式理解路线**，以大语言模型为代表。它不再显式抽取特征，而是把整篇作答作为上下文输入，依据评分量规（rubric）在语义层面整体理解、对照并直接以自然语言产出分数与逐句评语。其优点是对同义表达、部分正确、推理路径的把握是质的跃升，且反馈可读、面向改进；其代价是"给分的理由"本身也是生成物，黑箱性更强，且分数会随提示（prompt）与解码温度（temperature）漂移。

需要强调，二者并非替代关系。当前工业级产品往往是"显式特征 + 生成式理解"的混合体：用判别式模型或专用引擎保证可靠性与可审计性，用生成式模型补足语义理解与反馈生成。理解这条技术分野，是读懂后文"一致性高不等于效度高"这一核心张力的前提。就作用机理而言，生成式评价可归纳为六类。

6.2.1 自动评分（客观题与半结构化题）

对于客观题及答案空间相对收敛的半结构化题目（填空、简答、公式题、判断说明等），模型通过语义匹配与标准答案对照完成判分。相较于传统基于关键词或正则规则的自动批改，生成式模型能识别同义表达、部分正确与推理路径，从而覆盖更宽的作答形态。举例而言，

一道"用自己的话解释光合作用"的简答题，正则规则只能匹配预设关键词，遇到学生用等价但不同措辞的表达便会漏判；生成式模型则能判断语义是否等价，并对"答对了机理但用词不规范"给出部分分与具体提示。

在中国的规模化实践中，科大讯飞将中高考智能阅卷、口语评测、作文批改视为"同源技术"，其英语听说评测系统可实现自动化考试与评分，已推广到 20 多个省市的中高考英语口语考试[1];其智慧考试解决方案宣称可实现"一份试卷 15 秒内批改"，把主观题批改的边际时间压到极低[13]。这类系统的价值在于把海量、重复、标准明确的判分工作从人力中释放出来。

但客观题自动评分并非"零风险"的低垂果实。其风险集中在两处：一是**"边界作答"**——表达合理但偏离标准答案、或标准答案本身覆盖不全时的误判，尤其在开放性简答与理科证明题中，学生的非标准但正确的解法可能被判错；二是**标准答案质量的传导**——自动评分把命题者预设的标准答案作为唯一裁准，若标准答案本身有误或不完备，误差会被系统性放大到每一份作答。因此，即便是自动评分这一相对成熟的环节，也需要保留人工抽检与异常复核机制，而非全盘托管给机器。

6.2.2 作文与主观题评阅

这是生成式人工智能相较前代技术能力跃升最显著的环节。模型可依据预设评分量规，对作文、简答、论述等开放性作答从内容、结构、语言、逻辑等多维度给出分数与评语，并定位具体问题句段。其价值不仅在于"给分"，更在于生成可读的、面向改进的反馈。

关于其信效度，近两年已积累一批可核验的实证结论，且结论呈"谨慎乐观、条件依赖"的形态：

- **与人类评分的一致性可达甚至超过人际水平，但依赖提示工程与温度设置。** 一项针对英语学习者写作、比较 GPT-4、GPT-3.5 与 Claude 2 的研究发现，GPT-4 表现最佳，展现出优异的评分者内部信度（intra-rater reliability）与良好的效度；在"标准 + 样例参照"的评

分提示下，GPT-4 的评分准确性较 GPT-3.5、Claude 2 分别提升约 112% 与 114%[3]。研究同时指出，把温度（temperature）调低会使模型更趋确定性、显著提升重复测量的一致性，说明提示策略与解码参数是影响可靠性的关键变量[3]。

- **在高等教育 EFL 作文评阅中信度很高，但需锚定量规。** Yavuz 等（2025）在《British Journal of Educational Technology》上以 15 名 EFL 教师评分者与 ChatGPT、Bard 对同一组作文按五维量规（语法、内容、组织、风格与表达、机制）评分，发现经微调的 ChatGPT 模型在 10 次重复测量中信度极高（组内相关系数 ICC=0.972，标准差为 0.00）[4]。这提示：当评价被牢牢锚定在明确的量规上、并控制随机性时，生成式模型可以给出高度稳定的分数。
- **多维评分中与人类判断的对齐程度因维度而异。** 一项利用 LLM 做多维写作评估的研究发现，模型在语言形式类维度（语法、词汇、机制）上与人类评分的一致性较高，而在内容、论证与思想深度等高阶维度上一致性明显下降——这与"模型更擅长评判语言表层、更难评判语言之下的思维"这一贯穿全章的判断相互印证[3]。这提示产品设计不应给出单一总分了事，而应按维度报告分数与置信度，把机器可靠的维度与需要人工把关的维度区分开来。
- **传统专用引擎已树立可对照的基准线。** 作为参照系，ETS 的 e-rater 引擎在 GRE 议论文/论证文任务上与人类评分的相关约为 r≈0.78/0.79、加权 kappa 在 0.73–0.77 之间；在 TOEFL 独立/综合写作任务上相关约 r≈0.75/0.73、加权 kappa≈0.70[5]。更被反复引用的一个结论是：在 TOEFL 独立与 GRE Issue 任务上，e-rater 与单个人类评分者的一致性甚至高于两名独立人类评分者之间的一致性[2]。但 ETS 自身的研究报告也明确警示——一致性统计本身并不能证明"测的是什么"，高一致性仍可能伴随构念代表性不足（construct underrepresentation）或构念无关变异（construct-irrelevant variance）[5]。这一警示对生成式模型同样成立，是理解 6.4 节风险的关键。

需要补充一点方法论上的注意：上述研究的信效度指标（QWK、ICC、相关系数、加权 kappa）大多在受控的研究数据集上取得，且高度依赖提示设计、量规明确度与作答样本的分布。把它们直接外推到某个具体产品在真实课堂中的表现是不严谨的——同一模型换一套量规、换一个学段、换一种母语背景的学生群体，一致性可能显著下降。因此本章列举这些数字，意在刻画"能力上限与条件依赖性"，而非为任一产品背书。综合看，"作文与主观题评阅"是生成式评价能力最强、也最需要人机协同约束的环节：它能稳定给分、能生成可读评语，但"评分一致性"不等于"评分有效"。

6.2.3 口语与多模态评测

口语评测是评价智能化中技术链条最长的一类：需要先由自动语音识别（ASR）把语音转写为文本，再对发音、流利度、词汇、语法、内容与篇章连贯等子构念分别打分。国际上，多家考试机构已部署专用引擎——ETS 的 SpeechRater、剑桥的 CASE、Pearson（PTE）的 Ordinate、以及多邻国英语测试（DET）的 Duo Speaking Grader[6]。以 DET 的公开评分说明（2024）为例，其自动评分覆盖人类量规中的全部六个子构念（内容、篇章连贯、词汇、语法、流利度、发音），口语能力子分由"经典评分的口语"与"基于项目反应理论（IRT）的朗读"两部分聚合而成[7]。在信度层面，SpeechRater 分数与人类评分的相关约为 r=0.57–0.68[6]——明显低于书面作文引擎与人类的一致性，反映口语构念（尤其是"内容/交际有效性"）比书面语更难被机器可靠捕捉。

口语评测的效度风险有其特殊性。第一，**误差沿链条传导**：ASR 的转写错误会直接污染下游所有维度的评分，且对口音重、母语背景特殊的学生，ASR 错误率更高，从而把技术误差转化为公平问题。第二，**构念覆盖不均**：机器对发音、流利度、语速等"可声学量化"的维度评得较准，而对"内容是否切题、交际是否有效、观点是否有说服力"这类高阶维度评得较弱——这正是 SpeechRater 与人类相关只有 r≈0.57–0.68 的深层原因[6]。第三，**朗读题与开放题的**

鸿沟：DET 之所以把朗读题用 IRT 处理、把开放口语用经典评分，正是因为朗读题答案收敛、易于自动化，而开放表达题构念复杂、难以稳定评分[7]。

在中国，口语评测是最成熟的规模化落地之一。科大讯飞已为全国 64 个地市中考及 13 个省市高考口语计分考试或加试提供口语智能评测服务[1];以广东为例，作为高考英语的一部分，计算机化英语听说考试（CELST）采用基于机器学习、由人工评分校准集训练而来的自动评分引擎[8]。值得注意的是，此类高利害口语考试通常仍保留人工复评或抽检环节，机器评分并非唯一裁准——这与国际高利害写作考试"机器 + 人工"双评的做法一脉相承。随着 AI 眼镜、教育机器人等多模态终端的成熟，口语评测正从"读一段、答一题"的静态测评，向"边做实验边讲解""小组协作中的口头贡献""情境化角色扮演"等难评价维度延伸，语音、视觉、动作、协作痕迹被同时纳入证据来源（详见本院同期《AI 眼镜教育应用蓝皮书 2026》《教育机器人发展白皮书 2026》）。这一演进拓宽了评价的构念覆盖，但也把 ASR 误差、多模态对齐、隐私采集等风险叠加放大，越是复杂的构念，越应坚持人机协同而非机器独断。

6.2.4 过程性评价

依托对学习过程数据（作答轨迹、修改历史、交互对话、协作记录、击键日志等）的建模，模型可将传统上难以量化的过程表现转化为可观测、可反馈的指标，支撑形成性评价从"结果打分"向"过程刻画"迁移。近期研究把击键日志（keystroke logging）经 S-记法定性分析后视为一种学习分析，用以让写作过程中的修改范围、次序与整合"可见化"，从而支撑面向过程的形成性反馈[9]。以 Khan Academy 的 Writing Coach 为代表的产品，正沿此方向落地：它强调"过程与成品同样重要"，在提纲、起草、修改各阶段给予引导式反馈，并向教师提供班级/个体层面的反馈汇总、各写作阶段耗时报告与对话记录，使教师无须逐份翻阅上百份草稿历史即可掌握进展、并对疑似抄袭给出提示[10]。

过程性评价是生成式技术相对于传统"一次考试一个分"的终结性评价最具想象空间的方向，因为它把评价的证据来源从"最终成品"扩展到"完整过程"。以写作为例，一份终稿只能看到结果，而击键日志与草稿历史能揭示：学生是一气呵成还是反复推敲、修改集中在遣词造句还是谋篇布局、遇到卡点时如何求助。这些过程证据对形成性反馈的价值，往往超过一个终稿分数。Khan Academy Writing Coach 的产品逻辑正是把"耗时报告 + 对话记录 + 阶段反馈"作为教师洞察的核心，使教师无须逐份翻阅上百份草稿历史即可掌握进展[10]。

但必须清醒看到，过程性评价的学理仍不成熟。多项 2025 年的综述指出，用学习分析支撑形成性评价的证据依旧稀疏，形成性评价难以被稳定操作化，且当分析结果与形成性评价模型不能良好对齐时，教师往往不愿信任并据此行动[11]。其瓶颈有三：一是**数据完整性**——过程数据往往残缺、噪声大，跨设备、跨平台难以拼接为完整轨迹；二是**推断不确定性**——从"改了很多次"到"认真"或"吃力"之间的推断链条脆弱，容易过度解读；三是**隐私与合规**——过程数据采集最为密集、最贴近学生行为，一旦越界即构成对学习者的过度监控。换言之，"能采集过程数据"距离"能给出有效的过程性判断"之间仍有相当距离，过程性评价当前更适合作为教师判断的辅助线索，而非独立的评分依据。

6.2.5 学习诊断与反馈

在评分之上，模型可归因错误、定位薄弱知识点、推断误解成因，并生成个性化的矫正建议与后续学习路径，使评价从"测量"走向"干预"，直接反哺学习场景。这是"评—教—学"闭环中评价反哺学习最直接的一环：一个分数只告诉学生"考了多少"，而一次好的诊断能告诉学生"错在哪个知识点、为什么会这样错、下一步该练什么"。

中国头部教育科技企业的学习硬件正是这一机理的集中承载。好未来的九章大模型（MathGPT）是国内首个专为数学打造且首批通过备案的教育大模型，以解题和讲题算法为核心，具备数学自动解题与复杂应用题批改、语文英语作文批改、AI 分步骤讲题等能力，并

已落地"学小伴"App 与学而思学习机 Xpad;其"数学随时问"面向小学初中题目主打即问即答（此为厂商宣称，具体覆盖率[待补：需第三方核实]）[12][13]。网龙旗下的 101 教育 PPT—AI 助教则从课堂侧切入，可根据学生课堂数据有针对性地布置适合其个人水平的习题并进行在线批改，把诊断—反馈嵌入日常教学流程[18]。这些产品的共同逻辑，是把评价结果即时转译为下一步的学习动作，缩短"测—反馈—练"的循环。

诊断—反馈的核心局限是**归因准确性**。评分错了，影响的是一个分数；归因错了，影响的是学生接下来的整条学习路径。若模型把"粗心导致的计算失误"误判为"概念不清"，就可能给学生推送大量本不需要的概念补习，既浪费时间又打击信心；反之亦然。更隐蔽的是，生成式模型的归因往往措辞笃定、看似有据，容易让学生和家长过度信任。因此诊断—反馈类产品尤其需要透明地呈现"这一归因的依据是什么、置信度如何"，并保留教师对高影响判断的复核权。

6.2.6 AI 辅助命题

在评价的上游，模型可依据知识点、难度与题型要求辅助生成试题、干扰项与评分量规，并对题目质量（区分度、覆盖度、表述规范性）作初步审查，缩短命题周期。这一环节把生成式技术"内容生产"的本行能力与评价的专业约束结合起来，价值直接：命题长期是高耗时、高门槛的专业工作，一套高质量试题需要反复打磨题干、设计有效干扰项、控制难度与区分度。

近来这一环节出现了较强的实证支撑。一项迄今规模较大的现场研究覆盖美国 91 个大学班级、约 1700 名学生（其中 71 个班使用 AI 生成考卷、20 个统计学班使用标准化 AP 题目），采用双参数逻辑斯蒂（2PL）IRT 模型评估，发现 AI 生成题目在难度与区分度上与高利害标准化测验题目相当——AI 生成题目平均区分度 $\bar{\alpha}\approx1.3$、对照标准化题目 $\bar{\alpha}\approx1.2$，AI 生成考卷的最大测验信息量 $I_{max}\approx3.85$（信度 $R\approx0.79$）优于标准化对照 $I_{max}\approx2.61$（$R\approx0.72$）[14]。另有

比较研究发现，在专家评审下 LLM 生成题目有时甚至被优选，但结论一致强调：**人类审核对最终质量不可或缺**——LLM 生成题偶尔会出现"多个正确答案""废弃干扰项"或事实性错误，且题目须经过多阶段自动化筛选（评估、去重、难度与适切性判断）才能入库[14][15]。

对中国基础教育而言，AI 辅助命题正与命题改革方向叠加：2026 年中考命题被要求大幅压减机械记忆类试题比例、强化情境化与素养导向[19]，这类"重情境、重迁移、轻记忆"的题目恰恰最耗费命题人力，也最适合用生成式技术辅助批量生产候选题、再由学科教师精选打磨。但命题的偏差风险不容忽视：模型可能在情境设定、人物姓名、文化背景中引入刻板印象或地域偏见，也可能生成看似合理实则超纲或偏离课标的题目，因此"机器生成候选、教师专业终选"应是这一环节的基本工作范式。

下表概览六类机理的输入、输出与主要局限。

机理类别	主要输入	主要输出	关键局限	可核验证据（示例）
自动评分	客观/半结构化作答	分数、对错判定	边界作答误判	讯飞规模化中高考评测[1]
作文与主观题评阅	开放性文本作答	分数、多维评语	评分一致性≠效度、可解释性弱	GPT-4 内部信度高、依赖提示[3];ICC=0.972[4];e-rater κ≈0.70–0.77[5]
口语与多模态评测	语音/多模态作答	子构念分、发音反馈	ASR 误差、内容构念难测	SpeechRater r≈0.57–0.68[6];DET 六子构念[7]
过程性评价	学习过程/击键数据	过程指标、画像	数据完整性、隐私、操作化不成熟	击键日志学习分析[9];证据仍稀疏[11]
学习诊断与反馈	作答与错误数据	归因、矫正建议、路径	归因准确性	MathGPT 批改+分步讲题[12][13]
AI 辅助命题	知识点、量规	试题、干扰项、量规	题目质量与偏差、需人审	91 班/1700 生现场研究[14][15]

6.3 产品形态

从 2024 到 2026，评价类产品的技术底座正从"对话式单点工具"转向"智能体 + 多模态 + 端侧"的组合形态。这一演进有三条可观察的线索：其一，从"单轮问答式批改"走向"可规划、可调用工具、可多步推理"的智能体（agent）架构，使评价能在诊断—讲题—再练之间自主编排；其二，从"纯文本"走向"手写识别 + 公式解析 + 图像理解 + 语音评测"的多模态，使评价能覆盖真实作答的完整形态；其三，部分推理下沉到端侧（学习平板、AI 眼镜等），以降低时延、保护隐私并支持离线场景。

就市场体量而言，中国 AI+教育整体规模测算差异较大，需谨慎引用。Grand View Research 测算 2024 年中国 AI 教育市场约 5.09 亿美元、2030 年达约 28.46 亿美元（2025—2030 年 CAGR 约 31.6%）[16];国内券商与咨询口径普遍更高，如 2025 年中国 AI+教育市场规模超 700 亿元、预计 2030 年近 3000 亿元的表述[17]。两类口径相差一个数量级以上，根源在于统计边界、币种与"AI+教育"外延并不一致——前者仅计狭义 AI 软件、以美元计，后者常把硬件、内容、服务乃至教育信息化整体经费纳入、以人民币计。**为避免混淆，本蓝皮书不对二者做换算或合并。** 需特别提示：智能评价作为其中的细分赛道，尚缺乏独立、权威的规模拆解，[待补：智能评价细分市场规模/产品数量的权威独立来源]。就产品结构而言，当前评价类产品大致可归为四类，以下逐一说明并给出真实案例与关键指标。

6.3.1 智能评阅

面向作业、考试与作文批改场景，提供自动判分、评语生成与错因标注，主要用户是教师，价值主张是把重复性批改从教师工作量中剥离。典型代表包括：科大讯飞面向中高考的智能阅卷与英语听说评测系统，宣称可实现"一份试卷 15 秒内批改"[1][13];好未来九章大模型（MathGPT）承载的数学作业/试卷批改、复杂应用题批改与语文英语作文批改能力，已内置

于学而思学习机 Xpad[12][13];网龙 101 教育 PPT 的"AI 助教"提供作业个性化定制、在线批改与语音评测等功能，该软件累计服务全国 970 万教师、装机量超 3576 万，用户覆盖全国 32 个省级行政区[18];国际侧则有 Khan Academy Writing Coach，面向议论文、说明文与文学分析类作文，在结构组织、论证支撑、引言结论、风格语气等方面提供即时、具体、可执行的反馈，并于 2025 年 2 月起面向教师免费开放[10]。

多模态能力正成为该类产品的分水岭。中国基础教育的作答大量以手写形式存在——纸质作业、手写试卷、草稿演算——因此能否稳定识别手写答卷、解析数学公式与几何图形、处理拍照上传的图像，直接决定产品在真实课堂的可用边界。仅能处理规整电子文本的评阅工具，在中小学场景的适用面相当有限。该类产品的关键指标是**评分一致率**（与人类评分的一致程度）与**错因定位准确率**（能否把错误精确定位到具体知识点与句段）；但目前厂商公开披露的、经第三方验证的一致率数据仍然稀缺，[待补：主流智能评阅产品的公开评分一致率/错因定位准确率的独立测评数据]。

6.3.2 学情诊断

在班级与个体两个层面聚合作答数据，输出知识点掌握度、薄弱环节与学情报告，服务教师精准教学与教研。它与智能评阅的区别在于：评阅关注单份作答的"对错与好坏"，学情诊断关注跨作答、跨时间聚合后的"群体与个体规律"。科大讯飞智慧考试解决方案在批改之上提供考后学情反馈；101 教育 PPT—AI 助教可实时反馈课堂数据、追踪课后数据并据此有针对性地个性化布置习题、进行在线批改[18];好未来学习硬件以"诊断—讲题—再练"闭环为核心[12]。

此类产品与国家评价改革方向高度契合：《深化新时代教育评价改革总体方案》明确提出建立基于大数据和人工智能支持的教育评价机制，开展多维度的过程评价、增值评价与综合评价[19]，学情诊断正是"增值评价"（关注学生一段时期内的进步幅度而非绝对分数）在产品侧

的落点。其关键指标是**诊断覆盖度**（能否覆盖课标要求的知识点与能力点）与**掌握度估计的稳健性**（对同一学生在不同题目上的掌握度判断是否一致可信）。风险在于：学情报告一旦呈现为精确的数字与图表，容易造成"精确的错觉"——把基于有限作答的粗略推断包装成确定结论，误导教师的教学决策。因此学情诊断应清晰标注其估计的不确定性区间，而非只给点估计。

6.3.3 自适应测评

依据实时作答动态调整题目难度与路径，以更少题量估计能力水平并即时反馈。它把成熟的计算机自适应测验（CAT）与项目反应理论（IRT）范式，与生成式命题的题库供给能力结合：CAT/IRT 负责"该考多难的题、如何用最少题量逼近真实能力"，生成式命题负责"随时供给足够多、参数合格的候选题"，二者互补正好解决自适应测验长期受制于题库规模与曝光控制的瓶颈。

多邻国英语测试是其成熟代表：其口语能力子分由经典评分与基于 IRT 的朗读题聚合，评分覆盖内容、篇章连贯、词汇、语法、流利度、发音六个子构念[7]。前述覆盖 91 个班、约 1700 名学生的现场研究进一步证明，生成式题库可达到与标准化测验相当的难度与区分度（AI 生成考卷 I_max≈3.85、信度 R≈0.79）[14]，为"生成式供题 + 自适应选题"的组合提供了实证底气。其关键指标是**测量效率**（达到目标信度所需题量）与**能力估计的标准误**（估计的精确度）。风险在于生成式供题若缺乏严格的题目参数标定与曝光控制，可能引入难度失准、题目泄题或内容偏差，反而损害自适应测验赖以成立的题库质量前提。国内自适应评测能力多内嵌于学习硬件与在线练习系统，独立成品的产品名单与经披露的信度指标仍不充分，[待补：国内自适应测评独立产品名单与信度指标来源]。

6.3.4 能力画像

跨科目、跨时间聚合评价数据，构建面向核心素养或能力框架的学习者画像，支撑综合素质评价与生涯发展。它是四类产品中聚合层级最高、也最贴近高利害决策的一类：智能评阅评单份作答、学情诊断评一个学科、自适应测评估一次能力，而能力画像试图跨科目、跨学段刻画一个学生的整体素养轮廓。

这一形态与国家政策强绑定：《深化新时代教育评价改革总体方案》要求通过信息化手段客观记录学生品行日常表现与突出表现，并将其作为综合素质评价的重要内容，同时将数字素养纳入综合素质评价[19]。因此能力画像类产品多以区域/学校为部署主体，而非面向个人销售，其数据往往进入学生升学、评优等环节。也正因如此，它的风险最为尖锐：**其一，构念代表性**——"核心素养""综合素质"这类构念本身抽象、边界模糊，用可量化的指标去逼近它，极易发生"测得到的替代测不到的"（用出勤、竞赛、可打分行为代替真正的品格与创造力）；**其二，公平性**——画像若纳入课外活动、竞赛经历等资源依赖型指标，会系统性地有利于资源丰富的家庭；**其三，可解释与可申诉**——当画像影响升学时，学生与家长有权知道"这个评价基于什么、能否申诉"。其关键指标是指标效度、画像可解释性与数据合规。代表性产品与部署情况，[待补：综合素质评价/能力画像平台产品名单与效度证据来源]。

下表对四类产品形态作横向对照。

产品形态	核心功能	主要用户	代表产品/厂商	关键指标
智能评阅	自动判分、评语、错因	教师	讯飞智能阅卷[1];MathGPT/学而思学习机[12];101 教育 PPT AI 助教[18];Khan Writing Coach[10]	评分一致率、错因定位准确率[待补：公开一致率数据]
学情诊断	掌握度、学情报告	教师/教研员	讯飞智慧考试[1];101	诊断覆盖度、掌握

			教育 PPT[18]	度估计稳健性[待补]
自适应测评	动态选题、能力估计	学生/教师	多邻国英语测试（国际参照）[7]	测量效率、能力估计标准误[待补]
能力画像	素养聚合、画像	学校/区域	[待补：产品名单]	指标效度、可解释性、合规

四类产品形态呈现一条清晰的"聚合层级递增、利害权重递增、验证难度递增"的谱系：从智能评阅（单份作答、低利害、相对可验证），到学情诊断与自适应测评（学科级、中利害），再到能力画像（跨科跨时、高利害、最难验证）。这条谱系恰好对应本章的核心张力——聚合层级越高、利害越大，构念越抽象、效度与公平越难保证，而正是这些环节最需要审慎。因此，产品成熟度与部署利害应当匹配：在智能评阅这类相对成熟、可核验的环节可以规模化推进，而在能力画像这类直接影响升学的环节，则应以人工为主、机器为辅，并把可解释与可申诉作为硬性前提。这一判断构成下一节风险讨论与第 6.5 节发展建议的直接依据。

6.4 效度与公平风险

智能评价的技术可行性并不等同于教育适切性。上一节反复出现的一个反差——"一致性可以很高，但一致性不等于有效"——正是本节的核心。以下五类风险需被清醒对待，尤其是在高利害（high-stakes）场景。

其一，构念效度：一致性不等于测对了。 这是本节最根本、也最容易被"高一致性"数字掩盖的风险。所谓构念效度，指评价是否真的测到了它声称要测的东西（如"写作能力""数学推理"），而非测到了与之相关但并非目标的替代物。ETS 的研究报告早已明确，自动评分与人类评分的高一致仍可能伴随构念代表性不足（construct underrepresentation，只测到了目标能力的一部分）或构念无关变异（construct-irrelevant variance，把不该计入的东西计入了分数）[5]。生成式模型的这一风险更隐蔽：它能写出流畅、看似有据的评语，却可能是在对表层语言特

征（长度、句法复杂度、词汇丰富度）而非对"论证是否成立、内容是否真实、思想是否深刻"打分。围绕自动作文评分的经典质疑正源于此——批评者（如 Perelman 对 NAPLAN 自动评分的分析）指出，此类系统与作文长度高度相关、无法评估意义与论证、无法核查事实准确性、并可被结构完整但语义荒谬的文本（如故意用生成器拼凑的"胡话"作文）骗取高分[20]。当把评分权交给一个"看形式给分"的系统时，评价的构念效度即被侵蚀，而这种侵蚀在日常低利害使用中往往看不出来，只在被刻意攻击或用于高利害决策时才暴露其代价。

其二，公平性：对特定语言背景的系统性偏差。 训练数据与语言风格偏差可能对特定方言、母语背景、表达风格或弱势群体的作答产生系统性偏差，放大既有教育不公。近期实证显示，自动作文评分对英语学习者（ELL）等子群体存在被引入偏差的风险，且当某些人群在训练数据中被低估或缺失时，表征偏差会与训练数据中既有的历史与测量偏差交互，放大对弱势群体的不公[21]。在德语学习者作文上的对比研究、以及针对小学阶段 ELL 的效度研究，均记录到跨人群的评分差异[22]。

对中国基础教育而言，这一风险有其本土形态：口音与方言影响 ASR 识别质量，城乡与区域差异影响学生的表达风格与话语习惯，进而可能让口语与作文自动评分对农村、少数民族地区、方言口音较重的学生系统性偏低。一项对广东四城 AI 评分的比较研究即提示，自动评分结果在不同地区之间存在差异，值得进一步分层核验[8]。这意味着"平均一致性达标"可能掩盖"对某些学生系统性偏低/偏高"的问题——一个总体 QWK 很高的系统，完全可能对特定子群体不公。因此公平性必须按子群体（母语、方言、城乡、社会经济背景）分层评估，把"分组公平性诊断"作为高利害评价上线前的必检项，而非只看总体指标。

其三，可解释性与可申诉性。 分数与评语的生成过程难以完全追溯，"为何给这个分"缺乏可核验依据，削弱评价的公信力与申诉可能。生成式模型的黑箱性质使其比 e-rater 这类显式特征模型更难解释——后者至少能追溯到具体的语法/结构微观特征[2]，而前者的"理由"本身也是生成物，可能与真实的内部计算并不一致。缺乏可解释性直接侵蚀高利害评价的正当程序。

