Week 07 · LLM 的记忆问题「很快」就不再是问题了？

2026.02.15

> 本周，智谱发布基础模型 GLM-5；Anthropic 完成 300 亿美元 G 轮融资，投后估值达 3800 亿美元.....

要闻解读

① LLM 的记忆问题「很快」就不再是问题了？

引言：当前，智能体正经历范式转变，从高效的单任务执行模式，逐步向动态环境下的持续自适应、能力演化与经验积累模式转型。在此背景下，AI Memory 作为核心基石，赋能智能体保持行为一致性、做出理性决策并实现高效协作。在长期探索中，AI Memory 已经分化为「Agent Memory」与「LLM Memory」两条截然不同的演进路径。

OpenClaw 的「长效记忆」为何不代表「AI 拥有持久记忆」？

1. 开源项目 OpenClaw（原名 Clawdbot 及 Moltbot）在 2026 年初引起了一阵病毒式流行。在相关讨论中，在相关讨论中，OpenClaw 的核心竞争力在于它被视为「有手的 Claude」，能够跨越数周乃至数月的会话维持持久性记忆，通过学习用户的偏好、语气及特定工作流，使 AI 从「对话窗口」转变为真正「懂你」的数字雇员。

   ① 自 2025 年底发布以来，OpenClaw 项目在 GitHub 上的热度极速攀升，在 2026 年 2 月已突破 19 万颗星。[1-1]

2. 在此热度下，AI 社区对 OpenClaw 项目热议的核心不仅在于能够执行复杂的跨平台操作，更在于其展示出的「长效记忆」能力是否代表「AI 拥有持久记忆」的未来是否即将来临。

① 伴随 LLM 及 Agent 应用的突破和广泛应用，AI 的记忆（Memory）问题被视为推动更高阶智能演进的核心瓶颈，围绕改善 AI 记忆力的研究已成为 LLM 相关研究中最受关注的前沿方向之一。

② 在 2025 年先后涌现了诸多有关改善 AI Memory 的探索，如 Meta 团队的「SMF」工作、谷歌提出的「Nested Learning」范式及 HOPE 模型、MIT 的「BEYOND CONTEXT LIMITS」工作等（详见 Pro 会员通讯 2025 Week 46 &Week 50）

③ 学术界 AI Memory 的关注同样在不断提升。以顶会的研究主题为例，ICLR 2026 专门设立了「MemAgents」研讨会，旨在为智能体构建一个能够支持单样本学习和长程一致性的底层记忆基底。[1-11]

3. 伴随学术界对 Memory 架构和机制的持续探索，以及 OpenClaw 在工程上的「巨大成功」，长期以来被探讨的 AI 记忆问题开始形成边界，并分化「LLM Memory」与「Agent Memory」两条演进路径。[1-4]

① LLM Memory 构成了预测的底层计算机制，存在两种具体形态：嵌入在预训练模型权重中的参数化记忆，以及通过上下文窗口管理的运行时记忆。LLM Memory 是基础的计算载体，其优先级在于在有限的窗口内保证即时生成的准确性，而非维持连贯的自主行为。

② Agent Memory 则在前者的基础上延伸为系统性支撑自主行为的功能流程。它不再仅生成孤立的文本，而是协调感知、规划、行动的循环过程，使系统能够拆解并执行复杂任务。

4. 在 Agent 领域，或应用相对垂直的「Vertical Agent」语境下，记忆（Agent Memory）不再是科学难题，而是一个可以通过场景拆解、针对性构建得到解决的工程问题。[1-2] [1-3]

① 通过将数据组织为过程性、陈述性、元认知等不同格式，Agent Memory 让系统能够从历史经验中学习，这一层级通过实现反思和策略优化，推动数据从静态记录向动态 「经验」 转变，使 Agent 能够基于过往结果演化自身行为。[1-4]

5. 相对于 Agent Memory 的繁荣，LLM Memory）仍存在诸多挑战，如「稳定性-塑性困境」（Stability-Plasticity Dilemma）。当尝试通过微调将新信息注入参数时，模型往往会丢失旧的、同样重要的知识 。[1-4]

AI Memory 的研究视角在如何变化？

1. AI Memory 的核心价值不止于缓解大语言模型的上下文窗口有限、交互无状态等技术瓶颈，更被视为一种变革性赋能工具，推动人工智能系统从通用工具升级为具备自适应、协作能力的以人为中心的智能体。[1-4]

2. 伴随对 AI Memory 的持续探索，研究者开始从多样的视角审视这一 AI 能力的基石，并对其理论依据，运作机理及边界深入探索与迭代。

   ① 2025 年 4 月，华为诺亚方舟实验室的「From Human Memory to AI Memory」从人类认知科学中的记忆理论出发，为理解 LLM Agent 的记忆机制提供了一个类比框架。[1-5]

   ② 2025 年 12 月，哈工大、鹏城实验室和复旦等机构的「AI Meets Brain」则将人脑记忆机制与 Agents Memory 统一审视，并对其存储结构展开剖析。[1-6]

3. 2026 年初，北邮百家 MemoryOS 团队与华为的研究者发表了新的综述「Survey on AI Memory」，整合认知心理学与神经科学模型，明确了 AI 领域中 「Memory」 的概念边界。

   ① 该综述将「Memory」的概念边界分为三层，LLM Memory 主要强调计算核心，提供预测引擎；Agent Memory 聚焦于功能流程，负责管理面向任务的执行过程。

   ③ 在 LLM 与 Agent 之上，「AI Memory」是最宏观的认知概念，涵盖了人工认知的生物学灵感与终身学习的终极目标，聚焦于终身演化与经验积累，统领前两者。

4. 在明确 AI Memory 层级边界后，研究者进一步提出了 4W 记忆分类体系，以 When（生命周期）、What（类型）、How（存储）、Which（模态） 四个正交核心维度对 AI 记忆系统进行分类，每个维度对应

AI 记忆的一项基础特征，同时结合认知科学、计算机工程等领域理论，对各维度下的细分类型展开了详细界定与阐释。

① When（生命周期维度）聚焦记忆的存在时间与存续时长，从瞬时到跨会话持久存储，对应记忆在 AI 智能体系统中的时间跨度特征；

② What（类型维度）基于认知科学理论，按记忆捕获的信息类型 / 知识性质划分，反映记忆在智能体行为中的功能角色；

③ How（存储维度）探究记忆的表示形式与存储技术实现方式，区分模型内部的隐式存储与外部的显式存储；

④ Which（模态维度）按记忆处理的信息格式 / 模态划分，分为单模态与多模态，体现 AI 记忆对不同类型信息的处理能力。

5. 此外，北邮和华为的研究者梳理了单智能体-多智能体记忆架构以及记忆评估方法，并拆解了当前 AI 记忆发展所存在的架构冲突、理论方法缺口、安全与运维复杂性三大核心瓶颈。

① 「架构冲突与系统限制」涵盖 LLM 有限的上下文窗口与长时经验海量积累的矛盾；参数化记忆更新的计算成本高、易发生灾难性遗忘的问题；多模态记忆的异构信息融合与统一表征难题。

② 「理论与方法学缺口」强调当前对记忆的研究偏重时间维度，忽视对象与存储形式维度，概念碎片化；评估体系缺乏高阶能力指标（如泛化性、稳健性）；多智能体记忆共享缺乏成熟的信息分区、同步与一致性理论。

③ 「安全风险与运营复杂性」涉及不同场景。单智能体场景中，用户数据存储需平衡个性化与隐私保护，易存在敏感信息推理风险；多智能体协作场景中，静态权限设计无法适配复杂环境，易引发效率瓶颈与数据不一致。

表：AI 记忆的 4W 分类法与技术体系。[1-4]

近期工作在如何探索 LLM Memory 和 Agent Memory？

1. 伴随业界 AI Memory 的重视，以及对推动 LLM 向持续学习与自主智能演进的目标，2026 年初，研究者对 LLM Memory 的探索聚焦模型原生上下文与知识存储机制，Agent Memory 的相关工作则面向多轮交互、长期个性化与多智能体协作中的记忆管理。

2. DeepSeek 与北大的研究者在年初发布了「Engram」工作，提出了「条件记忆」（conditional memory）的稀疏性维度，以补充现有 MoE 模型的条件计算范式，以解决 Transformer 架构缺乏原生知识查找原语、需通过低效计算模拟检索的问题。[1-7]

   ① 该工作设计了「Engram」可扩展查找模块，借鉴经典 N-gram 嵌入思想，实现 O（1） 时间复杂度的知识查找，让模型能基于输入局部模式快速调用静态知识，无需反复通过深层计算重建高频、模板化信息。

   ② 经测试，在同等参数与 FLOPs 条件下，27B 规模的 Engram 模型在知识类任务（MMLU+3.4. CMMLU+4.0）、通用推理（BBH+5.0、ARC-Challenge+3.7）及代码数学领域（HumanEval+3.0、MATH+2.4）均显著超越 MoE 模型；长上下文任务中，Multi-Query NIAH 准确率从 84.2% 提升至 97.0%。

   ③ 机制分析表明，Engram 将浅层网络从静态重建任务中解放，让深层专注复杂推理，等效增加模型有效深度，同时其确定性检索特性支持存算解耦，记忆表可部署于 CPU/SSD，大幅降低 GPU 显存依赖。

3. 针对 LLM Memory 的能力评估问题，盛大集团陈天桥团队近期提出了提出了用于评估 LLM 长期交互式记忆能力的「EverMemBench」基准测试，以解决现有基准测试多聚焦二元、单主题对话，难以捕捉现实场景复杂性的问题。[1-8]

   ① EverMemBench 将记忆能力分解为三个核心评估维度，分别是细粒度回忆、记忆意识和用户画像理解。细粒度回忆侧重对具体事实的精准提取；记忆意识用于测试模型能否识别历史信息与新场景的关联性并合理运用；用户画像理解关注模型从长期对话中挖掘用户隐含习惯和特质的能力

② 研究者基于 EverMemBench 评估了多种长上下文语言模型（包括 Gemini-3-Flash、GPT-4.1-mini、LLaMA-4-Scout 等）和记忆增强系统（如 Zep、Mem0、EverMemOS 等）的性能，并指出当前 LLM 的记忆缺陷。

③ 实验结果表明，多跳推理在多参与方场景下性能急剧下降，表现最佳的 Gemini-3-Flash 准确率仅为 26.51%，核心原因是多参与方场景下相关信息分散在不同发言者、群组和时间点；时序推理仍是未解决的难题，现有模型难以处理信息的版本语义和演变逻辑，最优模型准确率不足 50%。

④ 此外，研究发现记忆意识受限于检索质量，现有基于相似度的检索方法无法弥合查询与隐含相关记忆间的语义鸿沟，导致记忆增强系统性能显著低于全上下文模型。

4. Agent Memory 方面，广东智能科学与技术研究院、东京科技大学和香港理工的研究者近期参考大脑认知机制设计了多智能体记忆框架「BMAM」，通过模拟海马体 - 新皮层的双记忆系统，结合前额叶皮层的任务协调逻辑，解决多 Agent 协作中的记忆管理、时序一致性与长程推理问题。[1-9]

