研究所
分析师：肖承志
SAC 登记编号：S1340524090001
Email：xiaochengzhi@cnpsec.com
研究助理：冯昱文
SAC 登记编号：S1340124100011
Email：fengyuwen@cnpsec.com

近期研究报告

《反转风格显著，小市值回撤——中邮因子周报 20250622》 - 2025.06.23
《关注基本面支撑，高波风格占优——中邮因子周报 20250615》- 2025.06.16
《结合基本面和量价特征的 GRU 模型》 - 2025.06.05
《Claude 4 系列发布，谷歌上线编程智能体 Jules——AI 动态汇总 20250526》 - 2025.05.27
《谷歌发布智能体白皮书，Manus 全面开放注册——AI 动态汇总 20250519》- 2025.05.20
《证监会修改《重组办法》，深化并购重组改革——微盘股指数周报 20250518》 - 2025.05.19
《通义千问发布 Qwen-3 模型，DeepSeek 发布数理证明大模型——AI 动态汇总 20250505》 - 2025.05.06
《基金 Q1 加仓有色汽车传媒，减仓电新食饮通信——公募基金 2025Q1 季报点评》 - 2025.04.30
《泛消费打开连板与涨幅高度，ETF 资金平铺机器人、人工智能与芯片——行业轮动周报 20250427》 - 2025.04.28
《国家队交易特征显著，短期指数仍交易补缺预期，TMT 类题材仍需等待——行业轮动周报 20250420》- 2025.04.21

金工周报

Anthropic 公开多智能体构建全流程，MiniMax 推出推理模型 M1——AI 动态汇总 20250623

Anthropic 公开多智能体构建全流程

Anthropic 近期公开的多智能体系统构建全流程，揭示了其如何通过协同多个 Claude 智能体实现复杂研究任务的高效处理。这一架构的核心在于将开放性问题的动态探索能力与并行化计算优势结合，同时通过精细的工程化设计解决协调复杂性、状态管理等挑战。

MiniMax 推出推理模型 M1

MiniMax 推出的全球首个开源大规模混合架构推理模型 M1，标志着人工智能领域在高效推理与长文本处理技术上的重大突破。这一模型的发布不仅重新定义了开源模型的性能边界，更通过架构创新与算法优化，实现了成本与效率的双重革命。

A.Lawsen 发表评论文章反击苹果对于大模型的观点

A. Lawsen 于 6 月 15 日发布博客《When Your Joke Paper Goes Viral》并发布相应的评论文章对苹果 Shojaee 等人之前关于大型推理模型（LRMs）在复杂规划谜题上存在"准确性崩溃"现象的研究提出了系统性反驳。作者 Lawsen 通过严谨分析揭示，原始研究的结论主要源于实验设计缺陷而非模型本质性推理能力局限，其核心论点围绕三个关键维度展开。

谷歌发布 Gemini 2.5 Flash-Lite

谷歌于 2025 年 6 月推出的 Gemini 2.5 Flash-Lite，是其 AI 模型家族中定位最为极致的轻量化产品，标志着大模型技术向高吞吐、低延迟场景的深度渗透。这一模型并非简单的能力阉割版，而是通过架构优化与成本控制策略，将旗舰级能力下放至经济型层级的典范之作。

风险提示：

以上内容基于历史数据完成，在政策、市场环境发生变化时存在失效的风险；历史信息不代表未来。

目录

1 AI 重点要闻 4
1.1 Anthropic 公开多智能体构建全流程 4
1.2 MiniMax 推出推理模型 M1 7
1.3 A.Lawsen 发表评论文章反击苹果对于大模型的观点 10
1.4 谷歌发布 Gemini 2.5 Flash-Lite 12
2 企业动态 15
2.1 谷歌曝光 AI 路线图，承认 Transformer 存在根本缺陷 15
2.2 OpenAI Codex 推出新功能，可一次生成多个方案 17
3 AI 行业洞察 19
3.1 哈佛大学开源 AI 训练数据集“Institutional Books 1.0” 20
4 技术前沿 22
4.1 MIT：沉迷 GPT 会导致大脑神经链接减少 47% 22
5 风险提示 25

图表目录

图表 1： Research 系统架构概览 4
图表 2： Clio 嵌入图：Research 功能的主要使用方式 4
图表 3： 多智能体系统架构 5
图表 4： M1 主流基准跑分 7
图表 5： 17 个主流测试集上 M1 表现 9
图表 6： When Your Joke Paper Goes Viral 10
图表 7： 针对苹果的评论文章 10
图表 8： Gemini 2.5 Flash-Lite 测评 12
图表 9： Gemini 2.5 家族对比 14
图表 10： Logan Kilpatrick 展示 Gemini 模型未来 15
图表 11： Codex 使用界面 18
图表 12： Institutional Books 1.0 项目 20
图表 13： 论文 Your Brain on ChatGPT 22
图表 14： Alpha Band 动态传递函数 EEG 分析 23
图表 15： 推理中大脑参与百分比只有 10%左右 23

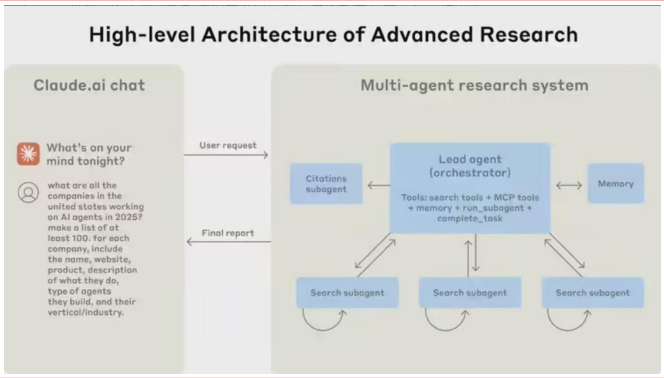
1 AI 重点要闻

1.1 Anthropic 公开多智能体构建全流程

Anthropic 近期公开的多智能体系统构建全流程，揭示了其如何通过协同多个 Claude 智能体实现复杂研究任务的高效处理。这一架构的核心在于将开放性问题的动态探索能力与并行化计算优势结合，同时通过精细的工程化设计解决协调复杂性、状态管理等挑战。

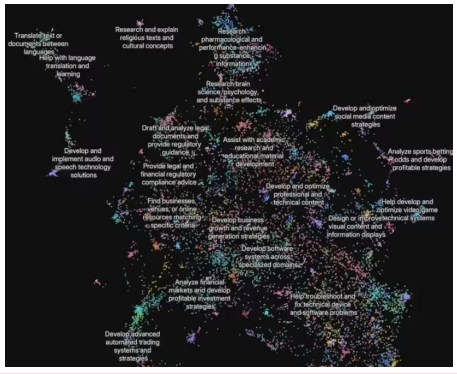
传统单智能体在处理开放式研究任务（如商业机会挖掘、医疗方案分析）时面临根本性局限：路径依赖性强且上下文窗口有限。Anthropic 通过数据证明，采用“领导-工作者”模式的多智能体系统（主智能体为 Claude Opus 4，子智能体为 Claude Sonnet 4）性能比单智能体提升 90.2%。其本质是通过并行消耗约单体的 15 倍 token 扩展算力，尤其适合需要“广度优先”探索的场景，如跨领域信息整合或动态调整研究策略的任务。但该架构对高价值任务的依赖性显著，因成本高昂不适用于简单查询或强依赖共享上下文的任务。

图表 1：Research 系统架构概览



资料来源：Anthropic，中邮证券研究所

图表 2：Clio 嵌入图：Research 功能的主要使用方式



资料来源：Anthropic，中邮证券研究所

系统采用分层协作的“协调者-工作者”（Orchestrator-Worker）架构，分为以下几个阶段实现闭环：

1. 任务解析与规划 ：用户提问触发主智能体（Lead Researcher）创建，其首要任务是将模糊需求转化为可执行计划，并将计划存储于外部记忆系统，避免因上下文截断导致信息丢失。

2. 动态任务分配：主智能体根据复杂度生成子智能体（Subagents），每个子任务需明确目标、输出格式、工具边界（如禁止重复搜索同一来源），避免早期版本中"50 个子智能体处理简单查询"的失控问题。