其四，被"刷分"与应试异化（构念无关变异的行为版本）。 一旦评分规则被逆向摸索，学习者可能针对模型偏好而非真实能力优化作答；当评分系统与作文长度、特定句式模板高度相关时，"应试策略"就会取代"真实能力"，出现"堆砌长句、套用模板、罗列高级词汇即得高分"的异化。前述 NAPLAN 案例中，批评者用刻意生成的、结构完整但语义荒谬的文本骗取高分，正是对这一风险最尖锐的演示[20]。更棘手的是**双向作弊**：学生用生成式工具代写作答，评价方再用生成式工具评分，评价可能退化为"AI 写、AI 判"的空转，使"以评促学"异化为"以评应付"。Khan Academy Writing Coach 内置抄袭提示、并把过程数据作为佐证[10]，正是对这一风险的产品级回应；但工具侧的检测远不能替代评价设计侧的稳健性——真正的解法是让评价任务本身更难被"表演"，例如强调情境化、过程性与口头解释。

其五，数据隐私与合规、过度替代教师专业判断。 过程性评价对学习行为数据的采集最为密集（击键、时长、修改轨迹、对话记录），其合规风险也最高；能力画像一旦被用于升学等高利害决策，滥用风险随之放大。此外，若产品在设计上默认"以机器结论为准"、把机器分数直接前置为默认值，会导致教师专业判断被隐性让渡——这既是伦理问题，也是效度问题：教师对语境、意图与个体差异的判断，恰是当前模型最难替代的构念部分。自动化偏见（automation bias，即人倾向于过度信任自动系统的输出）会进一步放大这一让渡，使"辅助"在事实上变成"主导"。

概言之，这五类风险有一条共同的技术根源：生成式模型擅长对"语言表层"作出高一致的判断，而教育评价所要测量的往往是"语言之下"的思维、内容与素养。这道缝隙决定了一个基本判断——**技术越是能给出高一致的分数，越要警惕它是否在测对的东西。** 尤其是当评价被制度赋予高利害用途时，这道缝隙会被放大为系统性的教育后果。技术能力与评价目标之间的这道缝隙，是所有发展建议必须正视的前提。

6.5 发展建议

面向"评—教—学"闭环的可信构建，结合上文机理与风险，提出如下建议。这些建议指向不同责任主体：分级分类与人机协同是产品设计与学校应用层面的准则，评价质量基准与第三方测评需要政府与专业机构牵头，数据治理需要监管与厂商共同前置。一个总的原则是——智能评价的推进节奏，应与其效度和公平的可核验程度相匹配：在能被验证的地方大胆用，在尚不能被验证的地方审慎用。

- **坚持人机协同、教师终裁，按高低利害分级分类。** 应按高/低利害区分应用边界：低利害场景（日常作业反馈、练习性口语测评、写作过程辅导）鼓励规模化应用以减负增效；高利害场景（升学考试、能力认定）严格限定模型为"辅助建议方"，最终判定权保留于教师，并保留人工复评。国际高利害实践的通行做法亦是"机器 + 人工"双评而非机器单独定分——e-rater 在 GRE 与 TOEFL iBT 写作中即与人类评分联合使用[2][5]，可作为分级设计的参照。
- **建立评价质量基准与第三方独立测评。** 应推动形成面向教育评价的可靠性、公平性、可解释性评测基准与公开数据集，使产品质量可比较、可监督。基准须超越"总体一致性"这一单一指标，纳入：分子群体的公平性诊断（按母语、方言、社会经济背景分层）[21][22]、抗"刷分"稳健性（对结构完整但语义无效文本的抵抗力）[20]、以及构念效度证据（分数与外部效标的关系）[5]。目前该类面向中国基础教育评价的独立基准与权威机构，[待补：国内评价质量基准/测评机构/公开数据集来源]。
- **强化可解释与可申诉机制。** 应要求评分附带依据（对应量规条目、定位到具体句段），并提供人工复核与申诉通道。评语应可回溯到量规，而非仅呈现一段生成文字；对高利害结果，须保留原始作答、模型版本、评分参数以支持复核。

- **锚定量规、控制随机性以提升可靠性。** 实证表明，把评分牢牢锚定在明确量规、并降低解码温度，可显著提升评分的可重复性（如经微调 ChatGPT 的 ICC 达 0.972）[4];提示策略（标准 + 样例参照）对准确性影响巨大[3]。产品侧应把"量规对齐 + 低温度 + 版本固定"作为高利害评分的工程默认，并公开其可靠性指标。
- **数据治理与隐私保护前置。** 过程性数据的采集、存储与使用须遵循最小必要与合规要求；能力画像用于综合素质评价时，须符合《深化新时代教育评价改革总体方案》关于客观记录、防止滥用的要求[19]，具体规范衔接第 7 章"治理与安全"场景。
- **与多模态、端侧演进协同，拓展难评价维度。** 应结合 AI 眼镜、教育机器人等新型终端的多模态感知能力（详见本院同期蓝皮书），把口语、实操、协作、表达等长期"难评价"的维度纳入评价形态；但口语等构念的机器信度（SpeechRater 与人类相关仅 r≈0.57–0.68[6]）提示，越是复杂构念越应坚持人机协同，避免把技术可行性误读为评价有效性。
- **提升教师与命题者的 AI 评价素养，防范自动化偏见。** 智能评价不是让教师退出，而是要求教师承担更高阶的角色——从"逐份打分的执行者"转为"评价证据的审辑者"：既要会用工具高效获取机器判断，也要有能力识别机器判断的失效点、抵抗过度信任自动系统的自动化偏见。应把 AI 评价素养纳入教师专业发展与命题培训，明确"机器给出候选、教师作出裁决"的协作范式；对 AI 辅助命题，须由学科教师对题目的科学性、课标符合度与文化适切性作专业终选[14][15]。这与教育部推进数字化赋能教师发展的方向一致[19]。

综上，智能评价的价值不在于用机器替代教师的专业判断，而在于把教师从重复性评分中释放出来，将生成式人工智能的规模化反馈能力，转化为面向每一个学习者的、及时而个性化的学习支持。本章反复出现的一条主线是：生成式模型能给出高一致的分数，但"一致"不等于"有效"，"能评"不等于"该评"。智能评价的健康发展，取决于技术能力、教育规律与治理约束三者的协同校准——尤其是在高利害环节，审慎与可核验，应始终优先于效率与规模。

本章参考来源

1. 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（含"29 省市高考听说、132 地市中考听说、累计服务超 4500 万人次""64 地市中考及 13 省市高考口语评测""推广至 20 多省市中高考英语口语考试"等表述）· 科大讯飞 · 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（e-rater 与单个人类评分者一致性在 TOEFL Independent 与 GRE Issue 上高于两名人类评分者之间的一致性）· 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 内部信度优、准确性提升约 112%/114%；温度与提示策略为关键变量）· 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 名人类评分者 + ChatGPT/Bard，五维量规；微调 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》及相关 e-rater 效度研究（GRE r≈0.78/0.79、κw 0.73–0.77；TOEFL r≈0.75/0.73、κw≈0.70；一致性统计不足以证明构念效度）· 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（SpeechRater/CASE/Ordinate/Duo Speaking Grader 概览；SpeechRater 与人类相关 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（口语六子构念：内容/篇章连贯/词汇/语法/流利度/发音；口语子分 = 经典评分 + IRT 朗读聚合）· 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 为高考英语组成部分，采用由人工评分校准集训练的机器学习自动评分引擎）· 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（击键日志 + S-记法作为学习分析，使写作修改过程可见化）· https://www.frontiersin.org/journals/education/articles/10.3389/feduc.2026.1773621/full
10. Khan Academy《Meet Khan Academy Writing Coach》《Writing Coach》产品与帮助页 · Khan Academy · 2024–2025（过程与成品并重、逐阶段引导反馈、班级/个体反馈汇总、耗时报告、抄袭提示；2025-02-27 面向教师免费开放）· 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（用学习分析支撑形成性评价的证据仍稀疏、操作化困难、结果与模型不对齐时教师不愿信任）· https://onlinelibrary.wiley.com/doi/full/10.1111/jcal.70056

12. 好未来九章大模型（MathGPT）介绍与备案信息（国内首个专为数学打造且首批过备案的教育大模型；数学自动解题/复杂应用题批改、语文英语作文批改、AI 分步骤讲题；落地学小伴 App 与学而思学习机 Xpad;"数学随时问"面向小初即问即答，覆盖率[待补]）· 好未来 / 中华网 / 极客公园 · 2023–2024 · https://m.tech.china.com/digi/digi/20240419/202404191508210.html ; https://www.geekpark.net/news/337563
13. 《批改一张试卷不超过 15 秒，AI 赋能教育突破"不可能三角"》· 南方都市报 · 2024 · https://m.mp.oeeee.com/a/BAAFRD0000202411251028303.html
14. Isley 等 "Assessing the Quality of AI-Generated Exams: A Large-Scale Field Study." *arXiv* · 2025（91 个班/约 1700 名学生；2PL-IRT；AI 生成题区分度 ᾱ≈1.3 vs 标准化 1.2；AI 考卷 I_max≈3.85/R≈0.79 优于标准化 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 生成题经专家评审偶被优选，但人类审核对最终质量不可或缺；存在多正确答案/废弃干扰项问题）· https://onlinelibrary.wiley.com/doi/10.1111/ijsa.70021
16. Grand View Research《China AI in Education Market Size & Outlook, 2025–2030》（2024 约 5.09 亿美元，2030 约 28.46 亿美元，CAGR≈31.6%）· Grand View Research · 2025 · https://www.grandviewresearch.com/horizon/outlook/ai-in-education-market/china
17. 多鲸/券商口径 AI+教育市场规模（2025 年中国 AI+教育超 700 亿元、2030 年近 3000 亿元等，口径与外延不一，需谨慎比较）· TopMarketing 转载 / 知乎行研 · 2025 · https://itopmarketing.com/info20459

18. 网龙 101 教育 PPT 与 AI 助教产品信息（AI 助教：作业个性化定制/在线批改/语音评测；课堂与课后数据实时反馈并个性化布置习题；累计服务 970 万教师、装机量超 3576 万）· 网龙 101 教育 PPT 官网 / 多知网 · 2018–2024 · https://ppt.101.com/news/06202018/20180620153929242.shtml ; http://www.duozhi.com/company/201804247108.shtml

19. 中共中央、国务院《深化新时代教育评价改革总体方案》及教育部答记者问（建立基于大数据和人工智能支持的教育评价机制；多维度过程评价/增值评价/综合评价；信息化手段记录并纳入综合素质评价、数字素养纳入综合素质评价）· 2020 印发/2024 解读 · https://www.eol.cn/zhengce/jiedu/202410/t20241013_2636778.shtml

20. Perelman, L. 关于自动作文评分与 NAPLAN 的分析（AES 与作文长度高度相关、无法评估意义与论证、可被结构完整但语义荒谬文本骗取高分）· 澳大利亚教育工会（AEU）等 · 约 2017–2018 · 摘要经 WebSearch 检索确认（原 PDF 链接已失效，结论以搜索返回与既有文献为据）[待补：稳定可访问的原文链接]

21. "Assessing fairness in finetuned scoring models with demographically restricted training data" 与相关综述（ELL 等子群体被引入偏差风险；训练数据低估/缺失导致表征偏差与历史/测量偏差交互放大不公）· *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）与 "Validity of automated essay scores for elementary-age English language learners: Evidence of bias?"（跨人群评分差异证据）· ACL Anthology / *Studies in Educational Evaluation* · 2024 · https://aclanthology.org/2024.bea-1.18/ ; https://www.sciencedirect.com/science/article/abs/pii/S1075293524000084

第 7 章　治理与安全（新增场景）：合规、隐私与学术诚信

7.1 场景定位：从"能力叙事"转向"责任叙事"

在本蓝皮书前六章确立的"赋能教学—支持学习—支持教研—智能评价"四场景之上，治理与安全构成第五个、也是贯穿全局的场景。它不是与前四者并列的又一类应用，而是决定前四类应用能否合法、可信、可持续落地的**约束层与前提层**。2024 版旧报告以对话式大模型为主线，将合规与伦理作为"发展建议"的一个段落附于末尾；2026 年产品形态已转向智能体化、多模态、端侧化，治理议题的复杂度随之量级上升——一个能自主规划、调用工具、跨会话记忆并可驱动可穿戴硬件（详见本院《AI 智能眼镜教育产业蓝皮书 2026》）的教育智能体，其数据流、决策链与责任边界远非"一问一答"的合规框架所能覆盖。因此本章将治理与安全独立成章，作为产品评估与采购决策的**否决性维度**。

本章的基本判断是：2026 年制约生成式人工智能教育产品规模化的首要瓶颈，正从"模型能力不足"转向"治理能力不足"。**能力—合规落差**（capability–compliance gap）——即产品的技术能力已经跑在教育场景所需的合规、隐私与诚信保障之前——是本章的核心分析线索。这一落差在数据上有直接印证：美国民主与技术中心（CDT）2024—2025 学年的全国代表性调查显示，85% 的教师与 86% 的学生报告在过去一学年使用过 AI，但不足半数的师生表示学校向其提供过任何 AI 相关培训或指引；仅十分之一的教师报告接受过"当怀疑学生的 AI 使用有害其身心时如何应对"的培训（CDT, 2025）。能力的普及速度远超治理机制的建立速度，这是本章一切讨论的现实起点。

7.1.1 治理对象的三层结构

- **数据层**：学习者的身份、行为、生物特征（语音、面部、注视）、心理状态推断等数据的采集、存储、流转与再利用。
- **模型层**：基座模型与教育垂类模型的训练数据来源、偏见、幻觉、可解释性，以及智能体的记忆与工具调用留痕。
- **应用层**：产品在真实教学场景中的使用边界，包括对未成年人的特殊保护、师生知情与选择权、以及人机责任划分。

三层之间存在耦合：端侧化虽在数据层缓解了原始数据上云的隐私压力，却在模型层引入了本地模型难以审计的新问题；多模态在应用层丰富了交互，却在数据层显著扩大了敏感数据的采集面。本章下述四大主题——多法域合规（7.2）、隐私与数据保护（7.3）、学术诚信（7.4）、责任与人机协同（7.5）——正是沿这三层交叉展开，并在 7.6 汇聚为可采购、可比较的治理成熟度分级。

7.1.2 为何治理在 2026 年独立成场景

将治理与安全从"发展建议"提升为独立场景，源于三个结构性变化，它们共同放大了旧框架的失效。

其一，**产品形态从工具跃迁为主体**。对话式产品是被动的"问答工具"，其数据流是"用户输入—模型输出"的单次闭环，责任链清晰。而智能体（agent）具备自主规划、工具调用、跨会话记忆三项能力后，它开始以"准主体"身份代替师生执行多步操作——查询学情数据库、调用批改接口、修改学习计划、向家长推送信息。每一次工具调用都是一次独立的数据处理与决策行为，旧框架针对"一次问答"设计的告知—同意机制无法覆盖这条动态展开的行为链。

其二，**采集边界从"输入框"扩展到"物理空间"**。文本产品只处理用户主动键入的信息；多模态与端侧硬件（可穿戴、常开摄像头/麦克风）则被动采集物理环境中的语音、图像、面部与注视，且波及非用户第三方。数据的"主动提供"退化为"被动捕获"，同意的有效性随之动摇。

其三，**治理成本从"事后补救"前移为"准入门槛"**。2023—2025 年间，我国《暂行办法》《标识办法》、欧盟 AI 法案、美国 COPPA 修订相继落地，合规从"上线后可整改的软约束"变为"上线前必须满足的硬门槛"：备案、安全评估、内容标识、白名单审核，任何一项缺失都可能导致产品无法进入市场或校园。这一变化使治理不再是产品的"加分项"，而是"通过项"——正是本章将其列为采购否决性维度的根本原因。

7.2 合规框架：多法域叠加下的产品义务

生成式人工智能教育产品面对的是一个多法域、多层级、快速演进的规则体系。一款既可能面向国内 K12 与高校、又可能出海欧盟或北美的产品，须同时满足数条相互独立却时有冲突的合规链。本节按"中国规则体系—欧盟—美国—国际组织"的顺序梳理其结构，文号、年份与生效时间均以本次检索到的官方与权威来源为准。

7.2.1 我国规则体系的分层

我国对生成式 AI 教育产品的规制，呈现"上位法—专门规则—行业准入"三层叠加结构。

上位法层：《中华人民共和国个人信息保护法》（2021 年 8 月 20 日十三届全国人大常委会第三十次会议通过，2021 年 11 月 1 日起施行，即 PIPL）确立了处理个人信息的基本框架。其中与教育产品直接相关的关键条款包括：第 28 条将"不满十四周岁未成年人的个人信息"明确纳入**敏感个人信息**范畴，与生物识别、医疗健康、行踪轨迹等并列；第 29 条要求处理敏感个人信息应取得个人的**单独同意**，法律、行政法规规定应取得书面同意的从其规定；第 31 条要求处理不满十四周岁未成年人个人信息的，应当取得未成年人父母或者其他监护人的同意，

并制定专门的个人信息处理规则（全国人大常委会，2021）。这意味着面向小学与初中低年级的教育产品，凡采集面部、语音、注视等生物特征或对未成年人做心理推断的，几乎必然触发敏感信息+未成年人的双重高强度合规义务。

在未成年人保护的行政法规层面，《未成年人网络保护条例》（国务院令第 766 号，2023 年 9 月 20 日国务院第 15 次常务会议通过，2024 年 1 月 1 日起施行）是我国首部专门、综合性的未成年人网络保护立法，就网络信息内容规范、个人信息保护、网络沉迷防治等作出系统规定，要求个人信息处理者严格设定未成年人个人信息的访问权限、开展合规审计，并将网络素养教育纳入学校素质教育内容（国务院，2023）。该条例与 PIPL 第 31 条形成互补：PIPL 确立"监护人同意+专门处理规则"的个人信息处理底线，条例则进一步就内容适龄、沉迷防治、网络欺凌治理等作出场景化要求，教育产品须同时满足两者。此外，《数据安全法》（2021 年 9 月 1 日施行）就数据分类分级与安全保护义务提供了上位依据，教育数据中涉及规模化未成年人信息的部分，可能触发更高等级的数据安全保护要求 [待补：教育数据是否被地方目录列为重要数据的具体认定]。

专门规则层： 《生成式人工智能服务管理暂行办法》（国家网信办、国家发展改革委、教育部、科技部、工信部、公安部、广电总局七部门联合公布，2023 年 7 月 13 日发布，2023 年 8 月 15 日起施行）是我国生成式 AI 的核心专门规则。其对教育产品的关键义务包括：训练数据来源合法、尊重知识产权、采取措施提升训练数据质量；对具有舆论属性或社会动员能力的服务开展安全评估并履行算法备案手续；对图片、视频等生成内容按《互联网信息服务深度合成管理规定》进行标识；以及"采取有效措施防范未成年人用户过度依赖或者沉迷生成式人工智能服务"（国家网信办等，2023）。截至 2025 年 12 月 31 日，累计已有 748 款生成式人工智能服务在国家网信办完成备案、435 款生成式人工智能应用或功能完成登记，仅 2025 年全年即新增 446 款服务备案（国家网信办，2026）——备案数量的高速增长本身即是该规则强约束力的直接体现。

内容标识规则近年进一步细化。《人工智能生成合成内容标识办法》由国家网信办、工信部、公安部、国家广电总局四部门联合发布，自 2025 年 9 月 1 日起施行；配套的强制性国家标准 **GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》**（2025 年 2 月 28 日发布，2025 年 9 月 1 日与《标识办法》同步实施）确立了显式标识与隐式标识的技术基线：显式标识指在文本起始/末尾/中间、音图视频等位置添加人可感知的文字提示或通用符号；隐式标识则包括在文件元数据中写入标识信息，以及在生成内容中添加数字水印等（国家网信办等，2025；GB 45438-2025）。对教育产品而言，这意味着 AI 生成的作业讲解、范文、习题、语音朗读等内容均须依法标识，且传播平台负有核验元数据隐式标识、检测显式标识或生成痕迹的义务。

教育行业准入层： 通用 AI 规则之上，教育行业叠加了专门的进校准入规则。2025 年 5 月，教育部基础教育教学指导委员会发布《中小学生成式人工智能使用指南（2025 年版）》与《中小学人工智能通识教育指南（2025 年版）》。前者确立了**分学段使用规范**：小学阶段禁止学生独自使用开放式内容生成功能，教师可在课内适当使用辅助教学；初中阶段可适度探索生成内容的逻辑性分析；高中阶段允许结合技术原理开展探究性学习。指南要求各校建立健全生成式 AI 工具**"白名单"制度**，经严格审核评估、仅允许符合教育场景需求且数据安全合规的工具进入校园，并构建覆盖数据安全、伦理审查、内容监管和风险防控的全链条保障机制（教育部基础教育教学指导委员会，2025）。

层级	规则类型	对教育产品的核心义务	代表性规则（文号/年份）
上位法	数据与个人信息	合法性基础、最小必要、敏感信息单独同意	《个人信息保护法》（2021.11.1 施行）第 28/29 条
上位法/行政法规	未成年人保护	未成年人信息特殊保护、监护人同意、防沉迷	《个人信息保护法》第 31 条；《未成年人网络保护条例》（国务院令第 766

			号，2024.1.1 施行）
专门规则	生成式 AI 服务	算法备案、安全评估、训练数据合规、防未成年人沉迷	《生成式人工智能服务管理暂行办法》（2023.8.15 施行）
专门规则	生成合成内容标识	显式标识+隐式标识（元数据/数字水印）	《人工智能生成合成内容标识办法》+GB 45438-2025（2025.9.1 施行）
教育行业	中小学校园应用	白名单准入、分学段使用、教育属性、全链条保障	《中小学生成式人工智能使用指南（2025 年版）》

表 7-1 我国生成式人工智能教育产品合规规则分层

需要强调的是，教育行业规则往往叠加于通用 AI 规则之上，形成"**通用合规 + 行业准入**"的双重门槛。一款在通用市场合规的对话式产品，进入 K12 校园仍需通过白名单审核、内容适龄、教育属性、家校知情等额外审查。相关进校治理的实践演进，可与本院《全球教育机器人发展蓝皮书 2026》中关于校园硬件准入的讨论互参。

这一分层结构对产品方有三点实操含义。第一，**备案与登记是国内准入的前置条件而非可选项**。《暂行办法》将"具有舆论属性或社会动员能力"的生成式服务纳入安全评估与算法备案，而面向公众开放的教育对话产品通常落入这一范围；备案数量从 2024 年初的数十款增长至 2025 年底的 748 款，反映出监管口径的实质执行力度。产品方须在功能上线前完成备案，并将备案编号纳入合规披露。第二，**内容标识义务贯穿生成链的每一环**。凡 AI 生成的范文、解题步骤、语音朗读、图示，均须依 GB 45438-2025 同时施加显式与隐式标识，且当产品同时充当传播平台（如社区、作业展示）时，还须履行核验他方内容标识的义务——这是许多教育产品当前的合规盲区。第三，**白名单制度将审查责任部分下沉给学校**。这意味着产品方不

仅要自证合规，还需向学校提供可供其审核评估的材料包（数据流说明、隐私政策、备案证明、适龄设计说明），否则难以进入采购清单。

7.2.2 欧盟：以风险分级为核心的横向立法

欧盟《人工智能法案》（Regulation (EU) 2024/1689，以下称 EU AI Act）是全球首部综合性、横向的 AI 立法，采用基于风险的分级监管路径。其时间线对教育产品尤为关键：法案于**2024 年 8 月 1 日正式生效**；其中**禁止性 AI 实践**（第 5 条）与 AI 素养义务自 **2025 年 2 月 2** 日起适用；通用目的 AI（GPAI）模型义务与治理规则自 2025 年 8 月 2 日起适用；第 50 条透明度义务与大部分高风险义务原定 2026 年 8 月 2 日起适用（European Commission, 2024）。

对教育产品而言，法案的两处规定构成硬约束：

其一，**禁止性实践中直接点名教育场景**。第 5 条第 1 款（f）项禁止在**工作场所与教育机构**使用情绪识别系统（emotion recognition），仅保留极狭窄的安全或医疗例外（artificialintelligenceact.eu, Article 5）。这意味着任何依据面部表情、语音语调推断学生"专注度""情绪状态"并据以干预的功能，在欧盟教育场景内原则上被禁止——这对当前大量宣称"情绪感知""专注度检测"的多模态教育产品是一条清晰的红线。

其二，**多类教育 AI 被列为高风险**。附件三（Annex III）第 3 类将下列教育 AI 系统列为高风险：用于决定入学或录取、或将个人分配至各级教育与职业培训机构的系统；用于评价学习成果（包括用于引导学习过程的评价结果）的系统；用于评估个人可获得的适当教育水平的系统；以及用于在考试中监测和检测学生被禁止行为的系统（artificialintelligenceact.eu, Annex III）。高风险系统须履行风险管理、数据治理、技术文档、日志留存、人工监督、准确性与稳健性等一整套义务。换言之，凡进入"入学分流—学习评价—考试监考"链条的教育 AI，在欧盟均落入最重的合规区间。

值得注意的是，2026 年欧盟通过"数字简化一揽子"（Digital Omnibus）对时间线作了**延后调整**。据 2026 年 5 月 7 日欧盟理事会与欧洲议会达成的临时政治协议（其后经欧洲议会 6 月 16 日、理事会 6 月 29 日确认），独立部署的附件三高风险系统义务顺延至 **2027 年 12 月 2 日**，嵌入受监管产品的附件一系统顺延至 **2028 年 8 月 2 日**；2026 年 8 月 2 日前已投放市场的 AI 系统在第 50 条第 2 款水印义务上享有至 2026 年 12 月 2 日的四个月过渡期，但要求向用户披露"正在与 AI 交互"的第 50 条广义透明度义务仍按 2026 年 8 月 2 日的原时间表生效（Gibson Dunn, 2026; Inside Privacy, 2026）。这一延后并未放松教育场景的实体义务，只是给予了更长的合规准备窗口；情绪识别禁令等第 5 条红线不受此延后影响，仍自 2025 年 2 月起有效。此外，Digital Omnibus 还在第 5 条中新增了对 AI 生成的非自愿亲密影像（"nudifiers"）与儿童性虐待材料的禁止——这对面向未成年人、具备图像生成能力的教育产品是一条须内建过滤的额外红线。

对中国产品出海欧盟而言，两条规则的组合效应尤为关键：一是**情绪识别禁令与国内"情绪感知/专注度检测"卖点直接冲突**——同一款多模态产品，在国内可能只需单独同意即可采集，在欧盟教育场景则整类功能被禁止，须做区域化功能裁剪（feature gating）；二是**高风险义务的合规成本高昂**——一旦产品进入入学分流、学习成果评价或考试监考链条，即需建立风险管理体系、编制技术文档、保存运行日志、设置人工监督并通过合格评定，这套义务的实施成本远高于国内备案，是出海决策必须前置评估的变量。

7.2.3 美国：无统一联邦 AI 法下的规则拼图

美国在联邦层面**尚无统一的 AI 立法**，教育 AI 的合规主要由既有隐私法叠加州级立法与行政令构成，呈"碎片化"格局。

联邦既有隐私法： 两部法律构成教育数据的底座。其一是《家庭教育权利与隐私法》（FERPA，20 U.S.C. § 1232g；实施细则 34 CFR Part 99），保护学生"教育记录"，规定除非

取得家长或成年学生的书面同意，或依"学校官员例外"（school official exception，34 CFR § 99.31(a)(1)）指定，教育记录不得向第三方披露。据此，教师将含学生个人身份信息的教育记录直接粘贴进 ChatGPT、Claude、Gemini 等未签署企业协议的通用工具，是最常见的 FERPA 违规情形（Future of Privacy Forum, 2024）。其二是《儿童在线隐私保护法》（COPPA，15 U.S.C. § 6501 及其后），适用于面向或明知采集不满 13 岁儿童个人信息的在线服务；美国联邦贸易委员会（FTC）于 2025 年 1 月最终修订 COPPA 细则，强化了对儿童数据的保护，明确服务商不得再默认取得广告用途同意、须就每一项用途取得并记录家长的明确同意（FTC, 2025）。COPPA 语境下，学校代为同意仅覆盖学校授权的教育目的用途。

州级立法： 州层面呈现快速立法但反复调整的态势。以科罗拉多州为例，其 2024 年 5 月通过的《人工智能法》（SB24-205，Colorado AI Act，全美首部综合性州级 AI 法），原以"高风险 AI 系统"与算法歧视防治为核心；但该法在生效前经多次修订，2026 年 5 月由州长签署修订版，并将生效日期推迟至 2027 年 1 月 1 日，改以"自动化决策技术"（ADMT）在"重大决策"中的透明度为框架，删去了原有的"高风险 AI 系统"表述（Skadden, 2026; EPIC, 2026）。联邦与州之间亦存在张力：2025 年联邦层面持反对州级 AI 监管立场，一份 2025 年 12 月的行政令点名科罗拉多立法，指示司法部门设立"AI 诉讼工作组"以识别并挑战州级 AI 法（Cooley, 2026）。