① BMAM 框架采用模块化设计，包含海马体启发的情景记忆模块、新皮层启发的语义记忆模块，以及负责记忆整合与检索的协调模块，实现工作记忆与长期记忆的分层管理、并行更新与精准检索。

② 该工作提出了 「灵魂侵蚀（soul erosion）」 诊断视角，其定义为时间连贯性、语义一致性和身份保全三个维度的综合退化，将 Agent 记忆失效模式与架构设计关联。

③ BMAM 针对 「灵魂侵蚀」每种模式设计了专用解决方案，如通过 StoryArc 时间线索引解决时间侵蚀，借助海马体到颞叶的整合机制稳定事实以应对语义侵蚀，利用杏仁核启发的显著性标记保护身份相关信息，避免被临时上下文覆盖。

④ 经测试， BMAM 在 LoCoMo 基准上准确率达 78.45%，跨会话任务集成准确率 52.6%，优于现有多 Agent 记忆系统。消融实验则证

实了海马体式情景记忆是核心性能驱动因素，移除后准确率下降24.62%，其他子系统则针对特定失效模式提供补充支持。

5. 人民大学信息学院，MemTensor 和上海高级算法研究院等机构的研究者在近期工作中探究了长期个性化对话系统中记忆噪声积累、人格不一致、无限对话与有限上下文冲突的问题，提出「Inside Out 」框架，以显式结构化的 PersonaTree 作为长期记忆核心，实现无界交互下的用户个性化状态维护。 [1-10]

   ① 「Inside Out 」框架借鉴认知科学中记忆的功能分工特性，将记忆分解为情景记忆、语义记忆、显著性感知和控制导向四大功能专用子系统，而非单一无结构存储，各子系统在不同时间尺度上协同工作。

   ② 该工作基于生物心理社会模型构建分层记忆 schema，设计迭代式树更新机制，同时训练轻量级 MemListener 模型，通过过程奖励强化学习将非结构化对话流转化为 ADD、UPDATE、DELETE 等可执行树操作，实现记忆的动态演化与精准维护。

   ③ 推理阶段采用自适应生成 pipeline，快速模式直接用 PersonaTree 保障低延迟响应，代理模式则在树结构约束下按需引入细节，平衡效率与个性化深度。

   ④ 实验验证，PersonaTree 在抑制上下文噪声、保持人格一致性上全面优于全文连接与现有个性化记忆系统，且小模型 MemListener 即可实现与大模型相当的记忆操作性能，开创 「小模型维护记忆、大模型负责生成」 的高效协同范式，为长期个性化对话 Agent 提供了结构化、可解释的记忆解决方案。

② Anthropic 如何用「经济原语」衡量 AI 影响？

引言：头部 AI 公司 Anthropic 从 2025 年 2 月 10 日启动了《Anthropic Economic Index》经济指数系列研究，通过对 Claude 的使用数据进行分析，以观察 AI 经济带来的影响。在 2026 年初，Anthropic 研究组该系列的第四篇报告中分析了 Claude 的后台匿名对话记录、匿名调用日志等数据，从 AI 能力跃迁、价值重构、发展失衡、职能演变等维度展开实证分析，指出

AI 已经开始从技术工具升级为能够影响经济结构、任务分工与整体生产力的驱动因素。

Anthropic 发现了哪些 AI 对经济的影响？

1. Anthropic 于 2026 年 1 月中旬发布经济指数系列第四篇报告，基于 Claude 平台数据，采用大规模数据采样、隐私保护分析方法，分析其地理差异、任务特征及生产力与劳动力市场影响，并提出「经济原语」框架，为 AI 对经济的影响提供可量化视角。

   ① 聚焦 AI 对经济的影响，报告数据围绕 2025 年 11 月（Opus 4.5 发布前）的 Claude 使用数据展开，具体基于 100 万条 Claude.ai 对话和 100 万条 1P API 记录，覆盖全球国家和美国各州的使用细分，呈现 AI 使用的现状特征，也预测未来扩散趋势与经济影响。

2. 该报告以此前发布的第三份经济指数报告为基础，采用新周期数据开展同类分析，从 Claude 用户在使用行为、地域分布和任务类型等维度展开分析，发现任务使用仍高度集中、增强型使用量超过自动化重新主导 Claude.ai、全球不均衡的一系列关键趋势。

   ① 此前，2025 年 9 月的第三篇经济指数报告以「Uneven geographic and enterprise AI adoption」为主题，从演变趋势、地域差异与企业应用等维度讨论了 AI 的领域高度集中、地域分布不均以及企业端深度自动化的特征。

   ② 该次报告在任务使用方面，Claude.ai 的前 10 大常见任务占总使用量的 24%，较上一份报告微升，API 端达从 28%提升至 32%。修改软件纠错等计算机与数学类任务仍是核心，占 Claude.ai 使用量的 34%、1P API 的 46%。

   ③ 2025 年 11 月，Claude.ai 的增强型使用（用户学习、迭代任务、获取反馈）占比升至 52%，自动化使用（少交互完成任务）降至 45%，扭转了 8 月自动化占优的局面，但 API 端仍以自动化使用为主。

④ 从美国各州看，AI 使用量扩散速度极快，全球层面，使用量与人均 GDP 强相关（人均 GDP 每增 1%，Claude 使用量增 0.7%），高低收入国家的使用差距无缩小或扩大迹象。

3. 在第四篇报告中，Anthropic 研究组新提出的「经济原语」指标，从「任务复杂性」、「人机技能匹配」、「使用场景」、「AI 自主性」、「任务成功率」五个维度分析观察，发现 AI 应用场景和采用效率存在明显的分化特征与地域差异，并与任务复杂度、人机技能水平高度相关。

   ① 「经济原语」是在第三篇报告的研究基础上进一步拓展提出的新量化指标，依托这套更细化的分析维度，研究者得以对 AI 的地域使用分布、普及情况及应用效率展开更深度的实证分析。

   ② 研究者发现，Claude 的使用随收入水平呈现分化，任务复杂度越高成功率越低，人机教育水平高度绑定进一步放大了 AI 收益不均，同时全球 AI 使用分布不均，C 端与企业 API 的使用模式也存在显著差异。

AI 在「经济原语」视角下证如何影响全球经济？

1. Anthropic 研究组基于「经济原语」框架，提出了「用途」、「国际和国内区域」、「教育水平」、「制度环境」、「人类技能」 五个分析视角，并围绕 AI 深度嵌入全球经济这一趋势，开展了系统性、结构化的观察与分析。

2. 关于 AI 在日常场景中的「用途」，报告通过大规模统计发现 Claude 整体使用以工作为主，但由于不同地区的人均 GDP 、支付能力、网络条件等因素，Claude 的 AI 采用率和应用多样性呈现出低收入国家偏向学习提问基础、高收入国家用途丰富的差异。

   ① 在成熟市场（高收入地区），AI 采纳已跨越工具化阶段，表现出明显的普适化与生活化，从核心工作流（Work）扩展至生活管家、创意协作等个人日常（Personal）领域。

② 相比之下，新兴市场（低收入地区）的使用高度聚焦于教育（Coursework）场景。这种模式表明欠发达地区正利用 AI 弥补教育资源缺口，尚未向广泛的深层经济活动渗透。

3. 针对 AI 在不同「国际和国内区域」应用情况的观察，基于经济原语框架的数据分析发现，在国家之间、美国州层面，Claude 的使用与 GDP 、任务时间、AI 自主性、工作用途等变量在不同层级表现出的相关性不一致，无法判定各变量与 Claude 使用行为存在直接因果关系。

   ① 全球国家层面，AI 使用指数（AUI）与无 AI 辅助时的人类任务完成时间、AI 决策自主权均呈负相关；而美国州层面上述相关关系未具备统计显著性，可能受样本量限制。

   ② 美国州层面，AI 使用指数与工作用途呈正相关，但该相关关系在全球国家层面并不成立，上述变量与 AI 使用指数仅存在相关性，并非必然因果关系。

4. 为了观察各地区「教育水平」对于 AI 任务成功率的影响，从国际与美国各州的统计数据发现教育年限与 AI 的任务成功率的关联并非绝对，传统教育对 AI 使用的直接贡献可能被高估。

   ① 教育年限越长的地区，AI 任务成功率反而越低。这归因于高教育人群倾向于尝试极高复杂度、长周期的任务（如系统重构），这类任务天然较低的成功率拉低了平均分。

   ② 在同等环境下（如美国州内），高教育水平表现出对成功率的正向驱动，高技能用户能通过更优策略提升中等难度任务的交付质量。

   ③ 一旦控制 GDP 或任务难度等变量，教育的独立作用便显著消失，意味着教育是通过连接更复杂的经济活动来「间接」影响 AI 效能。

5. 而对于各地区不同的「制度环境」对 AI 应用效能的影响，报告指出 AI 竞争的下半场不仅是技术的竞争，更关乎制度设计与产业生态的适配程度。唯有完善的制度与产业环境，合理界定 AI 的自主决策边界，才能通过赋予 AI 「自主性」来释放其核心价值。

① 在国家层面，较高的 AI 采用率通常依赖于高频的人类干预，而非技术本身的独立突破，宏观领先不代表技术自主性强。

② 在微观层面，由于 AI 应用更聚焦于具有地方特色和高专业壁垒的垂直领域，经济密度（如人才聚集、产业专业化）才是支撑 AI 迈向高自主性的实质性底座。

③ 资本密集地区不仅 AI 采用度更高，其应用也更具深度，资源缺失是低收入地区采纳 AI 的核心阻碍。

6. 此外，为探究「人类技能」与 AI 性能的关系，报告从高阶认知、协作统筹与问题建模等维度展开分析，发现 AI 输出质量高度依赖用户交互逻辑，且提示技能（Prompting）存在分布不均。而人类提示深度与 AI 回应深度的强相关性，会进一步导致不同区域的 AI 应用价值的差距。

① 在国家层面二者相关系数达 0.925，美国州层面为 0.928，表明 Claude 呈现高度动态响应特征，其表达与专业深度并非固定，仅在用户输入高复杂度提示时才会激活相应高水平能力。

② 与语言风格相对固定的模型不同，Claude 的回应复杂度直接由用户提示方式决定，用户输入的具体性、逻辑清晰度与背景完整性，直接影响模型对复杂指令的理解精度与最终输出的专业深度。

③ 高教育水平地区的用户往往能展现出更强的「提示语工程」技能，通过更具结构化的提问引导 AI 产生高质量反馈，最终在宏观上体现为不同地理区域间 AI 实际产出价值的失衡。

Claude 用户如何映射 AI 对劳动力市场的影响？

1. 该报告最后从时间节约、有效覆盖率、任务成功率与自主性等方面，分析其在不同任务类型中的动态特征，并从「杠杆失衡」、「价值外溢」、「职能置换」与「瓶颈制约」四个特征维度，探讨 AI 如何重塑未来生产力格局。研究发现，人工智能对经济的影响并非统一，并在生产力结构、产业分工与区域发展等层面形成差异化影响。

