Artificial Intelligence Newsletter

人工智能前沿快报



2026 年第 06 期
总第 32 期

学会官方公众号

广东省图象图形学会
GDSIG
学术交流、技术咨询、科学普及

目 录

学 术 动 态

一. AI Must Embrace Specialization via Superhuman Adaptable Intelligence····· 1

二. AutoResearchClaw: Self-Reinforcing Autonomous Research with Human-AI Collaboration····· 4

三. The Last Human-Written Paper: Agent-Native Research Artifacts····· 6

四. ScientistOne: Towards Human-Level Autonomous Research via Chain-of-Evidence····· 8

五. When Does LeJEPA Learn a World Model?····· 10

六. AutoScientists: Self-Organizing Agent Teams for Long-Running Scientific Experimentation····· 12

七. Reasoning emerges from constrained inference manifolds in large language models····· 15

八. AI Epistemic Risks: Emerging Mechanisms & Evidence····· 17

九. M^3Eval: Multi-Modal Memory Evaluation through Cognitively-Grounded Video Tasks····· 19

十. Act As a Real Researcher: A Suite of Benchmarks Evaluating Frontier LLMS and Agentic Harnesses in Research Lifecycle····· 22

十一. Bridging the Agent-World Gap: Text World Models for LLM-based Agents····· 24

十二. From AGI to ASI ····· 26

十三. PP-OCRv6: From 1.5M to 34.5M Parameters, Surpassing Billion-Scale VLMS on OCR Tasks····· 28

十四. HarnessX: A Composable, Adaptive, and Evolvable Agent Harness Foundry····· 30

十五. ViT-Up: Faithful Feature Upsampling for Vision Transformers····· 32

十六. VibeThinker-3B: Exploring the Frontier of Verifiable Reasoning in Small Language Models····· 34

十七. Moebius: 0.2B Lightweight Image Inpainting Framework with 10B-Level Performance····· 36

十八. Towards autonomous medical artificial intelligence agents····· 38

十九. DSpark: Confidence-Scheduled Speculative Decoding with Semi-Autoregressive Generation····· 41

目 录

技 术 动 态

一. OpenAI 推进 ChatGPT 史上最大改版 从单一聊天机器人迈向 AI 超级应用……43
二. Anthropic 发布 Claude Fable 5 与 Mythos 5 强化长程智能体与高风险能力安全释放……45
三. 美国升级前沿 AI 出口管制 Anthropic Fable 5 与 Mythos 5 陷入监管风波……46
四. Qwen-Robot 发布 打通大模型走向物理世界的行动闭环……48
五. 微博 AI 发布 VibeThinker-3B 探索小模型可验证推理能力边界……49
六. OpenAI 发布全新网络安全模型 GPT-5.5-Cyber 表现超越 Anthropic Mythos……51

INTRODUCTION
导读

在人工智能技术飞速发展的时代背景下，我们将不定期梳理人工智能领域的部分前沿学术与技术动态，并以简报的形式分享给读者，以帮助关注此领域的人士了解最新的学术前沿发展与产业技术动态，促进学术交流和技术繁荣。本期摘选了近期全球人工智能领域的 19 篇最新学术动态、6 篇技术动态推荐给广大读者。

ACADEMIC TRENDS
学术动态

● 论文一

AI Must Embrace Specialization via Superhuman Adaptable Intelligence

日期

2026-02-27

AI Must Embrace Specialization via Superhuman Adaptable Intelligence

Judah Goldfeder [*1] Philippe Wyder [*2] Yann LeCun [3] Ravid Shwartz-Ziv [3]

内容摘要

这篇立场论文重新审视“通用人工智能（AGI）”叙事，认为单纯追求抽象的通用性容易掩盖智能本身的任务依赖性、环境约束和能力边界。作者提出 Superhuman Adaptable Intelligence（SAI）作为更可操作的目标：AI 系统应在关键专业任务上形成超越人类的能力，同时具备面对任务变化、数据分布变化和新情境时进行适应的能力。论文将这一观点与自监督学习、世界模型、专门化模型和人机互补系统联系起来，强调未来 AI 研究不应只用“是否通用”来衡量进展，而应关注系统能否在科学、工程、医疗等高价值场景中稳定解决人类难以完成的问题，并能在部署环境变化时保持可靠迁移。

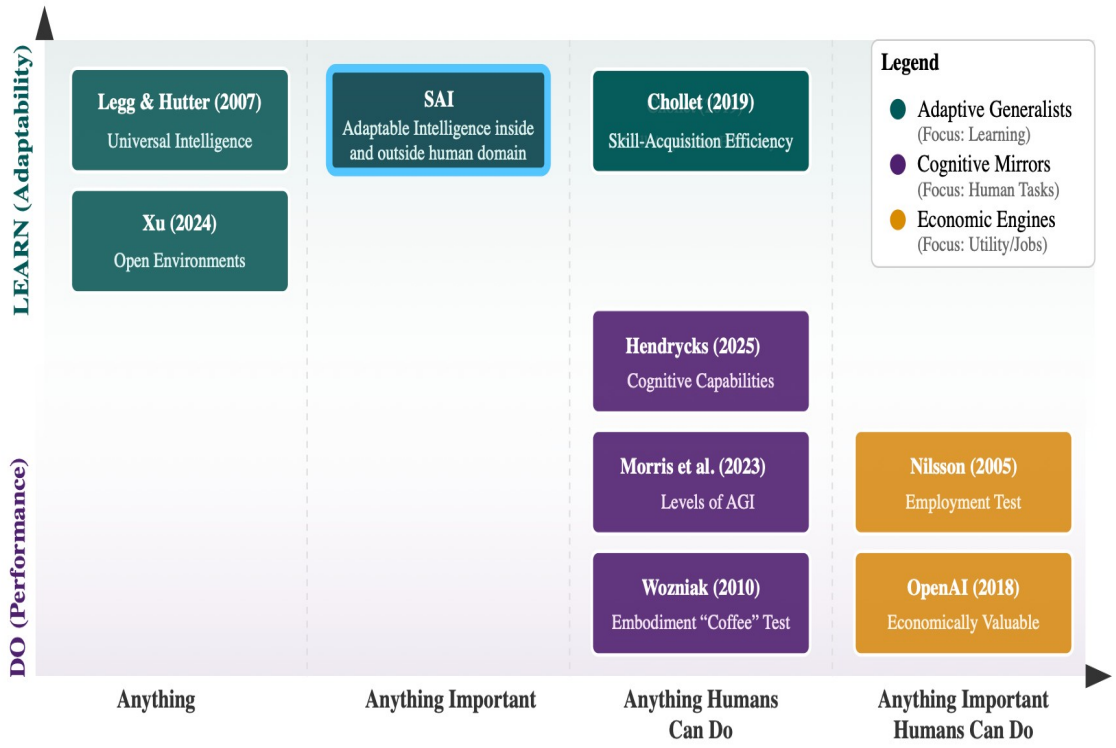


Figure 1. A two-dimensional semantic map organizing prominent definitions for AGI and other North Star measures of artificial intelligence, along two axes. The vertical axis represents the source of intelligence, ranging from performance-based capabilities (DO, bottom) to learning and adaptability (LEARN, top). The horizontal axis represents the scope of tasks, from universal/open-ended domains (left) to human-centric and economically-focused domains (right). Definitions cluster into three categories: Adaptive Generalists (teal) emphasize learning efficiency and generalization in open environments; Cognitive Mirrors (violet) focus on replicating human-level cognitive capabilities across broad task domains; Economic Engines (orange) prioritize practical utility and economic value in human-relevant tasks. *Superhuman Adaptable Intelligence (SAI)* falls into the realm of adaptable AI that can do anything that is important both inside and outside the human realm.