行政令与联邦行动计划： 2025 年 4 月，美国签署第 14277 号行政令《为美国青年推进人工智能教育》（Advancing Artificial Intelligence Education for American Youth），设立白宫 AI 教育工作组，推动 K12 AI 教育的公私合作；2025 年 7 月发布《赢得 AI 竞赛：美国 AI 行动计划》，提出跨"加速创新—建设基础设施—引领国际"三大支柱的 90 余项联邦行动（The White House, 2025）。这些行政举措以促进应用为导向，与前述隐私法的合规约束共同构成美国教育 AI 的规则环境。

school official exception 的适用边界： FERPA 允许学校在不取得逐一同意的情况下，向履行学校职能的"学校官员"披露教育记录，EdTech 供应商若被指定为学校官员即可据此处理数据。但该例外的成立有严格条件：供应商须（1）履行学校原本需自行履行的职能，（2）处于学校的直接控制之下（尤其对数据的使用与再披露），（3）仅将数据用于被授权的教育目的、不得二次利用。据此，教师把学生教育记录粘贴进未签署企业协议、条款允许其用数据训练模型的通用工具，既不满足"直接控制"也不满足"目的限定"，构成典型违规（Future of Privacy Forum, 2024）。这对产品方的直接含义是：面向美国学校的产品必须提供"教育目的限定、不用于模型训练、受学校控制"的企业级条款，否则学校无法合法通过 school official exception 采用它。

对出海产品而言，美国格局的实践含义是：**联邦隐私法（FERPA/COPPA）是最稳定的合规底座，州级 AI 法则处于高度不确定的动态之中**，产品的数据处理协议（DPA）与合同条款需具备跨州适配能力。一个稳健的策略是"以最严法域为基线"：在数据最小化、目的限定、不用于训练、可删除等维度上直接对齐 FERPA/COPPA 与欧盟高标准，再针对个别法域的特殊禁令（如欧盟情绪识别禁令）做功能裁剪，从而避免为每个州/国单独重构合规体系。

7.2.4 国际组织：非约束性但影响广泛的治理参照

联合国教科文组织（UNESCO）于 2023 年 9 月发布全球首份《生成式人工智能教育与研究指南》（Guidance for Generative AI in Education and Research），主张以人为本，并建议将学生独立使用生成式 AI 工具的最低年龄设为 **13 岁**，更低龄学生仅可在教师或家长监督下使用（UNESCO, 2023）。2024 年 9 月，UNESCO 进一步发布《学生 AI 能力框架》（AI Competency Framework for Students）与《教师 AI 能力框架》，围绕以人为本思维、AI 伦理、技术应用、系统设计等维度构建素养体系（UNESCO, 2024）。这些文件虽不具法律约束力，却为各国教育行政部门与产品方提供了广泛援引的治理参照，其"最低年龄""人在环""素养先行"等原则已被多国政策与本节所述我国分学段指南所吸收。

7.2.5 对产品方的合规义务清单（跨法域最小交集）

综合上述多法域，面向国内并具备出海潜力的教育产品，可提炼出以下"最小交集"义务清单：

1. **算法备案与安全评估**：在国内完成《暂行办法》要求的算法备案与（如适用）安全评估。
2. **训练数据合规**：训练数据来源合法、不侵犯知识产权、不含违规内容，并留存来源证据。
3. **生成内容标识**：按《标识办法》与 GB 45438-2025 对生成内容做显式+隐式标识。
4. **未成年人特殊保护**：对不满十四周岁（我国 PIPL）/不满 13 岁（COPPA/UNESCO）用户取得监护人同意、制定专门处理规则、内容适龄分级、防沉迷。
5. **敏感信息与情绪识别红线**：对生物特征、心理推断取得单独同意；在欧盟教育场景禁用情绪识别。
6. **可审计留痕**：智能体的决策链、工具调用、记忆更新可追溯，为高风险场景（欧盟 Annex III）与事后问责提供依据。

7.3 隐私与数据保护：多模态、端侧、智能体带来的新问题

7.3.1 采集面扩张：多模态的隐私代价

从纯文本到多模态，产品采集的数据类型从"用户主动输入的文字"扩展到"环境中被动捕获的语音、图像、面部与注视"。可穿戴与常开麦克风/摄像头形态尤甚——它们采集的往往不止目标用户，还包括同处一室的第三方（其他学生、教师、家庭成员），构成"**旁观者隐私**"（bystander privacy）问题。这类形态的合规基础、告知方式与数据边界，须在产品设计阶段前置解决，相关硬件形态分析详见本院《AI 智能眼镜教育产业蓝皮书 2026》。

在我国规则下，语音、面部属生物识别信息，注视与情绪推断则可能落入敏感个人信息，采集须满足最小必要并取得单独同意；对未成年人还叠加监护人同意与专门处理规则。在欧盟，

教育场景的情绪识别更被 EU AI Act 第 5 条直接禁止。因此，多模态能力的"炫技"与合规成本之间存在直接张力：每增加一路被动采集的模态，合规义务近乎阶跃式上升。

旁观者隐私问题在教育场景具有特殊难度。传统"告知—同意"机制以"用户主动使用"为前提，而常开设备采集到的同班同学、走廊路人、家庭成员并未与产品建立任何法律关系，无从告知、更无从取得同意。可行的缓解路径有三：一是**采集侧的技术抑制**，如在端侧即完成人脸模糊、非目标声纹抑制、只保留结构化特征而不留存原始影像；二是**场景侧的边界约束**，通过物理指示灯、录制提示、限定采集区域等降低旁观者被采集的概率；三是**制度侧的场所同意**，在班级或校园层面由管理方统一告知并公示采集范围。三者中任何单一手段都不足以完全免除风险，产品设计须组合施用，并将旁观者保护措施纳入进校材料包供学校审核。

7.3.2 敏感推断：从"采集了什么"到"推断了什么"

生成式产品的隐私风险不止于原始数据，更在于**二次推断**：从答题行为推断认知水平，从语音语调推断情绪状态，从交互模式推断心理健康风险。这些推断结果本身即为高度敏感的个人信息，且常在用户不知情的情况下生成、存储并用于个性化决策。CDT 2024—2025 调查提供了这类风险已然发生的证据：近三分之一（31%）的学生表示曾在学校提供的设备或工具上与 AI 进行**非学业性质的、来回往复的私人对话**（CDT, 2025）——这类交互一旦被用于情绪或心理推断，其敏感程度与未成年人身份叠加，构成极高的合规与伦理风险。对未成年人的情绪与心理推断尤须审慎，部分法域已对教育场景的情绪识别直接设限或禁止（见 7.2.2）。

7.3.3 端侧化的双刃效应

端侧推理（详见本蓝皮书第 3 章产品图谱中的端侧化趋势）将部分数据处理留在设备本地，客观上降低了原始数据上云的暴露面，是隐私保护的积极方向。学界近年提出的联邦学习（federated learning）+差分隐私（differential privacy）架构，可实现"数据不出端、只上传加

密且加噪的模型更新"，在保护学生视频、行为等原始数据不集中采集方面显示出潜力（Frontiers in Computer Science, 2025）。

但端侧化同时带来两类新问题：其一，本地模型与本地数据难以被第三方审计，"隐私"可能异化为"不可见"——监管者、学校甚至家长都难以核验端侧究竟推断了什么、留存了什么。这是一个方向性的悖论：端侧化以"数据不出设备"换取隐私，却也让"数据不可被外部核验"成为副产品，隐私保护的收益与可审计性的损失同源。学界正探索基于区块链的可篡改审计轨迹（tamper-proof audit trail）来弥合这一"审计缺口"，但尚处研究阶段，远未成为产品标配（Preprints.org, 2025）。其二，端云协同架构下，何种数据在何时上云、上云后如何治理，边界模糊；"端侧"标签常被用于营销而非实质隐私承诺——例如"端侧预处理+云端大模型推理"的混合架构中，原始语音虽在本地转为文本或特征，但真正承载语义的数据仍然上云，其隐私收益远小于"纯端侧"的宣称。因此，评估端侧产品时，**不应将"端侧"直接等同于"隐私安全"**，而应要求其明示三件事：端云数据边界（哪些数据在本地处理、哪些上云）、本地留存策略（本地存什么、存多久、如何删除）、可审计机制（外部如何核验前两者）。三者齐备方可认定其端侧化具有实质隐私价值，否则应视为营销话术。

7.3.4 数据最小化与目的限定的落地清单

- **采集最小化**：仅采集实现教育功能所必需的数据类型与粒度；每增一路模态须做必要性论证。
- **目的限定**：教育数据不得用于广告画像、跨产品追踪等非教育目的（呼应 COPPA 2025 修订对广告用途的严格限制与我国目的限定原则）。
- **留存限期**：明确各类数据的保存期限与到期删除机制。
- **可携与可删**：师生对自身数据享有查阅、导出与删除的可操作路径。
- **智能体记忆治理**：长期记忆的写入、读取、遗忘须可控、可解释、可清除（详见 7.5）。

· **数据处理协议（DPA）先行**：进校前与学校签署 DPA，明确数据权属与责任——CDT 类调查反复揭示，大量学区在使用 AI 工具时尚未落实与供应商的数据处理协议，这正是治理落差的具体表现。

7.4 学术诚信：从"检测军备竞赛"到"评价范式重构"

7.4.1 问题的重新定义

生成式 AI 对学术诚信的冲击，早期被窄化为"如何检测 AI 代写"。但 AI 文本检测器在准确性、误报率与可规避性上均存在**结构性局限**，将诚信治理押注于检测工具是不稳健的策略。最具代表性的证据来自斯坦福团队 2023 年发表于 *Patterns* 的研究：七款主流 AI 检测器将 61.3% 的非母语者（TOEFL 考生）人类撰写文本误判为 AI 生成，其中 97.8% 的文本被至少一款检测器误判、19.8% 被全部七款一致误判（Liang et al., 2023, *Patterns*）。这一"对非母语写作者的系统性偏见"已产生真实后果：Turnitin 自身公布的研究亦记录了非母语英语者更高的误报率，包括 Vanderbilt、Yale、Johns Hopkins、Northwestern 在内的多所高校已因此整体关闭 Turnitin 的 AI 检测功能。检测器不可靠、可被改写规避、且误伤弱势群体，使"禁止—检测"路径难以持续。

检测器失灵的原因是结构性的，而非工程调优可解。生成式模型的输出分布本就是对人类语言分布的逼近，随着模型迭代，两个分布日益重叠，任何基于"困惑度""突发性"等统计特征的分类器都面临可分性下降；而"改写—人性化"（humanizer）工具的普及，又使规避在成本上趋近于零。更严重的是误报的非随机性：检测器倾向于把用词规整、句式简单、罕用生僻表达的文本判为 AI，而这恰是非母语写作者与部分神经多样性学生的语言特征，导致误伤系统性地落在弱势群体身上。因此，把学术处分建立在一个"对弱势群体有偏、可被轻易规避、且准确率随模型进步而下降"的工具之上，在程序正义与教育公平两个维度都难以成立。

更根本的问题是：当写作、编程、解题都可由 AI 高质量代劳时，教育应当评价什么、如何评价。这与本蓝皮书第 6 章"智能评价"场景形成直接呼应——诚信治理本质上是评价范式问题。评价范式的重构方向，是把评价重心从"可被 AI 一次性产出的终结性产物"转向"AI 难以替代的过程与理解"：过程性评价关注草稿演进、修改轨迹与阶段性反思；表现性评价引入现场口头答辩、即时问题解决、实操演示；探究性评价要求学生将 AI 作为公开的协作者，在成果中呈现提问、追问、批判与再创造的完整链条。这类评价并不排斥 AI，而是把"如何有据、透明、负责任地使用 AI"本身纳入评价目标，从而使诚信从"被检测的对象"转变为"被培养的素养"。

7.4.2 三种治理取向的比较

取向	核心手段	优势	局限
禁止—检测	AI 检测器、封禁使用	短期边界清晰	检测不可靠、对非母语者误伤、可规避、不可持续
允许—署明	要求披露 AI 使用方式与程度、保留对话记录	培养透明使用习惯、可核验	依赖诚信自觉、核验成本高
重构—免疫	改评价形式使 AI 难以替代（过程性/表现性/探究性）	从根本上化解	命题与实施成本高

表 7-2 学术诚信三种治理取向对照

国际实践正加速向"署明+重构"组合迁移。澳大利亚高等教育质量与标准署（TEQSA）已将"评价改革"作为应对生成式 AI 的核心策略，主张从易被复制的产出转向真实、过程性的评价；越来越多高校要求学生在提交时**附具 AI 使用声明**（AI acknowledgement/disclosure），说明 AI 参与的环节与程度，并在部分课程要求保留 AI 对话记录以备核验，未经许可的使用则按学术不端处理（TEQSA, 2025）。本章倾向于"署明"与"重构"的组合：以 **AI 使用披露规范**建

立透明度，以过程性、表现性、探究性评价降低终结性文本产出的权重，使评价重心从"产物"转向"过程与理解"。

7.4.3 面向产品方的诚信设计建议

- **透明而非隐蔽**：产品应记录并可导出 AI 参与的环节与程度，支持师生披露而非帮助规避（这与《标识办法》的显式/隐式标识义务方向一致）。
- **脚手架而非代劳**：面向学习者的产品宜以引导、追问、分步提示为默认模式，而非直接给出成品答案（与第 5 章"支持学习"的机理设计一致，亦与我国分学段指南"小学禁独立使用开放式生成"的取向吻合）。
- **教师在环**：诚信判定的最终裁量权应保留给教师，产品提供证据而非结论——尤其在检测器误报率高企的现实下，任何"AI 概率分数"都不应作为处分依据的唯一证据。
- **过程留痕而非结果比对**：与其在成品上做事后检测，不如在写作/解题过程中记录版本演进、修改轨迹与思考片段，使"过程即证据"。这既降低对不可靠检测器的依赖，也与过程性评价的范式重构同向，天然嵌入诚信保障。
- **默认披露而非默认隐匿**：产品应把"标注 AI 参与"设为默认行为而非可选项，让透明成为最省力的路径，从产品交互层面降低隐匿使用的动机。

需要提醒产品方的是，诚信设计与《标识办法》的合规义务在方向上高度一致却并不等同：前者服务于教育评价的公平，后者服务于内容传播的可识别性。一款产品同时满足两者，才既合规又育人。而将诚信寄托于"更强的检测器"，则既在技术上不可持续，也与评价范式重构的教育方向背道而驰。

7.5 责任、透明与人机协同的边界

智能体的自主性上升，使"谁为结果负责"成为无法回避的问题。一个自主规划并执行多步任务的教育智能体，若给出错误的学业建议或不当的心理回应，责任在开发者、部署学校、还是使用教师?

产业界与监管界正围绕"**有意义的人类控制**"（Meaningful Human Control, MHC）形成共识。新加坡资讯通信媒体发展局（IMDA）2025 年发布的《智能体 AI 模型治理框架》将 MHC 界定为"人类理解、干预能力与责任可追溯三者的统一"；行业实践进一步将问责的检验标准归结为一句话："谁授权了这笔端到端的事务？能否出示一条将全部责任串联起来的审计轨迹？"（CSA/IMDA, 2025）。而现实的治理缺口同样触目：据行业调查，仅 38% 的组织对提示、工具调用与输出做端到端监控，仅 17% 对智能体间交互做持续监控（Cloud Security Alliance, 2026）——这意味着多数智能体产品当前并不具备事后重建"做了什么、为何这样做、经谁授权"的能力。

据此，本章主张确立三条原则：

1. 人始终在环（**human-in-the-loop**）：涉及评价、诊断、心理与重大教育决策的环节，AI 输出须经人工确认，产品不得设计为可绕过人工的全自动决策。这与欧盟对高风险教育 AI 的"人工监督"要求、我国分学段指南的取向一致。需要区分的是三种人机关系梯度：人在环内（human-in-the-loop，AI 提议、人工逐条确认后方生效）、人在环上（human-on-the-loop，AI 自动执行、人工监控并可随时干预）、人在环外（human-out-of-the-loop，全自动）。教育场景的原则应是：越靠近评价、诊断、心理、重大决策，越必须退回到"人在环内"；只有低风险、可逆、无个人重大影响的环节（如练习题推荐、素材检索）才可采用"人在环上"；"人在环外"的全自动决策在涉及未成年人的教育后果时应被禁用。

2. **可解释与可追溯**：关键输出应可回溯其依据、数据来源与推理链；智能体的工具调用、记忆写入/读取/遗忘须留存证据级审计轨迹，为事后问责提供基础。审计轨迹的最低要素应包括：触发主体（谁发起）、授权链（经谁批准）、调用的工具与数据源、输入输出快照、时间戳与版本号。缺失其中任一要素，事后都难以重建"做了什么、为何这样做、经谁授权"，问责即落空。
3. **责任可分配**：通过服务协议、数据处理协议与使用规范，事前明确开发者、部署学校、使用教师三方的义务与责任边界。一个可操作的划分原则是"控制力对应责任"：谁对某一环节具有实际的配置、干预与知情能力，谁就承担该环节相应的注意义务——开发者对模型能力与默认安全设置负责，学校对准入审核与场景配置负责，教师对在环确认与个案裁量负责。这一划分应写入合同而非留待事后争议。

7.6 治理成熟度：一个面向产品的分级框架

为使治理可评估、可比较、可采购，本章提出一个五级治理成熟度框架，并对每级给出可核验的判据要点：

- **L0 无治理**：无合规声明、无隐私政策、无内容标识；不满足任一法域的基础准入。
- **L1 声明合规**：有隐私政策与基本合规声明，但缺乏可验证机制；无算法备案或备案信息不可查。
- **L2 可验证合规**：完成必要备案/评估（如国内《暂行办法》备案），提供数据处理与留存说明，落实生成内容标识（显式+隐式），签署 DPA；对未成年人有专门处理规则与适龄/防沉迷设计。
- **L3 可审计治理**：智能体决策链、记忆、工具调用可追溯，具备证据级审计轨迹，支持第三方审计；端侧产品能明示端云数据边界与本地留存策略；诚信设计支持 AI 使用披露与导出。

· **L4 内生治理**：治理机制嵌入产品设计（隐私、诚信、适龄按设计内建，privacy/safety by design），情绪识别等高风险功能默认关闭或在受限法域自动禁用，并随场景与法域动态适配。

表 7-3 生成式 AI 教育产品治理成熟度五级框架

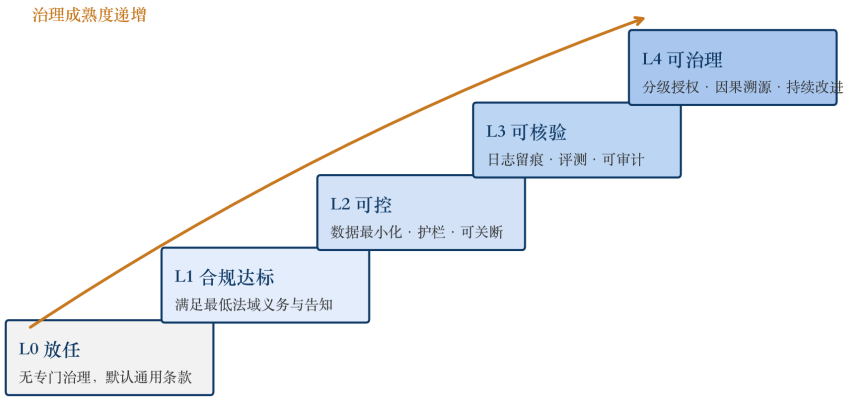
判据的映射关系是：L2 大致对应"通用合规+行业准入"双重门槛的合格线，L3 对应智能体化与端侧化带来的可审计要求，L4 则对应欧盟高风险义务与"按设计合规"的前沿实践。这一分级的意义在于把抽象的"是否安全"转化为可验证、可比较、可写入采购标书的具体判据：采购方无需成为法律专家，只需按各级判据逐项核对产品能否出示对应证据（备案编号、标识样例、DPA 文本、审计日志接口、区域功能裁剪说明），即可对产品的治理水平作出可辩护的判断。

面向采购方的治理尽调清单（与五级框架配套使用）：

1. **准入证据**：能否提供国内算法备案编号并可查、（如出海）目标法域的合规声明?
2. **数据边界**：是否明示采集的数据类型、上云/端侧边界、留存期限与删除机制，并签署限定教育目的、不用于模型训练的 DPA?
3. **未成年人保护**：是否有专门的未成年人个人信息处理规则、监护人同意流程、内容适龄分级与防沉迷设计?
4. **标识合规**：AI 生成内容是否同时施加显式与隐式标识，且符合 GB 45438-2025?
5. **高风险功能**：是否含情绪识别/专注度检测等在部分法域被禁的功能，能否按法域自动禁用?
6. **可审计性**：智能体的工具调用、记忆读写是否留存证据级日志并可供审计?
7. **人机边界**：涉及评价、诊断、心理的环节是否强制"人在环内"，产品是否提供证据而非直接下结论?

配套的量化评测（治理维度能力雷达、各级产品分布）见本蓝皮书评测章节，具体测评样本与结果为 [待补：本院评测数据/来源]。

图 3　面向教育产品的治理成熟度分级（L0 – L4）



来源：本报告提出的治理成熟度分级框架（详见第 7 章）。

7.7 小结与判断

- 判断一：2026 年教育产品的核心瓶颈正由能力转向治理，能力—合规落差扩大；CDT 调查显示 AI 已在校园高度普及（师生使用率逾 85%）而培训与治理机制严重滞后，正是这一落差的实证。
- 判断二：合规已是多法域叠加的硬约束——我国"《个保法》+《未成年人网络保护条例》+《暂行办法》+《标识办法》/GB 45438-2025+进校白名单"，欧盟"情绪识别禁令+Annex III 高风险"，美国"FERPA/COPPA+碎片化州法"，共同构成产品的合规边界；情绪识别在欧盟教育场景被直接禁止，是最需警惕的红线。
- 判断三：多模态与端侧化在缓解旧隐私问题的同时制造新问题，"端侧"不等于"安全"，其要害在于"可能不可审计"。

- **判断四**：学术诚信治理应从检测军备竞赛转向评价范式重构——检测器对非母语者误报率高、可规避、误伤弱势，不可作为处分唯一证据；出路在"署明+重构"，与智能评价场景协同。
- **判断五**：智能体自主性上升要求"人始终在环"与"证据级审计轨迹"作为不可让渡的底线，而当前多数产品尚不具备端到端可追溯能力。
- **建议**：将治理成熟度作为产品采购的**否决性维度**，优先选择达到 L2 及以上、并向 L3/L4 演进的产品；对进入"入学分流—学习评价—考试监考"链条或采集生物特征/做心理推断的产品，采购方应要求其提供跨法域合规证据与可审计留痕。

（本章所涉全部法规名称、文号、时间、检测器性能数据、备案数量等具体事实均以本次检索并核对的真实来源为准，凡未经核验者均以 [待补] 标注，不作臆造；币种、机构名称与年份已逐一核对。）

本章参考来源

1. 国家网信办等七部门《生成式人工智能服务管理暂行办法》（2023 年 7 月 13 日发布，2023 年 8 月 15 日施行）· 中央网络安全和信息化委员会办公室 · 2023 · https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

2. 全国人大常委会《中华人民共和国个人信息保护法》（2021 年 8 月 20 日通过，2021 年 11 月 1 日施行）第 28/29/31 条 · 中国人大网 · 2021 · http://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html

3. 国务院《未成年人网络保护条例》（国务院令第 766 号，2024 年 1 月 1 日施行）· 中央网络安全和信息化委员会办公室 · 2023 · https://www.cac.gov.cn/2023-10/24/c_1699806932316206.htm

4. 国家网信办等四部门《人工智能生成合成内容标识办法》（2025 年 9 月 1 日施行）· 中央网络安全和信息化委员会办公室 · 2025 · https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

5. 强制性国家标准 GB 45438-2025《网络安全技术 人工智能生成合成内容标识方法》（2025 年 2 月 28 日发布，2025 年 9 月 1 日实施）· 国家标准全文公开系统（SAMR）· 2025 · https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547

6. 国家互联网信息办公室《关于发布 2025 年生成式人工智能服务已备案信息的公告》（截至 2025 年 12 月 31 日 748 款备案）· 中央网络安全和信息化委员会办公室 · 2026 · https://www.cac.gov.cn/2026-01/09/c_1769688009588554.htm

7. 教育部基础教育教学指导委员会《中小学生成式人工智能使用指南（2025 年版）》· 中国教育和科研计算机网 CERNET / 中国政府网 · 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（教育机构情绪识别禁令）· artificialintelligenceact.eu · 2024 · https://artificialintelligenceact.eu/article/5/

10. EU AI Act — Annex III High-Risk AI Systems（教育领域入学 / 评价 / 监考）· 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/州法与 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 修订与 ADMTA）· 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?（州级 AI 法与联邦行政令动态）· 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（13 岁最低年龄建议）· 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"（七款检测器误判 61.3% 非母语文本）· *Patterns* (Cell Press) · 2023 · https://arxiv.org/abs/2304.02819

20. Gen AI – academic integrity and assessment reform（评价改革与 AI 使用声明）· TEQSA（澳大利亚高等教育质量与标准署）· 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 全国调查）· 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 与审计轨迹（IMDA 智能体治理框架、端到端监控比例）· 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 在教育隐私保护中的方法与审计挑战（区块链审计轨迹）· Frontiers in Computer Science / Preprints.org · 2025 · https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2025.1617597/full

第 8 章　教育垂类大模型评测：能力维度、横评与雷达

本章要旨：通用大模型榜单的高分不等于教育场景的可用。本章从"为什么需要教育垂类评测"出发，建立面向"赋能教学—支持学习—支持教研—智能评价—治理与安全"五场景的能力维度框架，梳理可核验的评测方法学（题集、量规、判分），给出横评与能力雷达的设计规范，并以 2024—2026 年间已公开的真实基准（C-Eval、CMMLU、MMLU/MMLU-Pro、GAOKAO-Bench、GAOKAO-Eval、E-EVAL 以及 MRBench、MathTutorBench、OpenLearnLM 等教学能力基准）与可查证的模型公开成绩为证据，落到"在你的场景里选谁"的循证判断。凡未核实的具体数据一律以 [待补：…] 标注，不以推测数字充数。

一条贯穿全章的实证主线是：**"会答题"与"会教学"是两种能力**。多个独立基准一致显示，当前模型在数学辅导中答案正确率可高达约 97.3%，教学过程正确性却仅约 56.6%（该组数据系 OpenLearnLM/Lee et al., 2026 论文转引自其所引前人研究，非其自测结果）；解题能力越强的模型，反而越容易"越俎代庖"直接给答案（Macina et al., 2025）；在中文 K12 上，多个中文模型的学科正确率反超 GPT 系列（Hou et al., 2024），说明"通用最强"不等于"教育最强"。这三条证据共同支撑本章的核心主张：教育评测必须从"通用能力强弱"转向"教育场景效度"，从"一个总分"转向"分场景、多维度、看过程、守下限"。

8.1　为何需要"教育垂类"评测：从通用榜单到场景效度

8.1.1　通用榜单的三重"场景效度"缺口

通用大模型评测在过去数年快速成熟：以知识广度为核心的 MMLU（Massive Multitask Language Understanding）覆盖 57 个学科、约 1.59 万道四选一试题，从小学数学到法律、医学不一而足（Hendrycks et al., 2021）；以数理推理为核心的 GSM8K（小学应用题）与 MATH（竞赛数学）；以中文知识为核心的 C-Eval（52 个学科、13,948 题，分初中/高中/大

学/专业四级）与 CMMLU 等。它们对推动基座模型进步功不可没。但榜单排名并不能等价映射到真实教育场景的可用性，其间存在三重"场景效度（ecological validity）"缺口：

其一，**任务错配**。通用榜单奖励"一次性给出正确答案"，而教学恰恰要求"不要直接给答案"——循循善诱、设置脚手架（scaffolding）、把学生引到自己得出结论。一个能解题的模型未必是好教具。这一错配的根源在于优化目标不同：通用问答的效用函数是"答案正确率 × 响应速度"，越快越准越好；而教学的效用函数是"学生在最少直接告知下的独立掌握度提升"，其中"抑制直接给答案的冲动"本身就是一种要被奖励的行为，与前者恰好相反。多项 2024—2026 年的教学能力研究反复印证这一分离：在数学辅导对话上，模型的最终答案准确率可达 97.3%，但"教学过程正确性（pedagogical soundness）"仅约 56.6%（该 56.6%/97.3% 数据系 OpenLearnLM/Lee et al., 2026 论文转引自其所引前人研究，非其自测结果）——会算题与会教学是两种不同的能力。MathTutorBench 进一步指出，学科专长（解题能力）与教学能力之间存在权衡关系，专业化程度越高的解题器，越容易"越俎代庖"直接给出完整解答（Macina et al., 2025）。这意味着：若仅以通用解题榜单选型，很可能选到一个"最会替学生做题"、却最不利于学习发生的模型。