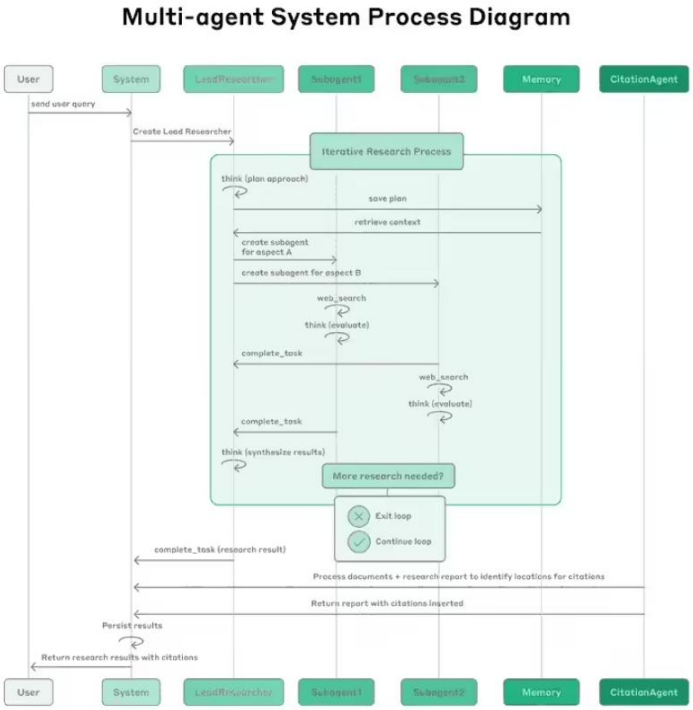
3. 并行执行与交错思考：子智能体独立执行网络搜索或工具调用，采用"交错思考"（Interleaved Thinking）策略——即在每次工具调用后暂停，评估结果可信度并调整下一步行动，类似人类研究者的反思过程。

4. 结果汇总与迭代：主智能体综合子智能体返回的中间结果，决定是否启动新一轮探索。这一阶段通过"记忆-检索"机制保持上下文连贯性，例如将关键信息压缩后存储于外部文件系统，减少 token 消耗。

5. 引用生成与验证：最终输出前，专门的"引用智能体"（Citation Agent）对所有结论进行溯源，确保权威信源优先（如学术 PDF 而非 SEO 优化的内容农场），并通过结构化标注提升结果可信度。

与传统检索增强生成（RAG）的静态检索相比，该架构通过多步动态搜索实现自适应信息获取，能够根据中间发现调整策略，形成更高质量的答案。

图表 3：多智能体系统架构



资料来源：Anthropic，中邮证券研究所

多智能体系统的核心挑战在于协调复杂性，Anthropic 通过三大技术创新实现可控性：分别是提示工程的八项法则、双层并行化加速 、评估方法的革新。

提示工程的八项法则 包括模拟智能体视角观察行为模式（如发现冗余搜索后优化终止条件）、主智能体任务分解的“四要素”（目标/格式/工具/边界）、以及“先广后窄”的搜索策略（从宽泛查询逐步聚焦）。特别值得注意的是，系统引入“工具测试智能体”自动重写模糊的工具描述，使后续调用效率提升 40%。

双层并行化加速 是指，主智能体并行启动 3-5 个子智能体，每个子智能体同时调用 3 个以上工具，使复杂任务耗时减少 90%。这种设计突破了同步执行的瓶颈（主智能体需等待所有子任务完成），但异步化仍面临状态一致性与错误传导的挑战。

评估方法的革新 层面，采用“小样本快速迭代+LLM 裁判+人工抽查”的三元评估法。例如，用 LLM 根据事实准确性、引用质量等维度打分（0.0-1.0），而人工测试则发现自动化评估的盲点（如信源偏好偏差）。

Anthropic 强调，多智能体系统的“最后一公里”工程化难度远超预期，需解决四大核心问题：

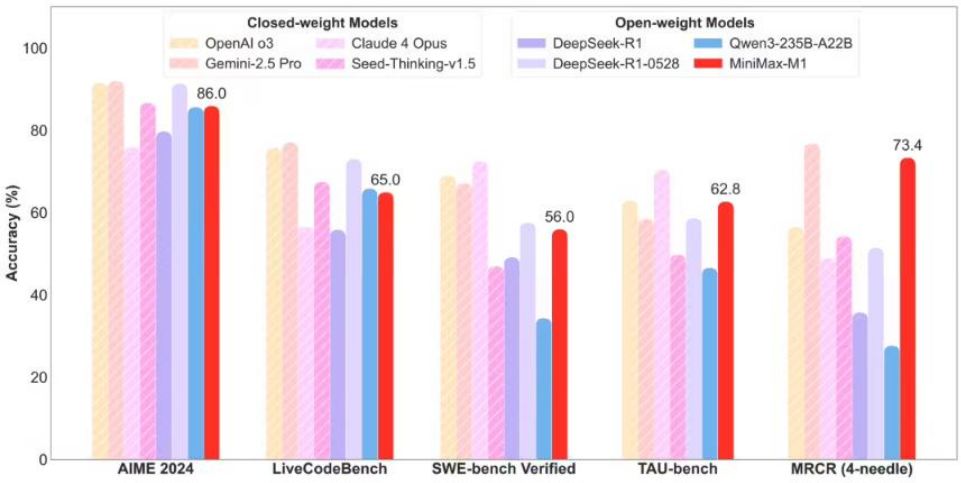
- 错误恢复机制：智能体的长时运行特性要求支持断点续执行，例如在工具调用失败时，系统需告知智能体并引导其调整策略，而非重启整个流程。
- 非确定性调试：通过全链路追踪系统监控决策模式（如工具调用序列），定位失败根因。例如某次故障被追溯为子智能体误解了"半导体短缺"的时间范围。
- 彩虹部署策略（Rainbow Deployments）：新旧版本并行运行并渐进迁移流量，避免中断正在执行的任务。
- 状态一致性管理：未来将探索异步执行以提升吞吐量，但需权衡结果协调复杂度与错误传导风险。

Anthropic 的多智能体系统展现了 AI 在复杂问题解决上的突破，其核心价值在于将人类研究策略编码为可扩展的架构（如动态任务分解、启发式搜索），并通过工程实践（如提示精细化、状态持久化）弥补了 LLM 的固有局限。尽管存在成本与协调复杂性的约束，该系统已在医疗决策、技术漏洞分析等场景中验证了其不可替代性——用户反馈表明，其能帮助发现传统方法难以捕捉的研究关联，节省数天工作量。这一架构的未来演进可能聚焦于异步协作优化与成本控制，进一步拓展多智能体系统的适用边界。

1.2 MiniMax 推出推理模型 M1

MiniMax 推出的全球首个开源大规模混合架构推理模型 M1，标志着人工智能领域在高效推理与长文本处理技术上的重大突破。这一模型的发布不仅重新定义了开源模型的性能边界，更通过架构创新与算法优化，实现了成本与效率的双重革命。

图表 4：M1 主流基准跑分



资料来源：MiniMax，中邮证券研究所

M1 的核心突破在于其独创的“闪电注意力”（Lightning Attention）混合架构，该架构将线性注意力与传统 Softmax 注意力机制深度融合，解决了 Transformer 模型在处理长序列时计算量平方级增长的固有瓶颈。传统模型如 DeepSeek R1 在处理 10 万 Token 生成任务时需消耗大量算力，而 M1 通过分块策略——局部上下文采用传统注意力、全局摘要采用线性注意力——将计算量压缩至竞品的 25%。这种设计使得模型原生支持 100 万 Token 的输入窗口，与谷歌 Gemini 2.5 Pro 持平，同时输出长度达 8 万 Token，创下开源模型纪录。

在模型结构上，M1 延续了稀疏化专家混合网络（MoE）路线，总参数规模达 4560 亿，但通过动态激活约 459 亿参数实现高效推理。这种设计源于 MiniMax 此前 MiniMax-Text-01 的技术积累，其通过 7.5 万亿 Token 的定向增强预训练（STEM、编程代码与复杂推理数据占比 70%）夯实基座能力，再经监督微调注入链式思考机制，最终通过 CISPO 强化学习算法完成能力跃迁。

在涵盖数学推理、代码生成、长文本理解等 17 项主流基准测试中，M1 展现出全面竞争力。尤其在长上下文任务 OpenAI-MRCR（1M）中，M1 以 73.4%准确率超越 Claude 4 Opus 等闭源模型，仅次于 Gemini 2.5 Pro 位列全球第二；在代理工具调用基准 TAU-bench 上，M1-40k 版本以 62 分超越所有开源模型及 Gemini 2.5 Pro，验证其作为智能体基座的潜力。值得注意的是，M1-80k 版本在多数测