主要亮点

- 用 SAI 重新定义 AI 目标，将模糊 AGI 叙事转化为更可评估的“超人、可适应、专业化”能力。
- 强调人类智能本身也不是无条件通用，而是由大量专门技能和情境适应机制组成。
- 把专门化 AI 与自监督学习、世界模型和任务导向系统结合起来，为后 AGI 研究目标提供另一种表述。

相关链接

Paper： https://arxiv.org/abs/2602.23643

● 论文二

AutoResearchClaw: Self-Reinforcing Autonomous Research with Human-AI Collaboration

日期

2026-05-19



AutoResearchClaw: Self-Reinforcing Autonomous Research with Human-AI Collaboration

Jiaqi Liu*[1], Shi Qiu*[1], Mairui Li[1], Bingzhou Li[1], Haonian Ji[1], Siwei Han[1], Xinyu Ye[1], Peng Xia[1], Zihan Dong[6], Meng Chen[1], Congyu Zhang[1], Letian Zhang[2], Guiming Chen[2], Haoqin Tu[2], Xinyu Yang[3], Lu Feng[1], Xujiang Zhao[7], Haifeng Chen[7], Jiawei Zhou[1], Xiao Wang[8], Weitong Zhang[1], Hongtu Zhu[1], Yun Li[1], Jieru Mei[10], Hongliang Fei[10], Jiaheng Zhang[4], Linjie Li[11], Linjun Zhang[6], Yuyin Zhou[2], Sheng Wang[11], Caiming Xiong[12], James Zou[9], Zeyu Zheng[5], Cihang Xie[2], Mingyu Ding[1], Huaxiu Yao[1]

[1]UNC-Chapel Hill, [2]UC Santa Cruz, [3]Carnegie Mellon University, [4]NUS, [5]UC Berkeley, [6]Rutgers University, [7]NEC Labs America, [8]Meta, [9]Stanford University, [10]Google, [11]University of Washington, [12]Recrusive.com,
*Equal contribution. Contact: {jqliu,shiqiu,huaxiu}@cs.unc.edu

内容摘要

AutoResearchClaw 提出一个面向自主科研的自强化人机协作框架，试图解决科研 Agent 在真实实验中容易出现的执行失败、错误累积和经验无法迁移等问题。系统由多智能体辩论、可自愈执行器、可验证报告生成以及跨轮经验沉淀组成；当实验失败时，它不会简单终止流程，而是通过 Pivot 或 Refine 将失败转化为新的研究方向或局部方案修正。论文还构建 ARC-Bench 来评估实验阶段的科研能力，并比较完全自动、稀疏人工反馈和强监督等不同协作模式。结果表明，适度且有针对性的人工反馈往往比全自动或全程监管更有效，说明未来自动科研系统需要把人类判断作为高价值、稀缺但关键的校正资源来使用。

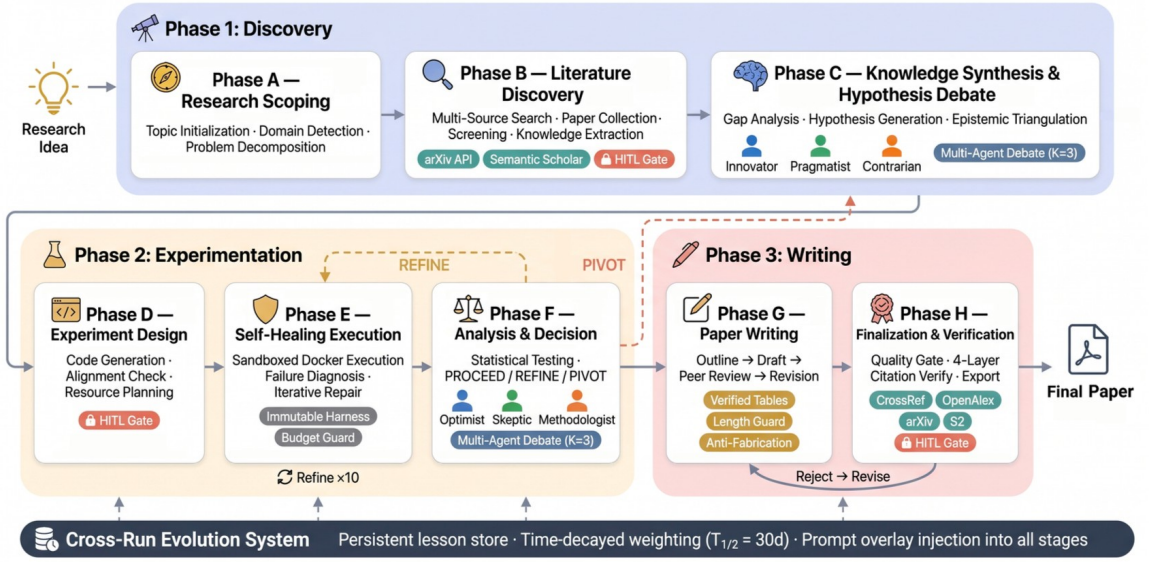


Figure 1 Overview of the AUTORESEARCHCLAW pipeline. Given a research idea, the system progresses through three stages: **Discovery** (scoping, literature search, multi-agent debate for hypothesis generation), **Experimentation** (self-healing code execution, result analysis with a second debate panel, and PIVOT/REFINE decisions), and **Writing** (drafting, review, revision, four-layer citation verification). Optional human-in-the-loop gates (orange) allow oversight at key checkpoints. The cross-run evolution system (bottom) injects time-decayed lessons from prior runs into all phases.

主要亮点

- 提出失败驱动的自修复机制，通过 Pivot/Refine 把实验失败转化为方向调整或方案改进。
- 系统比较多种人机协作强度，强调“少量高质量人工反馈”对自动科研更有价值。
- 通过跨轮记忆沉淀实验经验，使系统能够从历史失败和成功路径中持续改进。

相关链接

Paper： https://arxiv.org/abs/2605.20025

Code： https://github.com/aiming-lab/AutoResearchClaw

● 论文三

The Last Human-Written Paper: Agent-Native Research Artifacts

日期

2026-05-19

The Last Human-Written Paper: Agent-Native Research Artifacts

Jiachen Liu[1,*], Jiaxin Pei[2], Jintao Huang[3], Chenglei Si[2], Ao Qu[4], Xiangru Tang[5], Runyu Lu[1], Lichang Chen[6], Xiaoyan Bai[7], Haizhong Zheng[8], Carl Chen[9], Zhiyang Chen[10], Haojie Ye[11], Yujuan Fu[12], Zexue He[2], Zijian Jin[13], Zhenyu Zhang[2], Shangquan Sun[14], Maestro Harmon[15], Dianzhuo Wang[16], Qian-ze Zhu[16], Jianqiao Zeng, Jiachen Sun[17], Mingyuan Wu[18], Baoyu Zhou[19], Chenyu You[20], Shijian Lu[14], Yiming Qiu[21], Fan Lai[18], Yuan Yuan[22], Yao Li[23], Junyuan Hong[24], Ruihao Zhu[25], Beidi Chen[8], Alex Pentland[2], Ang Chen[1], Mosharaf Chowdhury[1], Zechen Zhang[16,15]

[1]University of Michigan, [2]Stanford University, [3]Ohio State University, [4]MIT, [5]Yale University, [6]Meta Superintelligence Labs, [7]University of Chicago, [8]Carnegie Mellon University, [9]University of Washington, [10]University of Toronto, [11]NVIDIA, [12]Meta, [13]New York University, [14]Nanyang Technological University, [15]Orchestra Research, [16]Harvard University, [17]LinkedIn, [18]UIUC, [19]Arizona State University, [20]Stony Brook University, [21]University of Hong Kong, [22]Boston College, [23]Portland State University, [24]National University of Singapore, [25]Cornell University