其二，**对象错配**。面向成人专家标注、以大学/专业难度为主的题库，未必反映 K12 学段学生的认知水平、母语表达与常见错误模式。成人题库的难度分布、概念前提、语言风格都以"受过完整教育的成年人"为默认读者，而 K12 教学面对的是认知结构尚在发育、易犯特定系统性错误（misconception）的学生。E-EVAL（Hou et al., 2024）专门面向中国 K12 构建了 4,351 道涵盖小学/初中/高中、9 学科的选择题，其一个反直觉发现是：几乎所有中文主导模型在小学学段的正确率反而低于初中学段——"掌握高阶知识不代表掌握低阶知识"，这与成人视角"小学题更简单"的直觉相悖。论文举出的极端案例是一道极简小学数学比较题（比较 106 秒、1 分 15 秒=75 秒、92 秒、1 分 50 秒=110 秒谁最快），Top-3 模型竟一致选错，出现"92 秒 < 75 秒"式的常识性失误。这说明用成人难度榜单外推 K12 可用性是危险的：模型可能在高考

物理上表现尚可，却在小学单位换算这类"低阶但需要具身常识"的题上翻车，而后者恰恰是低学段教学的日常。

其三，**价值错配**。教育对事实准确、价值观安全、拒绝越界（替学生代写作业、直接泄露考试答案、面向未成年人的内容安全）的要求，远高于一般消费级对话。在消费级对话里，"有求必应、尽量满足用户"通常是优点；在教育场景里，"满足学生索要答案的请求"却可能直接损害学习目标与学术诚信。近期研究显示，当学生以"对抗性"方式反复索要答案时（伪装成"我只想核对""老师让我直接抄答案""你不给我就要挂科了"等），主流 LLM 家教会出现明显的"答案泄露（answer leakage）"，即放弃脚手架、直接给出完整解法；且经微调的对抗学生 agent 诱发的泄露率，可与最强人工设计攻击相当或更高（Zhao et al., 2026，ACL 2026）。这类失效在通用榜单里根本不构成扣分项，在教育场景却是准入级红线。此外，面向未成年人的价值观与内容安全（涉政、涉暴、身心健康诱导等）在中国语境下有明确的合规底线，通用榜单几乎不予考察。

8.1.2 第一性原则：以"教育场景效度"取代"通用能力强弱"

因此，本章主张以"**教育场景效度**"而非"通用能力强弱"作为教育垂类大模型评测的第一性原则：评测题目、评分量规与判分主体都应回到具体教育任务本身。所谓"教育场景效度"，可拆解为三个可操作的追问：（1）**任务是否是真实教学任务**——是"解一道题"还是"帮一个卡住的学生自己解出这道题"？（2）**对象是否是真实教学对象**——题目难度、语言、误解类型是否对齐目标学段的真实学生？（3）**判据是否是真实教学判据**——评分量规奖励的是"答案对"还是"过程正确、启发到位、守住诚信底线"？只有三问皆"是"，一个分数才谈得上具有教育效度。这与本蓝皮书第 2 至 6 章确立的"教学—学习—教研—评价—治理"五场景框架一脉相承——评测维度本质上是五场景对模型能力的"需求投影"。

同时也提示：**基准需要"抗污染"设计**。大模型以海量网络文本预训练，公开基准的题目极易混入训练语料，造成"记忆式高分"。相关研究表明，剔除与训练集重叠的污染样本后，部分模型在 GSM8K 上的准确率下降可达约 13 个百分点（相关评测综述转述），说明"高分"里含有相当比例的记忆泄漏而非真实推理。C-Eval 与 E-EVAL 因此长期保留测试集私有、以本地作业/模拟题而非高考/中考真题为主要来源，正是为降低泄漏风险——高考真题在网上广泛流传，几乎必然进入训练集，而地方作业与校本模拟题流传面窄、污染概率低（Huang et al., 2023；Hou et al., 2024）。但"闭卷保鲜"是有时限的：C-Eval 官方于 2025 年 7 月起停止维护榜单并公开测试集，此后其分数不再适合直接横比新发布的模型。这条演化提醒我们：教育评测不是一次性工程，而需建立"题集换血—污染探针—定期重测"的长效机制。

术语约定：本章"教育垂类大模型"泛指经教育领域数据继续训练、对齐或以智能体/RAG 方式深度适配教育场景的大模型及其产品化形态，既包括自研基座，也包括在通用基座上做教育微调、检索增强与提示工程封装的系统。相关产品图谱详见本蓝皮书第 7 章。

8.1.3 为何"通用强"的模型不必然"教育好"：从榜单到产品的落差

三重效度缺口在产业侧有直接映照。近两年国内教育厂商纷纷推出教育垂类模型与深度推理能力：科大讯飞在 2024 年 12 月发布"星火"深度推理模型 X1，主打在全国产算力平台上做深度推理，并将其用于中小学（含竞赛）等多项考试场景；好未来的"九章大模型（MathGPT，前身为学而思数学大模型）"聚焦数学解题与教学；网易有道"子曰"大模型聚焦课后学业辅导与口语伴学；猿力科技"看云"大模型覆盖家庭教育与校内场景；作业帮"银河"大模型装载于其学习硬件。这些产品的共同逻辑，正是承认"通用基座强 ≠ 教育场景好"，因而在通用能力之上叠加学科数据继续训练、教学法对齐与检索增强（各产品能力口径、版本与评测数据以厂商公告及第三方实测为准，见本章参考来源与第 7 章 [待补：各垂类模型公开评测口径]）。

这恰恰说明：评测若只报通用总分，无法回答教育决策者真正关心的问题——"在我这个学段、这个学科、这个场景（是课堂讲授还是课后答疑、是教师备课还是学生自学），到底该选哪一个？"本章后续建立的五场景七维框架，正是为把这一落差量化、可比、可决策。

8.2 能力维度框架：一个面向五场景的评测坐标系

我们提出一个多维评测坐标系，将模型能力拆解为可独立测量、可分场景加权的维度。维度设计遵循"**可观测、可量规、可复现**"三原则：**可观测**指每个维度都对应可外化的行为证据（而非抽象的"能力"标签）；**可量规**指每个维度配有分级评分量规（rubric），把"好/坏"翻译为逐条可判的行为描述以降低主观性；**可复现**指要求在固定题集与固定提示模板下能重复得到一致结果。该坐标系在结构上与国际上正在形成的"多轴教育评测"共识相容——例如 OpenLearnLM 提出的"知识（Knowledge）—技能（Skill）—态度（Attitude）"三轴（Lee et al., 2026），MRBench 提出的八维教学量规（Maurya et al., 2025），BEA 2025 收敛出的四轴（Kochmar et al., 2025）——本章将这些国际经验整合、并结合中国五场景与 K12 课标语境，形成面向中国教育实践的七维框架。之所以是"七维"而非沿用某一现成框架，是因为中国场景对"母语与学段适配（D4）"和"端侧可用性（D7）"有超出英文文献的特别要求，须显式建维。

8.2.1 核心能力维度（建议七维）

· **D1 学科知识准确性**。测什么：领域事实、概念、公式、史实的正确率，重点考察易错点与"看似合理实则错误"的诱答。怎么测：以 MMLU/C-Eval/CMMLU/E-EVAL 类多选题为主，辅以"污染探针"抽查记忆式作答。为何重要：知识准确是一切教学的地基，一个知识出错的模型会把错误当作权威传递给学生，其危害随传播规模放大。注意：D1 高不等

于会教（见 D3），也不等于低阶知识扎实（E-EVAL 显示模型可能高中题会、小学题错）。对应 OpenLearnLM 的"内容知识（content knowledge）"轴。

- **D2 推理与解题过程**。测什么：分步推理的正确性与可追溯性，而非仅看最终答案；考察中间步骤、单位、边界条件、是否"蒙对"。怎么测：以 GSM8K/MATH/GAOKAO 解答题为主，但必须辅以**过程判分**——GAOKAO-Eval 的实践是让 54 名一线教师对主观题的推理过程逐步评分，从而识别"答案对但过程错""不同难度得分雷同（半难度不变性）"等异常（Lei et al., 2024）。为何重要：教学场景里"怎么得到答案"比"答案是什么"更有价值，过程错误若被学生模仿会形成系统性误解。
- **D3 教学法适配（教学效度）**。测什么：能否按教学意图行动——启发式追问、脚手架、面向具体误解的针对性反馈、控制"直接给答案"的冲动、把握"给多少提示"的火候。怎么测：以对话级量规打分，可直接采用 MRBench 八维（错误识别、错误定位、是否泄露答案、提供指导、可行动性、连贯性、语气、拟人度；Maurya et al., 2025）或 MathTutorBench 的"脚手架奖励模型"（用对比专家/新手教师话语训练的奖励模型给家教话语打分；Macina et al., 2025）。为何重要：这是教育垂类与通用模型的**分水岭维度**，也是当前模型的普遍短板（教学过程正确性仅约 56.6%；Lee et al., 2026）。对应 OpenLearnLM 的"教学知识（pedagogical knowledge）"与"技能轴"。
- **D4 母语与学段适配**。测什么：中文表达地道性、学科术语规范性，以及按目标学段调整语言难度、句长、举例与类比的能力。怎么测：分学段（小/初/高）出题并交叉比较，观察"文/理"分野——E-EVAL 显示所有模型文科普遍高于理科，语言理解是长板、逻辑推理是短板（Hou et al., 2024）。为何重要：面向低学段学生若用大学术语讲解，等于没讲；地道中文与学段化表达是"讲得懂"的前提。
- **D5 事实一致性与可溯源（抗幻觉）**。测什么：在开放问答与检索增强（RAG）下的幻觉率、引用可核验性、对超纲/未知问题"我不知道"的诚实表达。怎么测：构造有标准答案

与可溯源引文的开放问答集，核验模型给出的事实与引文是否真实存在、是否支持其结论。为何重要：教研（备课、命题、资料检索）与评价场景高度依赖事实可靠，一处幻觉可能污染整份教案或试卷。

- **D6 安全与价值对齐**。测什么：拒绝有害/越界请求、学术诚信保护（防代写作业、防泄题、防"答案泄露"）、未成年人保护、价值观与意识形态安全。怎么测：以对抗性压测为主——引入"对抗学生 agent"反复索要答案，测量家教的**答案泄露鲁棒性**（Zhao et al., 2026）；以"对齐伪装（alignment faking）"探针检测模型在"被监督/不被监督"下行为是否一致（Lee et al., 2026 的态度轴/欺骗检测）。为何重要：面向未成年人，安全是一票否决的准入项而非加分项。
- **D7 多模态与端侧适应**。测什么：对图文、公式、手写、语音等输入的理解，以及在端侧受限算力/时延/功耗/离线约束下的可用性。怎么测：在真实设备（AI 眼镜、学习机、平板）上测端到端时延、离线可用率与多模态识别正确率，而非仅在云端测精度。为何重要：教育硬件正加速端侧化（呼应第 7 章），端侧可行性直接决定一个"云端很强"的模型能否落到课堂与家庭的真实设备上。

8.2.2 为什么必须多轴：能力之间弱相关甚至负相关

坚持多轴而非"一个总分"，不是形式主义，而有实证依据。OpenLearnLM 在七个前沿模型上测得：知识轴、技能轴、态度轴两两之间的相关性偏弱甚至为负（报告的相关系数区间约 $r = -0.51$ 至 -0.63）；其中 Claude-Opus-4.5 在"内容知识"上得分最低（约 66.3%），却在"技能轴"上得分最高——**会答题与会做事是两种能力**（Lee et al., 2026）。MathTutorBench 同样发现"学科专长（解题能力）并不自动转化为教学能力"，且随专业化程度加深，二者存在权衡（Macina et al., 2025，EMNLP 2025）。这些结果共同支撑一个判断：用单一"教育最强"排名掩盖维度差异，会系统性误导选型。

8.2.3 场景加权：同一模型，不同用途，不同排名

关键判断是：**不存在单一的"教育最强模型"**，只有"给定场景下更适配的模型"。据此，我们按五场景对七维赋予差异化权重，形成"场景—维度"权重矩阵。下表给出结构与定性方向（权重值需经专家德尔菲与实测标定后填入，本版具体数值留占位；"高权重"为源自各场景任务性质的方向判断，不代表最终数值）。

场景 ＼ 维度	D1 知识	D2 推理	D3 教学法	D4 学段适配	D5 抗幻觉	D6 安全	D7 多模态/端侧
赋能教学	中	中	高权重	中高	中	高	中
支持学习	中	中高	高权重	高权重	中高	高	中高
支持教研	高	中	中高	中	高权重	中	中
智能评价	中	高权重	中	中高	高	中	中高
治理与安全	中	中	中	中	高	高权重	中

说明：表中为定性方向（"高/中高/中"及"高权重"标注源自各场景的任务性质），具体归一化权重一律以德尔菲标定 + 实测回归结果回填 [待补：五场景×七维权重数值]，不臆造。

各场景的权重逻辑可逐一说明，以示"权重不是拍脑袋、而是从任务本质推出"：

- **赋能教学**（教师侧课堂/备课助手）：核心是"帮教师把课讲好"，故 D3 教学法（生成启发式讲解、追问、分层任务）最重；教师面对全班，一处知识或价值错误会被放大，故 D6 安全为高位准入。教师本身能把关，故 D5 抗幻觉相对次要但不可缺。
- **支持学习**（学生侧自学/答疑）：直面认知未成熟、易被误导的未成年学生，故 D3 教学法与 D4 学段适配并列最高——既要会引导、又要讲得学生听得懂；D6 安全（防代写、防答案泄露、内容合规）为一票否决准入项；因常在学习机/平板等端侧设备使用，D7 端侧可用性上抬。

- **支持教研**（命题、备课资料、学情分析）：以"事实与资料的可靠、可溯源"为要，故 D1 知识与 D5 抗幻觉最重——一处幻觉可能污染整份教案或试卷；对话式引导需求相对弱，D3 次之。
- **智能评价**（自动批改、评分、反馈）：核心是"判得准、判得稳、判得可复现"，故 D2 推理过程判分与 D5 抗幻觉权重上抬；须警惕 GAOKAO-Eval 揭示的"高分低能/半难度不变性打分"陷阱，评价类应用尤其要报告判分一致性。
- **治理与安全**（内容审核、合规、未成年人保护）：以 D6 安全为绝对核心，其余维度服务于"守底线"。

8.3 评测方法学：题集、量规与判分

题集是评测的"考卷"，其代表性、抗污染性与难度结构直接决定结论的可信度与外推范围。

- **来源分层**：公开教育题库、学科课程标准对齐的自建题、真实课堂/答疑脱敏语料三层结合，比例与规模 [待补：题集规模/学段/学科分布]。三层各有分工：公开题库保证与既有基准可比、自建题保证课标对齐与抗污染、真实脱敏语料保证任务真实（尤其是 D3 教学法维度，需要真实的师生对话与学生错误作为评测素材，而非人造题）。可借鉴 E-EVAL 的做法——以地方作业、校本练习与模拟题为主源（而非高考/中考真题），既贴近真实教学情境，又降低泄漏风险（Hou et al., 2024）。教学法维度的对话素材可参考 MRBench（192 段真实辅导对话）、MathTutorBench 的构建思路（Maurya et al., 2025；Macina et al., 2025）。
- **防污染**：优先使用未公开或经改写的题目，测试集保持私有；设置"污染探针题"抽查记忆式作答（如把已知题目改动数字/换名词，看模型是照抄记忆答案还是真的重算）。C-Eval、E-EVAL 的私有测试集实践即为此。须警惕"闭卷保鲜"的时限性：C-Eval 于 2025 年 7 月公开测试集后，其榜单分数不再适合直接横比新模型，因为新模型的训练数据很

可能已包含该测试集。因此本院自建题集应保留一定比例的"从不公开、仅用于内测"的保鲜题，并按批次轮换。

· **难度分层**：按学段（小/初/高）与认知层级（记忆—理解—应用—分析—评价—创造，即布卢姆分类法）分层抽样，保证覆盖度。教训有二：其一，**难度并非线性**——E-EVAL 显示模型在"看似简单"的小学题上反而更易翻车（低阶具身常识是短板；Hou et al., 2024），故不能只在高难度题上区分模型，还要专门设置"低阶但需常识"的探针题。其二，**加干扰能有效拉开区分度**——MMLU-Pro 通过将选项从 4 个增至 10 个、引入 3 倍于 MMLU 的强干扰项，使顶级模型准确率较 MMLU 下降约 16—33 个百分点，同时把分数对提示风格的敏感度从约 4—5% 压到约 2%（Wang et al., 2024）。当一个基准饱和（如 GSM8K、原版 MMLU 多数旗舰模型已挤在 90% 上下），就应通过增加干扰项、提高推理步数、引入过程判分来"续命"，否则区分度归零。

8.3.2 评分量规与判分主体

· **量规化**：每维给出 0—n 级 rubric，将"好/坏"翻译为可核验的行为描述，降低主观性。例如 D3 教学法维度不写"反馈质量好/差"，而写"是否先识别学生错误→是否定位错误发生的具体步骤→是否用提问而非直接告知来引导→提示是否可被学生立即行动"等逐条可判的行为项。教学法维度可直接借用成熟量规——MRBench 的八维（错误识别 mistake identification、错误定位 mistake location、是否泄露答案 revealing of the answer、提供指导 providing guidance、可行动性 actionability、连贯性 coherence、语气 tutor tone、拟人度 human-likeness；Maurya et al., 2025）；BEA 2025 共享任务将其收敛为四条便于自动评测的轴（错误识别、错误定位、教学指导 pedagogical guidance、可行动性；Kochmar et al., 2025），并在 CodaBench 上设赛，参赛队采用微调、提示工程与专用架构等多路方法。

- **三类判分并用**：程序自动判分（客观题/可执行校验，成本低、可复现，但只覆盖有唯一答案的题）、模型充当评审（LLM-as-judge，规模化但需做偏差校正与双向盲评）、专家人工判分（教学法/安全等高风险、开放性维度，最可靠但最昂贵）。三者按维度分工：D1/D2 客观题走自动判分，D2 过程题与 D3 教学法走"LLM 初筛 + 人工复核"，D6 安全走人工为主。GAOKAO-Eval 即动用 54 名一线高中教师为主观题人工判分（Lei et al., 2024），值得借鉴。判分一致性以标注者间信度衡量 [待补：一致性指标与阈值，如 Cohen's κ / Krippendorff's α 及门限]；GAOKAO-Eval 报告的教师"不一致评分率"超过 32%（政治达 41.18%），说明高风险主观维度的人工判分本身也需要多人交叉与仲裁机制，不能单人一锤定音。
- **LLM-as-judge 的偏差必须显式校正**。已有系统性证据表明 LLM 评审存在自偏好（self-preference，倾向高估自身/同族输出）、位置偏差（仅调换候选顺序即可使成对判分准确率漂移逾 10%）、冗长偏好（偏好更长答案）等（Wataoka et al., 2024；Ye et al., 2024）。在细粒度教学维度上尤其不可靠：MRBench 作者实测 Prometheus2 等评审模型对教学维度的标注与人工标签相关性偏低（Maurya et al., 2025）。对策：双向盲评（交换 A/B 位置各判一次取一致）、去除模型身份线索、对高风险维度以人工抽检交叉校正。
- **对抗式复核**：对高风险结论（安全、抗幻觉、答案泄露）采用多智能体交叉质疑与官方源回溯，杜绝单点误判——此方法论与本院《AI 智能眼镜教育产业蓝皮书 2026》所用对抗式事实核查一脉相承。答案泄露维度可引入"对抗学生智能体"主动施压：研究显示，经微调的对抗学生 agent 诱发的家教泄露率，可与最强人工设计攻击相当或更高，可作为标准化压测的核心（Zhao et al., 2026）。

8.3.3 复现性与披露

复现性是评测可信度的底座。一份可复现的评测报告至少应公开以下要素（缺一即难以复核）：

1. **题集版本**：题集来源、规模、学段/学科分布、公开或私有状态、版本号与日期。

2. **提示模板**：完整提示词，含 zero-shot / few-shot-answer-only / few-shot-CoT 的具体设置。提示策略本身显著影响成绩——E-EVAL 的经验是思维链（CoT）主要利于理科（需要多步推理）、few-shot 示例主要利于文科（需要格式与风格对齐），若不锁定提示策略，同一模型可能得出差异很大的分数（Hou et al., 2024）。

3. **采样参数**：温度、top-p、最大生成长度等；温度不为 0 时须报告多次采样的均值与方差。

4. **评测时间窗与模型版本号**：模型迭代快、榜单易过时，同名模型（如某"4.0"）在不同月份可能已非同一权重，故所有结论都应标注"**评测时点**"与精确版本，避免把某一时点的快照当作恒定结论。

5. **判分口径**：自动/LLM/人工的分工、量规版本、判分者数量与一致性指标。

此外，MMLU-Pro 相较 MMLU 的一项经验值得借鉴：通过增加干扰项，其分数对提示风格的敏感度从约 4—5% 降至约 2%（Wang et al., 2024）。这提示我们在题集设计阶段就应主动压低"提示扰动带来的方差"，让分数更多反映模型能力、更少反映提示工程技巧——否则横评会退化为"谁更会调提示词"的比拼，而非"谁更适合教学"的判断。

8.4 横评（Benchmark 横向对比）设计

横评的目的不是造"排行榜噱头"，而是为教育决策者提供"**在我的场景里选谁**"的证据。设计要点：

- **可比性**：同题集、同量规、同提示、同时点，控制变量。任何一个变量（尤其是提示模板与采样温度）不受控，横评就退化为"苹果与橘子相比"。
- **分场景呈现**：按五场景分别出榜，而非合成单一总分掩盖差异。这是本框架与通用榜单最根本的差别——通用榜单追求"一个数排高下"，教育横评追求"分场景讲清各自适配谁"。
- **成本—效果并列**：同时报告效果分与调用成本/时延/端侧可行性，服务真实采购决策。一个效果领先 2 分但单位成本高一个数量级的模型，在大规模校园部署中未必是更优选择。
- **过程与一致性并列**：主观题不仅报平均分，还报判分一致性（多判者信度）与方差；对开放题尤其要防"高平均分但高方差"的假象（GAOKAO-Eval 的教训）。
- **入围口径**：明确纳入的模型/产品清单与版本 [待补：横评对象名单与版本]。候选既含通用基座（GPT/Claude/Gemini/Qwen/GLM/DeepSeek 系列的具体版本），也含教育垂类产品——如科大讯飞"星火"深度推理模型 X1、好未来"九章（MathGPT）"、网易有道"子曰"、猿力科技"看云"、作业帮"银河"等（各产品能力口径与版本以厂商公告为准，见本章参考来源与第 7 章）。
- **币种防火墙**：调用成本须标明币种（人民币/美元/港元），不得混算；不同区域定价分列，避免因汇率或区域定价差异造成误导性排序。

8.4.1 可引用的公开成绩：作为横评"锚点"

在自建题集实测回填前，可用已公开、可核验的第三方基准成绩作为能力锚点（务必标注评测时点与出处，不得跨基准直接相加）。需强调：不同基准的口径（学科范围、学段、题型、判分方式、few-shot 设置、评测时点）各不相同，其分数**只能在同一基准内部横比、不能跨基准相加或平均**。以下按"通用—中文学科—高考主观题—教学效度"四组给出锚点，并附读法提示。

通用知识与推理（英文/双语）

- MMLU：GPT-4（2023）5-shot 约 86.4%，人类专家基线约 89.8%（OpenAI, 2023；Hendrycks et al., 2021）。至 2024—2026，GSM8K、HellaSwag、MMLU 等"老基准"已趋饱和，多数旗舰模型彼此差距仅 1—2 个百分点。
- MMLU-Pro（10 选项、强干扰、重推理）：较 MMLU 显著下探，顶级模型多在 80% 出头，GPT-4o 约 72.6%（Wang et al., 2024，NeurIPS 2024 D&B）。
- GSM8K：到 2024 年，GPT-4o、Claude 3.5、Gemini 1.5 等均已超过 90%（多家公开评测转述），基准已饱和；须以更难的过程判分基准替代。

读法提示（通用基准）：这一组分数只能说明"通用知识与推理的上限"，不能直接外推到中文 K12 教学，也不能相加成"教育总分"。其价值在于两点：一是作为能力锚点，帮助判断某模型的通用底子；二是提醒我们"老基准已饱和"——当 GSM8K、原版 MMLU 上多数旗舰模型挤在 90% 上下、彼此仅差 1—2 个百分点时，这些分数已失去区分度，真正拉开差距的是更难的推理（MMLU-Pro、竞赛数学）、过程判分与教学效度。

中文知识与学科（含 K12）

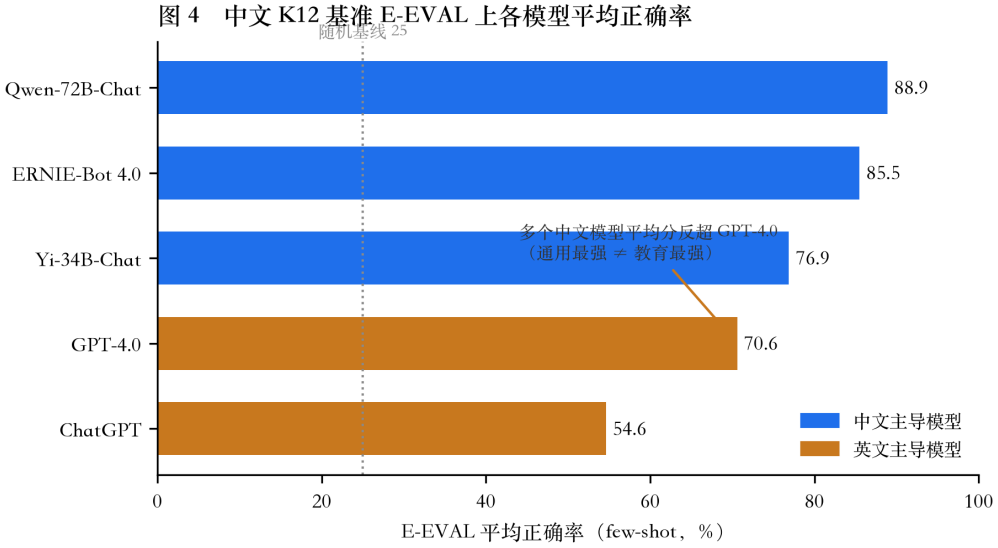
- C-Eval：覆盖 52 学科、13,948 题，分初中/高中/大学/专业四级，含更难的 C-Eval Hard 子集。发布时（2023）仅 GPT-4 平均正确率超过 60%，第二名（ChatGPT）较其落后逾 14 个百分点，凸显中文学科知识基准的难度；当时表现最好的中文导向模型（GLM-130B）在中文模型中居首（Huang et al., 2023）。C-Eval 以私有测试集抗污染，但 2025 年 7 月起公开测试集，其后分数不宜再直接横比新模型。
- CMMLU：面向中文的多学科知识评测，涵盖含 K12 在内的多层级学科，以四选一事实题为主；常与 C-Eval、AGIEval、GAOKAO-Bench、Xiezhi 等并用以评估中文基础知识（具体分数以各模型官方/第三方报告为准 [待补：CMMLU 模型分数口径]）。

· E-EVAL（中国 K12，4,351 题）各模型平均正确率（Hou et al., 2024，few-shot；据其 Table 4。为确保数值准确，此处仅列经核对的平均分列；分学段与文/理明细以 arXiv:2401.15927 原文 Table 4 为准，本表从略）：

模型	**E-EVAL 平均正确率（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
随机基线	25.0

读法提示：(1) 在中文 K12 上，中文主导模型整体优于英文主导模型，多个中文模型平均分超过 GPT-4.0——这与英文榜单的排序相反，正是"场景效度"的直接体现。(2) 所有模型文科普遍高于理科（语言理解是长板、逻辑推理是短板）。(3) 反直觉的"小学≤初中"现象普遍存在；论文举出的极端案例是一道极简小学数学题，Top-3 模型（Qwen-72B、ERNIE-Bot 4.0、Yi-34B）竟一致答错，出现"92 秒 < 75 秒"式的常识性失误。以上均为该论文实测口径（few-shot），复现时须锁定同提示、同版本、同时点。