试中优于 40k 版本，例如在 AIME 2024 数学推理测试中达到 86%，较 40k 版本提升 2.7 个百分点，证明扩展计算资源可稳定提升性能。

图表 5：17 个主流测试集上 M1 表现

Tasks	Leading Close-Weights Models				Open-Weights Models			Our Models	
	OpenAI-o3	Gemini 2.5 Pro (06-05)	Claude 4 Opus	Seed-Thinking-v1.5	DeepSeek-R1	DeepSeek-R1-0528	Qwen3-235B-A22B	MiniMax-M1-40k	MiniMax-M1-80k
Extended Thinking	*100K*	*64K*	*64K*	*32K*	*32K*	*64K*	*32K*	*40K*	*80K*
Mathematics									
AIME 2024	91.6	92.0	76.0	86.7	79.8	91.4	85.7	83.3	86.0
AIME 2025	88.9	88.0	75.5	74.0	70.0	87.5	81.5	74.6	76.9
MATH-500	98.1	98.8	98.2	96.7	97.3	98.0	96.2	96.0	96.8
General Coding									
LiveCodeBench *(24/8~25/5)*	75.8	77.1	56.6	67.5	55.9	73.1	65.9	62.3	65.0
FullStackBench	69.3	–	70.3	69.9	70.1	69.4	62.9	67.6	68.3
Reasoning & Knowledge									
GPQA Diamond	83.3	86.4	79.6	77.3	71.5	81.0	71.1	69.2	70.0
HLE *(no tools)*	20.3	21.6	10.7	8.2	8.6*	17.7*	7.6*	7.2*	8.4*
ZebraLogic	95.8	91.6	95.1	84.4	78.7	95.1	80.3	80.1	86.8
MMLU-Pro	85.0	86.0	85.0	87.0	84.0	85.0	83.0	80.6	81.1
Software Engineering									
SWE-bench Verified	69.1	67.2	72.5	47.0	49.2	57.6	34.4	55.6	56.0
Long Context									
OpenAI-MRCR *(128k)*	56.5	76.8	48.9	54.3	35.8	51.5	27.7	76.1	73.4
OpenAI-MRCR *(1M)*	–	58.8	–	–	–	–	–	58.6	56.2
LongBench-v2	58.8	65.0	55.6	52.5	58.3	52.1	50.1	61.0	61.5
Agentic Tool Use									
TAU-bench *(airline)*	52.0	50.0	59.6	44.0	–	53.5	34.7	60.0	62.0
TAU-bench *(retail)*	73.9	67.0	81.4	55.7	–	63.9	58.6	67.8	63.5
Factuality									
SimpleQA	49.4	54.0	–	12.9	30.1	27.8	11.0	17.9	18.5
General Assistant									
MultiChallenge	56.5	51.8	45.8	43.0	40.7	45.0	40.0	44.7	44.7

* conducted on the text-only HLE subset.

资料来源：MiniMax，中邮证券研究所

软件工程场景的 SWE-bench 测试中，M1 虽以 56%略低于 DeepSeek-R1-0528 的 57.6%，但显著领先其他开源模型。这种“长板突出、短板不短”的特性，使其在分析 400 页技术文档或十万行代码库时，能保持连贯推理而无需分段处理，为复杂生产力任务提供新范式。

M1 最令行业震撼的莫过于其极致的成本控制能力。通过 CISPO 算法优化重要性采样权重，其强化学习阶段仅需 512 块 H800 GPU 运行三周，总成本 53.47 万美元，较 DeepSeek R1 的 500-600 万美元训练成本下降一个量级。该算法在 AIME 测试中收敛速度较字节跳动 DAPO 快一倍，通过避免裁剪低概率但关键的推理 Token（如 however、wait 等），维持了策略更新的稳定性。

推理阶段的效率优势更为显著：生成 8 万 Token 仅需 DeepSeek R1 约 30%算力，10 万 Token 时进一步降至 25%。这种效率直接转化为 API 定价的竞争力——128k-1M Token 档位输入成本 2.4 元/百万 Token,仅为同类长文本服务的基准价。

MiniMax 此次不仅开源模型权重与技术报告，更与国家超算平台、vLLM 等框架合作优化部署体验。其官网提供不限量免费试用，同时推出阶梯式 API 定价：0-32k Token 档位输出价格仅 8 元/百万 Token，显著低于 DeepSeek R1 同档服务。这种激进策略旨在快速构建开发者生态，截至发布后 8 小时，GitHub 仓库即获 440 星，虽暂落后于通义千问等成熟项目，但凭借长文本处理等差异化能力，有望吸引垂直领域开发者。

M1 的发布标志着混合注意力架构首次在千亿参数规模实现生产级应用，挑战了 Transformer 优化的主流范式。行业观察人士指出，当全球聚焦于 Transformer 微调时，MiniMax 通过线性注意力与 MoE 的融合，开辟了效率优先的新技术路径。其影响已超越技术层面：53 万美元的训练成本证明，通过算法创新可打破“算力-资本”壁垒，倒逼行业从参数竞赛转向价值创造。

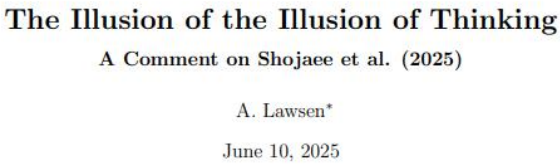
未来挑战仍存，如模型泛化能力在量子计算等新兴领域的表现、创意生成保守倾向，以及对高性能硬件的依赖。但 M1 已为 AI 民主化树立标杆——当百万 Token 上下文不再是天价服务，当开源模型性能比肩闭源巨头，技术普惠的进程将加速重构。正如 MiniMax 技术报告所述，这场与 DeepSeek 的“中国 AI 双雄之争”，最终赢家将是整个产业。

1.3 A.Lawsen 发表评论文章反击苹果对于大模型的观点

A. Lawsen 于 6 月 15 日发布博客《When Your Joke Paper Goes Viral》并发布相应的评论文章对苹果 Shojaee 等人之前关于大型推理模型（LRMs）在复杂规划谜题上存在"准确性崩溃"现象的研究提出了系统性反驳。作者 Lawsen 通过严谨分析揭示，原始研究的结论主要源于实验设计缺陷而非模型本质性推理能力局限，其核心论点围绕三个关键维度展开。

图表 6：When Your Joke Paper Goes Viral　　**图表 7：针对苹果的评论文章**

资料来源：A.Lawsen，中邮证券研究所



资料来源：A.Lawsen，中邮证券研究所

首先，文章指出原始研究未能识别模型对输出限制的主动认知。在河内塔等实验中，模型明确表现出对上下文窗口限制的自我意识，例如输出中会出现"为避免内容过长，我将在此停止"等表述。这种截断行为被错误归类为推理失败，实则反映模型对工程约束的合理响应。作者通过数学建模证明，若按字符级精度评估长序列输出，任何模型都会因统计必然性呈现"崩溃"，但这种评估方式忽视了模型动态调整输出策略的智能特性。

其次，文章揭露原始研究存在基准测试的根本性缺陷。在传教士与野人问题中，研究者设置了数学上无解的 N≥6 场景，却将模型无法解决不可解问题的表现记为失败。这种评估方法不仅无法检验真实推理能力，反而暴露出自动化评分系统对问题本身可解性验证的缺失。更严重的是，此类错误结论可能误导后续研究对模型能力的认知。

第三，作者通过替代性表征实验验证了模型的实际能力。当要求模型输出河内塔问题的 Lua 递归函数而非完整移动步骤时，包括 Claude-3.7-Sonnet 在内的主流模型在 N=15 高复杂度场景下均表现出近乎完美的准确性。这证实模型的"失败"源于输出格式约束，而非算法理解缺陷。文章进一步指出，原始研究采用的"组合深度"指标存在概念混淆——河内塔虽需指数级移动步骤，但单步决策复杂度仅为 O(1)；而积木世界等真正具有高分支因子的任务才更能检验规划能力。

在方法论层面，本文批判了当前 AI 评估体系中普遍存在的三个误区：将输出长度约束误判为认知局限、忽视问题本身的可解性验证、以及用表面指标替代计算复杂性分析。作者提出，优质评估应区分"无法解决"与"选择不穷举"的本质差异，这需要设计能够捕捉模型元认知能力的测试框架。

值得注意的是，本文的论证过程本身也经历了迭代完善。初稿中关于模型输出模式的假设经社区反馈后修正，最终版本更精确地反映出模型实际采用"直接输出最终步骤"而非"枚举中间状态"的策略。这种学术严谨性强化了核心论点：

当评估方法不能准确反映模型行为机制时，任何关于能力极限的结论都可能成为自我实现的预言。

文末提出的四项改进方向——设计约束感知的评估、验证基准可解性、采用真实验算复杂度指标、开发多元解决方案表征——不仅针对原始研究的不足，更为整个 AI 推理能力评估领域提供了建设性框架。其根本主张在于：当前紧要问题并非"LRM 能否推理"，而是"我们的评估方法能否区分推理与输出约束"。这一洞见对突破 AI 能力认知的边界具有深远意义。