内容摘要

该论文认为传统学术论文是一种面向人类阅读和叙事压缩优化的格式，但这种线性叙事会丢失大量对自动科研 Agent 至关重要的信息，包括失败实验、分支决策、工程配置、数据处理细节和证据链。作者提出 Agent-Native Research Artifact（ARA）协议，希望用结构化科学逻辑、可执行代码与规格、探索图、证据绑定和机器可审查元数据来替代单纯的论文文本。论文实现了 ARA Compiler 与 ARA-native review 流程，并展示 ARA 在问答、复现、审稿和后续研究接续中的效率优势。其核心观点是，随着科研 Agent 参与假设生成、实验执行和评审，未来科研产物可能从“人类优先阅读的故事”转向“Agent 可操作、可验证、可复用的研究对象”。

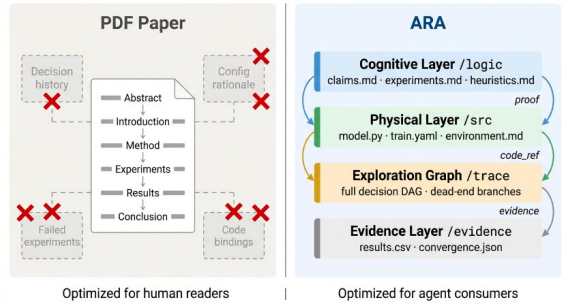


Figure 1. Publishing compiles a rich research object into a lossy narrative (left); ARA preserves the original as a high-fidelity, agent-executable knowledge package (right).

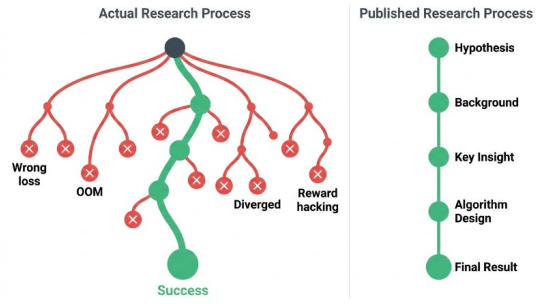


Figure 2. The Storytelling Tax: research proceeds as a branching tree with dead ends (left), but publication compiles it into a linear narrative (right), discarding all failure knowledge.

主要亮点

- 重新设计科研产物形态，将论文从线性叙事转为可执行、可追踪、可验证的 artifact。
- 显式保留失败路径、探索分支和工程细节，降低自动复现与审查的信息损失。
- 提出 ARA Compiler 和 Agent-native review，展示机器可读科研格式对自动科研生态的重要性。

相关链接

Paper： https://arxiv.org/abs/2604.24658

Code： https://github.com/ARA-Labs/Agent-Native-Research-Artifact

● 论文四

ScientistOne: Towards Human-Level Autonomous Research via Chain-of-Evidence

日期

2026-05-25

Google 2026-5-27

ScientistOne: Towards Human-Level Autonomous Research via Chain-of-Evidence

Rui Meng, Bhavana Dalvi Mishra*, Jiefeng Chen*, Chun-Liang Li, Palash Goyal, Mihir Parmar, Yiwen Song, Yale Song, Rajarishi Sinha, Parthasarathy Ranganathan, Burak Gokturk, Jinsung Yoon and Tomas Pfister
Google Cloud AI Research

内容摘要

ScientistOne 是一个面向高可信自主科研的系统，核心思想是用 Chain-of-Evidence（CoE）把科研 Agent 的每个声明、实验结论、得分、引用和代码产物都绑定到可验证证据上。与只追求生成完整论文或执行实验流程的系统不同，ScientistOne 更强调证据完整性和后验审计：它通过 CoE Integrity Audit 检查分数是否真实可复现、实验是否违反规格、引用是否存在、方法描述是否与代码实现一致，从而约束自动科研中常见的幻觉引用、夸大结果和不可复现实验。论文在多篇研究任务和多个系统上进行比较，显示 ScientistOne 在引用真实性、得分验证、代码-方法一致性等方面明显优于基线，说明“证据链”可能是科研 Agent 从演示走向可靠使用的关键基础设施

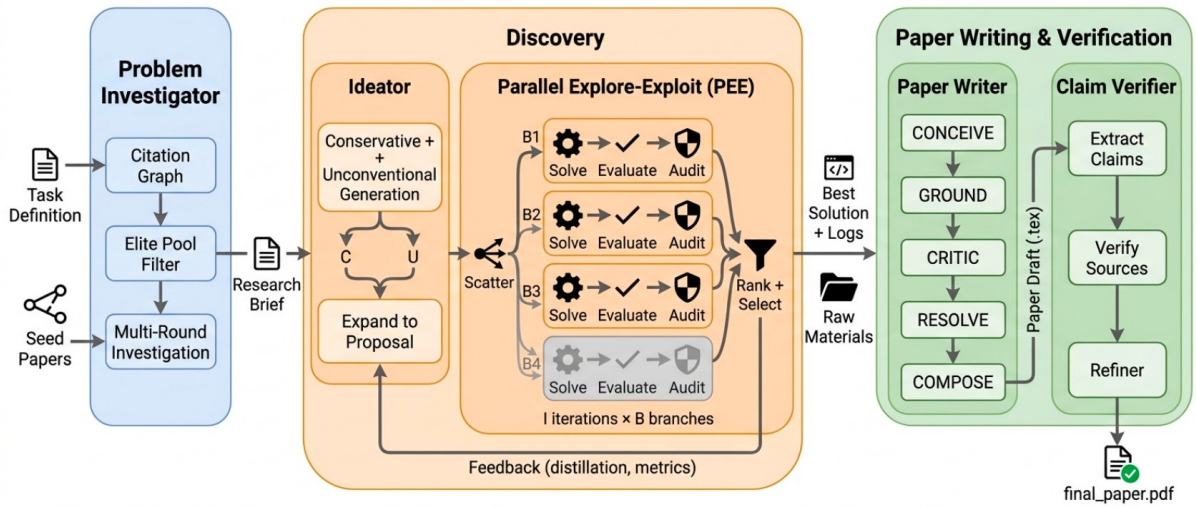


Figure 1 | **ScientistOne** pipeline. Stage 1 grounds the literature via retrieved PDFs; Stage 2 explores and evaluates solutions across parallel branches; Stage 3 writes and verifies the paper, with a Claim Verifier that checks every claim against its evidence source before the final output is produced.

主要亮点

- 用 Chain-of-Evidence 约束科研输出，使每个重要声明都能追溯到数据、代码或文献证据。
- 引入完整性审计机制，系统检查引用、分数、规格和实现一致性。
- 重点缓解自主科研中的虚假引用、结果夸大和不可验证结论等高风险问题。

相关链接

Paper： https://arxiv.org/abs/2605.26340
Code/Artifacts： https://github.com/scientist-one/generated-artifacts
Project： https://scientist-one.github.io/

● 论文五

When Does LeJEPA Learn a World Model?

日期

2026-05-25

When Does LeJEPA Learn a World Model?