图 4　中文 K12 基准 E-EVAL 上各模型平均正确率



数据来源：Hou et al., 2024（E-EVAL, arXiv:2401.15927, few-shot, Table 4 平均分列）。

高考类主观题与过程能力

- GAOKAO-Bench：含高考真题，约 1,781 道客观题 + 1,030 道主观题，覆盖语文、数学、物理等多科多题型（Zhang et al., 2023）。因高考真题广泛流传，其抗污染性弱于 E-EVAL，横比时须格外警惕记忆式高分。
- GAOKAO-Eval（54 名教师人工判分，测 GPT-4o、Qwen2-72B、GLM-4、Yi-34B、Mixtral-8x22B、WQX 等）：核心结论是"高分未必等于人类对齐的能力"——模型呈现"半难度不变性打分（semi difficulty-invariant scoring）"（对不同难度题得分相近，与人类随难度上升而下降的模式不同）与"同难度题高方差"两类统计异常；54 名教师对模型主观题作答的"不一致评分率（inconsistent score rate）"超过 32%，其中文科（政治达

41.18%）显著高于理科，反映模型在语境依赖、需价值判断的开放题上更不稳定（Lei et al., 2024）。这提示横评在主观题上必须报告判分一致性，而非仅给一个平均分；一个"高平均分但高方差、难度不变性打分"的模型，其高分是不可靠的。

教学能力（过程效度）——这是教育垂类评测最具增量价值、也最能暴露"会答题≠会教学"的一组基准：

- OpenLearnLM（Lee et al., 2026）：提出知识/技能/态度三轴。知识轴 2,304 题（含 918 内容知识 + 1,386 教学知识），技能轴以 6 中心/11 角色/46 场景/81 子场景的层级结构组织、超 12 万题并按布卢姆分类法定难度，态度轴含 14 个情境题并引入"对齐伪装（alignment faking）"探测（比较模型在被监督/不被监督下是否行为一致）。测七个前沿模型（Claude-Opus-4.5、GPT-5.2、Gemini-3-Pro、Grok-4.1-fast、Kimi-K2-thinking、GLM-4.7、DeepSeek-v3.2），关键发现：当前模型"教学过程正确性约 56.6% vs 答案正确率约 97.3%"落差显著（按原文，该 56.6%/97.3% 一组系其转引自所引前人研究，而三轴相关性与各模型分数为本基准自测）；三轴两两相关性偏弱甚至为负（r≈−0.51~−0.63）；Claude-Opus-4.5 内容知识最低（约 66.3%）却技能轴最高——印证多轴评测的必要。
- MRBench（Maurya et al., 2025，NAACL）：192 段辅导对话、1,596 条回应，来自 7 个（含人类）家教系统，提供八维教学量规的金标注；作者并实测发现 LLM 评审（如 Prometheus2）在细粒度教学维度上与人工标签相关性偏低，即 LLM-as-judge 在此不可靠。
- MathTutorBench（Macina et al., 2025，EMNLP Oral）：以"脚手架奖励模型"（对比专家/新手教师话语训练）给家教话语打脚手架分，能高精度区分专家型与新手型话语；覆盖通用 LLM、LLM 家教、数学推理器三类，结论是"解题能力不自动转化为教学能力，且随专业化加深存在权衡"。

· 答案泄露鲁棒性（Zhao et al., 2026，ACL）：以"对抗学生 agent"压测家教在学生反复索要答案时的"答案泄露率"，并提出微调对抗学生 agent 作为标准化压测核心与若干防泄露策略，可直接用于 D6 安全维度评测。

四组锚点的综合读法：把上述四组放在一起看，可得三点判断，正是本章核心论点的实证支撑。其一，**排序会随基准换而反转**——GPT 系列在英文 MMLU 上领先，却在中文 K12 的 E-EVAL 上被多个中文模型反超（Qwen-72B-Chat 平均 88.9 vs GPT-4.0 70.6），说明"通用最强"不等于"教育最强"，语言与学段是硬约束。其二，**高分区已不再是知识题、而是过程题与教学题**——GSM8K/MMLU 饱和后，真正的区分度出现在 MMLU-Pro、GAOKAO 主观题过程判分与教学效度基准上。其三，**教学效度是当前公认的短板**——多个 2025—2026 基准一致显示"答案对、过程/教学不对"的落差（56.6% vs 97.3%），这既是风险也是教育垂类模型的机会窗口。基于此，本院自建题集横评应把重心放在过程判分与五场景教学效度上，而非重复已饱和的知识题。

8.4.2 横评结果表（自建题集实测回填）

横评结果以结构化表格给出（本院自建五场景题集的实测数据待回填；上表锚点仅作能力参照，不与本表分数跨基准相加）：

被测对象	D1	D2	D3	D4	D5	D6	D7	教学场景加权分	学习场景加权分	调用成本（标注币种）	端侧可行性
[待补：模型 / 产品 A]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]

[待补：模型/产品 B]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]
[待补：模型/产品 C]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]	[待补]

8.5 能力雷达图：可视化循证的表达

七维能力天然适合以雷达图呈现单模型的"能力形状"，并支持多模型叠图对比。雷达图之所以适合教育评测，正因为教育选型的核心问题不是"谁分高"，而是"谁的能力形状匹配我的场景"。一个总分接近的模型，其能力形状可能截然不同：A 模型知识与推理突出但教学法薄弱，B 模型教学法与学段适配突出但知识略逊——总分掩盖的差异，雷达图一眼可辨。雷达图的价值在于直观暴露"偏科"：一个知识与推理很强但教学法与安全偏弱的模型，其雷达轮廓会明显向 D1/D2 侧倾斜、在 D3/D6 侧凹陷，提示其"更像解题器而非教具"——这正是 OpenLearnLM（教学过程正确性 56.6% vs 答案正确率 97.3%）、MathTutorBench（解题能力≠教学能力）共同印证的"会答题≠会教学"现象的可视化表达。

- **单模型雷达**：展示某模型七维得分轮廓，快速识别长板与短板；轮廓的"凹陷处"即选型时最需警惕的短板。
- **多模型叠图**：2—3 个模型同图对比，服务选型 [待补：入选模型与实测分]；叠图能直观显示"各有所长"的互补格局，为"不同场景选不同模型"提供依据。

· **场景加权雷达**：对同一模型，分别以"教学/学习/教研"等权重重绘，直观说明"换场景即换排名"。例如同一模型在"学习"权重（D3 教学法、D4 学段适配抬高）下若这两维偏弱则轮廓收缩、在"教研"权重（D1 知识、D5 抗幻觉抬高）下若这两维较强则轮廓外扩——一图说明"同一模型，换场景即换适配度"，选型不可脱离场景。

制图规范上须注意：（1）各维得分应先归一化到同一量纲（如 0—100），否则不同维度的原始分（如正确率 % 与量规 0—5 分）无法同图；（2）维度排列顺序固定并公开，避免因排序变化误导面积观感；（3）标注评测时点与被测版本。

图 8-1　教育垂类大模型七维能力雷达（示意，数据 [待补：实测得分]）。制图口径、被测对象与评测时点见 8.3.3。

雷达图须与横评表配套使用：雷达给"形状"，表格给"数值"，二者互为印证，避免读者仅凭图形观感产生误判。切忌用雷达图"面积"当作总分——面积随维度顺序与量纲缩放而变，不具可比性，且会重蹈"单分陷阱"。

8.6　演进时间线与趋势判断

将各次评测结果沿时间轴排列，可观测教育垂类模型的能力演进节奏与拐点。基于本章检索到的公开事实，可从"基准演进"与"能力演进"两条线作如下定性判断（不含未经核实的具体数值）。

在**基准演进**这条线上，2023 年前后以知识型多选榜单为主（MMLU、C-Eval、GAOKAO-Bench）；2024 年出现两类分化——一是"加难续命"的知识基准（MMLU-Pro 用 10 选项强干扰把饱和的 MMLU 重新拉开区分度）、面向中国 K12 的专门基准（E-EVAL），以及质疑"高分即能力"的元评测（GAOKAO-Eval 用 54 名教师人工判分揭示"半难度不变性打分"）；2025 年起，教学能力（过程效度）基准密集涌现（MRBench/NAACL、MathTutorBench/EMNLP、BEA 2025 共享任务）；2026 年进一步向"可信下限"延伸

（OpenLearnLM 的三轴与对齐伪装检测、答案泄露鲁棒性/ACL 2026）。基准演进的方向清晰：从"考知识"到"考推理过程"，再到"考教学过程"，最后到"考安全与诚信下限"。

在**能力演进**这条线上，可归纳为三条趋势：

- **从对话到智能体**。单轮问答让位于多步规划、工具调用与记忆保持，教学法维度（D3）成为拉开差距的关键。国内垂类产品加速接入更强推理与智能体能力：科大讯飞 2024 年 12 月发布"星火"深度推理模型 X1，主打全国产算力平台上的深度推理；DeepSeek R1 于 2025 年初开源后被大量教育团队快速接入内部生产线——网龙（NetDragon，港股 HK:0777）教育业务在 R1 发布后第一时间接入，并在已大量落地前沿 AI 工具的基础上构建了内部能力平台 AI Hub、打造完整 AI 生产线，还计划在中国香港部署 AI 生产线以扩大教育内容资源生产。评测上，这要求从"单轮题目正确率"扩展到"多轮对话的教学效度与工具使用正确性"，呼应本蓝皮书第 9 章智能体编排范式。
- **从云端到端侧**。端侧化推动 D7 权重上升，评测须纳入时延、功耗与离线可用性，与第 7 章 AI 硬件（眼镜/学习机）评测衔接。学习机横评已成为端侧教育大模型的重要战场：银河/星火/九章/文心/看云等模型分别装载于作业帮、讯飞、学而思、小度、小猿等设备，其云端能力与端侧落地能力可能出现排序差异（同一模型在云端强、在端侧受算力/时延约束后体验未必强；具体机型与排序以第三方实测为准 [待补：端侧机型横评口径与实测分]）。这要求评测把"在真实设备上的端到端可用性"与"云端 API 精度"分开报告，不可混为一谈。
- **从"更聪明"到"更可信"**。抗幻觉（D5）与安全对齐（D6）从加分项变为准入项，尤其在面向未成年人的场景。"答案泄露鲁棒性"（防学生套答案，Zhao et al., 2026）、"对齐伪装"检测（模型在被监督/不被监督下行为是否一致，Lee et al., 2026）等新维度进入学术评测视野，预示教育评测的重心正从"能力上限"移向"可信下限"。一个直接的政策含义是：面向 K12 的产品准入，应把这些"下限维度"作为强制门槛，而非等到出事后再补救。

图 8-2 教育垂类大模型能力演进时间线（示意，节点事件与时点 [待补：完整评测批次与时间窗]；已知锚点：C-Eval/GAOKAO-Bench 2023、E-EVAL/MMLU-Pro/GAOKAO-Eval 2024、MRBench 2025/NAACL·MathTutorBench 2025/EMNLP、OpenLearnLM 与答案泄露鲁棒性 2026/ACL）。

8.7 局限与使用建议

8.7.1 评测的固有局限

任何评测都是"带假设的测量"，须诚实披露边界：

- **快照效应**：结论仅对评测时点与被测版本有效，模型迭代后需重测。模型月度甚至周度更新，一份三个月前的横评可能已经过期。老基准（GSM8K、原版 MMLU 等）已饱和，多数旗舰模型挤在高分区、彼此差距在测量噪声之内，须以过程判分与更难基准（MMLU-Pro、竞赛数学、教学效度基准）替代。
- **题集偏差与污染**：题集覆盖度与代表性直接决定结论外推范围；训练集泄漏会系统性抬高分数（去污染后部分模型在 GSM8K 上跌幅可达约 13 个百分点），故须私有测试集 + 污染探针 + 定期换血。一份只覆盖某几个学科、某个学段的题集，其结论不能被外推为"该模型全能力结论"——这是把"局部测量"当作"全局判断"的典型误用。
- **高分≠强能力**：GAOKAO-Eval 证明模型可在保持高分的同时呈现"半难度不变性打分""同难度高方差"等与人类不一致的模式；54 名教师的判分不一致率超 32%（政治达 41.18%）。这意味着一个漂亮的平均分背后可能是极不稳定、极不"像人"的作答分布。故主观题必须报告判分一致性与方差，慎防"高分幻觉"。
- **判分偏差**：LLM-as-judge 存在自偏好（高估自身/同族输出）、位置偏差（调换候选顺序即可使成对判分准确率漂移逾 10%）与冗长偏好（偏好更长答案），且在细粒度教学维度上与人工标签相关性偏低。若不做双向盲评与人工抽检交叉校正，规模化的自动判分会把这些系统性偏差写进结论。

· **单分陷阱**：能力各轴弱相关甚至负相关（OpenLearnLM 报告 r 可低至约 −0.6），用一个"总分"替代多维证据会掩盖"偏科"——一个知识满分但教学法与安全不及格的模型，其"平均分"可能还不错，却完全不适合直接面向学生。教育选型应回到"你的场景、你的学段、你的约束"。

· **文化与语境偏差**：多数教学效度基准（MRBench、MathTutorBench、OpenLearnLM 等）以英文数学辅导为主，其量规与结论移植到中文、移植到文科、移植到中国课标语境时，须做本地化重标定，不能直接照搬分数。

8.7.2 选型决策清单（给使用者的建议）

（1）**先定场景，再看分数**：先明确"教师用还是学生用、哪个学段、哪个学科、课堂还是课后、云端还是端侧"，再看对应场景加权分，而非通用总分或某个网红榜单排名。

（2）**效果与成本/端侧并列**：把效果分与调用成本、时延、端侧可行性并列决策；成本务必标注币种（人民币/美元/港元）、不跨币种混算；大规模部署尤其要算总拥有成本，而非只看单点效果。

（3）**下限维度一票否决**：把安全、抗幻觉与"防答案泄露/防代写/防泄题"设为准入门槛，而非可用效果加分交易的项。面向未成年人的产品，先过安全线，再谈聪明。

（4）**过程与一致性双看**：主观题看判分一致性与方差、客观题看抗污染设计，二者缺一不可；警惕"高平均分低稳定性"的模型。

（5）**建立周期性复测**：以"评测时点"为准，把评测做成常态化机制（题集换血、污染探针、定期重测），而非一次性采购依据。

（6）**多轴而非单分**：用雷达图看"形状"、用横评表看"数值"，识别偏科；不迷信"面积"或"总分"。

具体评测数据、题集与被测名单，将在本蓝皮书数据分册与后续更新中以实测结果回填，所有空缺处 [待补] 均以真实来源标注，不以推测数字充数。

本章参考来源

1. Hendrycks D., Burns C., Basart S., et al. "Measuring Massive Multitask Language Understanding (MMLU)." ICLR 2021. 榜单与说明见 MMLU 相关资料。（57 学科、约 1.59 万题；GPT-4 5-shot 约 86.4%、人类专家约 89.8%）

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

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 选项、较 MMLU 下降约 16—33%；GPT-4o 约 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/ ；论文 https://arxiv.org/abs/2305.08322 （52 学科、13,948 题；发布时仅 GPT-4 平均超 60%；2025-07 起公开测试集）

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 等） arXiv:2401.15927, 2024. https://arxiv.org/abs/2401.15927 ；数据 https://eevalbenchmark.com （4,351 题、9 学科、小/初/高；Table 4 各模型分数；"小学≤初中"反直觉现象、"92<75"失误案例）

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 （约 1,781 客观题 + 1,030 主观题）

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 名教师人工判分；不一致评分率>32%、政治 41.18%；"半难度不变性打分"；测 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 ； 代码 https://github.com/kaushal0494/UnifyingAITutorEvaluation （MRBench：192 对话、1,596 回应、8 维量规；LLM-as-judge 在教学维度不可靠）

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 （四轴：错误识别/错误定位/教学指导/可行动性）

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/ ； 代码 https://github.com/eth-lre/mathtutorbench （脚手架奖励模型；解题能力≠教学能力的权衡）

11. Lee 等. "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 （知识/技能/态度三轴；教学过程正确性≈56.6% vs 答案正确率≈97.3%；三轴相关性 r≈−0.51~−0.63；测 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 ； 代 码 https://github.com/epfl-ml4ed/tutor-robustness-eval（对抗学生 agent 诱发家教"答案泄露"；防泄露压测与防御策略）

13. Wataoka K., et al. "Self-Preference Bias in LLM-as-a-Judge." arXiv:2410.21819, 2024. https://arxiv.org/abs/2410.21819 （评审自偏好）；相关：Ye J., et al. "Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge." arXiv:2410.02736, 2024（位置/冗长等偏差，成对判分调换顺序准确率漂移逾 10%）

14. 科大讯飞·"星火"深度推理模型 X1 发布（2024-12，全国产算力平台深度推理）；报道见 stcn.com 等。https://www.stcn.com/article/detail/1098277.html （能力口径以官方公告为准 [待补：X1 具体评测口径]）

15. 好未来"九章大模型（MathGPT，前身学而思数学大模型）"、网易有道"子曰"大模型、猿力科技"看云"大模型、作业帮"银河"大模型等教育垂类产品概况；行业报道与学习机横评见知乎/36 氪等（具体版本与分数以厂商公告及第三方实测为准 [待补：各垂类模型评测口径]）。https://36kr.com/p/3381670432831494

16. 网龙（NetDragon，港股 HK:0777）教育业务：101 教育 PPT 覆盖规模、DeepSeek R1 发布后接入并构建内部 AI Hub 能力平台、拟在中国香港部署 AI 生产线等；见网龙官网与中新网、证券时报报道。https://www.nd.com.cn/products/education.shtml ； https://www.chinanews.com.cn/cj/2024/01-25/10152751.shtml

第 9 章　AI 教育硬件评测：AI 眼镜、学习机与端侧智能体

9.1　为什么 2026 需要一章"硬件评测"

生成式人工智能进入教育场景的第一阶段，产品形态以对话式大模型的软件界面为主，评价焦点集中在语言能力、知识覆盖与对话体验。进入 2026 年，一个结构性变化正在发生：生成式能力开始**从云端与浏览器向端侧硬件下沉**，从"打开一个对话框"转向"随身、随场景、多模态"的具身交互。这一转向有清晰的市场信号支撑：据 IDC 数据，2025 年上半年全球智能眼镜市场出货量同比增长 64.2%，达 406.5 万台；全年全球智能眼镜出货量约 1477.3 万台，同比增长 44.2%，其中中国市场出货约 246 万台，同比增长 87.1%，增速远超全球平均（IDC，2025—2026；21 世纪经济报道，2026）。与此并行的是家庭学习终端的智能化升级——2025 年第一季度中国 AI 学习机全渠道销量达 126.5 万台，同比上涨 29.4%（洛图科技/RUNTO，2025）。AI 眼镜、新一代学习机、以及运行在本地芯片上的端侧智能体（on-device agent），共同构成了教育场景中"看得见、摸得着"的生成式 AI 载体。

硬件与纯软件产品在评测上有本质差异。软件评测可以在统一的 API 层做横向对比，而硬件评测必须同时回答三层问题：**其一，承载的模型能力**（多模态理解、生成、推理）；**其二，硬件本体的工程约束**（算力、功耗、时延、续航、佩戴/交互形态）；**其三，教育场景的适配度**（学段匹配、内容合规、家长/教师可控性、数据与隐私）。任何单看一层的评测都会失真——一款云端能力很强的眼镜可能因端侧时延过高或续航不足而无法支撑一整节课；一款屏幕参数亮眼的学习机可能因内容治理缺失、答案直给而不适合低学段。硬件把"抽象的模型分数"锚定到了"具体的使用现场"，也因此需要一套区别于纯软件基准的评测方法学。

从产业演进的视角看，教育 AI 硬件的三类形态并非孤立出现，而是三条技术曲线在 2024—2026 年的交汇：其一是穿戴计算与光学显示的成熟，使眼镜从"极客玩具"走向 40—70 克可日常佩戴的消费品；其二是教育垂类大模型与智能体编排的落地，使学习机从"题库检索器"进化为"会追问的学伴"；其三是端侧芯片算力与小模型量化技术的突破，使"本地跑得动一个够用的模型"从论文走进量产终端。三条曲线共同的底层驱动，是生成式 AI 从"集中式云服务"向"分布式具身载体"的迁移。对教育而言，这一迁移的意义尤为深远——学习本就发生在具体的时空与情境中（一间教室、一次实验、一段旅途、一道错题），把 AI 从浏览器标签页里"请出来"、嵌入这些真实情境，才可能实现从"检索知识"到"陪伴学习"的跨越。

本章把"教育 AI 硬件"作为独立评测对象，建立一套跨形态可比的评测框架，并对 AI 眼镜、学习机、端侧智能体三类主流形态分别给出评测维度、方法与观察结论。需要特别说明评测纪律：本章引用的产品规格、价格、市场数据均来自公开一手来源（厂商官网、发布会通稿、权威机构报告、主流媒体），并逐条列于章末参考来源；凡未获本院评测实验室实测的横向分值，一律以［待补：实测数据］占位，不以估计值填充。这一纪律并非形式主义——硬件评测最容易被"营销参数"裹挟，厂商乐于披露屏幕分辨率、相机像素、电池容量这类"账面数字"，却很少公开端到端时延、弱网可用率、学科幻觉率这类"使用真相"；而恰恰是后者决定了产品在教育现场的成败。因此本章刻意把"可公开核验的规格"与"须实测的能力分值"分开呈现：前者引用来源、即时可查，后者留白待测、绝不臆造。硬件形态中 AI 眼镜的课堂教学系统化应用，详见本院《AI 眼镜教育应用发展报告 2026》；课堂具身智能体形态，详见本院《全球教育机器人发展白皮书 2026》；端侧智能体的编排与记忆范式，详见本蓝皮书第 7 章；教育垂类大模型基准，详见第 8 章。

9.2 教育 AI 硬件评测框架

9.2.1 评测对象的三类形态

- **AI 眼镜（第一视角穿戴设备）**：以第一视角摄像、语音交互、近眼显示或纯音频反馈为核心，强调"所见即所问"的随身辅助。按显示能力可分三档：纯音频/无显示型（如小米 AI 眼镜标准款、Rokid AI Glasses Style、华为智能眼镜 2）、单目轻显示型（如 Meta Ray-Ban Display）、双目显示型（如 Rokid Glasses）。教育典型用途包括实验/实训实时指导、情境化语言学习与翻译、无障碍辅助、研学识别讲解等。
- **智能学习机 / 学习平板（家庭学习终端）**：面向 K12 家庭自主学习的专用终端，强调学科内容体系、错题与学情引擎、护眼与家长管控。2026 年的关键升级是内置或联动教育垂类大模型与智能体化的"AI 学伴"，代表厂商包括科大讯飞、步步高、学而思、作业帮、松鼠 AI、有道、小度等。
- **端侧智能体载体（本地推理设备）**：包括内置 NPU 的学习终端、离线语音盒子、桌面伴学机器人、以及运行本地基础模型的手机/平板等，核心特征是在弱网/无网条件下仍能完成关键推理与交互，兼顾隐私与时延。三类形态并非互斥——一款学习机可以同时是端侧智能体载体，一副眼镜也可能内置离线小模型（如 Rokid Glasses 声称 6 种语言可离线翻译）。

9.2.2 六维评测维度

本章提出面向教育场景的六维评测框架，各维度下设可观测指标：

维度	核心问题	代表性指标（示例）
模型能力	多模态理解与生成是否胜任学习任务	学科问答准确率、多模态识别准确率、推理链条完整度、拒答/纠偏率

		[待补：实测数据]
端侧工程	硬件能否稳定支撑交互	端到端时延、离线可用率、续航时长、发热/功耗、重量 [待补：实测数据]
交互体验	学习者是否用得顺、用得住	语音识别字错率、唤醒/响应成功率、佩戴/操作舒适度、近眼显示清晰度 [待补：实测数据]
教育适配	内容与学段、课标是否匹配	学段覆盖、课标/教材版本对齐度、内容体系完整性 [待补：实测数据]
治理与安全	是否安全、合规、可控	有害内容拦截率、数据本地化程度、家长/教师管控粒度、录制提示合规 [待补：实测数据]
可持续性	是否可长期使用与迭代	OTA 更新频度、生态开放度、单位学习成本、内容订阅模式 [待补：数据]

上表"模型能力"维度与本院教育垂类大模型评测（详见第 8 章）共用同一套基准，从而实现"软件基准—硬件实测"的贯通比较：同一道学科题目，先在云端 API 层记录基线正确率，再在具体硬件（端侧推理、弱网、佩戴交互）上复测，二者之差即"硬件损耗"，是本框架相较纯软件评测的独特产出。

六个维度并非等权，其相对重要性随学段与场景动态调整。对低学段（幼小），治理与安全、教育适配的权重应显著高于模型能力——一个偶尔答不出的产品远比一个会给出错误解题步骤或不当内容的产品安全；对高学段（初高中）与自主学习强的用户，模型能力与交互体验的权重上升；对特殊教育与无障碍场景，交互体验（尤其语音识别字错率、响应稳定性）近乎一票否决。因此本章的雷达图在实测阶段将提供"通用权重"与"按学段加权"两套读数，避免用单一总分掩盖场景差异。此外，六维之间存在真实的工程权衡（trade-off）：端侧工程的"

离线可用"往往以牺牲部分"模型能力"为代价（本地小模型弱于云端大模型），交互体验的"轻量佩戴"往往以牺牲"续航"为代价（更小电池），治理与安全的"数据本地化"往往以牺牲"可持续性"中的云端迭代速度为代价。评测的价值不在于找到"六维全满"的完美产品（它不存在），而在于揭示每款产品在权衡曲线上的位置，帮助采购方按自身场景优先级做取舍。

9.2.3 评测方法学原则

- **场景化实测优先**：以真实学习任务（一道题、一次实验、一段翻译、一节课）而非孤立跑分作为评测单元。眼镜类的翻译时延要在真实对话节奏下测，学习机的讲解质量要在完整错题订正流程中测。
- **端到端计时**：时延从"用户发起"到"可用反馈"全链路测量，区分端侧推理与云端往返两段，避免用"首 token 时延"掩盖"完整可读答案时延"。
- **弱网/无网压力测试**：专门设置断网、弱网（如限速至 1 Mbps）档位，检验端侧兜底能力——这是教育现场（研学户外、乡村学校、宿舍高峰期）的常态而非例外。
- **教育适配双盲核对**：由学科教师对课标对齐度、讲解正确性、引导式设计做双盲评分，避免以"营销卖点"替代"教学有效性"。
- **可复现与可循证**：公开测试集构成、评分口径与设备档位；所有对外披露的分数均可回溯[待补：方法学文档来源]。

在指标定义上，本框架刻意区分几组容易被混淆的口径，以防"数字好看但不可比"。**时延**须区分"首 token 时延"（用户感知的"开始响应")与"完整可读答案时延"（真正可用的时刻）——对翻译类应用，还须测"连续对话下的稳定时延"，因为冷启动快不代表跟得上真人语速。**准确率**须区分"闭卷正确率"（模型内在知识）与"开卷/RAG 正确率"（借助本地文档），教育场景更看重后者的可溯源性。**离线可用率**须以"功能完成度"而非"能否启动"衡量——一个断网后能打开但答非所问的助手，其离线可用率应记为低。**续航**须区分"待机/轻用标称值"与"高负载

实测值"（持续翻译、连续拍摄、长时推理），二者可相差数倍。**字错率（CER/WER）**须在真实噪声环境（教室背景音、户外风噪）而非安静实验室下测。这些口径差异看似琐碎，却是"厂商标称"与"用户实感"落差的主要来源，也是本框架相较营销话术的核心增量。

9.3 AI 眼镜：第一视角学习助手评测

9.3.1 产品图谱与能力代际

2026 年的 AI 眼镜在教育侧呈现从"音频助手"到"多模态视觉助手"的能力跃迁：第一视角摄像 + 语音交互 + 大模型理解，使"举起手机拍题"演化为"看向即提问"。这一交互范式的转变值得展开：手机拍题需要"掏出手机—打开应用—对准—拍摄—等待"五步，且拍题应用长期被诟病为"抄答案神器"；而眼镜的"看向即问"把交互压缩到近乎零操作，同时因为它看到的是"学习者正在看的真实场景"（一页书、一个实验、一段对话），天然更贴合"辅助理解"而非"替代作答"。但这枚硬币的另一面是隐私风险的陡增——第一视角意味着持续记录他人与环境，这是眼镜类相较手机的根本性治理挑战。就形态而言，市场大体分为纯音频/无显示型、单目轻显示型、双目显示型三档，重量、显示能力、续航与价格逐档变化，教育适配性也随之不同：无显示型胜在轻便与低干扰、适合语音问答与翻译播报，显示型胜在信息可视化（字幕、提词、步骤叠加）、但增加了重量、耗电与眩晕风险。下表汇总若干具代表性的公开机型规格，用于说明代际差异（规格取自厂商官网与发布通稿，横向能力分值以本院实测为准）。