2. 通过对比 Claude.ai 交互数据，研究发现 AI 使用存在「杠杆失衡」特征。任务复杂度越高，AI 加速效应越显著，但成功率随之下降，呈

现收益与风险失衡的非对称特征，进而拉大不同教育水平群体间的能力差距。

① 复杂任务中 AI 虽报错概率更高，但节省工时更多，扣除人工复核成本后，净生产力增益仍显著高于简单任务，成为高增益生产力杠杆。这使 AI 成为一种即便不完美也可使用的「高增益杠杆」。

② 高教育、高技能群体更擅长驾驭复杂任务，能够通过高质量提示、纠错与复核，将 AI 的不完美转化为高收益；而低教育群体多局限于简单任务，AI 带来的效率提升相对有限。

3. 通过对有效 AI 覆盖与职业渗透的分析，研究发现 AI 应用存在「价值外溢」特征。AI 优先作用于数据录入、放射诊断等职业中耗时最长、占比最高的核心任务，即便仅改造单一关键环节，也能显著提升整体工作效率，形成局部突破带动全局增益的外溢效应。

   ① 这类职业的 AI 任务覆盖比例并不高，但集中在占用工时间最大的核心环节，且成功率表现突出，仅通过单点替代即可实现整体效率大幅提升。

   ② 与全面改造工作流程不同，AI 无需覆盖全部任务，只需在最耗时、最核心的环节实现有效替代，就能产生显著的整体产出增益。

4. 随着高认知子任务被 AI 剥离，「职能置换」的特征日益显现。岗位内部任务结构持续重构，进而呈现去技能化与技能分化并存的特征，个体核心价值从专业执行转向对 AI 的统筹与驱动能力。

   ① AI 优先替代高教育要求的专业任务，使得部分岗位剩余工作偏向基础操作与复核，出现去技能化趋势，拉低了部分中端知识型岗位的技能门槛。

   ② 部分岗位因 AI 承担常规任务而转向更高阶的决策与协调工作，另一部分则聚焦低技能事务，职业技能结构呈现两极分化，最终重塑岗位的核心价值与能力要求。

5. 结合 AI 生产率影响与任务互补性分析，AI 应用呈现出「瓶颈制约」的特征。AI 自身效率提升无法单独带动整体生产率增长，其实际价值

空间受限于工作流中无法被替代的人工环节，致使技术优势难以充分转化为经济收益。

① 真实的生产率增长受限于工作流中的残余瓶颈，这与「O-Ring 理论」中流程效率由最慢环节决定的逻辑一致；若忽视人工验证成本、物理审批、真人谈判等非 AI 环节的限制，会显著高估 AI 对宏观生产率的实际贡献。

② AI 技术优势向经济价值的转化，关键取决于能否通过流程重构消除这些残余瓶颈，实现工作流各环节的协同高效，而非单纯依赖 AI 自身的速度提升。

参考信源

1. Anthropic Economic Index report: Uneven geographic and enterprise AI adoption - Anthropic

2. Anthropic Economic Index report: Economic primitives - Anthropic

速递

本周值得关注技术要事

MIT 和哈佛大学提出生成建模范式「漂移模型」

日期： 2 月 8 日

事件：MIT（含何恺明）与哈佛大学合作团队近日发布论文，提出一种生成模型范式 —— 漂移模型（Drifting Model），该范式以最小化生成样本漂移为核心训练目标，可让神经网络优化器直接驱动数据分布演化，从而实现高效图像生成。

要点速览

1. 在核心机制上，漂移模型摒弃了传统扩散模型的推理迭代过程。

   ① 其核心特征在于「推送」（pushforward）映射在训练过程中不断演化，由一个单次前向、非迭代的网络表示。训练过程被视为通过不断

更新网络，将先验分布（如高斯分布）映射为推送分布，并使其逐步逼近真实数据分布的演化轨迹。

2. 为驱动这一演化，该团队引入了「漂移场」（drifting field）概念以控制样本运动。

   ① 该漂移场依赖于生成分布与数据分布的差异，通过采样正样本（真实数据）和负样本（生成数据）来估计速度向量场。

   ② 训练目标最小化生成样本的漂移，这种机制使得目标分布中未覆盖的模态对当前分布产生引导作用，从而通过标准的迭代优化过程（如 SGD）推动底层分布演化，直至系统达到平衡态。

3. 实验结果显示，漂移模型在 ImageNet 256×256 上单步生成表现优异。

   ① 在潜空间生成协议下，该模型取得了 1.54 的 FID（图片质量指标），在单步生成方法中达到当前最优性能（SOTA），可与多步扩散模型竞争。

   ② 在像素空间生成协议下，模型以 1.61 的 FID 优于 GAN 及 StyleGAN-XL（FID 2.30），且在该任务的单步推理场景中，计算量仅为后者的约 1/18（87G FLOPs vs 1574G FLOPs）。

媒体报道

1. 模型「漂移」新范式，何恺明新作让生成模型无须迭代推理 - 机器之心

谷歌提出数学研究智能体「Aletheia」

日期： 2 月 12 日

事件：近期，谷歌 DeepMind 发布博客，推出了数学研究智能体「Aletheia」。该智能体由 Gemini Deep Think 驱动，可识别候选方案中的缺陷，能够承认自身无法解决的问题。

要点速览

1. 该智能体主要由 Gemini Deep Think 高级版、推理时 Scaling Law 和工具调用能力三项技术推动。

① 随着推理时计算量（inference-time compute）的增加，Gemini Deep Think 在 IMO-ProofBench 高级测试中的得分高达 90%。

② 且根据谷歌内部 FutureMath Basic 基准，即便从奥赛级别跨越到博士级练习题，Scaling Law 依然有效。

2. 在 「Aletheia」今年 1、2 月份完成的首批六篇论文中，共包括 0 人类、人类与 AI 协作和大规模半自主评估等类别。

① 零人类方面，1 月发布的论文「Eigenweights for arithmetic Hirzebruch Proportionality」完全由 Aletheia 生成，没有任何人工干预，研究了算术几何领域中被称为 「特征权重」（eigenweights）的若干结构常数。

② 人类与 AI 协作方面，2 月发布的论文「Lower bounds for multivariate independence polynomials and their generalisations」由人类与 Aletheia 合作完成，二者共同证明了相互作用粒子系统（独立集）的相关界限。

③ 大规模半自主评估方面，1 月发布的论文「Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems」中，该智能体针对 Bloom 的 「Erdős 猜想」 数据库中的 700 个开放问题开展评估，并自主解决了其中列出的四个未解难题。

④ 针对 Erdős-1051 问题，该智能体给出了自主解答，并推动了 1 月另一篇论文「Irrationality of rapidly converging series: a problem of Erdős and Graham」中相关成果的推广。

⑤ 此外，「Aletheia 」还在 1 月发布的两篇论文「Arithmetic volumes of moduli stacks of Shtukas」和「Strongly Polynomial Time Complexity of Policy Iteration for L_\infty Robust MDPs」中，提供了关键中间命题的证明。

媒体报道

1. 谷歌 AI 连发 6 篇数学论文！Gemini 攻入博士级科研，91.9%刷爆 SOTA - 新智元

微软和清华提出双向感知塑形框架 BiPS

日期： 2 月 8 日

事件：微软亚洲研究院与清华大学研究团队近日发布论文，提出双向感知塑形框架 BiPS（Bi-directional Perceptual Shaping）框架，以解决视觉-语言模型（VLM）在复杂任务中因捕捉错误视觉证据而导致的推理偏差问题。

要点速览

1. BiPS 的核心思路在于将以往推理阶段（Inference-time）通过外部工具（如裁剪、掩码）或生成潜在视觉 token 获取中间视觉线索的补救范式，转化为训练阶段（Training-time）的感知塑形策略。
2. 该框架设计了「一拉一推」的双向感知塑形机制。

   ① 其中，「拉」（证据保留视图）通过系统性剔除干扰元素、仅保留回答必需的视觉组件，强制模型将回答锚定在完整的证据链上，确保模型「看全」。

   ②「推」（证据消融视图）则精准移除决定答案的关键细微细节，通过最大化原始图像与消融视图的预测分歧（KL 散度），打破文本捷径，迫使模型依赖不可替代的局部视觉线索，确保模型「看准」
3. 在数据构建与训练流程上，研究团队以图表数据为核心训练载体。

   ① 由于图表具有高密度、细粒度的视觉元素（如折线、刻度、图例），且具备极强的程序化可控性，研究者能够自动化地构建证据保留与消融的对照视图。

   ② BiPS 使用约 1.3 万条图表样本对基础模型进行微调，无需人工标注视觉边界框或掩码，通过双向 KL 约束（一致性 + 分离）与 GRPO 框架的组合训练目标，从而实现跨任务的能力迁移。
4. 实验结果显示，BiPS 在 8 个主流视觉理解与推理基准测试中均实现了稳定提升。以 Qwen2.5-VL-7B 为基础模型，该框架带来了平均 7.3% 的准确率提升。

① 在强推理能力 Qwen3-VL-8B-Thinking 上，BiPS 同样表现优异，其中 CharXiv（真实图表理解）得分从 53.0 提升至 58.1，MathVision（数理逻辑推理）得分从 62.7 提升至 63.9。

媒体报道

1. AI 看图一本正经胡说八道？「一拉一推」让模型看得全又准丨微软 x 清华 - 量子位

面壁智能推出混合注意力架构 SALA

日期： 2 月 11 日

事件：面壁智能近日发布技术报告，提出融合稀疏与线性注意力的混合架构 SALA（Sparse Attention-Linear Attention），并推出基于该架构的文本模型 MiniCPM-SALA，用于解决 Transformer 架构处理超长上下文时面临的计算和内存瓶颈问题。

要点速览

1. SALA 架构为弥补 Transformer 全注意力机制的两大瓶颈设计，一是计算复杂度随序列长度呈二次方增长，二是自回归生成过程中的 KV-Cache 会随 token 数量激增占用大量内存。
2. 采用了 SALA 架构的 MiniCPM-SALA 模型保留了 Transformer 架构中每个注意力块后的 FFN 层，其中 25% 的层采用 InfLLM-V2 稀疏注意力，75% 的层采用 Lightning Attention 线性注意力，同时通过层选择机制确定稀疏注意力模块的位置。

   ① 其中 InfLLM-V2 无额外参数，可在稠密与稀疏模式间无缝切换，便于继承权重以实现训练的稳定初始化。

   ② Lightning Attention 与标准 softmax 注意力功能相近，能够减少参数适配的复杂度，同时具备更优的长度泛化能力。
3. MiniCPM-SALA 通过 HALO 架构转化、持续稳定训练、短衰减训练、长衰减训练和监督微调五个阶段完成了持续训练。

① 该训练过程中从短序列逐步扩展至长序列，启用稀疏注意力机制并适配不同的学习率与批次大小，最终实现模型通用能力与长上下文处理能力的平衡。

4. 性能评估方面，9B 参数的 MiniCPM-SALA 在知识、代码、数学推理等标准基准测试中平均得分为 76.53 分，与同规模开源模型性能相当，长上下文相关机制的融入未造成模型通用能力的显著退化。