1.4 谷歌发布 Gemini 2.5 Flash-Lite

谷歌于 2025 年 6 月推出的 Gemini 2.5 Flash-Lite，是其 AI 模型家族中定位最为极致的轻量化产品，标志着大模型技术向高吞吐、低延迟场景的深度渗透。这一模型并非简单的能力阉割版，而是通过架构优化与成本控制策略，将旗舰级能力下放至经济型层级的典范之作。

Gemini 2.5 Flash-Lite 继承了 2.5 系列的核心架构基因，包括稀疏 MoE 设计与百万 Token 上下文窗口支持，但通过动态计算资源分配实现了质的飞跃。其最显著的特点是采用"可控思考预算"机制，默认关闭深度推理功能以换取速度优势，开发者可通过 API 参数按需开启思考模式，在简单分类任务与复杂逻辑推理间灵活切换。这种设计使得模型在保持多模态处理能力的同时，将首 Token 响应时间压缩至 0.29 秒以下，生成速度最高可达 428 Token/秒，远超同类轻量模型如 Claude 4 Sonnet 的吞吐效率。

图表 8：Gemini 2.5 Flash-Lite 测评

Benchmark	Gemini 2.5 Flash-Lite Preview 06-17 Non-thinking	Gemini 2.5 Flash-Lite Preview 06-17 Thinking	Gemini 2.5 Flash Non-thinking	Gemini 2.5 Flash Thinking	Gemini 2.5 Pro Thinking
Input price ($/1M tokens no caching)	$0.10	$0.10	$0.30	$0.30	$1.25 ($2.50 > 200k tokens)
Output price ($/1M tokens)	$0.40	$0.40	$2.50	$2.50	$10.00 ($15.00 > 200k tokens)
Reasoning & knowledge Humanity's Last Exam (no tools)	5.1%	6.9%	8.4%	11.0%	21.6%
Science GPQA diamond	64.6%	66.7%	78.3%	82.8%	86.4%
Mathematics AIME 2025	49.8%	63.1%	61.6%	72.0%	88.0%
Code generation LiveCodeBench (UI: 1/1/2025-5/1/2025)	33.7%	34.3%	41.1%	55.4%	69.0%
Code editing Aider Polyglot	26.7%	27.1%	44.0%	56.7%	82.2%
Agentic coding SWE-bench Verified	31.6% (single attempt) / 42.6% (multiple attempts)	27.6% / 44.9%	50.0% / 60.0%	48.9% / 60.3%	59.6% / 67.2%
Factuality SimpleQA	10.7%	13.0%	25.8%	26.9%	54.0%
Factuality FACTS Grounding	84.1%	86.8%	83.4%	85.3%	87.8%
Visual reasoning MMMU	72.9%	72.9%	76.9%	79.7%	82.0%
Image understanding Vibe-Eval (Reka)	51.3%	57.5%	66.2%	65.4%	67.2%
Long context MRCR v2 (8-needle)	16.6% (128k (average)) / 4.1% (1M (pointwise))	30.6% / 5.4%	34.1% / 16.8%	54.3% / 21.0%	58.0% / 16.4%
Multilingual performance Global MMLU (Lite)	81.1%	84.5%	85.8%	88.4%	89.2%

Methodology

Gemini results: All Gemini scores are pass @1."Single attempt" settings allow no majority voting or parallel test-time compute; "multiple attempts" settings allow test-time selection of the candidate answer. They are all run with the AI Studio API with default sampling settings. To reduce variance, we average over multiple trials for smaller benchmarks. Aider Polyglot score is the pass rate average of 3 trials. Vibe-Eval results are reported using Gemini as a judge. Google's scaffolding for "multiple attempts" for SWE-bench includes drawing multiple trajectories and re-scoring them using model's own judgement. For Aider results differ from the official leaderboard due to a difference in the settings used for evaluation (non-default).

Result sources: Where provider numbers are not available we report numbers from leaderboards reporting results on these benchmarks: Humanity's Last Exam results are sourced from https://agi.safe.ai/ and https://scale.com/leaderboard/humanitys_last_exam, LiveCodeBench results are from https://livecodebench.github.io/leaderboard.html (1/1/2025 - 5/1/2025 in the UI). Aider Polyglot numbers come from https://aider.chat/docs/leaderboards/. FACTS come from https://www.kaggle.com/benchmarks/google/facts-grounding. For MRCR v2 which is not publically available yet we include 128k results as a cumulative score to ensure they can be comparable with other models and a pointwise value for 1M context window to show the capability of the model at full length. The methodology has changed in this table vs previously published results for MRCR v2 as we have decided to focus on a harder, 8-needle version of the benchmark going forward.

Input and output price reflects text, image and video modalities.

资料来源：Google，中邮证券研究所

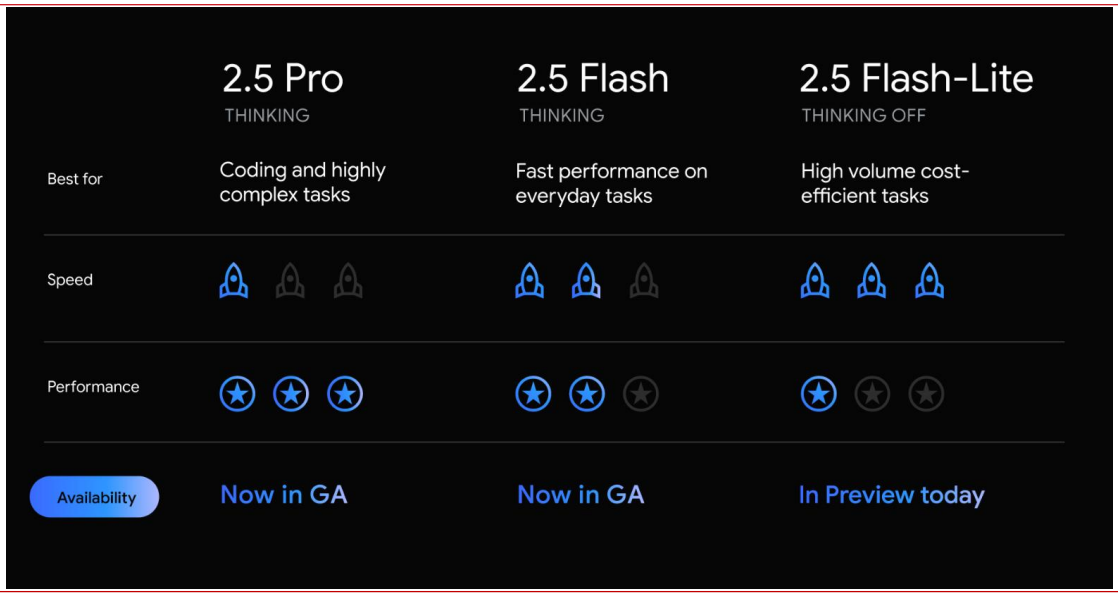
硬件适配层面，Flash-Lite 针对谷歌第七代 TPU（算力 4614 TFLOPs）进行专项优化，利用 Ironwood 芯片的并行计算特性，在处理 18K Token 长输入时仍能维持 160+ Token/秒的稳定输出。这种硬件协同设计使其在同等算力消耗下，较前代 2.0 Flash-Lite 实现 30%的能效提升，为高并发场景提供基础设施保障。

在谷歌构建的模型性能-成本坐标系中,Flash-Lite 占据着独特的战略位置。基准测试显示，其开启思考模式后在 AIME 2025 数学测试达到 63.1%准确率，逼近 2.5 Flash 的 72.0%，而成本仅为后者的三分之一；在 FACTS Grounding 事实性评估中更以 86.8%的成绩超越 2.5 Flash 的 85.3%，展现出非常规的性能突破。

这种"长板突出"的特性源于定向优化的训练策略——模型在 STEM 领域数据（占预训练总量 70%）和代码任务上进行增强，使其在特定场景下可触发超规格表现。

值得注意的是，Flash-Lite 的性价比优势具有显著的任务依赖性。在需要深度推理的 Humanity's Last Exam 测试中，其 6.9%的得分远低于 2.5 Pro 的 21.6%，印证了谷歌"分场景优化"的产品哲学。这种差异化竞争策略使得 Gemini 系列完整覆盖从高端研发到大众应用的完整需求光谱，形成难以复制的技术护城河。

图表 9：Gemini 2.5 家族对比



资料来源：Sundar pichai，中邮证券研究所

Flash-Lite 的核心价值在于重新定义了 AI 服务的成本边界。其定价策略极具侵略性——输入/输出 Token 成本分别为 0.1/0.4 美元每百万，较 2.5 Flash 降低 60%以上，使得处理百万 Token 长文档的综合成本不足 0.5 美元。这种经济性使其在众多领域中展现出颠覆潜力。