机型	显示形态	重量	芯片/影像	续航（标称）	参考价	关键能力
Rokid Glasses	双目单色 Micro-LED + 衍射光波导	49 g	高通骁龙 AR1 + 12MP 相机	约 4 小时（充电盒可循环补电）	2499 元	实时翻译（在线 89 种语言，声称 6 种可离线）、提词、AI 问答、拍

						照
Meta Ray-Ban Display	单目单色显示（右镜片，600×600、20° FOV、30–5000 nit）	69 g	配套 Neural Band 肌电腕带交互	约 6 小时（充电盒补至约 24 小时）；腕带约 18 小时	799 美元（含腕带）	显示叠加、手势/肌电交互、翻译、拍摄
小米 AI 眼镜	无显示（音频+摄像）	约 40 g（钛合金框）	12MP、105°广角，0.8s 抓拍	[待补：官方标称续航]	1999 元起（另有电致变色版）	拍照录像、问答、翻译、通话、支付
华为智能眼镜 2	无显示（音频）	[待补：官方重量]	音频为主	[待补：官方标称续航]	1699 元起	翻译、播报、健康监测、手机联动
Rokid AI Glasses Style	无显示（音频+摄像）	38.5 g	12MP 相机	[待补：官方标称续航]	[待补：官方定价]	接入 ChatGPT/Gemini、实时翻译、免提通话

2025 年国内 AI 眼镜进入"百镜大战"，华为、小米、OPPO、阿里（夸克）、百度（小度）、理想等厂商密集入局，产品线迅速从"音频耳镜"扩展到"带摄像"再到"带显示"（腾讯新闻/新浪财经，2025）。价格带也随之拉开——从数百元的基础蓝牙音频款（如小米 AI 眼镜最低款约 499 元），到千元级的拍照+翻译款（小米 1999 元起、华为智能眼镜 2 约 1699 元起、钛空/方框款 2299 元），再到两千元以上的显示款（Rokid Glasses 2499 元、Meta Ray-Ban Display 799 美元）。对教育采购而言，价格带与能力带的对应关系值得留意：并非越贵越适合教育，一副 2499 元的显示款若显示信息密度不足、续航撑不满一节课，其教育效用未必胜过一副专注翻译、续航更长的音频款——关键仍是"场景—能力"匹配而非"价格—档次"匹配。

需要澄清 Rokid 产品命名以避免常识错误：**Rokid Glasses**（2024 年 11 月 18 日于杭州 Rokid Jungle 发布、首发价 2499 元、约 49 g、双目单色 Micro-LED 衍射光波导显示、骁龙 AR1 平台、12MP 相机、约 4 小时续航配可循环补电的充电盒）是带显示的 AI+AR 消费级眼镜，获 CES 2025 相关奖项，2025 年第二季度开售；而 **Rokid AI Glasses Style**（约 38.5 g）是无显示的音频型款式，接入 ChatGPT/Gemini 与实时翻译。二者定位、形态、价格均不同，不可混为一谈。整体市场层面，多家机构预计 2026 年全球 AI 智能眼镜出货量将从 2025 年的约 600 万台增至约 2000 万台、市场规模从约 12 亿美元增长至约 56 亿美元，中美合计占近 80% 份额；IDC 预计 2026 年全球智能眼镜出货有望突破约 2368.7 万台（华尔街见闻/虎嗅、IDC，2025—2026）。需要提醒：上述为不同机构的市场"预测"，非"实际"数据，币种为美元，勿与人民币口径混算。教育垂直渗透的独立统计目前仍稀缺 [待补：教育场景出货/份额数据]。

网龙（NetDragon，港股 HK:0777）与 AI 眼镜的关联需准确表述：网龙于 2023 年 11 月向 Rokid 战略投资 2000 万美元，并签署为期五年的战略合作协议；2025 年初 Rokid 生态受资本市场关注时，网龙股价一度显著波动。网龙由此在其网络游戏、教育科技主业之外，切入 AI 眼镜及相关软件、内容赛道（金融界/智通财经，2025）。截至本章撰写，公开信息显示网龙与 Rokid 的合作以战略投资与生态协同为主，尚无检索可证实的、以网龙品牌独立发售的教育专用 AI 眼镜量产机型 [待补：如有网龙自有品牌眼镜机型及规格，需以官方通稿核实]。网龙自研教育产品线以软件与交互硬件为主——101 教育 PPT（累计装机量超 3576 万，内置"AI 助教"功能）、Promethean 交互显示屏、未来实验等（网龙华渔教育官网，2024）。凡涉及网龙眼镜的具体规格、价格、发售时间，均须以官方发布为准，不确定处一律留［待补］。

9.3.2 教育典型用例

- **实验与实训指导**：第一视角识别操作步骤，实时提示规范与安全风险，把"看示范—做操作"合并为"边做边被指导"。传统实验教学的痛点在于教师无法同时盯住每个工位，而眼镜的第一视角恰好把"每个学生正在做什么"变为可被 AI 监看与提示的信息流：滴定管读

数是否正确、酒精灯是否规范熄灭、电路连接是否短路，都可在操作瞬间给出提示。适用于理化生实验、职教实训（数控、烹饪、美容美发）、医护技能训练（心肺复苏手法、无菌操作）。此用例对手部动作的细粒度识别要求高，评测重点是抖动、遮挡、快速运动下的步骤识别稳定性与误报率——误报（把正确操作判为错误）比漏报更破坏学习信任。

- **情境化语言学习与翻译**：这是当前 AI 眼镜最成熟、最具规模化潜力的教育用例。语言习得研究长期强调"可理解输入"（comprehensible input）与真实语境的重要性，而翻译眼镜恰好把"看到实物即获得目标语标注"这一沉浸式条件低成本地提供给学习者。以 Rokid Glasses 为例，其宣称在线支持 89 种语言翻译、6 种语言可离线（中英日德法西，离线由自研 LLM 支撑），显示端可将译文以字幕形式实时叠加于视野（在线翻译引擎由微软提供），并支持看文本即译（菜单、路牌、文献）；无显示型（小米、华为、Rokid Style）则以语音播报译文，Rokid Style 更直接接入 ChatGPT/Gemini。对留学预备、外语沉浸、双语课堂、看菜单/路牌/原版文献即译等场景有直接价值。评测重点是真实对话节奏下的连续翻译时延、专有名词与学科术语（生物物种名、化学式、历史地名）的准确率、以及离线与在线的质量差——离线 6 种语言的翻译质量能否支撑真实交流，是判断"离线是不是噱头"的关键。
- **无障碍与特殊教育**：为视障学习者提供环境描述、文字朗读、障碍物提示，为听障学习者提供实时语音转字幕，第一视角形态天然贴合"随身、免手持、解放双手"的无障碍需求。相比手机需要举起、对准、点击，眼镜"看向即工作"的交互对肢体协调受限的学习者更友好。此场景对语音识别字错率、环境描述的准确性与安全性（如误判台阶）要求极高，且服务对象是最需要保护的群体，治理与安全权重最高。
- **研学与户外学习**：识别动植物、文物、地质地貌、天文场景，触发生成式讲解，把"到此一游"变为"到此一学"，让博物馆、自然保护区、历史遗迹成为"活的教材"。此场景恰是

网络最不稳定的场景（山野、地下展厅、境外无漫游），因而最凸显离线能力的评测价值——一副在没信号时就"失明失语"的研学眼镜，其宣传的博物功能形同虚设。

9.3.3 评测焦点与观察

· **多模态识别准确率**是眼镜类的核心门槛。第一视角图像受光照、抖动、遮挡影响显著，识别准确率与手机拍摄有系统性差距，须在真实运动/光照条件下测量而非静态摆拍 [待补：实测数据]。

· **端到端时延**决定课堂与对话可用性。翻译类应用尤其要测"跟得上说话节奏"的连续时延，而非单句冷启动时延；显示型设备还须评测近眼显示的清晰度（分辨率、FOV、亮度）与眩晕/疲劳风险——Meta Ray-Ban Display 的显示为 600×600、20° FOV、30–5000 nit，属"信息提示"而非"沉浸阅读"级别，教育内容的可读性需实测验证 [待补：实测数据]。

· **续航与佩戴舒适度**是"能否上一整节课"的现实约束。当前主流显示型眼镜标称续航多在 4—6 小时区间（Rokid Glasses 约 4 小时、Meta Ray-Ban Display 约 6 小时），高负载（持续翻译、连续拍摄）下会明显缩短；重量在 40—70 g 区间，长时间佩戴的鼻梁与耳部压力需纳入舒适度评分 [待补：实测数据]。

· **语音交互质量**是无显示型眼镜的主要交互通道，也是显示型的重要补充。评测须测唤醒成功率（含误唤醒率）、真实噪声下的语音识别字错率、以及多轮对话中的上下文保持能力。教室、地铁、户外的背景噪声会显著抬高字错率，而低学段学习者发音不标准、语速不稳，对识别鲁棒性提出额外要求——一个总是"听不懂孩子说话"的眼镜，功能再多也无法建立使用习惯。

· **隐私与合规**在眼镜形态下从"可选项"变为"准入项"。第一视角摄像涉及他人肖像与课堂录制，其敏感性远高于手机拍照——手机拍照是"显性动作"，旁人可感知；而眼镜摄像是"隐性常态"，被拍者难以察觉，这正是隐私争议的核心。评测须核查：是否有清晰的录制

提示（外部可见的指示灯/提示音，且不可被用户静默关闭）、影像与语音是否本地处理、数据留存与上传策略是否透明可查、能否满足我国《个人信息保护法》《数据安全法》《未成年人网络保护条例》（2024 年 1 月 1 日施行）《人脸识别技术应用安全管理办法》（2025 年 6 月 1 日施行）《个人信息保护合规审计管理办法》（2025 年 5 月 1 日施行）等要求。校园部署尤须注意：课堂录制涉及其他学生的肖像与声纹，须解决知情同意、数据最小化、留存期限与删除机制等问题，不能以"教学需要"一笔带过。眼镜类在课堂场景的系统化治理框架，详见本院《AI 眼镜教育应用发展报告 2026》。

9.4 智能学习机：家庭学习终端的智能体化评测

9.4.1 从"题库检索"到"AI 学伴"

学习机的核心竞争力正从"内容与题库的规模"转向"AI 学伴的智能体化能力"。所谓"智能体化"，指的是产品从"被动响应"（学生问一句、机器答一句）升级为"主动编排"：AI 学伴能围绕一个学习目标，自主拆解为讲解、提问、诊断、推荐、复盘等子任务，并根据学生的实时反应动态调整路径。这与第 7 章讨论的智能体范式一脉相承——差别在于，学习机上的智能体面对的是最需要耐心与准确性的用户群体，容错空间更小。2026 年的关键变化是：学习机普遍接入教育垂类大模型，并以**智能体编排**把"讲解—提问—诊断—推荐"串联成闭环，围绕学情记忆提供个性化路径。理想状态下，一台真正智能体化的学习机应能做到：学生做错一道题，它不只是给答案，而是诊断出背后的知识漏洞、追问以确认误解类型、推送针对性的微课与变式题、并在若干天后主动回测这个知识点是否已巩固——把离散的"答疑"编织成连续的"培养"。这一转向由 2025 年初 DeepSeek 开源模型的爆发直接催化——学而思、网易有道、希沃、小猿、高途、作业帮等头部机构密集宣布接入 DeepSeek（新京报，2025）。其中，学而思走"开源基座 + 行业数据后训练"路线，其自研"九章大模型"以 DeepSeek-V3 为基座之一，叠加教育专有数据二次训练，第四代学习机采用"九章大模型 + DeepSeek 双引擎"架构，2025

年 11 月全渠道销量宣称突破 15 万台、同比增长 130%；松鼠 AI 在自研智适应大模型之外亦接入 DeepSeek（占比约 10%），累计出货量宣称突破 20 万台（知乎问答/证券时报，2025，均为厂商口径，非独立审计数据）。

这一轮升级的本质，是学习机从"存储介质"向"服务介质"的转变。过去学习机的护城河是内容——谁买断了更多名师课程、谁的题库更全，谁就更强；如今护城河转向服务——谁的 AI 学伴更懂启发、谁的学情引擎更准、谁的多模态批改更快，谁才更强。这也解释了为何 2025 年在线教育巨头（作业帮、学而思、有道、小猿）能凭借模型与数据优势迅速切入硬件、并在销量榜上超越传统硬件厂商：当竞争焦点从"堆内容"转向"炼模型"，拥有教研数据与算法团队的一方占据了新的制高点。

市场格局方面，2025 年一季度作业帮、学而思、科大讯飞、步步高、小猿、小度为销量前六品牌，合计约 74.4% 份额（部分月度榜单中作业帮以约 31.8% 居首，学而思、小猿、科大讯飞分列其后）；深耕教育 28 年的步步高拥有超 1.8 万个线下售点，其线下渠道与售后网络仍是不可忽视的护城河。市场规模上，2024 年中国智能平板学习机市场规模约 270.72 亿元、同比增长约 48.27%，机构预计 2027 年将超 500 亿元（艾媒咨询/洛图科技，2024—2025）。需要提醒：上述销量、份额、市场规模口径不一（销量 vs 销售额、季度 vs 年度、全渠道 vs 线上、机构测算 vs 厂商自述），引用时须标明口径与来源，勿跨口径相加；厂商自述的累计出货、增长率尤须谨慎，宜标注"厂商口径、未经独立审计"。

9.4.2 评测维度的特殊性

- **学科内容体系与课标对齐**：区别于通用平板，学习机的价值高度依赖与课标、教材版本的对齐度。评测须核查是否覆盖多版本教材（人教、北师大、苏教等）、幼小初高全学段，以及内容更新是否跟随课标修订 [待补：课标对齐实测]。科大讯飞 T30 Ultra 宣称覆盖"幼

小初高全科"，并内置 188 本牛津大学出版社授权英文分级读物（15 级难度）（快科技，2024）。

- **错题与学情引擎**：评测记忆机制是否真正沉淀个体学情、推荐是否收敛于薄弱点，而非泛化推题。这是"智能体化"与"题库检索"的分水岭——真正的学伴应能跨会话记住"这个孩子在一元二次方程判别式处反复出错"，并据此编排后续路径 [待补：实测数据]。
- **护眼与使用管控**：屏幕护眼参数与家长管控粒度是家庭终端的硬约束。以科大讯飞 T30 Ultra 为例，其宣称采用"类自然光 + 微纳米类纸护眼屏"，3K 分辨率、120Hz 刷新、247 PPI、硬件级低蓝光，14.7 英寸屏、12GB+1T 存储、12000mAh 电池、四摄，配"星闪"（NearLink）AI 手写笔（超万级压感），首发价 11699 元（快科技，2024）；步步高、学而思等亦各有护眼方案。管控维度须评测使用时长限制、应用白名单、内容过滤、家长远程查看等粒度 [待补：管控项清单]。
- **AI 学伴质量（教学有效性）**：这是最难也最关键的维度。评测应看讲解正确性、启发式引导（而非直接给答案）、拒答与纠偏能力。学而思九章大模型宣称采用"苏格拉底式"讲解——不直接给答案，而是先分析知识点与题型，通过连续追问引导学生逐步推导（学而思公开资料，2025）。这类"引导式"设计是否真正促进思考，须以过程性指标（学生自主完成步数、独立正确率提升）而非"答对率"单独评价 [待补：实测数据]。

9.4.3 典型风险

学习机的智能体化带来三类突出风险，须在评测中重点核查。

一是**答案直给削弱学习过程**。若产品以"秒出答案""一拍即得"为卖点，会把学习机异化为"作弊器"——学习者绕过了思考，只留下抄写。这是"AI 辅助学习"最深的悖论：技术越便利，越可能剥夺学习所必需的"有效认知努力"（productive struggle）。评测应奖励"引导式而非替代式"的交互设计，把"是否促进思考过程"作为核心标尺，具体可测：讲解是否先追问再揭示、

是否在学生卡壳时给"脚手架式提示"而非直接答案、是否在学生答对后要求其复述理由。学而思九章大模型宣称的苏格拉底式讲解、以及部分产品设置的"引导模式/解答模式"双开关，是正向设计的样本，但其真实效果仍须以过程性指标实测验证。

二是**幻觉进入学科解题**。尤其在数理推理中，错误的中间步骤比错误的最终答案更具误导性，因为学习者会模仿其推理路径、内化错误方法。一道题最终答案碰巧对、但中间某步逻辑错误，对学习的危害甚至大于全错——它教会了学生一个"看起来对"的错误套路。高校与厂商的共识路线是"大模型 + RAG + 知识图谱"以抑制幻觉、对齐课程大纲：用检索把答案锚定到教材，用知识图谱约束推理路径的合法性（53AI 应用案例，2025）。评测须以学科正确率、推理步骤可追溯性（每步能否溯源到教材/公式）、错误自纠率为硬指标，并按学科分层——数理化的步骤幻觉危害最大，须单独重测。

三是**未成年人过度依赖与数据风险**。情感陪伴型、拟人化交互一方面提升黏性、缓解孤独，另一方面可能诱导低龄用户产生过度情感依赖、削弱现实社交与自主学习意愿；相关政策建议已提出，宜将高互动、高沉迷性质的交互纳入防沉迷机制（多方政策建议，2025）。同时，学习机沉淀了极为敏感的未成年人学情、行为、乃至面部/声纹数据，其数据治理须对标《未成年人网络保护条例》与《个人信息保护法》，把家长知情、数据最小化、留存与删除机制作为准入项。评测须核查：情感交互是否有边界与防沉迷设计、敏感数据是否本地化或加密、家长能否查看与删除孩子的数据画像。凡涉及低学段，上述三类风险的权重均应高于"模型有多聪明" [待补：实测数据]。

9.5 端侧智能体：本地推理、隐私与离线能力评测

9.5.1 为什么端侧化在教育场景尤其重要

端侧化（on-device）在教育中的价值不止于时延，更关乎**隐私与可用性**。未成年人数据高度敏感，本地处理天然缩小数据外泄面，契合《个人信息保护法》的最小化原则与"隐私设计"

（Privacy by Design）理念；校园与家庭网络条件参差，离线可用直接决定"关键时刻能不能用"——研学户外、乡村学校、宿舍网络高峰，都是端侧兜底的真实战场。2026 年，随着端侧芯片算力提升与小语言模型（SLM）、量化/蒸馏技术成熟，"本地跑得动一个够用的智能体"从概念走向落地。

硬件侧，高通骁龙 8 Gen3 宣称可运行百亿参数级模型、对 70 亿参数 LLM 每秒生成约 20 token；联发科天玑 9300 支持终端运行 10 亿/70 亿/130 亿参数模型（厂商公开资料/技术媒体，2024—2025）。模型侧，Apple 于 WWDC 2025 推出的第三代设备端基础模型约 30 亿参数，专为 Apple 芯片优化，随附 Foundation Models 框架，允许开发者直接、免费、离线调用设备端模型（Apple ML Research，2025）；开源侧的进展更为迅猛——Qwen3-4B 在 MMLU-Redux 上得分约 83.7（超过体量两倍于己的部分模型）、Phi-4-mini（3.8B）在完整 MMLU 上约 67%（约为 Llama 3.1 8B 的水平却仅用约一半内存）、Qwen2.5-7B 约 74.2（各团队技术报告，2025）。量化技术进一步压低部署门槛——Gemma 3 4B 经量化感知训练可从约 8GB（BF16）降至约 2.6GB（int4）、Gemma 3 1B 从约 2GB 降至约 0.5GB，质量损失控制在数个百分点内（Google，2025）；4bit 量化（Q4_K_M）可将某 13B 模型内存占用从约 13.2GB 降至约 4.8GB、端侧推理时延从约 4200ms 优化至约 980ms（技术媒体实测，2025）。这些数字合起来说明一件事：一个具备初高中学科问答与讲解能力的模型，已经可以在高端手机与新一代学习终端上本地运行，而不必事事上云。对教育而言，这不仅是"更快"，更是"更可控"——模型、数据、日志都留在设备内，家长与学校对"孩子和什么在对话、说了什么"拥有了物理层面的掌控力。

9.5.2 端侧智能体的技术范式

- **小模型 + 检索增强（RAG）**：以本地知识库补足小模型的知识边界，把教材、错题、笔记、课堂讲义作为可检索记忆。这既解决小模型"知识不足"，又缓解"幻觉"——答案有本地文档可溯源。教育场景中，RAG 让一个 3—7B 的本地模型也能"背下这一学期的课本"。

- **本地记忆与个性化**：在设备侧沉淀长期学情画像，减少云端画像依赖，把"这个孩子的薄弱点"留在本地。这与 9.4 的学情引擎在架构上同源，但强调"数据不出端"。
- **端云协同与降级策略**：常规任务端侧完成、复杂任务上云、断网时优雅降级到本地能力。这背后是一套"路由"逻辑：设备先判断任务难度与隐私敏感度，简单/敏感的（如背单词、口算批改、含个人信息的问答）留在本地，复杂/无敏感的（如长文写作辅导、开放式探究）上云。Apple 的"设备端约 30 亿参数模型 + Private Cloud Compute"是这一范式的典型工程化表达——云端计算也承诺不留存、不训练用户数据，试图把云的能力与端的隐私承诺结合（Apple，2025）。相关智能体编排、RAG 与记忆的新范式，详见第 7 章。

端云协同也重塑了教育硬件的成本结构，这是"可持续性"维度不可忽视的一面。纯云方案的边际成本随调用量线性上升（每次问答都在烧 token），对高频使用的学习场景意味着持续的推理开支，最终或转嫁为订阅费；端侧方案把算力成本前置到硬件购置（一次性买单更强的芯片），之后本地推理近乎"零边际成本"。对学校大规模部署与低收入家庭长期使用而言，端侧的"一次投入、长期免费、且离线可用"具有独特的普惠价值——它让"用得起 AI"不再取决于持续的网络与订阅支出。评测"单位学习成本"时，须把硬件折旧、订阅费、流量费合并计算，而非只看裸机售价。

9.5.3 评测焦点

- **离线可用率**：断网条件下核心功能（学科问答、翻译、讲解、批改）的完成度，是端侧智能体最关键的差异化指标。眼镜类可参照 Rokid Glasses"在线 89 种、离线 6 种语言"的分级——离线能力的广度与质量必须单独标注，不能用"在线能力"掩盖 [待补：实测数据]。
- **端侧推理时延与功耗**：本地推理的速度与热/电代价。须区分 Prefill（读题）与 Decode（生成）两段，并测量持续推理下的机身温升与耗电，因为发热会触发降频、进而拖慢时延 [待补：实测数据]。

- **本地能力与云端的差距（硬件损耗）**：同一任务在纯端侧、端云协同、纯云端三档下的质量差，量化"离线到底损失了多少"。这是 9.2.2 所述"软件基准—硬件实测"贯通比较的落点 [待补：实测数据]。
- **隐私实测**：不能只看厂商声明，须以抓包、流量分析核查"数据是否真正本地化、是否存在隐性上传、上传了什么、上传到哪"。厂商声称"本地处理"与实际"仍上传日志/画像"之间的落差，只有实测能揭穿。评测方法可包括：断网状态下核心功能是否仍工作（若断网即失效，则"本地"存疑）、联网状态下的出站流量分析与目的地核查、以及对隐私政策文本与实际行为的一致性比对。这是把"隐私承诺"变为"隐私可证明"的关键一步，也呼应 2025 年《个人信息保护合规审计管理办法》推动的"从做过走向可证明"（律所合规综述，2025）[待补：实测方法/来源]。
- **模型能力与内容的持续更新**：端侧的代价之一是"模型固化"——本地模型不像云端可随时热更新，知识可能滞后、能力可能落后于最新云端模型。评测须关注 OTA 更新机制是否完善、本地知识库（RAG 语料）能否随教材修订同步更新、以及厂商对端侧模型的长期维护承诺。一款端侧智能体若停止更新，其知识与安全护栏都会逐渐"过期"，这是可持续性维度的隐性风险 [待补：OTA 更新频度数据]。

9.6 横向对比与能力雷达（循证可视化）

为使三类形态在同一坐标系下可比，本章以六维评测框架输出**能力雷达图**与**产品横评表**，并以**能力演进时间线**呈现代际变化。所有可视化的底层横向分值均来自本院评测实验室实测，未获实测前一律留白，不以估计值填充；下表规格类信息取自公开来源、可即时核验，能力分值则待实测。

形态	模型能力	端侧工程	交互体验	教育适配	治理与安全	可持续性
AI 眼镜	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]

智能学习机	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]
端侧智能体	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]	[待补：实测]

基于公开规格与用例，可先给出**定性的形态画像**（非量化，量化待实测）：AI 眼镜在"随身第一视角、情境翻译"上不可替代，但受续航（约 4—6 小时）、重量（40—70 g）、显示信息密度（提示级而非阅读级）约束，尚不适合承担长时系统化学习，其教育定位是"情境助手"而非"主力终端"；学习机在"学科内容体系、护眼、家长管控、错题学情"上最成熟，是家庭系统化学习的主力，14.7 英寸大屏与万元级配置（如科大讯飞 T30 Ultra）对标"一对一家教"，但答案直给与学科幻觉是其治理重点；端侧智能体不是独立品类而是一种"能力属性"，可附着于眼镜、学习机、手机之上，在"离线、隐私、低边际成本"上具结构性优势，是弱网与数据敏感场景的兜底，但本地小模型与云端仍存能力差，须靠 RAG 与端云协同弥合。

三类形态之间存在清晰的**互补而非替代关系**，这决定了选型不应是"三选一"，而应是"按场景组合"。一个理想的家庭学习生态可能是：学习机承担系统化的学科学习与错题订正（主力），AI 眼镜承担语言沉浸、研学识别与实验指导（情境补充），而端侧化能力贯穿两者、保障离线可用与数据不出端（底座）。三者共享的应是同一套学情画像与同一套治理策略——学生在眼镜上练的口语、在学习机上订正的错题，理应汇入统一的成长档案，而这套档案越是本地化、越是家长可控，就越符合未成年人数据保护的方向。当前市场的现实是三类产品各自为战、数据割裂，跨设备的统一学情与统一治理仍是空白，这既是痛点也是网龙等兼具软件、内容、硬件生态的厂商的潜在切入点 [待补：跨设备学情打通的落地案例]。

图 9-1　三类教育 AI 硬件六维能力雷达（横向分值来源：本院评测实验室，[待补：实测批次/日期]；规格来源：厂商公开资料）

图 9-2　教育 AI 硬件能力演进时间线（2024 Rokid Glasses 发布 → 2025 DeepSeek 催化学习机智能体化、AI 眼镜"百镜大战" → 2025 Meta Ray-Ban Display 与 Apple 设备端模型 → 2026 端侧化深化；来源见章末）

9.7 发现、结论与选型建议

主要发现（基于框架的定性判断，量化结论待实测补齐）：

1. 教育 AI 硬件的瓶颈正从"模型够不够聪明"转向"端侧工程能不能稳定承接"。当前三类形态的共同短板集中在时延、续航、离线三项：眼镜受续航与显示信息密度约束、学习机受幻觉与答案直给约束、端侧智能体受本地—云端能力差约束 [待补：实测佐证]。

2. 智能体化是三类形态的共同方向，2025 年 DeepSeek 开源是重要催化剂；但"引导式学习"与"答案直给"的设计分野，比参数规模更能预测教育价值——一款用苏格拉底式追问的学习机，可能比一款参数更大但秒出答案的产品更有教育意义。

3. 治理与安全在硬件形态下从"可选项"变为"准入项"，尤其是第一视角摄像的肖像/录制合规、未成年人数据的本地化、以及学科幻觉对低龄学习者的误导。2024—2025 年密集出台的《未成年人网络保护条例》《人脸识别技术应用安全管理办法》《个人信息保护合规审计管理办法》等，共同抬高了合规门槛。

4. 产品命名与关联关系必须精确核实：Rokid Glasses（显示款、2499 元）与 Rokid AI Glasses Style（音频款）不可混淆；网龙（HK:0777）对 Rokid 为 2023 年 11 月 2000 万美元战略投资 + 五年合作，其自有教育硬件以交互显示屏、软件（101 教育 PPT）为主，凡涉眼镜量产机型须以官方通稿为准。

选型建议（面向学校与家庭，方向性建议，不含具体品牌背书）：

- **按场景选形态**：随身情境辅助、语言沉浸、无障碍优先眼镜类；家庭系统化学习、错题订正、学科提优优先学习机类；隐私敏感、弱网/无网、数据不出端的场景优先端侧智能体（或具离线能力的上述设备）。
- **把"治理与安全"作为一票否决项**：内容拦截、数据本地化、家长/教师管控、录制提示合规，四者缺一不可；眼镜类还须额外核查第一视角摄像的肖像合规。