   ① 长上下文基准测试中，该模型平均得分为 38.97 分，在 128K 上下文长度的 RULER、NoLiMa 测试中，表现优于同规模基线模型，展现出在大上下文窗口下稳定的信息检索能力。

   ② 超长篇上下文测试中，该模型虽仅在 520K token 长度下完成训练，却能在无辅助技术的前提下，将上下文长度外推至 2048K token，在 1000K token 长度的 RULER 测试中，得分超过 80B 参数的 Qwen3-Next-80B-A3B-Instruct 模型，体现出较高的参数效率。

媒体报道

1. 9B 端侧开源模型跑通百万上下文，面壁全新稀疏-线性混合注意力架构 SALA 立功了！ - 量子位
2. 创新 Transformer！面壁基于稀疏-线性混合架构 SALA 训练 9B 模型，端侧跑通百万上下文 - AI 科技评论

原力灵机和阶跃星辰开源具身原生 VLA 模型 DM0

日期： 2 月 10 日

事件：原力灵机和阶跃星辰近日发布技术报告，开源了具身原生（Embodied-Native）VLA 模型 DM0，旨在解决现有 VLA 模型缺乏物理接地、易出现模块碎片化或灾难性遗忘的问题。

要点速览

1. 当前 VLA 研究多依赖互联网静态数据进行预训练，缺乏对物理交互所具备的动态、空间特性的学习，易导致模型出现模块碎片化或灾难性遗忘现象。

2. 针对上述挑战，研究者将具身感觉运动数据与语言、视觉数据置于同等重要的地位，融合网络语料、自动驾驶日志及机器人操作轨迹三类数据，使模型能够学习到兼具语义丰富性与物理可操作性的表征。

3. DM0 模型支持两种推理模式：一种可直接预测动作序列，另一种可先生成推理文本，再基于推理结果输出动作，通过概率分解实现推理与动作生成的有效衔接。

4. DM0 的架构由 VLM 和基于流匹配的动作专家两大核心部分构成。

   ① VLM 以 Qwen3-1.7B 为基础，搭配感知编码器（PE），可实现多模态感知与具身推理功能；基于流匹配的动作专家，则依托 VLM 的键值缓存生成连续控制动作。

5. DM0 的训练分为三个阶段，各阶段侧重点不同，分别为融合多领域异构数据、融入动作预测任务、聚焦目标具身平台。

   ① 预训练阶段侧重融合多领域异构数据，打造强多模态基础，使模型习得语义知识与物理先验。

   ② 中训练阶段侧重融入动作预测任务，结合多类型机器人数据，实现语言推理与物理动作的耦合，同时保留模型的通用能力。

   ③ 后训练阶段侧重聚焦目标具身平台，缩小机器人数据的范围，稳定跨模态对齐效果，最终得到可部署的专属策略模型。

6. 该模型的相关实验在 RoboChallenge 的 Table30 真实世界基准上开展，设置专业版（仅训练目标任务）和通用版（训练全平台任务）两种配置，将 DM0 与 GigaBrain-0.1、π_0。5 等主流模型进行对比测试。

   ① 实验结果显示，DM0-专业版（参数量 2B）的平均成功率达到 62.0%，较此前该基准上的最优结果（SOTA）提升超 10 个百分点；DM0-通用版的平均成功率和任务得分分别达到 37.3 和 49.08，显著优于 π_0。5 和 π_0模型，在精准操作、长时程推理类任务中表现突出。

媒体报道

1. 「具身原生」元年！专访原力灵机汪天才，解析具身智能的「PyTorch 时刻」 - 机器之心

达摩院推出具身大脑基础模型 RynnBrain

日期： 2 月 10 日

事件：阿里达摩院近期在 Github 和 HuggingFace 等平台上发布开源项目，推出具身大脑基础模型 RynnBrain，该模型旨在解决现有大模型在物理世界中缺乏三维空间感、不具备真实物理交互逻辑，且易产生幻觉式规划的问题，助力机器人更好地完成真实环境中的各类任务。

要点速览

1. RynnBrain 继承了达摩院此前 RynnEC 模型的物体与空间感知能力，并进一步研发出时空记忆和物理空间推理两大核心能力。

   ① 时空记忆能力可让模型构建涵盖空间、位置、事件、轨迹等多维度的三维认知表征，即便目标物体脱离当前视野，或任务执行过程被打断，模型也能精准回溯相关信息并继续完成任务。

   ② 物理空间推理能力则通过 「文本与空间定位交错」 的策略，将推理过程与物理世界的具体坐标强制绑定，有效减少模型的幻觉问题。

2. 该模型基于 Qwen3-VL 底座打造，采用达摩院自研的 RynnScale 架构优化训练效率，其中 MoE 架构的 RynnBrain-30B-A3B 仅需 3B 推理激活参数，性能即可超越 72B 规模的 Pelican-VL 模型。

3. 在性能表现上，RynnBrain 在 20 项具身相关任务（含自研 RynnBrain-Bench 的四大维度任务）中，有 16 项实现 SOTA。同时，该模型在导航、操作规划等下游任务中适配性良好，经微调后，其导航成功率和规划能力均实现显著提升。

   ① 其中 8B 版本在具身认知与定位相关任务中，综合表现全面超越 Gemini Robotics ER 1.5、Cosmos-reason2 等前沿模型，部分核心能力提升幅度超 30%。

4. 达摩院同步开源了 RynnBrain 的全套相关资源，该模型全系列包含 2B、8B、30B 等版本在内共计 7 个模型（涵盖 Dense 和 MoE 架构），开源内容包括模型权重、训推代码，以及全新的评测基准 RynnBrain-Bench。

① 该评测基准涵盖物体认知、空间认知、物体定位、具身点预测四大维度，填补了具身智能领域在时空细粒度任务评估方面的空白。

② 此外，达摩院还通过开放协议与接口实现模型、数据与机器人硬件的连接，同时并行推进 VLA 路线的技术探索。

媒体报道

1. 想让机器人春晚包饺子？阿里达摩院：别急，先把「大脑」优化一下 - 机器之心
2. 阿里达摩院开源具身大脑基模：3B 激活参数性能超越 72B，转身就忘事的机器人有救了 - 量子位
3. 机器人长出 800 个心眼？阿里达摩院开源具身新大脑，硅谷又坐不住了 - 新智元

蚂蚁发布离散扩散语言模型 LLaDA2.1

日期： 2 月 11 日

事件：蚂蚁集团、浙江大学、西湖大学和南方科技大学近日发布技术报告，提出离散扩散语言模型 LLaDA2.1。该模型采用了可纠错编辑（Error-Correcting Editable，ECE）机制，缓解了自回归模型解码速度和生成质量之间的权衡问题。

要点速览

1. 该研究者提出了一种「草稿-编辑」（Draft-and-Edit）范式，将传统的 M2T（Mask-to-Token）扩散过程和 T2T（Token-to-Token）编辑机制相结合。

   ① 具体而言，模型在每个时间步会识别两个活跃更新集合：一是置信度超过掩码阈值的解掩码集合，二是当前 token 与候选 token 差异超过编辑阈值的编辑集合，通过联合应用这两种操作实现动态状态演化。

   ② 该机制允许模型在并行生成草稿的同时，能够回溯并修正已生成的内容，从而有效解决了并行解码中常见的局部不一致性问题。

2. 该模型采用了双模式解码策略，支持极速与质量两种模式。

   ① 极速模式通过降低 M2T 置信度阈值实现激进式并行生成，依赖后续的 T2T 编辑进行错误修正，适用于代码生成、快速迭代等场景。

   ② 质量模式则采用保守阈值策略，确保初始生成质量，适用于正式文档和学术写作等高精度任务。

3. 研究者在 100B 规模的扩散模型上成功实施了大规模强化学习训练。针对扩散模型块状采样导致序列级似然难以计算的问题，研究者提出了基于证据下界（ELBO）的 EBPO（ELBO-based Block-level Policy Optimization）算法。

   ① EBPO 算法通过向量化似然估计和专门的梯度稳定机制，提升了模型的推理精度和指令遵循能力。

4. 在 33 项基准测试中，该模型展现出良好的性能表现，Flash 版本（100B）在编码任务 HumanEval+、BigCodeBench、LiveCodeBench 中的 TPS 分别达到 891.74、801.48、663.39，Mini 版本的峰值 TPS 达 1586.93。

   ① S Mode 下两个版本模型吞吐量大幅提升，性能出现小幅下降；Q Mode 下，Mini 和 Flash 两个版本的基准测试表现均超越前代模型 LLaDA2.0，仅伴随小幅效率损耗，多块编辑机制还能对推理和编码任务的性能实现进一步优化。

媒体报道

1. 里程碑时刻！100B 扩散语言模型跑出 892 Tokens /秒，AI 的另一条路走通了 - 机器之心
2. 小众架构赢麻了！通过编辑功能让 100B 扩散模型飙出 892 tokens/秒的速度！ - 量子位

小米开源 VLA 模型 Xiaomi-Robotics-0

日期： 2 月 12 日

事件：小米近日发布论文，发布并开源 VLA 模型 Xiaomi-Robotics-0，其核心优化目标为实现模型的高性能与实时流畅执行，技术突破主要集中在精心设计的训练流程与部署策略上。

要点速览

1. 该模型采用混合 Transformer 架构，由 Qwen3-VL-4B-Instruct 视觉语言模型和扩散 Transformer（DiT）组成，总参数量为 4.7B，可实现对双臂机器人的端到端控制。

   ① 其中，前者负责处理视觉与语言输入，后者基于 VLM 的 KV 缓存和机器人本体感受状态，通过流匹配生成动作块。

2. 模型训练分为预训练和后训练两个阶段，预训练采用两步法，后训练阶段围绕推理延迟问题，优化模型的异步执行能力展开优化。

   ① 预训练阶段中，研究者先联合约 2 亿步的跨模态机器人轨迹数据（含 338 小时乐高拆卸、400 小时毛巾折叠的自研遥操作数据）与超 8000 万条视觉-语言数据训练 VLM，通过 1:6 的比例采样两类数据，同时采用 Choice Policies 范式，使 VLM 学习预测动作块。

   ② 随后冻结 VLM，基于流匹配损失对 DiT 进行训练，引入 Beta 分布采样、adaLN 层和 SINK token，以提升训练效果与动作生成的稳定性。

3. 实验结果显示，该模型在 Libero、CALVIN、SimplerEnv 三大仿真基准测试中均达到当前 SOTA，Libero 基准的平均成功率为 98.7%；在 CALVIN 基准的 ABCD→D、ABC→D 拆分任务中，连续完成 5 项任务的平均长度分别为 4.80 和 4.75。

   ① 在真实机器人的乐高拆卸、毛巾折叠两项任务中，乐高拆卸任务中各方法的成功率相近，该模型的吞吐量显著优于 π0.5 等基线模型；毛巾折叠任务中，其吞吐量达到 1.2 件/分钟。

   ② 同时，该模型保留了原始 VLM 的核心能力，在 10 项视觉-语言基准测试中，表现优于多数 VLA 模型，在具身推理任务上的性能略优于 Qwen3-VL-4B-Instruct。