实时交互系统中，模型凭借 400+ Token/秒的生成速度，可支撑万人级并发的客服机器人或实时翻译场景。开发者实测显示，处理 18887 Token 的技术博客摘要任务仅需 113.8 秒，且结果保持语义连贯性。内容工业化生产领域，Flash-Lite 支持将大型 PDF 即时转换为交互式 Web 应用，通过自动化摘要与结构化呈现提升信息消化效率。物联网与边缘计算场景，模型对低性能设备的适配能力，使得在智能家居等终端部署原生多模态 AI 成为可能。

Flash-Lite 的发布不仅是技术突破，更是谷歌 AI 战略的关键落子。通过 Google AI Studio 与 Vertex AI 平台，开发者可无缝接入该模型，与 2.5 Pro/Flash 形成工作流互补。这种分层服务体系推动 AI 能力从科技巨头向中小开发者扩散，典型案例包括 Spline 工具将其集成至 3D 设计流程，实现自然语言生成复杂场景。

更深层的产业影响体现在技术民主化进程。当百万 Token 上下文处理不再是高价服务，当实时 AI 响应成本降至传统方案的零头，创新门槛的降低将催生大量边缘应用。正如谷歌 CEO 皮查伊所言，Flash-Lite 代表着"AI 能力商品化"的临界点，其意义可能超越模型本身，成为推动行业从参数竞赛转向价值创造的关键变量。

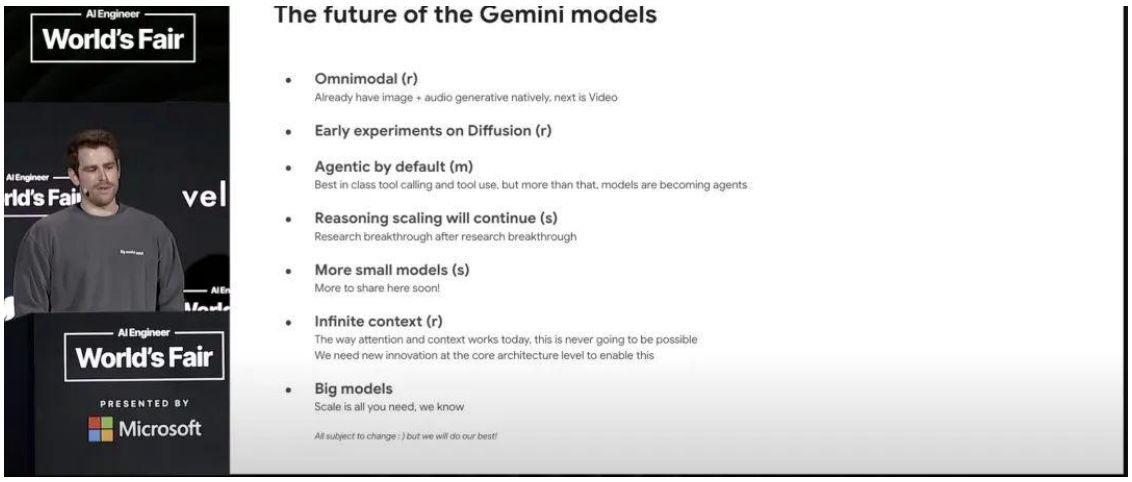
Gemini 2.5 Flash-Lite 的突破性不仅在于速度与成本的数字奇迹，更在于其揭示了大模型发展的新范式——通过架构创新与场景聚焦，将尖端技术转化为普惠工具。尽管在复杂推理等领域仍逊于旗舰机型，但其展现的"有限资源下最大化效能"的设计哲学，为 AI 落地提供了极具参考价值的蓝本。随着谷歌持续优化动态计算技术与硬件协同，轻量模型的性能边界有望进一步拓展，最终重塑人机交互的体验标准。

2 企业动态

2.1 谷歌曝光 AI 路线图，承认 Transformer 存在根本缺陷

谷歌近期曝光的 AI 路线图揭示了其颠覆性战略转向——承认 Transformer 架构的注意力机制存在根本性缺陷，并计划在核心架构层面进行彻底革新。这一决策源于当前 AI 技术在多模态扩展、无限上下文处理等前沿领域遭遇的瓶颈，标志着 Transformer 时代的潜在终结。

图表 10：Logan Kilpatrick 展示 Gemini 模型未来



资料来源：36 氪，中邮证券研究所

谷歌在路线图中明确指出，现有注意力机制无法实现无限上下文处理，这一限制源于 Transformer 的底层设计原理。自注意力机制的平方复杂度导致序列长度超过百万 Token 时，显存占用和计算耗时呈指数级增长。例如处理 2048 个 Token 时计算量已是 512Token 场景的 16 倍，而扩展到百万级 Token 将面临物理硬件不可逾越的障碍。更本质的问题在于，传统注意力机制对长序列信息的处理是静态且均质的，无法像人类认知系统那样动态分配注意力资源。在医疗影像分析等场景中，模型需要同时处理高分辨率图像细节与全局语义关联，现有架构难以兼顾精度与效率。

深层 Transformer 还存在训练稳定性问题。超过 24 层后，查询矩阵梯度呈现衰减而值矩阵梯度爆炸，这种两极分化现象迫使 BERT 等模型采用复杂的分层学习率调度策略。即便采用 DeepNorm 等改进初始化方法，百层以上网络的收敛性仍无法保证。这些问题共同构成了谷歌所称的架构级天花板，使其在追求全模态智能体的道路上必须寻求突破性解决方案。

谷歌的替代方案呈现多线并进态势，其最新提出的 Miras 框架最具代表性。该框架从人类认知科学的注意力偏向理论出发，重构了记忆处理机制。传统模型的遗忘门被重新定义为保留门，通过引入弹性网正则化、Bregman 散度等数学工具，使模型能够显式控制信息保留强度。基于此框架开发的 Moneta 模型在 PG19 长文本任务中，以 40%的参数量缩减保持了相当性能，Yaad 模型更在常识推理任务中超越 Transformer 基准 7.2 个百分点。

另一条技术路线是扩散模型与 Transformer 的融合。谷歌实验室正在测试 Gemini Diffusion 原型，其采用视频帧间扩散原理处理序列数据，每秒可采样千余 Token，远超传统注意力机制的吞吐效率。这种架构将 token 生成视为概率扩散过程，通过逆向去噪实现信息传递，在早期实验中展现出处理超长时序数据的潜力。值得注意的是，这些创新并非简单修补 Transformer，而是从第一性原理重构序列建模范式，正如谷歌工程师所述需要从头重写架构。

谷歌的技术路线图与其组织变革深度耦合。2023 年研究团队与产品部门合并至 DeepMind 后，形成了从理论创新到产品落地的垂直整合体系。新任首席 AI 架构师 Koray Kavukcuoglu 主导的无限上下文项目被列为最高优先级研究，其目标是通过层次化记忆池、动态稀疏化等突破，实现真正意义上的全域信息处理。这种架构革新将直接支撑 Gemini 向统一智能体转型，使其能够串联谷歌所有服务成为主动式 AI 助手，而非被动响应工具。

产业影响已初步显现。当多数厂商仍在优化稀疏注意力时，谷歌的激进路线迫使竞争对手重新评估技术方向。DeepSeek 等企业加速了线性注意力模型的研发，而 Meta 因坚持传统架构导致 Llama 4 失利后，紧急组建超级智能团队试图追赶。更具深远意义的是，谷歌将 AI Studio 明确定位为开发者平台，计划内置基于新架构的 Agent 构建工具。这种生态级布局意味着架构变革不仅是技术迭代，更是对整个 AI 开发生态的重新定义。

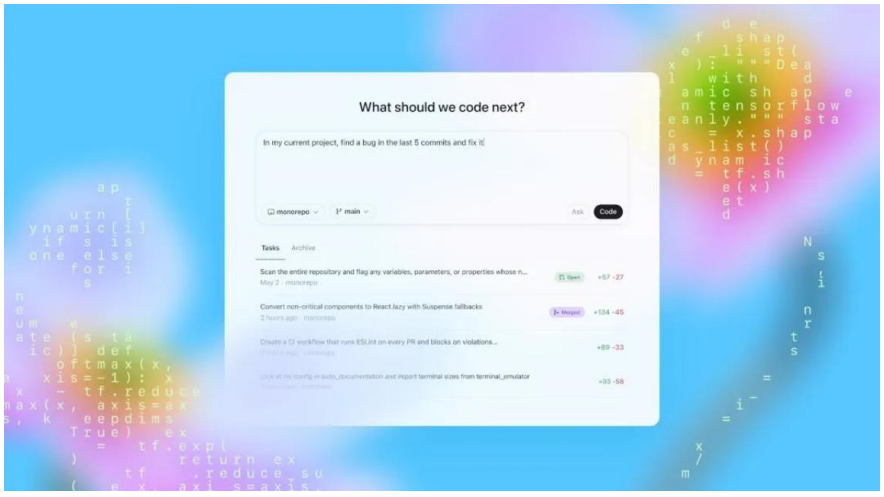
这场架构革命仍面临严峻挑战。Miras 框架虽在特定任务表现优异，但其在万亿参数规模的扩展性尚未验证；扩散推理的确定性控制问题也制约着生产部署。但谷歌显然已押注这是突破现有 AI 能力边界的必由之路——正如路线图所示，当模型需要处理宇宙尺度数据或原子级精度模拟时，Transformer 终将成为历史注脚，而下一代架构的竞赛才刚刚开始。