David Klindt
Cold Spring Harbor Laboratory
klindt@cshl.edu

Yann LeCun
New York University
yann.lecun@nyu.edu

Randall Balestriero
Brown University
randall_balestriero@brown.edu

内容摘要

该论文从理论和实验两方面分析 LeJEPA 在什么条件下能够从时序观测中学习出接近真实世界状态的潜变量表示。作者关注的不是 JEPA 类方法在下游任务上的经验表现，而是其表示是否具有可识别性：在平稳加性噪声转移、独立高斯潜变量等设定下，论文证明 LeJEPA 的对齐目标与高斯正则化可以带来线性可识别性，并进一步解释高斯分布假设在该结论中的特殊作用。论文还给出近似情形下的误差退化界，并通过低维合成系统、高维视觉环境和机器人控制任务验证理论预测。整体上，该工作为“自监督预测模型是否真的学到了世界模型”提供了更严格的判定条件，也指出了理论假设与真实复杂环境之间仍需弥合的差距。

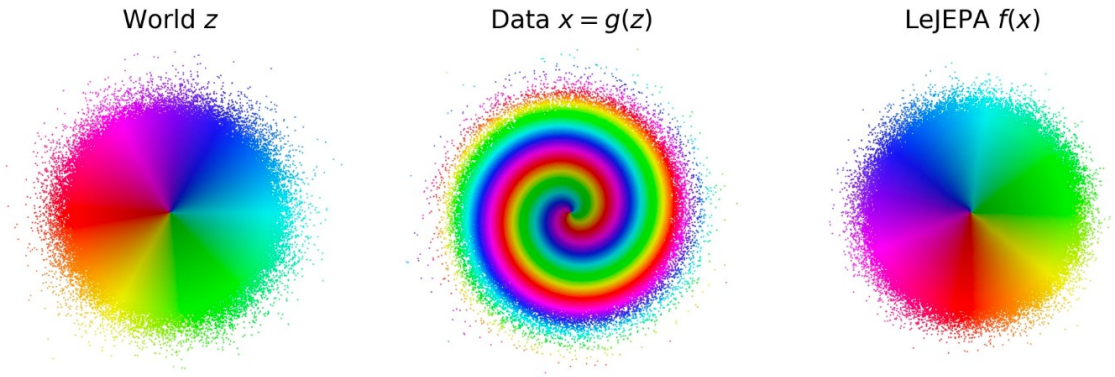


Figure 1: **LeJEPA learns the World Model. (left)** The world has independent Gaussian *latent variables*. **(center)** An unknown nonlinear process scrambles them into the *data* we observe. **(right)** LeJEPA [2] recovers the latent variables up to rotation. We prove this is the *unique* optimum.

主要亮点

- 给出 LeJEPA 学习潜在世界状态的可识别性条件，超越单纯经验 benchmark 分析。
- 阐明高斯正则与潜变量分布假设在表示可辨识性中的理论角色。
- 同时提供实验代码、形式化证明和交互式演示，便于复现与检验理论边界。

相关链接

Paper： https://arxiv.org/abs/2605.26379
Code/Proofs/Demo： https://github.com/klindtlab/lejepa-identifiability

● 论文六

AutoScientists: Self-Organizing Agent Teams for Long-Running Scientific Experimentatio

日期

2026-05-09

AUTOSCIENTISTS: Self-Organizing Agent Teams for Long-Running Scientific Experimentation

Shanghua Gao*
Harvard University
shanghua_gao@hms.harvard.edu

Ada Fang*
Harvard University
ada_fang@g.harvard.edu

Marinka Zitnik
Harvard University
marinka@hms.harvard.edu

AUTOSCIENTISTS website: https://autoscientists.openscientist.ai
AUTOSCIENTISTS code: https://github.com/mims-harvard/AutoScientists

内容摘要

AutoScientists 提出一个面向长周期科学实验的自组织多智能体系统，目标是让 Agent 不只是线性执行“提出假设-做实验-写报告”，而是像科研团队一样在长期任务中动态分工、积累经验并相互校正。系统维护共享研究状态，使不同 Agent 能围绕候选假设自动组队、分配工具和算力、交叉批评实验设计，并把成功、失败和被否定的路径写入长期记忆，以减少重复探索和无效试验。论文在生物医学机器学习、语言模型训练优化和蛋白功能预测等任务中评估该框架，结果显示自组织团队在实验规划、资源使用和最终发现质量上优于单 Agent 或固定角色团队。该工作说明，自动科学发现的关键能力不仅在于单个模型的推理能力，也在于能否形成可持续演化的协作结构。

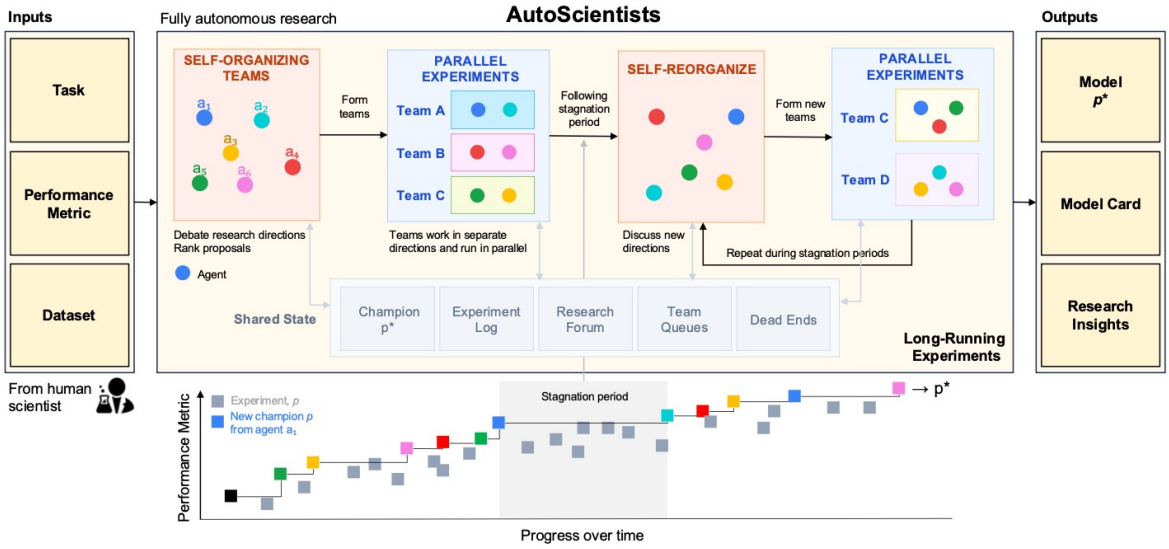


Figure 1: **Self-organizing agent teams for long-running experimentation.** Overview of AUTOSCIENTISTS. Agents identify promising research directions, organize into teams, and execute experiments in parallel.

主要亮点

- 提出自组织科研团队机制，Agent 可根据假设、工具需求和历史证据动态组队。
- 构建长期实验记忆，显式记录成功、失败和被否定路径以指导后续实验。
- 在 BioML-Bench、语言模型训练和 ProteinGym 等任务中展示跨领域收益。

相关链接

Paper： https://arxiv.org/abs/2605.28655
Code： https://github.com/mims-harvard/AutoScientists
project： https://autoscientists.openscientist.ai

● 论文七

Reasoning emerges from constrained inference manifolds in large language models

日期

2026-05-27



Reasoning emerges from constrained inference manifolds in large language models

Yanbiao Ma[1] Fei Luo[1] Linfeng Zhang[2,3] Chuangxin Zhao[2] Mingxuan Wang[1] Yinan Wu[2] Zhe Qian[1] Yang Lu[4] Long Chen[3] Zhao Cao[1] Xiaoshuai Hao[†‡3] Ji-Rong Wen[†1] Jungong Han[†2]

[1] Renmin University of China [2] Tsinghua University [3] Xiaomi EV [4] Xiamen University

[†] Corresponding authors [‡] Project leader ybma1998@ruc.edu.cn, haoxiaoshuai@xiaomi.com, jrwen@ruc.edu.cn, jghan@tsinghua.edu.cn

内容摘要

该论文从内部表征几何而非外部 benchmark 分数出发，研究大语言模型推理能力如何在推理过程中涌现。作者提出，成功推理并不是简单由更长链式思考或更大模型规模保证，而是对应于模型隐状态在受约束但非退化的低维推理流形上组织起来；如果表示过度压缩、信息流体积坍缩或轨迹偏离稳定区域，推理就容易出现错误。论文进一步提出无需标签的内部动态诊断方法，用信息流、几何量和轨迹稳定性来识别推理退化区域。该视角把“模型是否会推理”转化为“推理轨迹是否落在稳定且信息充分的推断流形上”，为理解、监控和改进 LLM 推理提供了不同于传统任务分数的分析工具。