媒体报道

1. 小米的首代机器人 VLA 大模型来了！丝滑赛德芙，推理延迟仅 80ms 丨全面开源 - 量子位

腾讯发布端侧小模型 HY-1.8B-2Bit

日期： 2 月 10 日

事件：腾讯混元团队发布技术报告，推出了面向消费级硬件场景的端侧小模型 HY-1.8B-2Bit，该模型通过 2 比特量化感知训技术进行压缩，等效参数量降至约 0.3B，，模型大小减少至原始精度模型的 1/6，突破了传统 INT4 量化的极限。

要点速览

1. 在技术实现上，研究者采用了数据优化、弹性拉伸量化以及训练策略创新三个方法来提升量化后模型的全科能力。

   ① 数据优化注重筛选高质量训练数据，弹性拉伸量化则是动态调整量化范围以适应不同权重分布，训练策略创新针对极低比特量化设计的特殊训练流程。

   ② 在该设计下，模型保留了原版的全思考能力，可根据任务复杂度灵活切换长/短思维链推理模式，为不同复杂度的任务提供相应深度的推理过程。

2. 部署方面，研究者提供了 gguf-int2 格式的模型权重与 bf16 伪量化权重，并已在 Arm 计算平台上完成适配，可部署于启用 Arm SME2 技术的移动设备（如 Apple M4 芯片、vivo x300 等）。

   ① 在 MacBook M4 芯片上的测试显示，固定线程数为 2 时，首字时延在 1024 输入内能够保持 3～8 倍 的加速，生成速度对比原始模型精度实现至少 2 倍 稳定加速。

   ② 在 天玑 9500 芯片上，对比 HY-1.8B-Q4 格式，该模型首字时延加速 1.5～2 倍，生成速度加速约 1.5 倍。

3. 当前该模型的能力仍受限于监督微调（SFT）的训练流程以及基础模型本身的性能与抗压能力。针对这一问题，混元团队未来将重点转向强化学习与模型蒸馏等技术路径，以进一步缩小低比特量化模型与全精度模型之间的能力差距，为边缘设备上的大语言模型部署开拓更广阔的应用前景。

媒体报道

1. 0.3B 参数，600MB 内存！腾讯混元实现产业级 2Bit 量化，端侧模型小如手机 App - 量子位
2. 主打一个快！腾讯开源 0.3B 端侧模型，手机耳机都能跑 - 智东西

Waymo 发布世界模型 Waymo World Model

日期： 2 月 7 日

事件：Waymo 近日发表博客，推出了基于 DeepMind Genie 3 打造世界模型 Waymo World Model，可实现大规模、高真实度的自动驾驶仿真，以及生成多类型高保真传感器数据。

要点速览

1. 该模型支持创建可交互的三维场景，且针对自动驾驶系统的特定需求进行了适配调整，能够生成包括摄像头图像、激光雷达点云在内的多传感器数据。
2. 该模型架构支持通过语言提示、驾驶输入或场景布局实现控制，工程师可依据该控制功能调整仿真参数；同时，该模型可模拟现实中难以大规模复现的极端事件。例如特定天气条件、特殊交通状况等。
3. 该模型采用预训练与后训练相结合的方式：先在大规模多样化视频数据上开展预训练，以获取世界知识；再通过专门的后训练流程，将二维视频知识迁移至适配 Waymo 硬件的三维激光雷达输出中。
4. Waymo 将该模型应用于自动驾驶系统的测试与验证，使自动驾驶系统能够在虚拟环境中运行数十亿英里，演练各类复杂及边缘场景。

媒体报道

1. Waymo 联手 DeepMind 打造世界模型：基于 Genie 3，让自动驾驶「脑补」罕见场景 - 机器之心

本周值得关注国内要事

智谱发布基础模型 GLM-5

日期： 2 月 12 日

事件：智谱 AI 发布基础模型 GLM-5，该模型架构集成了 DeepSeek Sparse Attention（稀疏注意力）机制，可在降低部署成本的同时保障高效的上下文处理能力。

要点速览

1. GLM-5 总参数量为 744B，采用稀疏激活架构（40B 激活参数），预训练数据量提升至 28.5T tokens。在 Artificial Analysis 评测榜单中，该模型位列全球第四、开源模型第一。

2. 编程能力方面，GLM-5 在 SWE-bench-Verified、Terminal Bench 2.0 等基准测试中取得开源模型最优成绩，编程性能与 Claude Opus 4.5 对齐，同时该模型在工程任务中展现出长时自主执行能力。

   ① 经测试，该模型曾连续运行超 24 小时，完成 700 余次工具调用、800 次上下文切换，最终从零构建出可运行的 Game Boy Advance 模拟器。

3. GLM-5 发布当日，智谱港股（02513.HK）股价上涨约 26%，本周累计涨幅约 70%，市值达 1756 亿港元。

   ① 目前该模型已完成与七家国产芯片的适配，开发者基于 GLM-5 已开发出 3D 大富翁游戏、数字平行世界应用、学术探索工具等十余款应用，部分应用已申请上架应用商店。

4. 此外，近期智谱也发布并开源了 GLM-OCR 模型，其参数量为 0.9B。该模型具备通用文本识别、复杂表格解析及信息结构化提取三项核心能力。

① 通用文本识别能力可处理照片、截图、扫描件等多种输入形式，能识别手写体、印章、代码等特殊文字。

② 复杂表格解析能力能够处理合并单元格、多层表头等复杂结构，并输出 HTML 代码。

③ 信息结构化提取能力可从卡证、票据、表格中提取关键字段，并输出 JSON 格式。

媒体报道

1. GLM-5 真够顶的：超 24 小时自己跑代码，700 次工具调用、800 次切上下文！ - 量子位
2. 股价暴涨 32%！GLM-5 登顶全球开源第一，25 分钟一镜到底搓出完整系统 - 新智元
3. 智谱最强模型发布！编程对齐 Claude Opus 4.5，七家国产芯片已火速适配 - 智东西
4. 智谱开源 OCR！测完我把手机里的扫描软件都卸了..... - 量子位

字节跳动发布视频生成模型 Seedance 2.0

日期： 2 月 12 日

事件：字节跳动近日发布视频生成模型 Seedance 2.0，该模型已整合至豆包、即梦及火山方舟体验中心，其 API 服务预计于 2 月中下旬在火山方舟上线，面向企业开放使用。

要点速览

1. 该模型优化了物理规律遵循、长效一致性与生成可控性，在运动场景中的生成可用率达到业界 SOTA 水平，可呈现多主体复杂交互动作，特写镜头的物理逻辑与细节表现接近实拍效果。
2. 该模型支持文本、图像、音频、视频四种模态输入，可提取输入素材中的构图、运镜、声音等元素；具备脚本还原、主体一致性保持与分镜设计能力，新增视频编辑与时长延展功能。

① 音频方面采用双声道立体声技术，可支持背景音乐、环境音效及人物解说多轨并行输出，且能精准对齐画面节奏。

② 该模型可应用于商业广告、影视特效、游戏动画、解说视频等场景，需注意的是，使用真人图像/视频作为主体参考时，须经本人验证或取得相关授权。

3. 据 Seed 团队与影视领域专家联合评测，该模型综合表现处于行业领先水平，但在细节稳定性、多人口型匹配、多主体一致性、文字还原精度及复杂编辑效果等方面仍有优化空间。

媒体报道

1. 豆包视频生成模型 Seedance 2.0 发布，豆包、即梦接入 - 字节跳动
2. 「强到可怕！」字节 Seedance2.0 灰度测试爆火，黑悟空老板：AIGC 的童年结束了 - 智东西
3. Seedance 2.0 杀入豆包！海外网友翻墙跪求，国内用户免费用，附一手实测 - 智东西
4. 字节 Seedance 2.0 正式发布！评测全面碾压，马斯克惊呼发展太快 - 智东西

阿里巴巴发布图像生成基础模型 Qwen-Image 2.0

日期： 2 月 10 日

事件：阿里巴巴近日发布图像生成基础模型 Qwen-Image 2.0，该模型将生图与编辑两条并行研发支线整合至单一架构。

要点速览

1. 该模型尺寸实现轻量化优化，相较 200 亿参数的 1.0 版本显著减小，同时推理速度提升，支持 2K 分辨率输出。
2. Qwen-Image 2.0 在长指令遵循与文字渲染方面表现突出，可支持 1K token 的超长提示词，能够处理近千个中英文字词的复杂指令。

① 该模型可还原信息图、PPT、多宫格漫画等内容中的文字排版、格式与细节，对中文古籍、公式、数据表格等内容的还原错误率较低，且文字与图像的融合效果自然。

3. 该模型在图像生成质感方面也进行了优化，色彩饱和度更适中，画面效果更贴近实拍，同时具备图片编辑功能，可实现多图拼接、元素替换、风格调整等操作。

4. 在 AI Arena 盲测平台的评测中，Qwen-Image 2.0 在文生图（Text-to-Image）基准测试中位列第三，次于 Gemini-3-Pro-Image-Preview 和 GPT Image 1.5。

   ① 在单图编辑（Image Edit）基准测试中，该模型位列第二，仅次于 Gemini-3-Pro-Image-Preview，在图像生成的真实感方面，该模型表现稍逊于 Gemini-3-Pro-Image-Preview。

5. 目前，Qwen-Image 2.0 已在阿里云百炼开通 API 邀测，用户可通过 Qwen Chat（chat.qwen.ai）免费体验，后续还将上线千问 App。

媒体报道

1. 字节发完阿里发！Qwen-Image 2.0 火线出击 - 智东西
2. 中文版 Nano Banana 来了？Qwen-Image-2.0 炸场：1K 长文本硬吃，中文生图彻底不拧巴了 - 量子位

MiniMax 发布旗舰 agentic 模型 M2.5

日期： 2 月 13 日

事件：MiniMax 近期发布了旗舰 agentic 模型 M2.5 ，该模型在多语言任务 Multi-SWE-Bench 上位列第一，在 SWE-Bench Verified 评测集上的通过率（80.2%）接近 Opus 4.6 的 80.8%。

要点速览

1. 在智能体框架适配方面，该模型可接入 OpenClaw 等工具，执行本地文件整理、数据分析及演示文稿生成等任务。

   ① 据 MiniMax 内部披露数据，其 30%的业务任务由该模型自主完成，编程场景中 80%的新提交代码由该模型生成。

2. 该模型训练数据涵盖 Go、Rust、Python 等十余种编程语言及数十万个真实环境，采用 Process Reward（过程奖励）机制监控长链路任务执行质量。

   ① 在 GDPval-MM 办公场景评测中，该模型在 Word 排版、Excel 金融建模等任务上的平均胜率为 59.0%（该评测为 MiniMax 内部构建的 Cowork Agent 评估框架）。

3. 该模型激活参数量为 10B，在 SWE-Bench Verified 测试中得分 80.2%，与 Claude Opus 4.6 的 80.8%接近；在 Multi-SWE-Bench 多语言任务中得分 51.3%，高于 Claude Opus 4.6 的 50.3%。