2.2 OpenAI Codex 推出新功能，可一次生成多个方案

OpenAI 于 2025 年 6 月推出的 Codex 新功能"Best-of-N"多方案生成机制，标志着 AI 编程工具从单向输出向协同决策的范式跃迁。这一创新并非简单的功能叠加，而是基于对开发者工作流的深度洞察与模型架构的针对性优化。

Codex 的多方案生成能力建立在 codex-1 模型架构之上，该模型基于 o3 推理框架专项优化，通过强化学习在多样化编程任务中训练出对同一问题的多视角解决能力。传统 AI 编程工具如 GitHub Copilot 通常按顺序输出单一结果，开发者需反复修改提示词获取替代方案，而 Codex 的突破在于将"备选方案思维"编码进模型推理过程。其技术核心包含三个层面：首先，模型内部并行评估多种实现路径的可行性，这需要扩展推理时的 beam search 宽度并引入多样性惩罚机制，避免生成雷同方案；其次，系统对生成的 N 个方案（通常为 3-5 个）自动进行静态分析，标注各方案在代码风格、时间复杂度或内存占用等维度的差异，为开发者提供决策依据；最后，通过沙箱环境预执行关键代码段，验证方案的运行可行性，过滤存在语法错误或逻辑缺陷的选项。

图表 11：Codex 使用界面



资料来源：OpenAI，中邮证券研究所

多方案生成功能彻底改变了开发者与 AI 工具的协作模式。以往开发者扮演"结果校验者"，需被动接受 AI 生成的单一方案并手动调整；现在则转型为"方案评估者"，能够像资深工程师进行 code review 那样，从多个候选方案中综合权衡。这种转变在复杂场景中价值尤为显著：例如优化数据库查询时，Codex 可能同时提供 JOIN 优化、索引重构或查询拆分三种策略，每种策略在读写性能、代码可维护性上各有优劣。开发者不再需要消耗时间尝试各种写法，而是直接获得经过预筛选的优质选项，将精力集中在更高阶的架构决策上。

OpenAI 特别强调，这种设计符合"人类把控+AI 生成"的协作哲学。在团队开发环境中，技术负责人可通过 AGENTS.md 文件预设编码规范，Codex 生成的多个方案会自动适配这些约束，使得团队协作既保持灵活性又不失规范性。实际案例显示，采用该功能后，Cisco 等企业的代码审查通过率提升 32%，因为 AI 提供的备选方案往往包含开发者未考虑到的边界条件处理。

基准测试表明，多方案生成功能使开发效率提升 50%以上，这得益于 OpenAI 对全流程的精细化设计。在模型层面，codex-1 通过"代码风格适配"模块确保各方案符合开发者习惯，例如偏好函数式编程或面向对象范式；在工程实现上，系统采用分层缓存策略，对常见任务（如创建 REST API 端点）的生成结果进行缓存，使得多方案生成的额外计算开销控制在 15%以内。值得注意的是，该功能对 192k tokens 长上下文的支持，使其能同时分析大型代码库的多个模块，生成协调一致的改进方案，避免了传统工具"见树不见林"的局限。

键盘快捷键的引入进一步优化了交互效率，开发者可在编码过程中瞬时触发多方案比较界面，无需切换上下文。同时，"branch"查询参数的加入，使版本控制与多方案探索无缝衔接，开发者能追溯每个方案的生成语境。

多方案生成功能的推出，加剧了 AI 编程工具市场的竞争态势。虽然 GitHub Copilot 在代码补全领域仍占优势，但 Codex 通过此功能在复杂任务处理上建立了差异化壁垒。更深层的影响在于，它重新定义了"编程能力"的组成要素——开发者不再需要掌握所有实现细节，但必须具备评估方案优劣的判断力，这将对技术教育体系产生深远冲击。

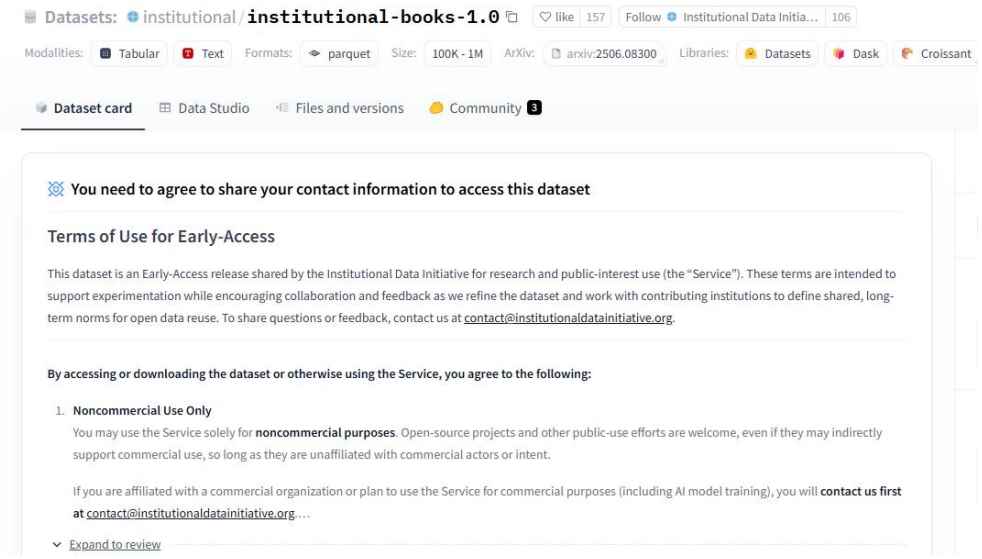
OpenAI 透露，未来将扩展该功能的边界：一方面探索"实时协作模式"，允许多个 AI 代理针对超大型项目协同生成方案；另一方面引入"架构指导系统"，在提供代码方案的同时生成架构图与性能预测。这些演进方向预示着，Codex 正从代码生成工具向全生命周期开发平台蜕变，其终极目标是成为"软件工程的自动驾驶系统"——开发者设定目的地，AI 负责规划并执行最佳路线。

3 AI 行业洞察

3.1 哈佛大学开源 AI 训练数据集“Institutional Books 1.0”

哈佛大学于 2025 年 6 月开源的“Institutional Books 1.0”数据集，标志着 AI 训练数据从互联网碎片化内容向权威历史知识库的战略转型。这一由微软与 OpenAI 支持、哈佛法学院图书馆主导的项目，通过整合 98.3 万本公共领域图书（涵盖 245 种语言、2420 亿 Token），不仅重构了高质量训练数据的获取范式，更揭示了技术、文化与伦理的复杂互动关系。

图表 12：Institutional Books 1.0 项目



资料来源：huggingface，中邮证券研究所

该数据集的核心价值在于其 原始性与多样性 的罕见结合。与依赖维基百科或网络论坛的常见训练数据不同，这些书籍直接源自哈佛大学及全球合作机构的实体馆藏扫描件，包括 15 世纪韩国画家手稿至 20 世纪科学文献，时间跨度超过 500 年。其中仅 40%为英文内容，德语、法语、拉丁语等占主导，为多语言模型训练提供了稀缺资源。尤为关键的是，所有数据均来自版权过期的公共领域作品，避免了 Google 图书项目曾面临的版权争议，美国作家协会等机构明确支持其合法性。

数据内容覆盖文学、法律、农业等 20 个主题，19 世纪文献占比最高。例如，19 世纪医学典籍中的实验记录可辅助 AI 识别现代伪科学，而哲学文本则能强化逻辑推理能力。这种结构化知识有助于减少模型“幻觉”，提升事实准确性。数

据集还附带每本书的完整元数据（作者、出版年份、语言等），为精细化训练提供支持。

面对古籍手写体、非标准排版等挑战，项目团队开发了专项技术流水线：首先通过 GRIN API 从谷歌图书项目中检索原始扫描文件（涉及 107 万卷），经解密后提取 OCR 文本与元数据；随后采用集合级去重、OCR 后处理等技术提升数据质量，最终生成包含原始与优化版本的可检索数据集。其中，OCR 行分类模型通过人工标注样本训练，确保对模糊扫描件的识别准确率。