- **优先引导式而非替代式**：以"是否促进思考过程"而非"是否快速给出答案"作为采购评价标尺，警惕"秒出答案"营销话术。
- **分级看待离线能力**：要求供应商明确标注"在线能力"与"离线能力"的分界（如翻译语种、可解题类型），不接受用在线能力掩盖离线短板。
- **要求可循证的评测证据**：向供应商索取公开、可复现的实测方法与数据（端到端时延口径、弱网档位、学科正确率测试集、隐私流量报告），而非孤立跑分与营销参数；对市场规模/份额类数据，核对口径与来源，警惕跨口径相加、厂商自述当实测、以及币种混算（美元预测勿与人民币规模混谈）[待补：评测清单来源]。
- **算总账而非算裸机价**：把硬件购置、云端订阅、流量、内容更新费用合并为"三年总拥有成本"再比较，尤其关注端侧方案"一次投入、离线可用、低边际成本"对大规模与低收入场景的普惠价值。

趋势展望： 展望 2026—2028，教育 AI 硬件有三条较为确定的演进线。其一，端侧化持续深化——随着更强的端侧芯片与更小而强的模型量产，越来越多的核心能力将"下沉"到设备本地，隐私与离线从卖点变为默认，云端更多承担复杂长任务与跨设备同步。其二，形态走向融合与协同——眼镜、学习机、手机之间的数据孤岛有望被"统一学情画像 + 统一治理策略"打通，具备软件、内容、硬件全栈能力的厂商（如网龙及其生态伙伴 Rokid）在这一融合中具备结构性机会，但前提是把跨设备的数据打通建立在"本地化、家长可控、合规可证明"的基础上。其三，治理从"事后合规"走向"设计内建"——随着《未成年人网络保护条例》《人脸识别技术应用安全管理办法》《个人信息保护合规审计管理办法》等法规的落地，隐私设计、防沉迷、内容护栏、录制合规将从"加分项"固化为"准入线"，倒逼产品在架构层面而非营销层面解决问题。对评测机构而言，这意味着评测的重心也将持续从"参数与跑分"转向"使用真相与治理证据"——这正是本章框架试图确立的方向。本章所有横向能力分值将在本院评测实验室完成实测后补齐并公开方法学，届时以可回溯、可复现的循证数据替换文中占位。

本章参考来源

1. 《49g Ultra-Light AR glasses with AI — Rokid Glasses》· Rokid 官网 · 2025 · https://global.rokid.com/products/rokid-glasses
2. 《Real-Time Translation Glasses | AI-Powered Smart Glasses》· Rokid 官网 · 2025 · https://global.rokid.com/pages/rokid-glasses
3. 《Rokid AI Glasses Style（Non-Display）38.5g Ultra-Light》· Rokid 官网 · 2025 · https://global.rokid.com/pages/rokid-ai-glasses-style
4. 《2499 元！Rokid Glasses 发布，AR 眼镜跑步进入消费时代》· 量子位 · 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. 《战略投资 ROKID，"纯正 AI 股"网龙(00777)打开全新想象空间》· 金融界/智通财经 · 2025 · https://m.jrj.com.cn/madapter/finance/2025/01/14153147428755.shtml
7. 《网龙投资 ROKID，掘金 AI 眼镜新风潮》· 搜狐/相关财经媒体 · 2025 · https://www.sohu.com/a/848866851_122004016
8. 《网龙将数字技术融入教育产品之中》· 数字中国建设峰会官网 · 2024 · https://www.szzg.gov.cn/2024/szzg/szfj/202405/t20240514_4823981.htm
9. 《网龙：布局 AI 新赛道，探索数字新征程》· 证券时报网 · 2024 · https://stcn.com/article/detail/1098277.html
10. 《101 教育 PPT 官网》· 网龙华渔教育 · 2024 · https://ppt.101.com/

11. 《New Meta Ray-Ban AI-Powered Display Glasses and Neural Band》· Meta 官网 · 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.《小米 AI 眼镜（产品页）》· 小米官网 · 2025 · https://www.mi.com/prod/xiaomi-ai-glasses

15.《华为 AI 眼镜规格参数》· 华为官网 · 2025 · https://consumer.huawei.com/cn/audio/ai-glasses/specs/

16.《AI 功能是噱头吗？测评小米、Rokid 等 10 款 AI 眼镜》· 南方都市报 · 2025 · https://m.mp.oeeee.com/a/BAAFRD0000202507191104492.html

17.《2026 年全球 AI 智能眼镜市场将达 56 亿美元，中美占 80% 份额》· 虎嗅（引 IDC/机构预测）· 2026 · https://www.huxiu.com/article/4857057.html

18.《巨头加速入局，AI 眼镜 2026 年打响新一轮排位赛》· 21 世纪经济报道 · 2026 · https://www.21jingji.com/article/20260114/herald/29460ed43ae258bde7c914301cb29cf6.html

19.《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》· 快科技 · 2024 · https://news.mydrivers.com/1/992/992671.htm

20.《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.《卖疯了的 AI 学习机，为何成为硬件三大赛道之一？》· 21 世纪经济报道 · 2025 · https://www.21jingji.com/article/20250729/herald/59577065a816e441b1b2c0e136eddbe3.html

22. 《艾媒金榜 | 2025 年中国智能平板学习机十大品牌》· 艾媒咨询 · 2025 · https://www.iimedia.cn/c880/106929.html

23. 《DeepSeek"来袭"，教育企业抢滩背后的博弈与思考》· 新京报 · 2025 · https://m.bjnews.com.cn/detail/1739588622168866.html

24. 《学而思发布第四代学习机，搭载九章大模型与 DeepSeek 双核架构》· 知乎（问答/厂商口径）· 2025 · https://www.zhihu.com/question/1904505543529313225

25. 《智启未来 学赋新生 2025 年中国 AI 学习平板市场洞察白皮书》· 洛图科技 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 端侧大模型技术分析：从技术原理到落地实操的深度拆解》· 龙腾亚太（人工智能技术与咨询）· 2025 · https://www.longtengyatai.com/info/264

32. 《2024 数据合规立法监管的回顾与 2025 年展望》· 上海市锦天城律师事务所 · 2025 · https://www.allbrightlaw.com/CN/10475/63efeb45f0af5dbe.aspx

33. 《2025 中国网络安全与数据保护年度回顾与 2026 年展望》· 安全内参 · 2025 · https://www.secrss.com/articles/86785

34. 《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. 《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》· 腾讯新闻 · 2025 · https://news.qq.com/rain/a/20250319A0774T00

36. 《AI 眼镜概念涨停潮：关注 Rokid 生态链概念机会》· 新浪财经 · 2025 · https://finance.sina.com.cn/stock/jsy/2025-02-20/doc-inemcpfh2390679.shtml

37. 《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》· 智通财经 · 2025 · https://cn.investing.com/news/stock-market-news/article-2678278

第 10 章　新范式与发展建议：智能体编排 / RAG / 记忆；政策标准、教师素养与公平；实施路线图

前九章分别就"赋能教学、支持学习、支持教研、智能评价、治理与安全"五大场景，对生成式人工智能教育产品的机理、图谱与典型形态做了循证考察。本章收束全书，转入两项收官工作：其一，从技术架构层面提炼支撑上述场景演进的三大新范式——智能体编排（Agent Orchestration）、检索增强生成（RAG）与长期记忆（Memory）——它们是 2026 年产品由"对话式单模型"走向"智能体化、多模态、端侧化"的共同底座；其二，从产业与政策层面给出面向多元主体的发展建议与分阶段可核验的实施路线图。

必须先厘清一个判断口径：本章所引产品与数据，均为本次编研经公开检索并核对的一手来源（厂商公告、官方文档、学术会议论文、政策原文），逐条列于章末「本章参考来源」。凡本次未能核实到可信来源的具体数量、份额、评测分数或时点，一律以 [待补：…] 占位，不以估计值或不确定专名替代——这既是本书自始至终坚持的循证纪律，也恰是本章第一节所论 RAG 范式在方法论层面的自我践行：把"来源可指认、结论可回链"从对产品的要求，同样施加于我们自己的写作。

10.1　三大技术新范式：从"会说话"到"会做事、可溯源、有记性"

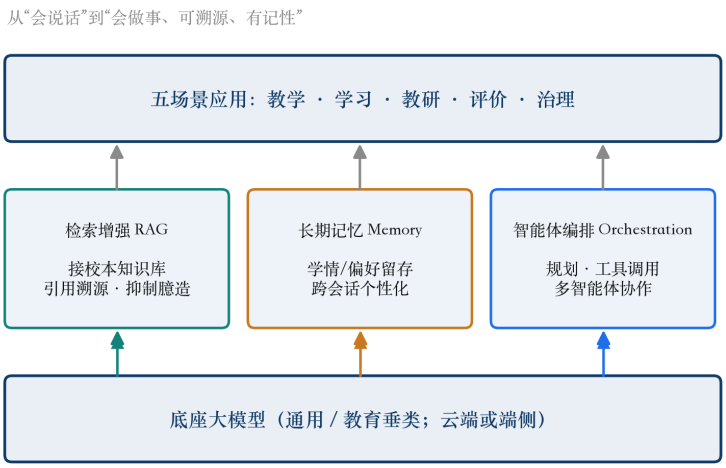
2024 年版报告以对话式大模型为主线，其能力边界集中于单轮或多轮的自然语言生成——模型是一个"应答器"：输入问题、输出文本，一次会话结束即遗忘上下文，无法自主调用外部资源，也无法为自己的断言指出出处。这一形态在"闲聊式答疑"上尚可用，但一旦进入真实教育任务，三重短板立刻显现：**不会做事**（无法把"出一份分层学案并对齐评价量规"这样的复合任务拆解并逐步完成）、**不可溯源**（无法说明某个结论来自哪本教材、哪条课标，教师

无从核验）、**没有记性**（记不住这个学习者上周错在哪、掌握到什么程度，"个性化"沦为一次性话术）。

2026 年的教育产品图谱正是沿着弥补这三重短板而外扩：模型被组织进一个**可规划、可调用工具、可溯源、可持续积累状态**的系统。我们把这一转变归纳为三条相互咬合的技术范式——编排让系统"会做事"，RAG 让系统"可溯源"，记忆让系统"有记性"。三者不是三个独立卖点，而是同一系统的三个面向；理解它们的机理与边界，是判断一款产品"是真智能体还是套壳对话框"的关键标尺。

这一转变的现实紧迫性，可从采用侧的加速度得到印证。据 RAND 公司调查，美国教师使用 AI 的比例在两个学年间从 25% 翻倍至 53%；Gallup 与 Walton 家庭基金会的调查显示，2024–25 学年有 60% 的 K-12 教师在工作中使用过某种 AI 工具；Pew Research 则记录到 2024 年有 26% 的美国青少年用 ChatGPT 完成学业，较 2023 年的 13% 翻倍（RAND、Gallup、Pew 2024–2025 转引自 tutorbase 汇编）。市场侧的多家机构测算虽口径不一（这正是本书对市场规模数字持审慎态度、不采信单一来源的原因），但一致指向高速增长——例如 Precedence Research 估算 2025 年全球"AI+教育"市场约 70.5 亿美元、Mordor Intelligence 估约 69 亿美元，量级相近而具体值有别，本书据此仅取"处于数十亿美元量级且年复合增速高"这一稳健判断，具体数值 [待补：市场规模统一口径]。采用面越广、投入越大，"产品到底靠不靠谱"的问题就越尖锐——而答案恰恰系于三范式的工程化质量。

图 6 生成式 AI 教育产品的新范式技术栈



来源：本报告分析框架（详见第 10 章）。

10.1.1 智能体编排：从单模型直答到多角色协作

机理：从"直答"到"感知—规划—行动—反思"的循环。 智能体编排指以大模型为推理内核（业界常称之为智能体的"大脑"），围绕一个持续运转的循环组织任务：感知环境与输入、规划下一步、调用工具执行、再对执行结果进行反思与修正。学界最有影响力的两个骨架，是 ReAct 与 Reflexion——ReAct（Reasoning and Acting，推理并行动）把"逐步思考"（思维链）与"落地动作"（调用工具、查询数据库、派生子智能体）交织起来，让模型不再一次性吐出答案，而是"想一步、做一步、看一步"；Reflexion 在此之上加入自我批评与记忆，让智能体能据上一轮的失败修正后续行为（EmergentMind、Medium 2024–2025）。2025 年的新工作进一步把"规划"与"执行"解耦：如 ReflAct 让智能体不只盯着"下一个动作"，而是持续对照任务目标反思自身的"信念状态"；分层式的规划—执行结构（把高层目标拆成子问题、每个子问题跑一个标准 ReAct 循环）则被用来抑制多步链条中的"概念漂移"，提升稳定性（arXiv:2505.15182 等 2025）。

理解这套循环，才能理解编排相较单模型直答的三重增益，以及它们各自服务的教育诉求：

- **任务可分解。** 把"出一份分层学案"拆成检索课标、诊断学情、生成分层任务、对齐评价量规、排版成稿等子步，每一步交给最擅长它的智能体或工具。这直接回应了教研工作"环节多、每环都要求专业"的现实。
- **过程可干预。** 每一步的中间产物（检索到的课标条目、生成的任务草稿）都可被教师或另一个智能体校验、修改、否决，而不是把一个黑箱结果一次性甩给用户。这为"教师在回路"提供了技术着力点。
- **结果可校验。** 引入独立的"评审"角色对交互质量打分，把原本隐性的教学质量判断显性化、可打分化，从而使自动化流程具备自我质检能力。

但增益从不免费。2025 年的综述明确指出：加入显式反思、分离规划与执行、强制目标复述等手段虽能系统性缓解上下文漂移、错误累积、幻觉与低效工具调用等常见失败模式，却也带来算力开销上升、提示变长、多智能体编排复杂度增加等代价（arXiv 2025）。这意味着编排不是"越多智能体越好"，而是一个在可靠性与成本之间的工程权衡。

工程底座已经成熟。 2025 年主流的编排框架相对完备：LangChain 提供模型、提示、记忆与工作流的抽象层，支撑多步与多智能体编排；微软于 2025 年 10 月 1 日以公开预览形式发布 **Microsoft Agent Framework**，将此前的 AutoGen（多智能体编排模式）与 Semantic Kernel（企业级 AI 治理）合并，标志着编排能力从"研究原型"走向"可运维的中间件"（Kubiya、TrueFoundry 2025）。这些通用框架为教育垂类产品提供了可复用的调度骨架，使中小团队无需从零搭建编排层。

教育场景中的典型编排形态，可从近两年的学术与产品实践归纳出三类——

- **导师—学生—评审三角色（自博弈式内容生产与质检）。** 一个智能体扮演学习者以暴露常见误解，一个扮演导师给出脚手架式引导，一个扮演评审对交互质量打分。三者反复对弈，既能批量生产贴近真实误解的教学内容，又能在生产环节内置质检闭环。其价值

不在于"更像老师"，而在于把"这段辅导好不好"这一原本靠人工抽检的判断，转化为可自动执行、可累积的评分信号。

- **按教学职能拆分的多智能体系统。** 将辅导任务分解到课程规划、对话教学、评价诊断等专职智能体上协同工作，是 2025 年智能辅导系统（ITS）研究的主流思路。ACM Web Conference 2025 上报告的 **GenMentor** 框架即先用微调后的大模型把学习者目标映射到所需技能，再据学习者的动态、多维画像调度学习路径，形成"目标—技能—路径"的编排链；同类工作还包括把"预览—分析—推理"流水线用于会话式学习诊断的多智能体协作（Wang et al., WWW 2025 Companion；arXiv:2503.11733 综述 2025）。
- **面向教师的备课编排（人在回路）。** 以教师为中心的多智能体系统把"生成分层学习单/差异化学案"拆成需求解析、内容生成、难度分层、格式排版等子任务，由不同智能体分工，教师在关键节点审核裁决。arXiv 上的 **FACET** 框架（2025）即以此支持教师规模化实施差异化教学，其定位被明确表述为"辅助教师"而非"替代教师"——这与本书主张的人机责任边界高度一致。

一个已规模化的商用范例：Khanmigo 的"约束式导师"编排。 可汗学院的 Khanmigo 是当前编排范式在 K-12 辅导侧落地最广的商用产品之一，其用户从 2023–24 学年试点的约 6.8 万增长到 2024–25 学年的逾 70 万（Khan Academy / reruption 2025）。就架构而言，它并非"把 GPT-4 直接暴露给学生"，而是一套受约束的编排：系统提示先"截获"学生意图、解析其当前所处的解题逻辑步，再把回复约束在"给提示、追问、指出具体语法/计算错误"上，而不直接改写或给出完整解答——业界称之为"以引导而非解答为首要指令的教学型智能体"（reruption、mlopsaudits 2025）。它还按成本与能力把任务路由到不同模型（GPT-4 承担辅导、轻量模型承担辅助任务、另有模型承担写作），并内置内容审核过滤以保证话题聚焦于教育、拒答不当问题；面向教师则提供批改、生成讲义、拟定讨论题等工具（Khan Academy 2025）。这一案例说明编排的工程重心不在"更强的模型"，而在"更严的约束与更清晰的分工"——把苏格拉

底式引导、道德边界与模型路由固化进编排层，才是它区别于普通聊天机器人的关键。当然，其在数学多步计算上的准确性直到 2025 年才通过持续更新得到改善，这也再次印证多步链条的可靠性需要长期打磨。

成熟度判断（保守）。 当前规模化落地者，多为**流程较固定、工具边界清晰**的教研与批改类任务：编排的每一步都有明确的输入产物与校验标准，失败可被及时拦截。而开放式、长程、需跨会话自主决策的"全自动教学智能体"仍以实验演示与小范围试点为主，其可靠性尚不足以在无人监管下进入课堂——多步链条中任一环节的错误会沿链放大（错误累积），且智能体"自主决定下一步"的能力越强，可解释性与可控性的挑战越大。因此本书主张：教育智能体编排应优先用于**有教师在回路、有明确校验点**的工作流，把"自主性"作为受控增量而非默认目标。相关性能对比与失败模式，参见第 [待补：评测章号] 章评测部分。

10.1.2 检索增强生成（RAG）：让回答"有据可查"

机理：把答案锚定在可指认的来源上。 RAG 在生成前先从受控知识库检索相关片段，将其作为上下文注入提示，使模型的输出锚定于**可指认的来源**而非纯参数记忆。学界把"给出结果来源、让用户知道信息从何而来"这一行为称为 grounding（接地/落地）：它通过把回答约束在检索到的文档上，显著降低大模型生成"听起来合理却错误"内容的倾向（Springer, *Business & Information Systems Engineering* 2025）。理解 RAG 的价值，需要区分两类幻觉——**内在幻觉**（输出与所提供的参考上下文不一致）与**外在幻觉**（输出无法由上下文验证）；RAG 主要治理的正是把无据之说当作既定事实的倾向，让"支持/未提及/矛盾/补充"这类对答案与证据关系的判断成为可能（arXiv:2505.04847 等 2025）。

对教育场景，这一范式回应了三个刚性诉求：

1. **事实性与可溯源。** 答案可回链至教材、课标或校本资料，教师可核验、学生可追问出处。对未成年人而言，"知道答案从哪来、并学会追问出处"本身即是一种可迁移的媒介素养训练，其教育价值不亚于答案本身。
2. **知识的可治理。** 知识库由机构策展，可增删、可审计，避免模型把陈旧或错误信息当作既定事实。教师可上传本校讲义使解释锚定于本课教材——Google 的 Learn About 即已开放"上传自有源文档以在其上接地解释"的能力，其课程内的解释可被约束在教师提供的材料上（Google I/O 2025）。
3. **垂类适配（低成本迁移）。** 无需重训模型即可接入学科语料与校本知识，把大模型的通用能力低成本迁移到具体学段与教材版本，显著降低垂类落地门槛。这是"一个通用大模型 + 一套校本知识库"能快速服务不同地区不同教材的技术前提。

演进方向：从"朴素拼接"到"工程化组合拳"。 2026 年的教育 RAG 正从"检索到片段就直接拼进提示"的朴素做法，走向一套模块化、可评测的工程组合。业界普遍认为 RAG 已完成从"可用"到"好用"的跃迁，关键在于模块化拆解、图结构化与智能体化编排（Synthimind、Data Nucleus 2025）——

- **混合检索（稠密+稀疏）。** 向量检索捕捉语义相似，关键词检索兜住术语与专名（如特定公式名、历史事件名），二者互补以弥补纯向量检索在精确匹配上的短板。
- **重排序（reranking）。** 在初检之后加入一个高精度的交叉编码器（Cross-Encoder），逐一为候选片段与问题的相关性打分，从初检的几十个候选中筛出真正相关的 Top-5 / Top-10 再喂给模型，以牺牲少量算力换取命中精度的显著提升。
- **图检索 / GraphRAG。** 把知识图谱的符号推理与稠密语义检索结合，支持可解释、上下文感知的多跳问答。教育领域已有面向课程问答的 **KA-RAG** 框架（*Applied Sciences* 2025），以"知识图谱 + 智能体化检索"的双检索策略回答课程相关问题，兼顾可解释性与语义覆盖——这对"概念之间有明确前后置关系"的学科（如数学、物理）尤为契合。

· **智能体化 RAG（Agentic RAG）。** 引入"规划"与"反思"，由智能体根据问题复杂度动态决定是否多步检索、是否调用外部工具（数据库、API）、是否自我纠错后再生成，使 RAG 从"一次性检索器"升级为会"想清楚再查、查不到就换路"的智能体。这正是编排范式向 RAG 的渗透，也印证三范式本就同源。

把上述环节串起来，一条 2026 年"够用"的教育 RAG 流水线大致是：学生提问"为什么二次函数配方要先提取二次项系数"→ 智能体判断这是需要教材依据的概念题，触发混合检索，从校本知识库同时取回向量相似片段与含"配方法""二次项系数"关键词的片段 → 交叉编码器对候选重排，选出与该问最相关的少数几段教材原文 → 模型在这些片段的约束下生成解释，并为每一步结论标注其所依据的教材段落编号 → 一个校验环节检查"生成的每句是否被检索片段支持"，把"未提及/矛盾"的句子拦下重生成。学生看到的不只是答案，还有"这段解释来自本册教材第 X 节"的可追问出处；教师则可一键核验。这条流水线中真正决定体验的，是知识库切分的颗粒度、重排的精度与校验的严格度——它们都属"策展与工程"而非"模型参数"。

可评测，才可信。 RAG 之所以能承载"循证"承诺，前提是它的忠实度（faithfulness）本身可被度量。2025 年围绕 RAG 幻觉与忠实度的评测方法快速成熟：出现了以句子级标签（"支持 / 矛盾 / 未提及 / 补充"，后两类归为"不忠实"）标注答案与证据关系的基准，以及用高能力模型作为"裁判"（LLM-as-Judge）评估检索片段相关性、引用忠实度与答案完整性的框架，部分方法在忠实度判定上与人工标注的一致性已达到相当高的水平（arXiv:2505.04847、ACL/EMNLP 2025 Industry 等）。这为教育采购方"要求供应商提供第三方可复现的忠实度评测"提供了可操作抓手。

RAG 还带来一个在教育里被低估的能力：**有依据地"承认不知道"**。当检索不到可靠支撑时，一个负责任的教育 RAG 系统应当拒答或明示"教材中未涉及"，而非硬编一个听起来合理的答案——2025 年已有专门在"弃答策略（abstention policy）"下评测 RAG 幻觉的工作（researchgate 2025）。对未成年人而言，一个会说"这个我不确定，我们一起查教材"的系统，

比一个永远自信给答案的系统更值得托付；把"知之为知之"内建为产品行为，本身就是一种诚实的教育示范。

这条工程化路径与本蓝皮书的循证纪律同源。 凡结论须有据，宁留占位不臆造——可以说，RAG 是把"编辑部的核查规范"内化进产品的一种技术表达。国内厂商亦沿此方向治理幻觉：科大讯飞公开表示，讯飞星火提出"基于多路径采样验证及事实性约束强化学习的幻觉治理技术"，在 RAG 等任务的回复可靠性上有显著提升（量子位、53AI 2025）。

教育落地的真实证据。 RAG 在高等教育辅导中的试点已有同行评议记录：*ScienceDirect* 2025 年一项针对 AI 辅导的高校试点研究，专门评估了 RAG 模型在辅导中的应用效果；Springer 亦收录了用 RAG 提升数字素养（digital literacy）的教学工作；面向课堂问答，学界还在系统比较向量检索与图检索的取舍，以确立最佳实践（arXiv, *Aligning LLMs for the Classroom with Knowledge-Based Retrieval* 2025）。这些研究共同指向一个对产业极具指导性的结论：教育 RAG 的差异化，主要来自**知识库策展质量与检索—引用工程**，而非底层模型的参数量——把资源投在"策展一套高质量、结构清晰、持续更新的校本知识库"上，往往比追逐更大的模型更划算。

这一判断对国内产业尤具现实意义。我国教材版本多、地区差异大、课标随年级螺旋上升，恰是"通用大模型 + 分地区分学段校本知识库"最能发挥价值的场景：与其为每套教材重训一个模型，不如策展一套结构化、带课标标签、可持续更新的知识库，让同一底座模型据不同知识库服务不同地区。换言之，RAG 把"教育的本地性"从模型训练的负担，转化为知识策展的资产——谁的知识库更权威、更新更勤、结构更利于检索，谁的产品就更可信。这也解释了为何 10.2.2 主张"工程化优先于参数竞赛"：在教育垂类，护城河更多在数据与工程侧。

10.1.3 长期记忆：让系统"记得住这个学习者"

机理：把无状态的应答器改造为有状态的学习伙伴。 记忆机制赋予产品跨会话、跨时段保留学习者画像、错题脉络、学习偏好与目标进度的能力，是"千人千面"个性化学习的技术前提。缺少记忆的产品，每一次对话都是"初次见面"，所谓个性化只能靠用户每次重新交代背景，既低效又难以累积。记忆在技术上可粗分为三层——

- **短期/工作记忆**：单次会话内的上下文窗口，服务于即时应答与话题连贯；
- **长期记忆**：跨会话沉淀的结构化学习者档案（知识点掌握度、易错点、学习节奏、目标进度），是"记得住这个学习者"的核心；
- **情节与语义记忆**：对具体学习事件的记录（情节，如"3 月 5 日在二次函数配方法上卡了三次"），及其抽象出的稳定知识状态（语义，如"该生对配方法尚未掌握"）。前者支撑可回溯的过程性评价，后者支撑对当前水平的稳定判断。

举一个具体场景可见三层记忆如何协同：某学生连续两周在"分式方程验根"上反复出错——**情节记忆**记下"11 月 3 日、11 月 7 日、11 月 12 日三次在验根步骤漏检增根"这些具体事件；系统据此抽象出**语义记忆**"该生尚未内化验根的必要性，属程序性遗漏而非概念性误解"这一稳定判断；下次辅导开场，**工作记忆**把这条语义结论载入当前上下文，于是导师不必重问就能直接在验根处加设脚手架。没有记忆的产品做不到这一点：它每次都从零开始，只能一遍遍教同样的通法，既浪费学生时间，也无从沉淀"这个学习者的成长曲线"。可见记忆之于个性化，不是锦上添花，而是使"精准"二字成立的前提。

工程实现上，MemGPT / Letta 一系工作把大模型类比为"操作系统"：由系统自动对膨胀的会话做摘要、把不常用信息移入可检索的数据库、并保存与编辑姓名/日期/偏好等细节，从而把"无状态的模式处理器"改造为"能跨长时段持续学习与适应的有状态智能体"（Letta / MemGPT 2025）。这一"操作系统"隐喻的价值在于：它把"记什么、记多久、何时调取、何时遗忘"变成可显式设计、可审计的策略，而非模型不可控的副产品——这恰是教育场景对记忆治理的刚

需所在。在教育侧，2025 年的综述与新工作明确把"可复用、参数化的学习者画像"作为突破口——单提示、单模型的旧路径难以规模化维持一致的学习者档案，而多智能体系统可维护一份跨交互一致、同时刻画心理与认知维度的画像（arXiv:2503.11733 综述；EMNLP 2025 Findings）。值得反复强调的是，好的个性化不应只盯认知维度：动机、情绪与自我概念等因素同样深刻影响学习成效，一个只记住"错题"却读不懂"畏难情绪"的系统，其个性化是残缺的。