4. M2.5 支持 PC 端、移动端及跨平台开发框架，可生成包含前端界面、后端逻辑及数据库的完整项目代码。

   ① 该模型推理吞吐量达 100 TPS（M2.5-lightning 版本），定价为每百万输入 Token 0.30 美元、每百万输出 Token 2.40 美元。

媒体报道

1. 1 美元时薪？这才是打工人的「梦中情模」 - 机器之心
2. 1 美金时薪雇个全栈替身，MiniMax M2.5 让打工人也能体验当老板的感觉 - 量子位
3. 一夜暴涨至 2100 亿！开源新王 MiniMax M2.5，革了 Opus 4.6 的命 - 新智元

中科曙光上线 3 万卡规模国产 AI 算力池 scaleX

日期： 2 月 6 日

事件：中科曙光近日宣布，由其提供的 3 套 scaleX 万卡（GPU 加速卡）超集群在国家超算互联网核心节点正式上线试运行。

要点速览

1. 该项目实现 3 万块 GPU 加速卡规模部署，投入实际运营的国产 AI 算力池正式形成，国产超集群从单点建设向规模化工程落地迈进。

2. scaleX 万卡超集群的核心优势体现在系统级工程能力上，可将大规模算力转化为稳定的生产产能。

   ① 其搭载的 scaleFabric 网络可实现 400Gb/s 超高带宽及低于 1 微秒的端侧通信延迟，且具备向十万卡、百万卡规模演进的扩展性。

   ② 此外，该集群通过高密设计、低 PUE 工程方案及智能调度系统，解决了大规模算力在网络、存储、散热及供配电层面的联动优化难题，实现了从「建得起来」到「稳得住、用得好」的升级。

3. 在生态兼容与应用层面，scaleX 采用 AI 计算开放架构，兼容 CUDA 等主流软件生态，且支持多品牌国产加速卡混合部署，能够降低用户的迁移与适配门槛。

   ① 目前，该集群已完成 400 余个主流大模型与世界模型的适配优化，并依托国家超算互联网对接了上千款应用。

媒体报道

1. 算力竞赛分叉：马斯克太空炼丹，中国 3 万 AI 卡同时点亮！ - 新智元
2. 全国最大国产 AI 算力池来了：部署超 3 万卡，上千款应用接入 - 量子位

Xmax AI 发布虚实融合实时交互视频模型 X1

日期： 2 月 9 日

事件：人工智能创企 Xmax AI 近日发布虚实融合实时交互视频模型 X1 及技术演示型应用「X-cam」。与以往 AI 视频生成多依赖预渲染、缺乏即时反馈的 「预制模式」 不同，该模型可通过手机摄像头，实现虚拟内容与现实世界的毫秒级实时交互。

要点速览

1. 针对视频生成的实时性与意图理解难题，Xmax AI 研发端到端流式重渲染视频架构，实现帧级别自回归 DiT（Diffusion Transformer）。

   ① 该架构通过多阶段蒸馏压缩与对抗训练，将单帧扩散采样速度提升百倍，将交互延迟控制在毫秒级。

2. 该模型采用统一交互架构，可同时理解三维空间关系与二维屏幕触控意图；配合自研的虚实融合数据合成管线，缓解了行业内交互数据稀缺的问题。

3. Xmax AI 团队基于 X1 模型的端侧实时生成能力，设计了「次元互动」「世界滤镜」「触控动图」「表情捕手」四项核心交互功能：

   ①「次元互动」：用户可从照片中提取虚拟角色并置于现实桌面场景，通过手势与角色完成具备物理反馈的实时交互。

   ②「世界滤镜」：可将摄像头拍摄画面实时转换为乐高、油画等风格，并保持人物动作实时同步。

   ③「触控动图」：用户可通过触屏拖拽，直接操控静态照片中的角色产生动态效果。

   ④「表情捕手」：实时捕捉现实物体特征，自动生成对应的动态 Emoji 表情包。

4. Xmax AI 核心团队成员来自清华大学 KEG 实验室、HCI 实验室，拥有字节跳动、华为、阿里巴巴等企业的产品落地经验。

   ① 创始人史佳欣为华为 「天才少年」 计划出身；联合创始人梁宸为香港科技大学（广州）助理教授、博士生导师；联合创始人翁跃庭为全栈工程师。

媒体报道

1. 童年的滚球兽「走进」现实？华为天才少年创业，全球首个虚实融合的实时交互视频模型来了 - 机器之心

科大讯飞发布星火 X2 大模型

日期： 2 月 11 日

事件：科大讯飞发布星火 X2 大模型，该模型基于全国产算力训练，在数学、推理、语言理解等维度测试中表现与 GPT-5.2、Gemini-3-Pro 持平。

要点速览

1. 技术层面，星火 X2 采用 293B 参数 MoE 稀疏架构，支持单台昇腾服务器量化部署；通过权重量化、低精度 KV Cache、VTP（虚拟张量并行）、分层通信等工程优化实现大 EP 并行推理，推理性能较前代 X1.5 提升 50%。

2. 星火 X2 在模型架构上沿用了星火 X1.5 的 MoE 稀疏架构，参数同样为 293B。在星火 X1.5 的基础之上，星火 X2 进行了训推采样校准强化学习算法、递归式高难数据合成方法、多阶段 RL 高吞吐采样方法（可实现训练效率提升 10%）和服务高性能部署优化算法等技术创新。

3. 实测结果显示，该模型在处理概率统计与逻辑推理类问题时，能够系统性拆解任务并规划求解路径，即便在快速生成模式下，也会呈现完整的思考链条。

   ① 例如处理次品率检验问题时，模型会先明确问题构成与适用解法，再分步计算得出概率值与所需样本量；解答真假城推理题时，会通过条件分析与情境枚举，推导出最优提问策略。

4. 目前，星火 X2 大模型可在讯飞星火网页版和 APP 体验，API 也已上线讯飞开放平台。

媒体报道

1. 神仙打架 +1！讯飞星火 X2 硬核亮相，行业深度全面升级 - 量子位
2. 单台昇腾服务器可跑！国产算力加持大模型升级，推理性能提升 50% - 智东西

字节上线图像生成模型 Seedream 5.0 Preview

日期： 2 月 10 日

事件：近日，字节跳动图像生成模型 Seedream 5.0 Preview 已落地剪映、CapCut 及字节 AI 创作平台小云雀，在即梦 AI 开启灰度测试，该模型的图片生成服务限时免费，支持 2K 和 4K 分辨率输出。

要点速览

1. 该版本继 2025 年 12 月 4 日上线的 4.5 版本后推出，新增检索生图功能，对提示词的理解精度有所提升，可生成细节更丰富的图像，同时支持用户对画面元素进行精确调整。

   ① CapCut 称该模型性能对标 Nano Banana Pro 且成本更低，目前向所有用户开放 20 次免费生成额度，美国地区后续上线。

2. 该模型的核心升级点包括优化提示词理解、细节呈现与文本渲染效果，新增笔刷控制和元素级编辑功能，多步逻辑推理及特定领域知识处理能力也得到了提升。

   ① 实测结果显示，该模型能准确呈现画面核心意象与高细节场景，但对部分精准限定条件的理解存在不足；针对含精准细节要求的复杂提示词，生成效果仍有瑕疵。

3. 该模型可解析 「静谧科技感」 等抽象描述，但与 4.5 版本相比无显著代际提升，联网检索功能的稳定性不足，优化方向主要集中在视觉美观度与生成风格多样性上。

   ① 与 Nano Banana Pro 横向对比中，该模型在知识型提示词的生成上步骤更详尽，但艺术表现力、中文文本渲染能力稍逊于前者，部分用户反馈其相比旧版本的改进幅度有限。

媒体报道

1. 刚刚，Seedream 5.0 预览版上线！字节又一新模型 - 智东西

具身智能创企「星海图」完成 10 亿元 B 轮融资

日期： 2 月 11 日

事件：具身智能创企「星海图」近日完成 10 亿元 B 轮融资，累计融资近 30 亿元，投后估值达百亿级别。

要点速览

1. 该轮投资方包括金鼎资本、北汽产投等产业资本，以及正心谷资本等机构，老股东凯辉基金、美团龙珠等持续追加投资。

2. 2026 年星海图将重点布局智能制造、物流、商业服务三大场景。

   ① 星海图创始人兼 CEO 高继扬判断，2025 年行业处于概念验证阶段，2026 年下半年将进入 「成果验证」 阶段，在特定任务上实现人工替代；且 2025 年下半年以来，开源具身模型数量增长约 5 倍，行业供应链也逐步成熟。

3. 该公司产品采用 「整机 + 智能」 软硬一体模式，已推出轮式双臂机器人 R1 系列，计划 2026 年 5 月发布灵巧手与双足人形机器人。

4. 2026 年星海图计划采集数十万小时多场景真机数据，并引入新型采集方法。

   ① 生态布局上，其已投资数据采集公司简智新创，2026 年计划将被投企业数量扩充至 15-20 家。

5. 此前，该公司孵化首席科学家许华哲离职后创立新企业，业务方向为具身智能 C 端应用，已获星海图种子轮投资。

媒体报道

1. 星海图完成 10 亿元 B 轮融资，成第四家估值百亿具身智能创企丨智能涌现独家 - 智能涌现
2. 星海图完成 B 轮 10 亿元融资，B 轮为什么会成为具身智能企业的一道门槛？- 甲子光年
3. 独家 | 星海图孵化首席科学家许华哲创业，新公司已获得星海图投资 - AI 科技评论

灵巧手创企临界点完成数亿元融资

日期： 2 月 11 日

事件：灵巧手创企临界点（AGILINK）近日完成数亿元融资，该轮融资由头部互联网大厂领投，BV 百度风投、云锋基金及 Synstellation Capital、均胜电子、龙旗科技、上汽金控等产业资本联合投资。

要点速览

1. 该公司前身为智元机器人灵巧手部门，于 2026 年 1 月 14 日完成注册独立，熊坤担任创始人兼负责人。

   ① 熊坤自 2018 年从港科大硕士毕业后，先后任职于腾讯 Robotics X 实验室、IDEA 研究院、汇川技术，于 2024 年 11 月加入智元。

2. 产品方面，临界点在 2025 年推出 OmniHand 及 OmniHand Pro，客户包括智元、灵初智能等，2026 年计划发布两款旗舰灵巧手产品。

   ① 该公司聚焦工程化与规模化，产品覆盖多自由度五指灵巧手及高性价比夹爪，应用于具身智能算法与数据采集、机器人集成、工业与科研场景。

3. 2026 年公司将加大算法投入，发布灵巧操作大模型及百万量级数据集，助力下游客户通过强化学习完成复杂任务。

媒体报道

1. 头部互联网大厂领投，「临界点」再获数亿元融资丨智能涌现独家 - 智能涌现

物理世界模型创企「极映科技」连续完成种子轮及天使轮融资

日期： 2 月 9 日

事件：物理世界模型公司「极映科技」近日宣布连续完成数千万元种子轮及天使轮融资。其中，种子轮由奇绩创坛投资，天使轮由元禾璞华领投，未来光锥跟投，远山资本担任独家财务顾问。该轮融资将用于底层架构研发与多物理场通用基础模型的迭代。