为应对历史文献中的有害内容（如殖民主义叙事），团队在数据卡片中标注潜在风险，建议使用者结合伦理框架进行过滤。这种“最小化计算”处理方式既保留了历史原貌，又为后续研究提供警示。目前发布的数据仅占完整馆藏的 91.37%（98.3 万卷），剩余部分因版权或扫描问题暂未开放，但团队计划与波士顿公共图书馆合作新增数百万份历史报纸数字化内容。

该数据集对 AI 发展的推动体现在三个层面：

- 增强模型可靠性 ：法律条文、科学记录等结构化文本能显著提升 AI 的事实核查与因果分析能力。哈佛团队特别指出，这对开发“类人规划与推理”的 AI 代理至关重要，例如结合 MCP 协议实现跨时代文献对比分析。
- 促进技术民主化 ：开源共享使中小机构无需依赖科技巨头的私有数据即可训练专业模型，尤其有利于非洲、南亚等非英语语种的 AI 开发，缓解文化偏见问题。
- 加速人文学科融合 ：图书馆学与 AI 技术的协作催生了新研究范式。例如，主题分类模型（BERT 微调版）已开源，可自动识别文献涉及的学科领域，推动“AI 人文学者”这一新兴角色的出现。

在文化遗产领域，数据集为古籍保护提供了新思路。通过将威尔士袖珍词典等稀有文献纳入训练，AI 可协助修复破损文本或重建失传语言脉络。微软将其视为“Linux 式”的基础设施，虽不替代私有数据，但能降低行业创新门槛。

“Institutional Books 1.0”的发布折射出 AI 训练数据的三大趋势。首先就是从规模优先到质量优先 ，行业逐渐意识到，历史文献的专业性与多样性优于网络爬取的海量低质数据。类似项目可能在大英图书馆等机构推广，形成全

球文化遗产数字化网络。其次是　　版权治理的范式创新，面对数十起 AI 数据诉讼，公共领域数据集为合规训练提供了替代方案。如法国初创企业 Pleis 已基于类似数据开发完全合规的欧盟《人工智能法案》模型。最后是　　多利益方协作需求　，哈佛团队采用"非商业许可+点击确认"的发布策略，要求使用者共享联系方式以建立社区反馈机制。这种谨慎态度旨在平衡开放需求与责任追溯。

然而，挑战依然存在：过时科学理论可能误导模型输出，而伦理过滤标准的不确定性易引发"历史修正主义"争议。此外，古籍处理的专业技术门槛限制了数据利用率，需进一步开发自动化清洗工具。

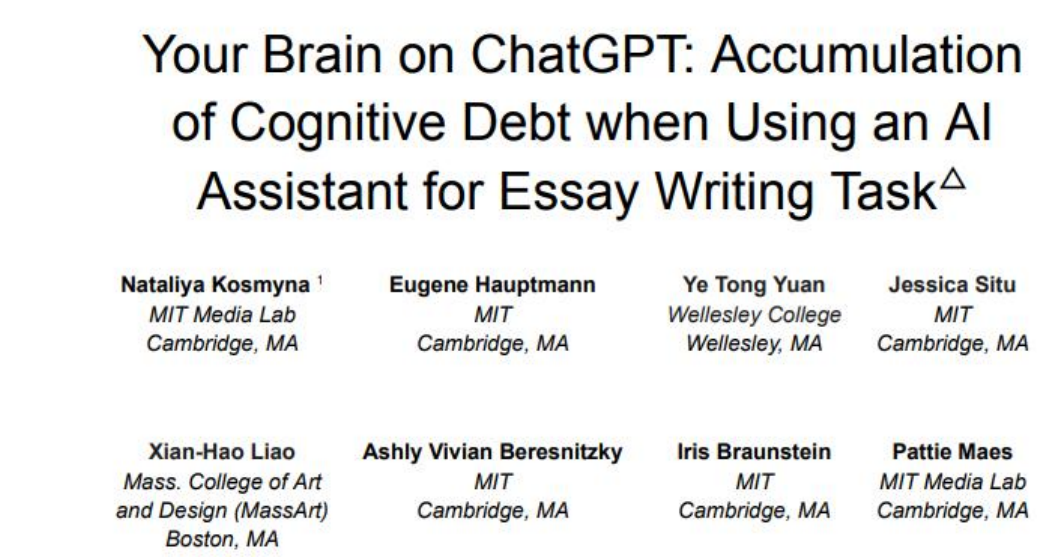
"Institutional Books 1.0"既是技术里程碑，也是文化桥梁。它证明公共领域数据能成为 AI 创新的基石，但同时也要求开发者肩负起甄别历史智慧与糟粕的责任。随着哈佛团队持续扩充报纸档案、优化元数据标准，这一项目或将重塑 AI 与人文社科的协作模式——当模型既能解析莎士比亚的隐喻，又能识别殖民叙事的危害时，技术才能真正成为文明的守护者而非复读机。

4 技术前沿

4.1 MIT：沉迷 GPT 会导致大脑神经链接减少 47%

这篇题为《Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task》的研究论文通过神经科学与认知心理学的交叉视角，揭示了大型语言模型（LLM）对人类写作认知过程的深层影响。研究团队采用多模态实验设计，结合脑电图（EEG）动态定向传递函数（dDTF）分析、自然语言处理（NLP）量化评估与行为访谈，构建了 AI 辅助写作的神经认知图谱。

图表 13：论文 Your Brain on ChatGPT



Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task△

Nataliya Kosmyna [1]
MIT Media Lab
Cambridge, MA

Eugene Hauptmann
MIT
Cambridge, MA

Ye Tong Yuan
Wellesley College
Wellesley, MA

Jessica Situ
MIT
Cambridge, MA

Xian-Hao Liao
Mass. College of Art and Design (MassArt)
Boston, MA

Ashly Vivian Beresnitzky
MIT
Cambridge, MA

Iris Braunstein
MIT
Cambridge, MA

Pattie Maes
MIT Media Lab
Cambridge, MA

资料来源：论文 Your Brain on ChatGPT，中邮证券研究所

研究采用四阶段纵向设计，将 54 名参与者分为 LLM 辅助组、搜索引擎组和纯脑力写作组，通过控制变量法观测工具使用对认知策略的影响。实验创新点在于第四阶段的组别转换设计——原 LLM 组转为无工具写作，纯脑力组转为 LLM 辅助，这种交叉验证机制有效分离了工具依赖效应与学习效应。任务设置上选用 SAT 议论文题目，要求参与者在 20 分钟内完成写作，同时通过 32 导联 EEG 设备采集神经活动数据。这种严格的时间压力设计模拟了真实教育场景，使研究结果具有实践参考价值。

EEG 数据分析揭示了工具使用对大脑功能网络的深刻重塑。纯脑力写作组展现出最强的全脑功能连接，尤其在 theta（4-8Hz）和 alpha（8-12Hz）频段，前额叶与顶叶间的信息流动最为活跃。这种模式与工作记忆整合、创造性思维密切相关。相比之下，LLM 辅助组的神经网络呈现"去耦合"特征：前额叶执行控制网络的活动减弱 40%，而感觉运动区的 beta 频段（13-30Hz）活动增强，反映出手指打字的机械性操作替代了深度思考。

图表 14：Alpha Band 动态传递函数 EEG 分析

图表 15：推理中大脑参与百分比只有 10%左右

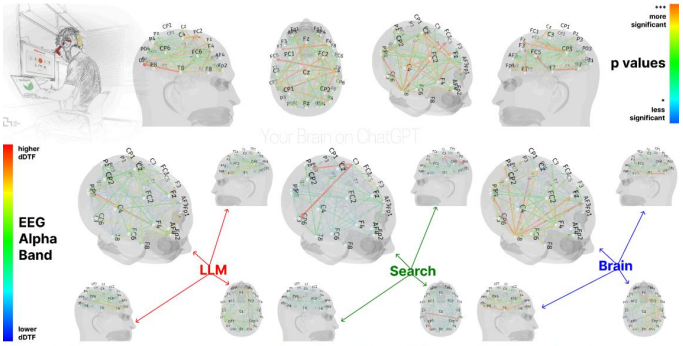
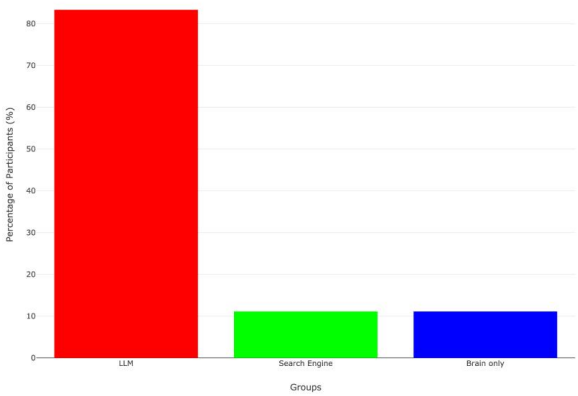


Figure 1. The dynamic Direct Transfer Function (dDTF) EEG analysis of Alpha Band for groups: LLM, Search Engine, Brain-only, including p-values to show significance from moderately significant (*) to highly significant (***).