产品化现状：记忆已成默认能力。 记忆已从研究概念进入商用产品的默认配置：OpenAI 于 2025 年 7 月 29 日上线的 ChatGPT **Study Mode**，其课程难度即"基于评估技能水平的问题**以及对既往对话的记忆**来因材施教"，并以自定义系统指令实现苏格拉底式引导（先追问、给提示、促自省，而非直接给答案）（OpenAI 2025）。Google 亦把 LearnLM 的画像与个性化能力注入 Gemini App 的定制测验等场景（Google 2025）。国内学习硬件同样把跨会话的学情沉淀作为核心卖点：科大讯飞 AI 学习机 2025 暑期发布会以"AI 1 对 1 精准学 / 答疑辅导 / 互动课"为三大核心功能，教师助手可"智能生成学情报告、快速诊断学情"，其个性化的前提正是对学习者状态的持续记录（量子位 2025）。好未来则把九章大模型嵌入学习机的"随时问、精准学、作文批改"等功能，同样依赖跨环节的学情沉淀（好未来 2025）。

风险提示（与治理章强绑定）。 记忆能力越强，越触及未成年人数据的**采集边界、最小必要、留存期限与被遗忘权**。UNESCO 2025 年《生成式人工智能教育与研究应用指南》即建议强制保护数据隐私、并对生成式 AI 工具的使用设置年龄限制（UNESCO 2025）；我国《中小学生成式人工智能使用指南（2025 年版）》亦按学段设置了明确的使用边界（教育部 2025，详见 10.2.1）。本书据此主张：**记忆不是越多越好，而应"够用、可控、可删"**——凡沉淀学习者画像，须与第 [待补：治理章号] 章"治理与安全"的红线严格绑定，把数据最小化、目的限定、可审计、可删除作为产品基线而非附加项。一个负责任的教育记忆系统，应当既"记得住学习者"，也"随时忘得掉"——被遗忘权必须有可实操的通道，而非停留在隐私政策的文字里。

10.1.4 三范式的协同：一个统一的产品底座

三者并非孤立特性，而应视为**同一系统的三个面向**：编排提供"做事"的骨架，RAG 提供"有据"的事实地基，记忆提供"有个性"的连续性。三者叠加，才构成一个既会做事、又可溯源、还记得住学习者的教育智能体；缺任何一环，产品都会退回某种残缺形态——有编排无 RAG 则"做得多但错得也多"，有 RAG 无记忆则"每次都要重新认识你"，有记忆无编排则"记得住却做不了复合任务"。

近两年的旗舰产品恰是三范式的同框范例。Google 在 I/O 2025 把 LearnLM 直接融入 Gemini 2.5，并在"五项教学法原则"上对比 Claude 3.7、GPT-4o、OpenAI o3 等竞品，结果显示 Gemini 2.5 Pro 在每一类别上均更受偏好；教育与教学法专家在多种学习场景中更偏好它，评价维度不止于答案准确性，还包括"是否给出恰当引导、是否纠正学生错误"等教学法要素（Google 2025）。其能力被注入 NotebookLM、Learn About、Gemini App 的定制测验等产品，并支持上传源文档做接地解释——这正是"编排（把学习拆成引导步骤）+ RAG（接地于源文档）+ 记忆/画像（据水平因材施教）"的合流。OpenAI 的 Study Mode 与之呼应，先以系统指令小步快跑收集真实学生反馈，再择机把该行为训练进主模型（OpenAI 2025）。可见头部厂商已把三范式当作产品底座而非可选功能——这也提示国内产品：竞争的下半场不在单点模型能力，而在三范式的工程化整合质量。

再叠加**多模态输入与端侧推理**，即支撑起"智能体化 / 多模态 / 端侧化"的 2026 产品形态。此处需补足前两个维度与三范式的关系：多模态让编排的"感知"环节从纯文本扩展到语音、图像、手写与第一视角视频，学生可以把一道手写错题拍下、把课堂讲解说出来，系统才能真正"看见"学习现场；端侧推理则回应了教育场景特有的隐私与网络约束——把轻量模型与部分记忆放在本地设备、只在必要时上云，既降低时延与带宽依赖，又把最敏感的未成年人数据留在本地，天然契合 10.1.3 所述"数据最小化、可控可删"的记忆治理取向。因此"端侧化"在教育语境里不只是性能选择，更是隐私与公平的架构选择：它让网络受限地区也能用上有记

性的个性化辅导，而不必把学习数据悉数上传。其在硬件端的具身化，可参见本院《AI-SLI 2026 AI 智能眼镜教育产业蓝皮书》与《AI-SLI 全球教育机器人发展白皮书 2026》——前者示范多模态感知与第一视角交互（本院关联企业网龙 HK:0777 已于 2023 年 11 月向 Rokid 战略投资 2000 万美元并签署五年战略合作协议，Rokid Glasses 于 2025 年第二季度开售，据 Omdia 数据 2025 年 Rokid 在带显示功能的 AI 眼镜细分市场排名全球第一；智通财经、证券之星 2025）；后者示范具身智能体在物理空间中的落地边界。软件三范式与硬件具身化互引，构成本院蓝皮书体系的完整闭环：软件解决"会想、会查、有记性"，硬件解决"能看、能听、能在场"，二者合流才逼近"随身、无缝、有据"的下一代教育交互。

10.2 面向多元主体的发展建议

在上述范式判断基础上，我们向政策与标准、产业与产品、学校与教师、研究与公平四类主体提出结构化建议。凡涉及尚未核实的具体指标、清单或文号者，均以占位标注，待据实补录。四组建议共享一条主线：让"新范式"从技术卖点转化为**可交付的可靠性与可治理的公平性**。

10.2.1 政策与标准：以"可用、可信、可治理"为纲

我国已在 2025 年形成推进 AI 教育的政策底座。教育部于 2025 年 5 月 12 日发布《中小学人工智能通识教育指南（2025 年版）》与《中小学生成式人工智能使用指南（2025 年版）》，形成"通识素养 + 使用规范"双轨体系：前者以素养培育为核心、螺旋式课程从小学的体验兴趣，到初中的技术原理理解，再到高中的系统思维与创新应用；后者按学段分级设置使用边界——小学阶段禁止学生独自使用开放式内容生成功能、教师可在课内适当使用辅助教学，初中可适度探索生成内容的逻辑性分析，高中允许结合技术原理开展探究性学习（教育部；中国教育在线、CERNET、多知网 2025）。同期发布的《中国智慧教育白皮书（2025 年 5

月）》披露，已有 23 个省级教育行政部门部署开展中小学人工智能教育，北京、广州等地已出台具体应用工作方案（教育部 2025）。在此底座上，建议——

- **建立教育 AI 产品的准入与评测规范。** 推动形成面向教育垂类的能力评测与安全评测标准，涵盖学科准确性、价值观安全、未成年人保护与可溯源性四个维度。可参照第三方评测的现有实践：中国信通院已开展教育大模型评估并设分级，学而思九章大模型为首批通过并获当前最高评级（4+ 级）者（好未来 2025）。评测应把 10.1.2 所述的"忠实度可复现测量"纳入必测项，避免把"看起来会引用"误当"引用可靠"。评测口径与制定主体的最终标准文号 [待补：标准文号/来源]。
- **明确数据与记忆的合规边界。** 就学习者数据的采集、留存、跨境流动与"被遗忘"给出可操作细则，落实最小必要与目的限定原则，与 10.1.3 所述记忆治理红线对齐。尤应对"记忆型产品"提出比一般应用更严的留存期限与删除响应要求。
- **衔接现行法规与国际共识。** 对内衔接上述两份《指南》与智慧教育白皮书，对外参照 UNESCO 2025《生成式人工智能教育与研究应用指南》关于数据隐私保护、使用年龄限制的建议（UNESCO 2025）；具体条款与施行日期以正式文本为准 [待补：政策文号/施行日期]。

政策工具的选择宜"软硬结合"：对准入与安全底线（价值观、未成年人保护、数据删除）宜用强制性标准，对快速迭代的技术形态（如具体的编排架构、记忆策略）宜留监管弹性空间、以指引与最佳实践引导，避免用僵化条文锁死尚在演进的技术。这与我国两份《指南》"分学段、留探索空间"的思路一脉相承——小学从严、逐级放开，本身即是把"年龄—能力—风险"匹配起来的精细化治理范例，值得在产品准入层面同样贯彻：面向不同学段的同一功能，其数据与自主性边界理应不同。

10.2.2 产业与产品：把"新范式"做成"可交付的可靠性"

· **以循证替代宣称。** 产品能力主张须有第三方可复现的评测支撑，避免夸大自动化程度。头部厂商已示范"先系统指令、小步收集反馈，再训练进主模型"的谨慎发布路径（OpenAI Study Mode 2025），而非一步宣称"全自动教师"。凡宣称"AI 一对一""精准学"者，应能提供可核验的成效证据与失败率数据。产品横评方法参见第 [待补：评测章号] 章。

· **RAG 与记忆的工程化优先于模型参数竞赛。** 如 10.1.2 所论，教育场景的差异化更多来自知识策展质量、检索与引用工程、记忆的可控性，而非单纯堆参数。国产实践亦印证这一取向：讯飞以"事实性约束强化学习"治理 RAG 幻觉（量子位 2025）；好未来以"九章大模型 + DeepSeek"双轮驱动，并把能力嵌入学习机的作文批改、随时问、精准学等具体功能，九章大模型入选 2025 全球智慧教育优秀案例（好未来 2025）——竞争焦点已从"谁的模型大"转向"谁的落地稳"。

· **端侧与多模态并重。** 面向隐私敏感与网络受限的校园环境，推进端侧推理与轻量化部署（如以轻量模型承担辅助任务、把敏感数据留在本地设备），既降本又护隐私。同时把第一视角感知、语音与图像等多模态输入纳入产品基线，与本院 AI 眼镜、教育机器人两条硬件主线协同，为"随身、在场"的学习交互做准备。

10.2.3 学校与教师：素养先行，人机协同

· **教师 AI 素养建设。** 将"提示工程、结果核验、伦理与数据意识、人机分工判断"纳入教师专业发展。教师应能识别并纠正 AI 的事实性与价值性错误，而非被动采信——这恰是 RAG 之"可溯源"与编排之"可校验"设计要服务的对象：技术把中间产物摊开，教师才能有效在回路。素养建设不应停在"会用工具"，而应达到"会判断工具何时不可信"的层次；对应的素养框架与研训学时 [待补：素养框架/来源]。

· **明确人机责任边界。** 评价与决策类环节坚持"AI 辅助、教师负责"，高利害判断（如学籍、升学、心理危机识别）不得完全交由模型。产品设计应把"教师最终裁决"固化为不可跳过的流程节点，如 FACET 一类以教师为中心的编排系统所示范（arXiv 2025）——把人机分工从"倡议"落实为"架构约束"。这里有一条容易被忽视的风险：当 AI 输出越来越流畅可信，人在回路可能退化为"橡皮图章"式的形式审核（automation bias，自动化偏见），教师因信任而放松核验。对此，产品与研训需双向发力：产品应主动暴露不确定性与来源（RAG 的可溯源在此又是抓手），研训则要专门训练教师"对流畅但可疑的输出保持警惕"的判断力——素养的高阶目标，正是"会判断工具何时不可信"。

10.2.4 研究与公平：把"数字鸿沟"纳入设计前置

· **弥合接入与使用的双重鸿沟。** UNESCO 2025 年警示："数字鸿沟正迅速演变为 AI 鸿沟"，截至 2024 年全球仍有近三分之一人口（约 26 亿人）无法接入互联网，女童、乡村人口、残障者与边缘群体尤受影响（UNESCO 2025）。生成式 AI 若缺乏公平设计，可能沿两条路径放大既有不平等：**接入鸿沟**（算力、终端、网络的有无）与**使用鸿沟**（教师能力、数字素养的高低）——即便设备到位，不会用、不敢用同样会造成落差。研究应把区域、城乡、校际的差距监测作为常态，防止技术红利"先富者先得"。相关本地公平数据 [待补：公平数据/来源]。

· **面向弱势群体的可及性设计。** 将无障碍、多语言与特殊教育需求纳入产品基线要求而非附加项——把学习内容与解释以多语言、可朗读、可放大、可调节难度的形式默认提供。可参照 NotebookLM 音频概览已支持 80 余种语言的实践（Google 2025），把"人人可用"作为设计前置约束，而非产品成熟后再补的"无障碍模式"。这里三范式同样能反哺公平：端侧化让网络受限地区可离线使用，多模态让识字困难或视障学生可用语音与图像交互，RAG 的多语言知识库让少数民族语言与母语非通用语的学生也能得到接地于本地教材的

解释。技术的普惠性不会自动发生，只有当"最不利处境的学习者也能用"被写进设计目标与验收标准，公平才可能落地。

· **把"减负"而非"增负"作为成效标尺。** 有研究（发表于 *Nature Scientific Reports*）发现，学生使用 AI 辅导相比课堂主动学习能在更短时间内学得更多（转引自 tutorbase 汇编 2025）——但"学得更多更快"不应异化为变相加压。研究与评价应把学习者的负担、动机与心理健康纳入成效指标，警惕把 AI 效率红利单向兑换成更多刷题、更长学时。以学习者为中心，意味着把"人是否更从容地成长"置于"指标是否更好看"之上。

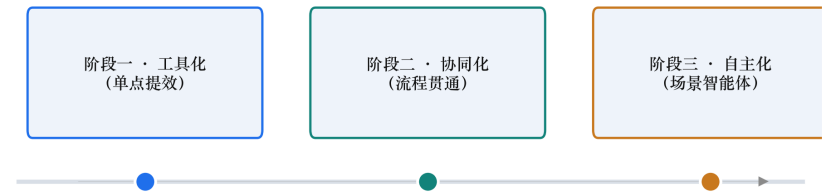
下表汇总四类主体的建议要点与优先抓手（未定项以占位标注）。

主体	核心建议	优先抓手	关键约束/参照
政策与标准	建立准入与评测规范、明确数据与记忆边界	教育垂类评测标准、数据合规与删除细则	两份《指南》、智慧教育白皮书、信通院分级评估、UNESCO 指南；标准文号 [待补]
产业与产品	以循证替宣称、工程化 RAG/记忆、端侧多模态	第三方忠实度评测、知识策展、端侧部署	谨慎发布路径（Study Mode）、幻觉治理（讯飞）；评测口径 [待补]
学校与教师	AI 素养建设、明确人机责任边界	教师研训、人机分工固化为流程节点	提示/核验/伦理素养、教师在回路（FACET）；素养框架 [待补]
研究与公平	弥合接入/使用双重鸿沟、可及性设计	差距监测、无障碍/多语言基线	AI 鸿沟（约 26 亿人未接入）、多语言可及；本地公平数据 [待补]

10.3 实施路线图：三阶段、可核验的推进节奏

为使建议可落地、可评估，我们提出"夯基—深化—协同"三阶段路线图。各阶段均设定性目标与可核验的里程碑；涉及具体时点、规模与考核指标者以占位标注，供实施主体据实填充。需强调，三阶段并非严格串行：允许在不同区域、不同学段并行推进，也允许发达地区先行进入第二、三阶段，同时为欠发达地区的"夯基"提供支持——路线图是逻辑先后，而非一刀切的时间表。

图 5 生成式 AI 教育产品落地的三阶段实施路线图



对话助手嵌入备课/答疑；建立数据与合规基线AG 接入校本知识库；人机分工；过程留痕与评测智能体编排 + 长期记忆；分级授权与可审计治理

里程碑均设可核验交付物与红线检查（详见正文）。

来源：本报告提出的实施路线图（详见第 10 章）。

10.3.1 第一阶段（夯基）：规范与基础设施

- **目标：** 明确评测与数据合规基线，完成教师 AI 素养的初步普及，摸清接入与使用两类鸿沟的底数。
- **里程碑：** 依据两份《指南》落地校本使用规范与分学段红线；对接第三方评测（如信通院教育大模型分级）确立本地采购的准入门槛 [待补：本地准入口径]；完成基础算力、终端与教师能力的公平性摸底 [待补：本地公平数据/来源]；开展首轮教师 AI 素养研训，重点覆盖"结果核验"与"数据意识"。
- **可核验标志：** 是否发布了校本/区域级使用规范；采购是否强制要求供应商提供第三方评测证据；教师研训覆盖率是否达到设定阈值 [待补]。

- **为何"夯基"不可跳过：** 采用侧的加速度（教师使用率两学年翻倍、青少年使用率翻倍，见 10.1 开篇）意味着"先用起来再谈规范"的窗口正在关闭——规范滞后于采用，就会积累难以回收的既成事实（数据已被采集、习惯已经养成）。夯基阶段的真正价值，是在规模化之前把评测标尺、数据红线与教师判断力立起来，避免"先污染后治理"。

10.3.2 第二阶段（深化）：范式落地与场景深耕

- **目标：** 将 RAG、记忆与智能体编排稳定嵌入赋能教学、支持学习、支持教研、智能评价四场景，形成可复用的参考架构，并把"教师在回路"固化进产品流程。
- **里程碑：** 建立跨场景的循证评测常态化机制，把答案忠实度、失败率、教师干预率纳入常规监测 [待补：机制/主体]；产出面向端侧与多模态的部署参考方案；在教研与批改等"流程固定、可校验"任务上先行规模化，再谨慎向长程辅导扩展。
- **落地次序建议：** 遵循"先易验、后难验"的稳妥次序——先在教研备课、客观题批改、有明确参考答案的答疑等"边界清晰、可校验"的任务上把三范式跑稳，积累失败案例库与教师干预数据；待评测机制成熟、可靠性可量化后，再谨慎向作文/主观题评价、长程个性化辅导等"高不确定、高利害"任务扩展。这一次序与 10.1.1 对编排成熟度的保守判断一致：把自主性作为受控增量逐步释放，而非一步到位。
- **可核验标志：** 是否有"教师在回路"的编排流程投入常态使用；RAG 答案是否默认可回链到校本知识库来源；是否建立了可复现的产品成效评测报告，且报告同时披露成功率与失败模式。

10.3.3 第三阶段（协同）：软硬协同与治理闭环

- **目标：** 实现"软件智能体—硬件具身"协同，在治理红线内规模化应用，接入与使用两类公平鸿沟显著收敛。

· **里程碑：** 与本院《AI-SLI 2026 AI 智能眼镜教育产业蓝皮书》《AI-SLI 全球教育机器人发展白皮书 2026》两条硬件主线形成互引与联评 [待补：联评口径]；建立数据与记忆的全生命周期审计闭环（采集—使用—留存—删除全程可追溯、可审计、可响应删除请求）；把可及性从"合规底线"提升为"设计前置"。

· **软硬协同为何是终局：** 软件三范式解决"会想、会查、有记性"，但学习发生在真实的物理与社交现场——课堂里的一次举手、实验台上的一个操作、同伴间的一次讨论。只有当感知与交互从屏幕延伸到第一视角眼镜、教育机器人等具身载体，AI 才能"在场"地理解学习情境，把编排的感知输入、RAG 的现场取证与记忆的连续画像真正落到学习现场。这也是本院以"软件蓝皮书 + AI 眼镜蓝皮书 + 教育机器人白皮书"三线并进的逻辑：三者互引联评，才能给出"随身、无缝、有据、有记性"的完整图景，而非各说各话的碎片。

· **可核验标志：** 是否具备学习者数据/记忆的"被遗忘"实操通道；软硬协同是否有可复现的联评报告；弱势群体的可及性指标是否纳入常态监测 [待补]。

10.4 结语

从 2024 到 2026，生成式人工智能教育产品的关键跃迁，不在模型"更能说"，而在系统"更会做事、更能溯源、更有记性"，并逐步走向多模态与端侧的具身形态。智能体编排、RAG 与长期记忆——"会做事、可溯源、有记性"——三范式合流，重构了产品的底座；本书以"五场景 + 三新范式"重构产品图谱，并以循证优先、宁留占位不臆造的研究纪律贯穿始终。

我们也清醒地看到范式的另一面：编排带来复杂度与不可控的自主性，RAG 的可靠性取决于知识库策展与忠实度评测，记忆越强则数据风险越高。因此三范式不是三张免检的王牌，而是三组需要被评测、被治理、被约束的能力。技术会持续迭代，但取向应保持稳定：**以学习者为中心、以教师为主责、以公平与安全为底线**。当模型越来越会"做事"，人对过程的可干

预、对来源的可核验、对数据的可控制，非但不应削弱，反而更需以制度与产品设计加以固化——这正是编排之"可校验"、RAG 之"可溯源"、记忆之"可删除"在价值层面的落点。

回望三范式，它们各自回应了一个古老的教育命题：编排对应"因材施教如何规模化"，RAG 对应"知识如何可信可溯源地传递"，记忆对应"如何真正认识并陪伴一个具体的学习者"。生成式 AI 没有发明这些命题，只是提供了迄今最有力也最需被审慎对待的工具。工具越有力，越考验使用者的价值定力：是把它用来把每个孩子看得更清、扶得更稳，还是用来把标准化的效率逻辑推向极致。本书的立场明确——三范式的意义，最终要以"是否让每一个学习者、尤其是最不利处境者，得到了更公平、更从容的成长"来检验。

愿本书为政策制定者、产业实践者、学校与研究者提供一份可核验、可延续的共同参照。技术的终点从不是更聪明的机器，而是更被看见、更被支持、更公平地成长的每一个学习者。

本章参考来源

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 2025 年 5 月技术报告）. 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（含忠实度句子级标签与 LLM-as-Judge 评测）. https://arxiv.org/html/2505.04847v2

12a. *Benchmarking Hallucination Evaluation for RAG Under an Abstention Policy*（弃答策略下的 RAG 幻觉评测，RAGAS/DeepEval/LLM-as-Judge）. ResearchGate, 2025. https://www.researchgate.net/publication/399331938

13. ReAct / Reflexion / ReflAct 智能体架构综述与原始工作：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*（Letta 框架与 MemGPT 长期记忆概念）. 2025. https://github.com/ksm26/LLMs-as-Operating-Systems-Agent-Memory

15. Kubiya. *Top AI Agent Orchestration Frameworks for Developers 2025*（含 Microsoft Agent Framework 2025-10-01 公开预览、LangChain 等）. 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. 教育部 / 中国教育在线. 《中小学生成式人工智能使用指南（2025 年版）》发布. 2025-05-12. https://www.eol.cn/zhengce/wenjian/202505/t20250512_2667831.shtml

18. CERNET（中国教育和科研计算机网）. 《中小学人工智能通识教育指南（2025 年版）》正式发布. 2025-05-13. https://www.edu.cn/xxh/focus/zc/202505/t20250513_2667990.shtml

19. 中华人民共和国教育部. 《中国智慧教育白皮书（2025 年 5 月）》（含 23 个省级部门部署 AI 教育）. 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*；及 *UNESCO convenes global leaders at Digital Learning Week 2025*（"AI 鸿沟"、截至 2024 年约 26 亿人未接入互联网、数据隐私与年龄限制建议）. 2023–2025. https://www.unesco.org/en/articles/unesco-convenes-global-leaders-digital-learning-week-2025-shape-inclusive-human-centred-futures-ai

21. 量子位. 科大讯飞"AI+教育"再提速：学习机功能升级（含"基于多路径采样验证及事实性约束强化学习的幻觉治理技术"、教师助手学情报告）. 2025-06. https://www.qbitai.com/2025/06/300819.html

22a. reruption / mlopsaudits. *Khanmigo: Khan Academy's GPT-4 AI Tutor*（架构：系统提示约束、模型路由、内容审核、教师工具；用户从约 6.8 万增至逾 70 万）. 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*（汇编 RAND 教师使用率 25%→53%、Gallup/Walton 60% K-12 教师、Pew 青少年 13%→26%、Nature Scientific Reports 学习成效）. 2025–2026. https://tutorbase.com/statistics/edtech-ai

22c. Precedence Research / Mordor Intelligence. *AI in Education Market Size*（2025 年全球市场约 69–70.5 亿美元、口径不一）. 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. 好未来（100tal）. 学而思九章大模型通过中国信通院教育大模型评估（4+ 级、最高评级）；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. 智通财经 / 证券之星. 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

附录 A 研究方法与数据口径

A.1 研究方法

本蓝皮书由 AI-SLI 采用"自建策展知识库 + AI 研制流水线"辅助研制：先建"赋能教学→支持学习→支持教研→智能评价→治理与安全"五场景分析框架，再对每一主题联网检索取证，仅依据检索到的真实一手资料撰写，每条具体事实绑定行内引用，并经人工审阅把关方向与红线。研究团队主导选题、框架、判断与最终结论；人工智能承担密集的检索、起草、制表与引用管理工作。产品图谱与形态判断以厂商公开资料、权威媒体报道与机构报告为据，避免以营销口径替代事实。

A.2 数据来源

数据以一手来源为主，包括：厂商官方产品文档与公告、各法域政策法规原文（如欧盟《人工智能法案》、美国 FERPA/COPPA、中国《生成式人工智能服务管理暂行办法》等）、国际组织与权威机构报告、公开评测基准与其可核验成绩、以及权威媒体报道。各章末"本章参考来源"逐条列出该章真实引用（标题·机构·年份·URL）。凡涉网龙（NetDragon，港股 HK:0777）等关联企业产品，均以其官方公开信息为准。

A.3 数据口径审慎原则

- **市场数据分口径**：不同研究机构对"生成式 AI 教育""教育科技（EdTech）"市场的定义、地域范围与统计年份不同，其规模数字不可直接横向比较或相加；引用时保留机构名、口径与发布年份。

· **币种防火墙**：人民币、美元、港元不混算；涉及金额一律标注币种与时点，不做隐性换算。

· **预测与实际区分**：凡机构预测/估计值一律标注"预测"与发布时点，与实际统计值区分呈现。

· **评测成绩可核验**：模型/产品评测分数仅采用可查证的公开来源；来源不明的分数不予采用。

· **占位而非臆造**：凡未能经检索核实的具体数字、政策文号、产品名单、评测成绩或文献，一律以 [待补：…] 标注，绝不臆造。本文件为研究版本，[待补] 处待多源核验后定稿。

A.4 局限与后续

生成式 AI 教育产品迭代极快，本版本对部分快速变化的规格、价格、评测成绩与市场规模保留 [待补] 占位，后续将以事实核查流水线补齐真实来源后形成定稿版。产品举例用于说明形态与机理，不构成任何选型推荐或投资建议。

附录 B 术语表

术语	释义
生成式人工智能（GenAI）	能够生成文本、图像、语音、代码等内容的人工智能技术，以大语言模型为代表。
智能体（Agent）	能理解目标、拆解任务、调用工具并连续执行多步任务的人工智能系统。
智能体编排（Agent Orchestration）	对多个智能体/工具的规划、调度与协作，使系统从"回答"走向"完成任务"。
检索增强生成（RAG）	在生成前检索外部知识库并将其作为依据，用于提升准确性、实现引用溯源、抑制臆造。
长期记忆（Memory）	跨会话保存学情、偏好与上下文，支撑持续个性化的机制。
多模态（Multimodal）	在统一模型中贯通文本、图像、语音、视频、代码等多种信息形态的能力。
端侧（On-device）	将推理能力部署于个人设备或本地终端，具备低时延、强隐私与离线可用特性。
教育垂类模型	面向教育场景，经学科语料、教学任务与安全对齐优化的专用大模型。
场景效度	评测所测能力与真实教育应用场景需求的一致程度。
过程性评价	面向学习过程（作答轨迹、修改、协作等）而非仅最终产物的评价方式。
护栏（Guardrail）	对模型输入输出施加的安全、合规与内容约束机制。
学术诚信	在人机协作学习中对原创性、署名与合理使用 AI 的规范与判断。

数据最小化	仅收集与处理达成目的所必需的最少个人数据的原则。
EU AI Act	欧盟《人工智能法案》，对 AI 系统按风险分级监管的法规。
FERPA / COPPA	美国《家庭教育权利与隐私法》/《儿童在线隐私保护法》。

附录 C　参考文献体例说明

本蓝皮书正文采用"随文标注 + 章末来源"体例：正文中对具体事实以括注或行文点出来源机构与年份，章末"本章参考来源"逐条给出可核验条目（标题 · 机构/作者 · 年份 · URL）。体例遵循以下约定：

- 仅收录研制过程中实际检索并核对过的来源；未经核实者不列入，相应事实在正文以 [待补] 标注。
- 优先一手来源（政策法规原文、厂商官方文档、机构原始报告、评测基准官方页面）；媒体报道作为补充并注明媒体名。
- 同一事实存在多方口径时，保留主要来源并在正文说明口径差异。
- 网址为便于核验的公开链接；如链接失效，以标题与机构名检索可复得。

（定稿版将统一转为编号引用并生成完整参考文献表。）

AI-SLI

2026 · AI 辅助研制 · 研究版