要点速览

1. 极映科技定位于构建工业级物理基础模型，旨在解决传统仿真软件（CAE）计算耗时长、使用门槛高及多物理场耦合难度大等问题。

   ① 与 Sora 等侧重视觉效果呈现的模型不同，极映科技通过自研架构，使模型直接学习偏微分方程等底层物理规律（如质量守恒、能量守恒），实现物理层面的真实映射。

② 其模型可将传统工业仿真反馈周期从「天」级压缩至「秒」级，响应速度显著提升，可应用于半导体、航空航天及具身智能等对精度要求较高的领域。

2. 该团队在物理仿真与软件研发领域拥有长期积累。创始人高鑫为迈阿密大学博士、密歇根大学博士后，拥有十年仿真与人工智能研究经验；联合创始人邱康曾任鹏城实验室算法工程师；联合创始人李福华为清华大学博士，曾任半导体企业研发高管。

3. 商业化方面，极映科技采用「仿真能力基础设施」模式，通过 API 调用或工业系统集成提供服务，目前已切入对仿真依赖度较高的半导体行业，并实现营收。

媒体报道

1. 独家对话极映科技高鑫：我们为什么要做一个比 Sora 难 10 倍的物理世界模型？- 甲子光年

大晓机器人完成天使轮融资

日期： 2 月 10 日

事件：大晓机器人近期完成天使轮融资，该轮资金将用于 ACE 具身全栈研发范式技术迭代、环境式数据采集体系建设、开悟世界模型 3.0（Kairos3.0）研发，以及具身超级大脑模组规模化应用，同时推进能源、交通、文旅等领域商业场景的拓展。

要点速览

1. 该次融资由蚂蚁集团领投，启明创投、金景资本、弘毅投资、联想创投、上海交大母基金菡源资产等机构跟投，原有股东商汤国香资本追加投资，投资方涵盖多种类型机构。

2. 该公司的 ACE 具身全栈研发范式于去年 12 月发布，以环境式数据采集为基础，通过多模态硬件采集多维度信息，为具身模型训练提供包含全交互要素的数据支撑。

3. 该公司的开悟世界模型 3.0（Kairos3.0）旨在建立跨机器人本体的统一世界理解框架，整合物理规律、人类行为模式与机器动作，着力解决具身智能领域数据稀缺与泛化难题，具备复杂场景交互生成及未来状态预测的能力。

4. 大晓机器人团队具备较强实力，核心成员包含两位全球知名 AI 科学家，初创团队成员涵盖南洋理工大学、香港大学等高校的 AI 领域研究者。

   ① 董事长王晓刚为商汤科技联合创始人，其谷歌学术总引用量超 14.5 万次，在 Research.com 发布的计算机科学家排名中，位列中国第 1、全球第 30；首席科学家陶大程为多院院士，论文被引次数超 17 万次，在 AI 领域有着重要贡献。

媒体报道

1. 蚂蚁投了一家上海具身智能公司 - 量子位

AI 智能运动穿戴品牌「苔源 MossCode」 完成数千万元天使轮融资

日期： 2 月 9 日

事件：近日，AI 智能运动穿戴品牌「苔源 MossCode」宣布完成数千万元人民币天使轮融资，由 XVC 和清流资本共同投资，投后估值约 1 亿美元。

要点速览

1. 该轮资金将主要用于产品研发团队扩张、实现产品稳定量产，并计划于 2026 年上半年正式启动欧美市场发售工作。

2. MossCode 定位于打造 AI Native 运动手表，区别于侧重数据记录的同类设备，其核心目标是构建一个能够长期理解用户状态、提供个性化策略指导的 AI 辅助工具。

① 产品核心逻辑为通过 AI 技术深度融合用户主观感受（RPE）与生理数据，建立个人运动能力上下文（Personal Fitness Context），助力用户在自身节奏范围内实现运动能力进阶。

3. 团队方面，MossCode 核心成员均来自 EcoFlow、OPPO、Apple、华为、Suunto、高驰等头部硬件企业及斯坦福等知名高校。

   ① 创始人倪若阳曾任职于红杉中国，后担任 EcoFlow 家庭储能事业部负责人；联合创始人兼 CTO 马潇曾负责 OPPO Find X 系列及传音相关产品工作，拥有深厚的消费电子与底层算力架构相关经验。

4. 市场竞争层面，MossCode 选择以欧美市场作为首发阵地，依托端侧算力优化及国产芯片供应链优势，在 Apple Watch 全智能定位与 Garmin 专业硬核定位之间，探索兼具全天候健康监测与专业运动指导功能的新一代智能穿戴产品形态。

媒体报道

1. 对话 MossCode：AI Native 的运动手表，估值 1 亿美金 - 极客公园

本周值得关注国外要事

Figure AI 发布 Figure 03 人形机器人

日期： 2 月 8 日

事件：Figure AI 近日正式发布第三代人形机器人产品 Figure 03。该机器人搭载了 Figure 团队新研发的具身智能大脑 Helix 02，在测试中实现了全身级自主控制，进一步提升了在复杂室内环境下的自主决策与动态协作表现。

要点速览

1. Helix 02 系统是基于单一神经系统构建的统一大脑，可实现机器人「行走与操作」的实时耦合。

   ① 相较于初代 Helix 采用的 System 1（脊髓）与 System 2（大脑）双层架构，Helix 02 进化为三层神经架构，在原有基础上新增了以 1kHz 频率运行的 System 0（小脑）基础层。

② 新增层级赋予机器人运动直觉，使其可将语义推理与动作执行深度整合，实现类似人类边走边作业的全身协作模式。

2. Figure 03 相比 Figure 02 完成了 90% 的制造组件优化，同时实现 9% 的减重，产品设计适配家庭与工厂大规模量产需求。该机型引入高灵敏度指尖触觉传感器（可感知低至 3 克的压力）与掌心摄像头，进一步提升操作感知能力。

   ① 依托 Helix 02 对 1000 小时人类动作数据的深度学习，Figure 03 具备跨尺度运动控制能力，可完成拧瓶盖、从遮挡处捏起药片及操作注射器等灵巧任务。

   ② 此外，该机型支持 2kW 无线感应快充，可实现全天候自主补能，保障持续作业能力。

3. Figure 03 与 Helix 02 的组合具备长时序自主规划能力，可在无人工干预情况下完成数分钟的复杂任务。

   ① 实测中，该机器人在执行装载洗碗机等家务时，动作平滑度较高，且具备故障隐式恢复机制。

媒体报道

1. 机器人成精了？Figure 03 下厨房，不经意关抽屉那一下，太像人了 - 新智元

Anthropic 上线 Opus 4.6 极速模式

日期： 2 月 8 日

事件：近期，Anthropic 上线 Opus 4.6 极速模式（Fast Mode），该模式下的模型响应速度提升至标准模式的 2.5 倍，定价同步上调 6 倍（输入/输出费用从 5/25 美元涨至 30/150 美元每百万 token），且完全独立于订阅额度计费。

要点速览

1. 该模式未改变模型权重与推理质量，用户可通过 Claude Code 或 VS Code 扩展输入/fast 指令切换该功能，界面显示闪电标识即表示激活。

2. 技术层面，该模型的上下文窗口扩展至 100 万 token（Opus 4.5 的上限是 20 万 token），MRCR v2 长上下文检索测试中得分 76%，可有效缓解长文本场景下的信息衰减问题；具备自主任务难度判断与自我纠错能力，能够在无人工干预的情况下优化推理路径。

3. 在模型性能方面，Opus 4.6 在多项权威基准测试中表现较好。Artificial Analysis Intelligence Index v4.0 涵盖 GDPval-AA、Terminal-Bench Hard、SciCode 等 10 项评估，该模型以 53 分位列综合排名第一，较 OpenAI 的 GPT-5.2（xhigh）高出两分。

   ① Arena.ai 盲测平台（真实人类盲测）数据显示，该模型在代码、文本、专家三个赛道均排名第一，代码竞技场得分 1576，较前代 Opus 4.5 提升 106 分；文本竞技场得分 1496；专家竞技场较第二名高出约 50 分。

   ② 在 GDPval-AA 知识工作性能评估中，其 Elo 得分为 1606，较 GPT-5.2 高出约 144 分，较前代 Opus 4.5 提升 190 分。

   ③ 在 Terminal-Bench 2.0 智能体编程测试中得分 65.4%，排名所有模型首位；ARC-AGI-2 抽象推理测试得分从 Opus 4.5 的 37.6% 提升至 68.8%，提升幅度较大。

媒体报道

1. 价格狂飙 6 倍！Claude 凌晨上线极速模式，网友集体破防：比 OpenAI 还黑 - 新智元

谷歌升级 Gemini 3 Deep Think 模式

事件：谷歌与科学家和研究人员近期展开合作，对 Gemini 3 Deep Think 进行了升级。Deep Think 将科学知识与日常工程实践相结合，更新后可分析图纸，对复杂形状进行建模，并生成用于 3D 打印的实体文件。

要点速览

1. 该模型此前已具备数学与编程基础，去年发布的版本曾在相关世界锦标赛中获得金牌，且可支持智能体开展研究级数学探索。此次更新在多项学术基准测试中取得了突破：

   ① 在「Humanity's Last Exam」中达到当前最优水平，无工具辅助情况下准确率为 48.4%。

   ② 经 ARC Prize 基金会验证，该模型在 ARC-AGI-2 测试中获得84.6%的成绩；在竞技编程平台 Codeforces 上取得 3455 Elo 的评分；在 2025 年国际数学奥林匹克竞赛中达到金牌水准。

2. 成本效益方面，更新后该模型在 ARC-AGI-1 测试中准确率为96.0%，单任务成本 7.17 美元；在 ARC-AGI-2 测试中准确率为84.6%，单任务成本 13.62 美元。

3. 该项目的参与者包括 2024 年 9 月加入 Google DeepMind 的姚顺宇（Shunyu Yao），其本科毕业于清华大学物理系，曾获清华大学本科生特等奖学金，具备深厚的物理学科背景。

4. 更新后的 Deep Think 模式已在 Gemini 应用中上线，目前 Google AI Ultra 订阅用户可以使用。此外，谷歌通过 Gemini API 向部分研究人员、工程师和企业开放了 Deep Think 的使用权限。

媒体报道

1. 刚刚 Gemini 上新模型，全球只有 7 人比它会编程，谷歌姚顺宇参与 - 机器之心
2. 清华传奇姚顺宇立功！全新 Gemini 一夜血洗编程，全球仅 7 人能赢它 - 新智元
3. 姚顺宇谷歌首秀，Gemini 新模型刷爆 SOTA：人类仅剩 7 人捍卫碳基编程 - 量子位

谷歌 Chrome 团队推出 WebMCP

日期： 2 月 11 日

事件：谷歌 Chrome 团队近期推出 WebMCP（Web 模型上下文协议），该协议可让 AI 智能体绕过传统 UI，直接与网站底层进行交互，目前该功能已在 Chrome 146 早期版本中通过功能标志开放体验。