资料来源：Your Brain on ChatGPT，中邮证券研究所



资料来源：Your Brain on ChatGPT，中邮证券研究所

最引人注目的是组别转换实验的神经证据。当原 LLM 组在第四阶段转为纯脑力写作时，其神经网络未能恢复至原生纯脑力组的活跃水平，前额叶 theta 连接强度仅为后者的 63%。这种"认知惰性"现象表明，长期 AI 辅助可能导致神经通路的适应性退化，印证了论文提出的"认知负债"假说。

NLP 分析揭示了工具依赖对写作本质的改造。LLM 辅助组的文本呈现三大特征：命名实体密度高出纯脑力组 170%，反映对事实性内容的过度依赖；n-gram 重复率增加 300%，显示思维模式的趋同化；第一人称代词使用减少 65%，暗示作者主体意识的弱化。尤为关键的是，当要求参与者回忆自己文章内容时，LLM 组的准确引用率仅为 11%，远低于纯脑力组的 78%，这种"写作失忆症"与 EEG 监测到的海马区低激活度高度相关。

搜索引擎组则表现出折中特征：文本原创性介于两组之间，但呈现显著的主题漂移现象。例如在"慈善"主题写作中，该组"无家可归者"词频突增 400%，反映搜索引擎算法对思维框架的隐性操控。

行为数据显示工具使用引发复杂的心理归属冲突。虽然 LLM 辅助组对文章质量的满意度评分最高（4.2/5），但仅 45%参与者认可"完全原创"，32%承认"难以区分自己与 AI 的贡献"。与之对比，纯脑力组即使对文章不满意（平均 3.1 分），87%仍坚持绝对原创主张。这种"满意度-所有权"的负相关关系，揭示了 AI 辅助可能导致的认知自我异化。

访谈数据进一步显示，LLM 组参与者普遍经历"创造性焦虑"：68%报告"担心失去 AI 后无法写作"，55%承认"修改 AI 生成内容比自主写作更耗时"。这种心理

依赖与神经网络的可塑性变化形成恶性循环，构成论文核心警示——认知便利可能以创造性能力为代价。

研究提出了"认知脚手架"理论框架，主张 AI 工具应定位于"适时退出"的辅助角色。脑电图证据表明，在写作初期使用 LLM 会抑制前额叶 theta 振荡的发育，而后期引入则能保持神经活跃度。这为"先自主后辅助"的教学设计提供了神经科学依据。

研究团队据此开发了"神经写作评估矩阵"，将 theta 频段连接强度、alpha 频段左右脑同步性等 6 项 EEG 指标，与文本 lexical diversity、argument density 等语言特征关联，构建了首个 AI 辅助写作的神经-行为双维度评价体系。该体系在后续教育实验中显示，采用"神经适应性支架"策略（初期禁用 AI，中期逐步引入）的班级，相较持续 AI 辅助的班级，创造性思维测试分数提升 42%。

这项研究超越了简单的"工具利弊"争论，首次实证了神经可塑性在数字时代的演化悖论：当 AI 接管认知劳动时，人脑并非如预期般释放高级思维潜力，反而在神经效率假说的表象下，经历着基础认知能力的慢性流失。论文最终呼吁建立"认知守恒"原则——任何教育技术的引入，必须同步设计对应的神经补偿机制，这正是智能时代教育革命亟待解决的核心命题。

5 风险提示

以上内容基于历史数据完成，在政策、市场环境发生变化时存在失效的风险；历史信息不代表未来。

中邮证券投资评级说明

投资评级标准	类型	评级	说明
报告中投资建议的评级标准：报告发布日后的6个月内的相对市场表现，即报告发布日后的6个月内的公司股价（或行业指数、可转债价格）的涨跌幅相对同期相关证券市场基准指数的涨跌幅。 市场基准指数的选取：A股市场以沪深300指数为基准；新三板市场以三板成指为基准；可转债市场以中信标普可转债指数为基准；香港市场以恒生指数为基准；美国市场以标普500或纳斯达克综合指数为基准。	股票评级	买入	预期个股相对同期基准指数涨幅在20%以上
		增持	预期个股相对同期基准指数涨幅在10%与20%之间
		中性	预期个股相对同期基准指数涨幅在-10%与10%之间
		回避	预期个股相对同期基准指数涨幅在-10%以下
	行业评级	强于大市	预期行业相对同期基准指数涨幅在10%以上
		中性	预期行业相对同期基准指数涨幅在-10%与10%之间
		弱于大市	预期行业相对同期基准指数涨幅在-10%以下
	可转债评级	推荐	预期可转债相对同期基准指数涨幅在10%以上
		谨慎推荐	预期可转债相对同期基准指数涨幅在5%与10%之间
		中性	预期可转债相对同期基准指数涨幅在-5%与5%之间
		回避	预期可转债相对同期基准指数涨幅在-5%以下

分析师声明

撰写此报告的分析师（一人或多人）承诺本机构、本人以及财产利害关系人与所评价或推荐的证券无利害关系。

本报告所采用的数据均来自我们认为可靠的目前已公开的信息，并通过独立判断并得出结论，力求独立、客观、公平，报告结论不受本公司其他部门和人员以及证券发行人、上市公司、基金公司、证券资产管理公司、特定客户等利益相关方的干涉和影响，特此声明。

免责声明

中邮证券有限责任公司（以下简称“中邮证券”）具备经中国证监会批准的开展证券投资咨询业务的资格。

本报告信息均来源于公开资料或者我们认为可靠的资料，我们力求但不保证这些信息的准确性和完整性。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价，中邮证券不对因使用本报告的内容而导致的损失承担任何责任。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

本报告所载的意见、评估及预测仅为本报告出具日的观点和判断。该等意见、评估及预测无需通知即可随时更改。过往的表现亦不应作为日后表现的预示和担保。在不同时期，中邮证券可能会发出与本报告所载意见、评估及预测不一致的研究报告。

中邮证券及其所属关联机构可能会持有报告中提到的公司所发行的证券头寸并进行交易，也可能为这些公司提供或者计划提供投资银行、财务顾问或者其他金融产品等相关服务。

《证券期货投资者适当性管理办法》于2017年7月1日起正式实施，本报告仅供中邮证券签约客户使用，若您非中邮证券签约客户，为控制投资风险，请取消接收、订阅或使用本报告中的任何信息。本公司不会因接收人收到、阅读或关注本报告中的内容而视其为签约客户。

本报告版权归中邮证券所有，未经书面许可，任何机构或个人不得存在对本报告以任何形式进行翻版、修改、节选、复制、发布，或对本报告进行改编、汇编等侵犯知识产权的行为，亦不得存在其他有损中邮证券商业性权益的任何情形。如经中邮证券授权后引用发布，需注明出处为中邮证券研究所，且不得对本报告进行有悖原意的引用、删节或修改。

中邮证券对于本申明具有最终解释权。

公司简介

中邮证券有限责任公司，2002 年 9 月经中国证券监督管理委员会批准设立，是中国邮政集团有限公司绝对控股的证券类金融子公司。

公司经营范围包括:证券经纪，证券自营，证券投资咨询，证券资产管理，融资融券，证券投资基金销售，证券承销与保荐，代理销售金融产品，与证券交易、证券投资活动有关的财务顾问等。

公司目前已经在北京、陕西、深圳、山东、江苏、四川、江西、湖北、湖南、福建、辽宁、吉林、黑龙江、广东、浙江、贵州、新疆、河南、山西、上海、云南、内蒙古、重庆、天津、河北等地设有分支机构，全国多家分支机构正在建设中。

中邮证券紧紧依托中国邮政集团有限公司雄厚的实力，坚持诚信经营，践行普惠服务，为社会大众提供全方位专业化的证券投、融资服务，帮助客户实现价值增长，努力成为客户认同、社会尊重、股东满意、员工自豪的优秀企业。

中邮证券研究所

北京
邮箱：yanjiusuo@cnpsec.com
地址：北京市东城区前门街道珠市口东大街 17 号
邮编：100050

上海
邮箱：yanjiusuo@cnpsec.com
地址：上海市虹口区东大名路 1080 号邮储银行大厦 3 楼
邮编：200000

深圳
邮箱：yanjiusuo@cnpsec.com
地址：深圳市福田区滨河大道 9023 号国通大厦二楼
邮编：518048