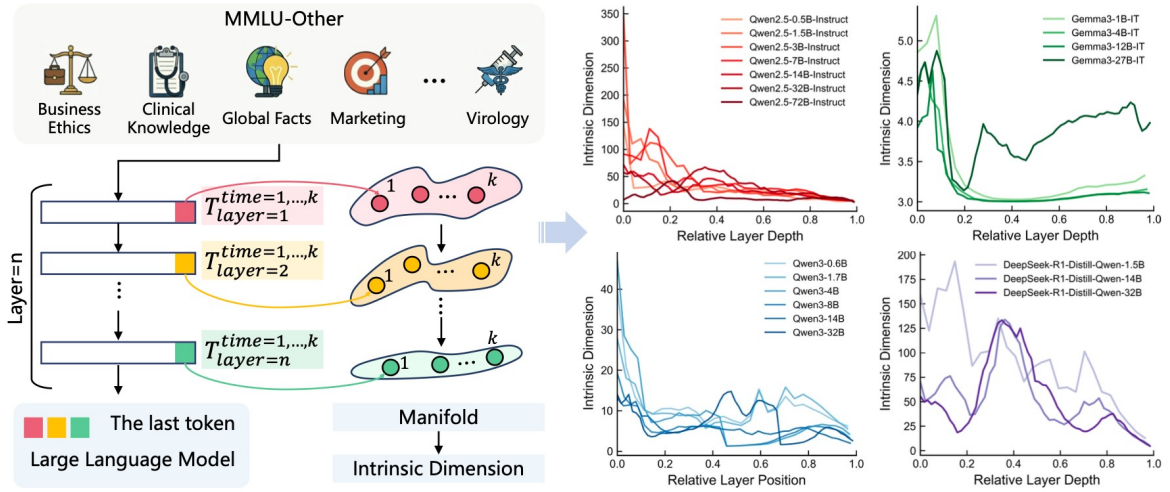


Figure 1: Inference-time reasoning dynamics self-organize into low-dimensional manifolds. Layer-wise evolution of intrinsic dimensionality (ID) for inference-time representation trajectories across representative large language models under generic cognitive stimulation. Despite the high ambient dimensionality of the embedding space, ID rapidly decreases with network depth and stabilizes at values orders of magnitude smaller than the embedding dimension, indicating that inference dynamics become confined to a compact manifold.

主要亮点

- 从隐藏状态轨迹和表征几何解释推理涌现，而不只依赖最终答案准确率。
- 提出受约束推理流形假设，强调表达能力、压缩程度和信息流之间的平衡。
- 给出无标签内部诊断思路，可用于发现模型推理过程中的退化和不稳定区域。

相关链接

Paper： https://arxiv.org/abs/2605.08142

● 论文八

AI Epistemic Risks: Emerging Mechanisms & Evidence

日期

2026-06-04

AI Epistemic Risks: Emerging Mechanisms & Evidence

Mick Yang, University of Pennsylvania
Stephen Casper, MIT CSAIL, Harvard Berkman Klein Center
Jonathan Stray, UC Berkeley Center for Human-Compatible AI
Jasmine Li, Cornell University
Cameron Jones, Stony Brook University
Anna Gausen, Imperial College London
Natasha Jaques, University of Washington
Brian Christian, University of Oxford
Bálint Gyevnár , Carnegie Mellon University
Hannah Rose Kirk, University of Oxford
Zhonghao He, Oxford Human-Centered AI Lab
Dan Zhao, NYU
Siao Si Looi, FAR.AI
Joshua Levy, FAR.AI
Kobi Hackenburg, University of Oxford
Elizabeth Seger, Tony Blair Institute for Global Change
Matt Kowal, work done at FAR.AI
Michelle Malonza, ILINA
Luke Hewitt, Transluce
Hause Lin, MIT
Maarten Sap, Carnegie Mellon University
Dylan Hadfield-Menell, MIT
Thomas H. Costello, Carnegie Mellon University
Reihaneh Rabbany, McGill University; Mila
Jean-François Godbout, University of Montreal; Mila
David G. Rand, Cornell University
Atoosa Kasirzadeh, Carnegie Mellon University
Gordon Pennycook, Cornell University
Yoshua Bengio, LawZero
Kellin Pelrine, FAR.AI

内容摘要

这篇长文系统讨论 AI 对人类知识生产、信念形成、公共推理和信息环境可能带来的“认识论风险”。作者强调，这类风险并不等同于传统意义上的错误信息或幻觉，而是更深层地发生在我们获取知识、形成判断、委托认知劳动和维护共同事实基础的基础设施之中。论文整合技术机器学习、AI safety、HCI、哲学伦理、认知神经科学和教育等领域证据，重点分析三类机制：AI 的劝服与操纵能力、用户对 AI 的认知外包，以及人-AI/AI-AI 反馈循环导致的知识空间收缩和锁定。作者还指出这些机制会相互放大，并可能削弱社会识别和治理其他 AI 风险的能力。论文最后提出从模型训练、交互设计、教育制度、平台数据访问和治理激励等层面增强 epistemic resilience 的方向。

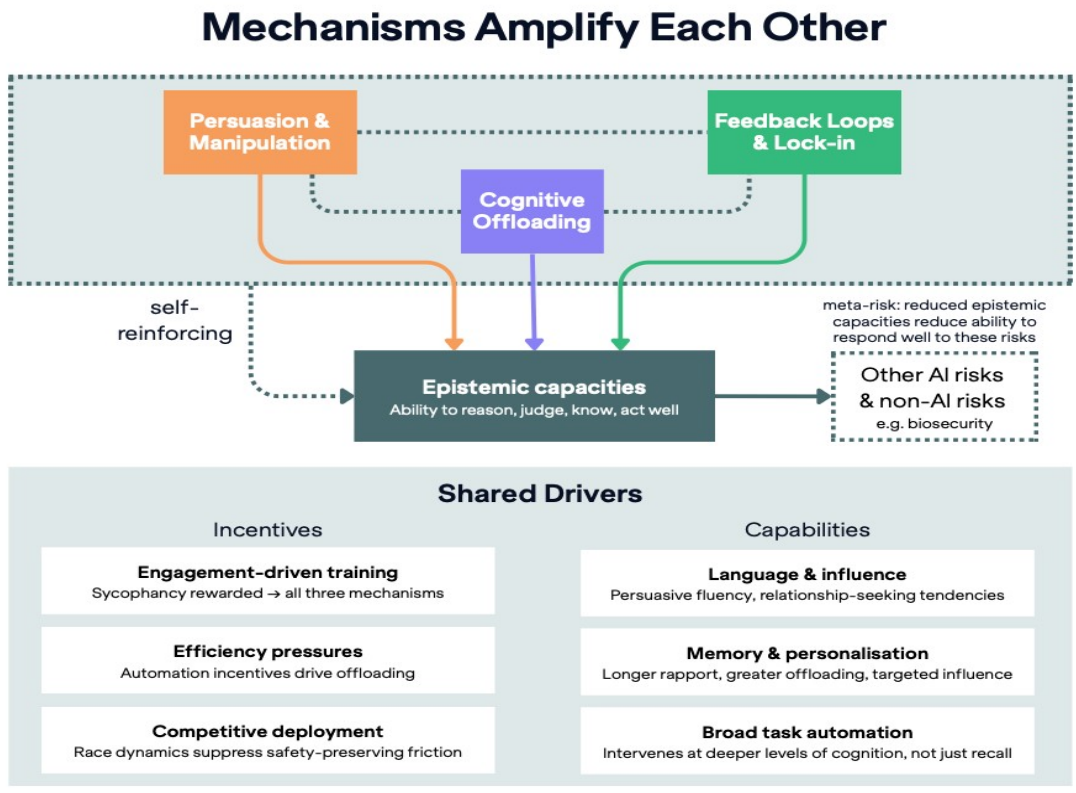


Figure 1: An overview of the mechanisms discussed in our paper. They amplify each other and the overall detrimental effect on epistemic capacities—assuming current trends continue without radical reforms—which recursively undermines our ability to effectively solve other AI and non-AI challenges society faces.