要点速览

1. 当前 AI 操作网页时，多依赖截图解析、DOM 抓取、模拟点击等方式，这类方式存在成本高、稳定性差、效率低等问题。WebMCP 则是将 AI 与网页的交互方式从视觉模拟转向底层直连。
2. 该协议下，通过一个 API：navigator.modelContext，AI 可直接与 Web 服务内核通信，无需通过模拟点击、页面解析或元素定位等方式实现交互。

   ① 例如在预订机票场景中，智能体可直接向航司网站发送相关指令，无需操作网站图形界面。
3. 该协议支持网站向 AI 开放标准化服务接口，以此替代屏幕抓取的传统方式。谷歌为开发者提供了两种接入方式：声明式 API 适配 HTML 表单类的标准操作，命令式 API 则用于处理需借助 JavaScript 实现更复杂、更动态的互动。

媒体报道

1. 谷歌 Chrome 深夜爆更，Agent 不用「装」人了！前端最后防线崩了？ - 新智元

Anthropic 完成 300 亿美元 G 轮融资

日期： 2 月 12 日

事件：Anthropic 近期完成 300 亿美元 G 轮融资，投后估值达 3800 亿美元，该次融资资金将用于前沿研究、产品开发及基础设施扩建。

要点速览

1. 该轮融资由 Coatue 与 GIC（新加坡主权财富基金）共同领投，D。E. Shaw Ventures、Dragoneer 等多家机构联合领投，安大略省教师退休基金会等机构参与投资。

① 该轮融资还包括 2025 年 11 月已公布的英伟达和微软的投资，即英伟达和微软承诺分别向 Anthropic 投资 100 亿美元和 50 亿美元。

2. 该公司当前年化经常性收入为 140 亿美元，近三年年均增速超 10 倍，目前 Claude 模型是唯一同时上线 AWS Bedrock、Google Cloud Vertex AI 及 Microsoft Azure Foundry 三大云平台的前沿 AI 模型。

   ① 以年化经常性收入计算，年消费超 10 万美元的 Claude 客户数量在过去一年增长了 7 倍，年消费超百万美元得 Claude 客户超 500 家，财富 10 强企业中有 8 家为其客户。

   ② Claude Code 的业务年化经常性收入已超 25 亿美元，企业订阅量自 2026 年初也增长了 4 倍，企业使用贡献已超总营收的一半。

3. 此外，该公司正在展开 IPO 前期工作，已聘请律师事务所开展前期准备工作，并与多家投行就上市事宜进行了初步接触。最快或于 2026 年上市。

媒体报道

1. 刚刚，一个 2.6 万亿 AI 独角兽诞生！英伟达微软押注，马斯克急了 - 智东西
2. Anthropic raises another \$30B in Series G, with a new value of \$380B - techcrunch

视频生成创企 Runway 完成 E 轮 3.15 亿美元融资

日期： 2 月 11 日

事件：美国视频生成企业 Runway 近日宣布完成 E 轮融资，金额为 3.15 亿美元，投资方包括英伟达、AMD、Adobe 等企业。该轮资金将用于下一代世界模型的研发，以及新产品及行业领域的拓展。

要点速览

1. 截至目前，该公司累计融资规模已达 8.15 亿美元。据知情人士透露，该轮融资完成后，Runway 估值或将达到 53 亿美元。

① 该公司上一轮融资于 2025 年 4 月完成，金额为 3.08 亿美元，软银、英伟达等企业参与投资，当时公司估值已超过 30 亿美元。

2. 2025 年 12 月，Runway 发布视频生成模型 Gen-4.5，该模型具备生成高逼真度影像的能力，可处理复杂场景及物理效果模拟任务。

   ① 在 Artificial Analysis 发布的文生视频模型排行榜中，Gen-4.5 排名第三，仅次于生数科技 Vidu Q3 Pro 和 xAI 的 grok-imagine-video，排名高于谷歌 Veo 3、OpenAI Sora 2 Pro 及快手可灵 2.5 Turbo 等模型。

3. Gen-4.5 发布十天后，Runway 推出通用世界模型 GWM-1，该模型以实现实时现实模拟为目标，具备交互性、可控性及通用性特征。

   ① 该模型包含三个版本，分别为 GWM Worlds（可探索环境）、GWM Avatars（对话角色）、GWM Robotics（机器人操作），Runway 计划将多领域动作空间整合至统一基础模型框架下。

4. 技术合作方面，Runway 于 2025 年 12 月与 AI 云服务商 CoreWeave 签署合作协议，旨在扩展其基础设施并扩大计算能力。

   ① 2026 年 1 月，该公司宣布正借助 NVIDIA Rubin 平台推进视频生成和世界模型技术。

媒体报道

1. 22 亿！黄仁勋苏姿丰联手，投了一家「世界模型」公司 - 智东西

OpenAI 解散使命对齐团队

日期： 2 月 12 日

事件：OpenAI 解散了使命对齐（Mission Alignment）团队，负责人 Josh Achiam 被任命为公司的首席未来学家（chief futurist），其余团队成员（六七人）已被全部重新分配至公司内部其他部门开展相关工作。

要点速览

1. Mission Alignment 团队于 2024 年 9 月成立，此前主要职责是向员工和公众传递公司使命、解读 AI 的影响，同时开展 AI 安全与可信开发相关工作，其宗旨是「确保通用人工智能造福全人类」。
2. 此次团队解散发生在 OpenAI 推进其 AI 战略的过程中。公司近期在模型能力迭代和产品发布方面步伐加快，包括推出新一代推理模型、扩展消费者产品线。
3. OpenAI 方面对此回应称，公司仍致力于 AI 安全研究，相关工作将分散至公司各团队持续开展。但原 Mission Alignment 团队成员的具体分配去向，以及这一调整对 AI 安全研究优先级的实际影响，目前尚不明确。

媒体报道

1. OpenAI disbands mission alignment team - techcrunch
2. OpenAI reportedly disbanded its Mission Alignment team - the verge

xAI 联合创始人吴宇怀宣布离职

日期： 2 月 10 日

事件：xAI 联合创始人之一吴宇怀近日在社交平台 X 发文宣布离开该公司。据媒体统计，到目前为止 xAI 12 人创始团队中已有 6 人离职，其中包括在吴宇怀离职 48 小时内相继宣布离开的联合创始人 Jimmy Ba。

要点速览

1. 吴宇怀是 xAI 五位华人创始成员之一，其余四人为戴子航、张国栋、杨格及 Jimmy Ba。
2. 吴宇怀的研究方向聚焦于构建具备推理能力的机器。求学期间主导或参与了 STAR 自训练推理模型、定理证明器 Alpha Geometry 及语言模型 Minerva 等项目，致力于研发具备自动化推理能力的 AI 系统。
3. 吴宇怀于 1995 年出生，博士后阶段加入谷歌，工作至 2023 年。同年，他与马斯克等 11 人联合创立 xAI。

① 吴宇怀在 2021 年获得斯坦福大学博士后学位，师从 CRFM 主任 Percy Liang 与 Jay McClelland 教授。求学期间，曾在谷歌 DeepMind AlphaGo 团队及 OpenAI 实习。

4. 到目前为止 xAI 12 人创始团队中已有 6 人离职，其中 5 人的离职发生在过去一年内。

   ① 其中，基础设施负责人 Kyle Kosic 于 2024 年年中跳槽至 OpenAI；随后，谷歌资深研究员 Christian Szegedy 于 2025 年 2 月离开；去年 8 月，Igor Babuschkin 离职并创办了自己的风投公司；而微软出身的 Greg Yang 则在上个月因健康原因离开 xAI。

媒体报道

1. 马斯克 xAI 再失联合创始人，12 人创始团队已有 6 人离场 -机器之心
2. 刚刚，又一位 xAI 华人离职！曾和马斯克并排坐发 Grok 3 - 智东西

《性能之巅》作者 Brendan Gregg 宣布加入 OpenAI

日期： 2 月 8 日

事件：系统性能优化领域的顶级专家、《性能之巅》作者、Linux 内核核心技术 eBPF 主要推动者 Brendan Gregg 近期宣布加入 OpenAI，将服务于 ChatGPT 性能团队，其直属上级为该团队负责人 Justin Becker。

要点速览

1. 系统管理领域的顶会 USENIX LISA 曾为 Brendan Gregg 颁发杰出成就奖，以表彰 Gregg 在该领域的突出贡献。

   ① Brendan Gregg 的技术贡献包括：著有《性能之巅》等性能工程参考书籍，开发了火焰图（Flame Graphs）及差分火焰图，作为主要贡献者推动了 eBPF 技术在 Linux 内核的应用，并长期维护了 Linux 生态里的 bcc 和 bpftrace 工具集。

   ② 此外，Gregg 还提出了 Off-CPU 分析方法，用于识别 I/O 等待导致的性能损耗；设计了延迟热力图，以暴露平均数值无法反映的延

迟分布异常；并建立了 USE 方法（利用率、饱和度、错误），以此来作为系统排查的框架化思路。

2. Gregg 在 2001 年至 2014 年间任职于 Sun Microsystems 及 Joyent，参与 DTraceToolkit 的开发工作；2014 年至 2022 年在 Netflix 担任高级性能架构师，负责处理大规模云架构下的复杂性能问题。

   ① 加入 OpenAI 前，他担任 Intel Fellow，致力于降低软件开发者理解硬件 PMU 数据的门槛。

3. 关于加入 OpenAI 的决策，Gregg 通过个人博客阐述了多重考量。他强调，OpenAI 存在超大规模集群上当日部署、即时生效的优化要求；同时认为，OpenAI 在性能改进方面不设限制的技术文化，与自己的工作方式高度契合。

   ① 从行业观察层面，他注意到 AI 工具已渗透至日常职业场景，进而推断后端系统负载将达到新的量级，而传统面向 CPU 和数据库的优化手段，难以应对 GPU 集群与神经网络带来的新挑战，因此需要建立专门针对大模型的工程方法。

媒体报道

1. 教科书《性能之巅》作者入职 OpenAI！迷弟总裁亲自欢迎 - 量子位

参考链接

[1-1] https://github.com/openclaw/openclaw

[1-2] https://www.usaii.org/ai-insights/vertical-ai-agents-explained-mechanisms-use-cases-and-adoption

[1-3] https://www.aalpha.net/blog/vertical-vs-horizontal-ai-agents/

[1-4] https://baijia.online/homepage/survey/Survey%20on%20AI%20Memory.pdf

[1-5] https://arxiv.org/pdf/2504.15965

[1-6] https://arxiv.org/abs/2512.23343

[1-7] https://arxiv.org/pdf/2601.07372

[1-8] https://arxiv.org/pdf/2602.01313

[1-9] https://arxiv.org/pdf/2601.20465

[1-10] https://arxiv.org/pdf/2601.05171

[1-11] https://iclr.cc/virtual/2026/workshop/10000792

以上是「2026 Week 07 · 会员要闻 」的全部内容。

没读过瘾 ？？？

机器之心为企业客户提供更具针对性的定制「研报订阅」服务，

点击 此处 填写表单进一步了解。