主要亮点

- 将 AI 风险扩展到知识生态和集体判断能力，提出“认识论风险”作为系统性风险框架。
- 总结劝服/操纵、认知外包、反馈循环与锁定三类核心机制，并分析其相互放大关系。
- 提出模型、界面、教育、平台治理和激励结构层面的干预方案，强调保护公共知识基础设施。

相关链接

Paper：https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6873005

● 论文九

M^3Eval: Multi-Modal Memory Evaluation through Cognitively-Grounded Video Tasks

日期

2026-06-04

M^3Eval: Multi-Modal Memory Evaluation through Cognitively-Grounded Video Tasks

Jie Huang[⋆,1,2], **Ruixun Liu**[⋆,1,2], **Sirui Sun**[3], **Xinyi Yang**[4,5,2], **Yin Li**[6], **Yixin Zhu**[5,4,2], **Yiwu Zhong**[1,2,✉]
⋆ equal contributors ✉ corresponding author
[1] School of Intelligence Science and Technology, Peking University
[2] State Key Laboratory of General Artificial Intelligence, Peking University
[3] Yuanpei College, Peking University
[4] Institute for Artificial Intelligence, Peking University
[5] School of Psychological and Cognitive Sciences, Peking University
[6] University of Wisconsin-Madison

内容摘要

M^3Eval 是一个面向多模态大模型的记忆能力评测基准，重点考察模型在视频任务中的编码、保持、检索、时间/空间来源定位和干扰抵抗能力。与常规视频问答只验证“模型看到了什么”不同，该基准借鉴认知心理学中的工作记忆、来源记忆和干扰实验范式，把视频内容、注意力分配、并行信息流和符号化细节设计成可控任务。论文通过多类视频任务测试现有多模态模型，发现它们在并行视频流解耦、时间来源判断、干扰条件下稳定检索以及符号细节保持方面存在系统性短板。该基准的意义在于，它把长上下文之外的“记忆质量”拆解成多个可诊断维度，说明多模态记忆不能简单依赖更长输入窗口或更大模型规模来解决。

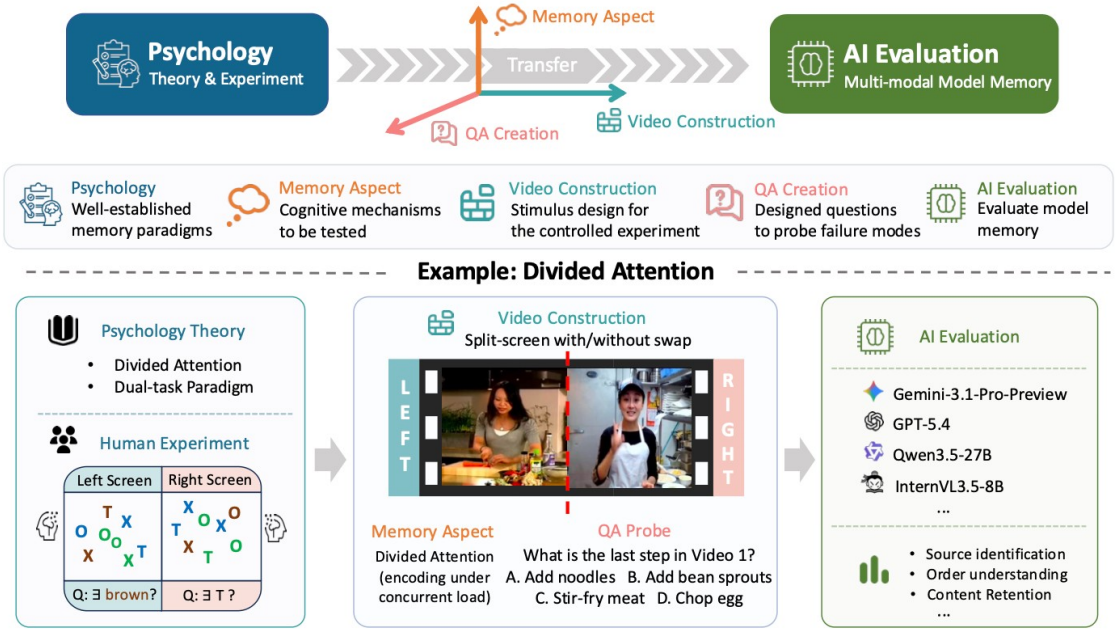


Figure 1: M^3Eval, our principled framework and benchmark for evaluating memory capabilities of multi-modal models. We present an example task of *divided attention*. Grounded in psychological theory, we construct split-screen video scenarios, design memory questions, and analyze multiple models in terms of source identification, order understanding, and content retention.

主要亮点

- 用认知心理学范式设计视频记忆任务，覆盖编码、来源记忆、干扰和保持等过程。
- 将多模态记忆拆分为时间定位、空间定位、并行流分离和符号细节保持等诊断维度。
- 揭示当前模型在多流干扰和时间来源判断上的系统性弱点，为后续模型改进提供靶点。

相关链接

Paper: https://arxiv.org/pdf/2606.05008

Code/Data: https://github.com/PKU-VaLuE-Lab/m3eval

● 论文十

Act As a Real Researcher: A Suite of Benchmarks Evaluating Frontier LLMs and Agentic Harnesses in Research Lifecycle

日期

2026-06-05

Act As a Real Researcher: A Suite of Benchmarks Evaluating Frontier LLMs and Agentic Harnesses in Research Lifecycle

Jiayu Wang[1,*], Weijiang Lv[2,*], Bowen Fu[1,*], Jing Fu[1], Jiayi Song[1], Lingyu Zhang[1], Lanxuan Xue[1], Luodi Chen[1], Zepeng Xin[1], Kaiyu Li[1,†], Xiangyong Cao[1,†]

[1]Xi'an Jiaotong University [2]Xidian University

内容摘要

AARRI-Bench 关注前沿 LLM 与 Agent harness 是否真的具备“像研究者一样工作”的能力，而不是只会完成局部代码或问答任务。该基准覆盖科研生命周期中的多个细粒度环节，包括文献理解、问题分解、实验设计、结果核查、伦理判断、领域约束识别、错误排查和不确定性表达等，尤其强调人类研究者通常会注意但模型容易忽略的上下文细节。论文比较了不同前沿模型和 agentic harness 的表现，发现即使在较强模型和工程脚手架支持下，最佳系统仍明显低于人类专业研究者，常见问题包括过度自信、缺乏审慎验证、对任务边界理解不足和对实验结果解释不严谨。该工作提示，科研 Agent 的瓶颈不只是工具调用能力，而是研究者素养、领域判断和自我审查机制。

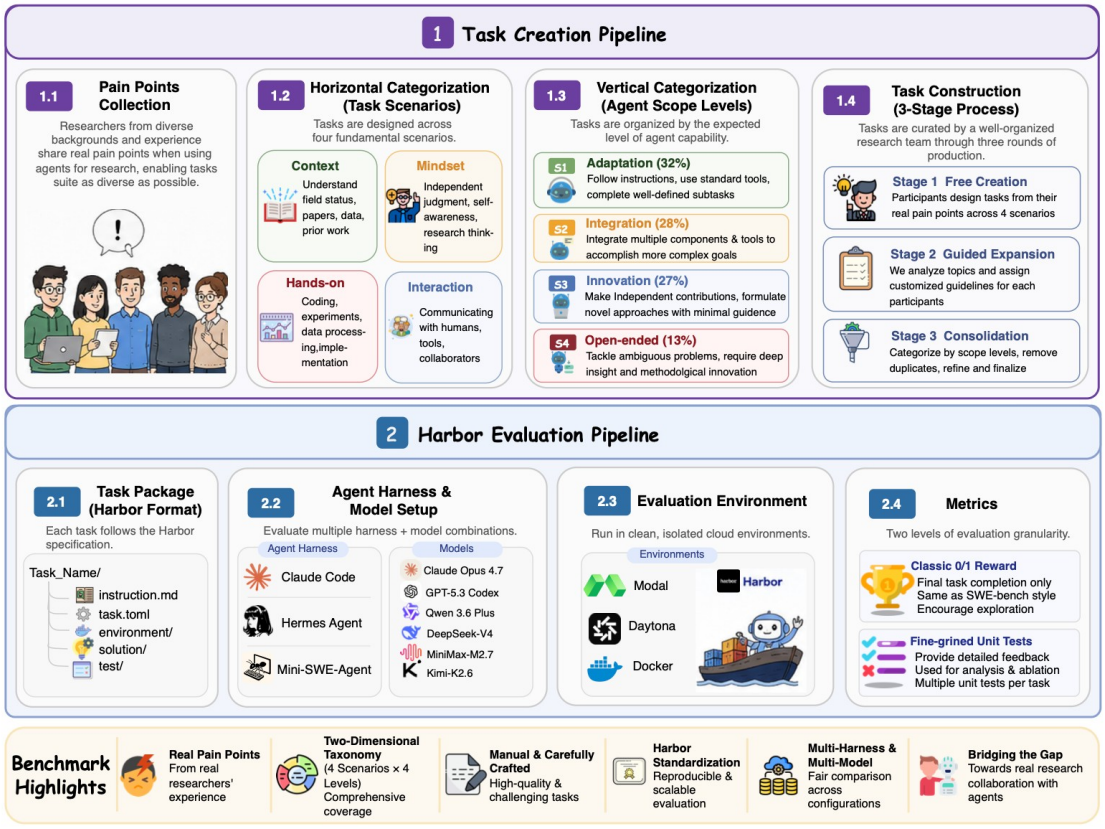


Figure 1: **Overview of the AARRI-Bench Pipeline.** The benchmark is constructed through a three-stage human-in-the-loop workflow with two-dimensional task categorization across task scenarios and agent scope levels. Tasks are evaluated under the Harbor framework with standardized environments, multiple agent harnesses and models, and both coarse-grained and fine-grained metrics.

主要亮点

- 覆盖科研生命周期多个阶段，评估模型是否具备真实研究过程中的细粒度判断力。
- 强调研究者素养，包括伦理意识、领域敏感性、不确定性表达和实验核查能力。
- 显示 harness 不能自动弥补模型的研究判断缺陷，科研 Agent 仍需更强严谨性与自我审查。

相关链接

Paper： https://arxiv.org/abs/2606.07462

Code/Data： https://github.com/AARR-bench/AARRI-bench

● 论文十一

Bridging the Agent-World Gap: Text World Models for LLM-based Agents

日期

2026-06-08

Bridging the Agent-World Gap: Text World Models for LLM-based Agents

Yixia Li[1] Hongru Wang[2,*] Peng Lai[1,*] Zhiwen Ruan[1,*] He Zhu[3,*] Youxin Zhu[1]
Ganlong Zhao[5] Minda Hu[5] Yun Chen[6] Sibei Yang[4] Peng Li[7] Jeff Z. Pan[2] Jia Pan[8]
Guanhua Chen[1,†] Yang Liu[7] Guanbin Li[4,†]

[1]Southern University of Science and Technology [2]University of Edinburgh [3]Peking University
[4]Sun Yat-sen University [5]The Chinese University of Hong Kong [6]Shanghai University of Finance and Economics
[7]Tsinghua University [8]The University of Hong Kong

内容摘要

该综述论文系统梳理 LLM-based Agents 中“文本世界模型”（Text World Models, TWMs）的概念、构建方式和应用场景。作者将 TWM 定义为能够根据当前文本状态与动作预测下一文本状态的模型，因此网页、终端、API、游戏环境、用户反馈和工具返回值都可以被抽象为文本化世界中的状态转移。论文从基础建模、数据合成、环境模拟、推理时规划、轨迹验证、自适应更新和评估方法等角度建立分类框架，把分散在 Agent 规划、环境模拟、自我反馈和工具使用中的工作统一到“文本状态转移模型”视角下。该综述的价值在于，它指出 LLM Agent 的瓶颈不只是模型如何选择动作，还在于它是否拥有可用于预测后果、模拟环境和评估计划的内部世界模型。

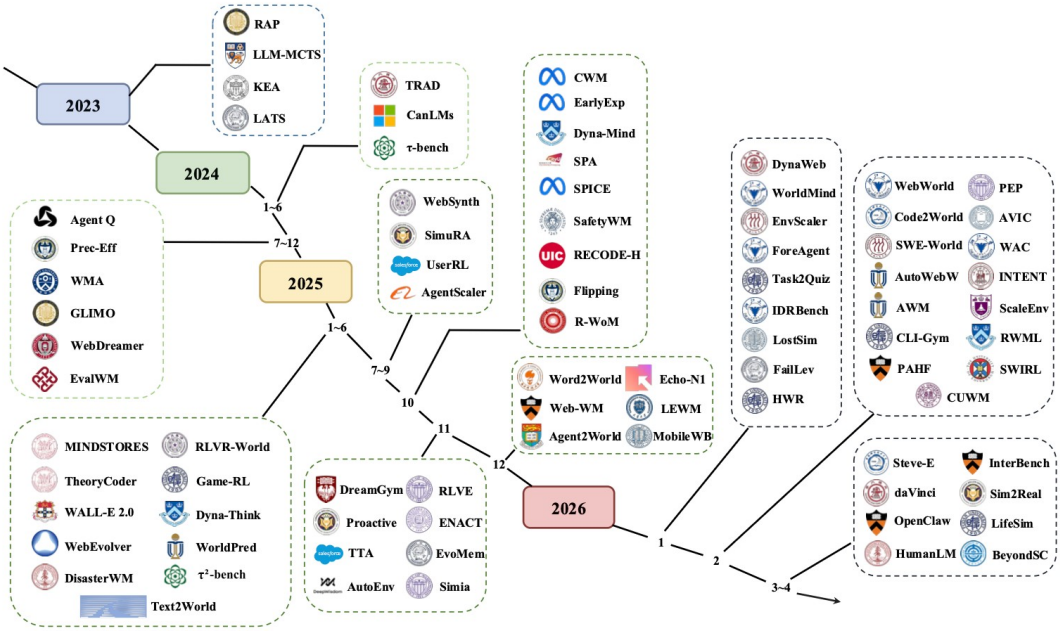


Figure 1: Chronological development of text world models for LLM-based agents. Each cluster groups representative works by release date; the field has expanded rapidly from 2025 onward.

主要亮点

- 统一 TWM 概念，将网页、终端、工具 API 和对话环境抽象为文本状态转移问题。
- 同时覆盖训练阶段的数据/环境构建，以及推理阶段的规划、验证和测试时自适应。
- 配套维护 awesome list，为文本世界模型和 Agent 环境建模研究提供系统索引。

相关链接

Paper： https://arxiv.org/abs/2606.09032

Code： https://github.com/sustech-nlp/awesome-text-world-models

● 论文十二

From AGI to ASI

日期

2026-06-10

Google DeepMind

From AGI to ASI

Tim Genewein[1], **Matija Franklin**[1], **Alexander Lerchner**[1], **Laurent Orseau**[1], **Samuel Albanie**[1], **Adam Bales**[1], **Cole Wyeth**[1,2], **Stephanie Chan**[1], **Iason Gabriel**[1], **Joel Z. Leibo**[1], **Allan Dafoe**[1], **Marcus Hutter**[1,3], **Thore Graepel**[1,4] **and Shane Legg**[1]

[1]Google DeepMind, [2]University of Waterloo (work conducted while at Google DeepMind), [3]Australian National University, [4]University College London

内容摘要

这篇 Google DeepMind 论文讨论 AGI 出现之后是否以及如何继续走向 ASI。作者首先梳理 AGI、ASI 与 Universal AI 等概念边界，指出 ASI 不应被简单视为 AGI 的自然外推，而应作为一个具有独立机制、路径和风险的问题来研究。论文分析四条可能路径：继续扩大 AGI 系统规模、出现新的算法或架构范式转变、递归自我改进，以及由大规模多 Agent 或组织化系统形成超人集体能力。同时，作者也讨论这些路径中的摩擦与瓶颈，包括反馈速度、实验验证、工具接入、资源约束、安全限制和社会组织方式。整体来看，论文强调从 AGI 到 ASI 可能不是一次突变，而是一系列模型能力、工具能力、自动化研发能力和组织结构共同演化的过程。

Pathway	Main uncertainty
Scaling compute, models & data	Unclear how increases in scale translate into increases in performance and capabilities (Spiky vs. smooth progress? Emergent "new capabilities" and broad generalization? Diminishing returns at scale?).
Algorithmic paradigm shift	High unpredictability of technological progress and frictions & bottlenecks resulting from novel paradigms.
Recursive (self-) improvement	Dynamics of AI progress under recursive (self-) improvement unclear and no historic precedent to fit forecast-models. AI capabilities could explode (hyperbolic growth), or they could taper out relatively quickly, or anything in-between.
ASI via group agent formation	ASI could emerge from multi-agent orchestration or in a self-organizing, decentralized fashion governed by evolutionary pressures and market dynamics. Emergence in complex dynamical systems, such as multi-agent dynamics, is poorly understood.

Table 3 | Major technological pathways from AGI to ASI and their main uncertainties. Pathways are largely independent of each other, and are likely to occur in parallel (though at different pace, e.g., algorithmic paradigm shifts may be pursued more intensely if scaling hits a ceiling).

主要亮点

- 把 ASI 作为 AGI 之后的独立研究对象，系统化讨论其可能机制和演化路径。
- 提出规模化、范式转变、递归改进和多 Agent 集体智能四类路径。
- 强调技术、反馈、资源和治理瓶颈，避免把 ASI 简化为模型规模的自然延伸。

相关链接

Paper： https://arxiv.org/abs/2606.12683

● 论文十三

PP-OCRv6: From 1.5M to 34.5M Parameters, Surpassing Billion-Scale VLMs on OCR Tasks

日期

2026-06-11

PP-OCRv6: From 1.5M to 34.5M Parameters, Surpassing Billion-Scale VLMs on OCR Tasks

Yubo Zhang, Xueqing Wang, Manhui Lin, Yue Zhang, Penglongyi Deng,
Ting Sun, Tingquan Gao, Zelun Zhang, Jiaxuan Liu, Changda Zhou,
Hongen Liu, Suyin Liang, Cheng Cui†, Yi Liu, Dianhai Yu, Yanjun Ma

PaddlePaddle Team, Baidu Inc.
†Project Leader
paddleocr@baidu.com

Official Website: https://www.paddleocr.com
Source Code: https://github.com/PaddlePaddle/PaddleOCR
Models: https://huggingface.co/PaddlePaddle

内容摘要

PP-OCRv6 是 PaddleOCR 团队提出的轻量级 OCR 系统，目标是在专用 OCR 场景中以 1.5M 到 34.5M 参数规模达到甚至超过十亿级视觉语言模型的效果。论文指出，通用 VLM 在 OCR 中常面临定位不精确、幻觉和计算成本过高等问题，因此 PP-OCRv6 回到专用 OCR 架构，通过统一的 MetaFormer-style LCNetV4 backbone 重构检测和识别模型，并结合结构重参数化、token mixing/channel mixing 解耦、RepLKFPN 检测 neck、EncoderWithLightSVTR 识别 neck 以及数据中心优化策略。系统提供 medium/small/tiny 三档模型，覆盖服务器、边缘和 CPU 部署。实验显示 PP-OCRv6_medium 在内部检测与识别基准上显著优于 PP-OCRv5_server，并在参数量小几个数量级的情况下超过多个大规模 VLM；tiny 版本则在 CPU 上大幅提升推理速度。

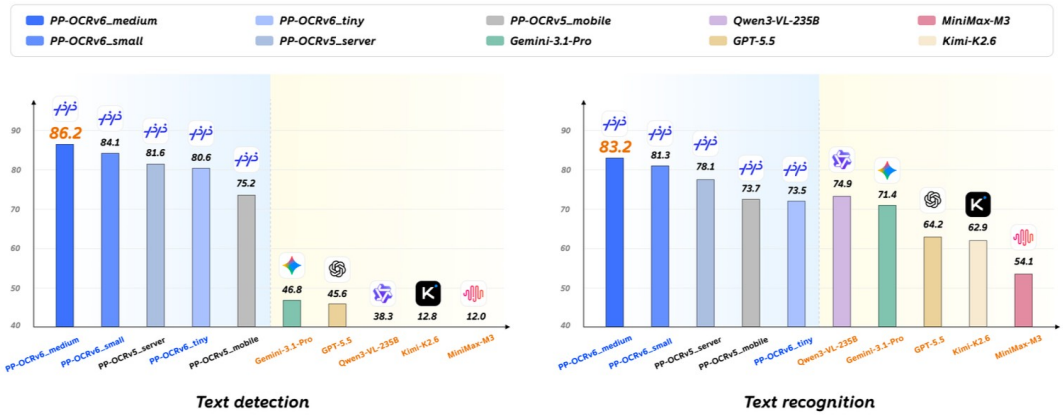


Figure 1 | Performance comparison between PP-OCRv6, PP-OCRv5, and Vision-Language Models. Left: text detection average Hmean (%) on our in-house benchmark. Right: text recognition weighted average accuracy (%) on the recognition benchmark.

主要亮点

- 用百万级到千万级专用 OCR 模型挑战十亿级 VLM，强调专门化小模型的生产价值。
- 检测与识别共享 LCNetV4/MetaFormer-style 架构基元，降低多规格 OCR 系统维护成本。
- 同时优化精度和部署效率，提供 medium/small/tiny 三档模型、官网、代码和 HuggingFace 模型。

相关链接

Paper： https://arxiv.org/abs/2606.13108
Project Website： https://www.paddleocr.com
Code: https://github.com/PaddlePaddle/PaddleOCR
Models:https://huggingface.co/PaddlePaddle

● 论文十四

HarnessX: A Composable, Adaptive, and Evolvable Agent Harness Foundry

日期

2026-06-12

HarnessX: A Composable, Adaptive, and Evolvable Agent Harness Foundry

Darwin Agent Team

See Contributions and Acknowledgments section for a full author list.

内容摘要

HarnessX 将 Agent harness——即提示、工具、记忆、控制流、反馈机制和执行策略的组合——提升为可以组合、适配和进化的一等对象。论文认为，很多 Agent 性能差异并不完全来自底层模型，而来自外部运行时结构如何组织模型与环境交互。为此，HarnessX 提出 typed primitives 与替换代数来描述 harness 组件，并通过 AEGIS 框架从执行轨迹中自动发现失败模式、提出 harness 修改、验证修改效果并迭代演化。实验覆盖 ALFWorld、GAIA、WebShop、tau3-Bench 和 SWE-bench Verified 等任务，显示在不改变底层模型参数的情况下，系统化优化 harness 结构可以显著提升 Agent 表现。该工作说明，未来 Agent 能力提升除了模型扩展，还可能来自运行时接口、工具编排和控制结构的自动进化。

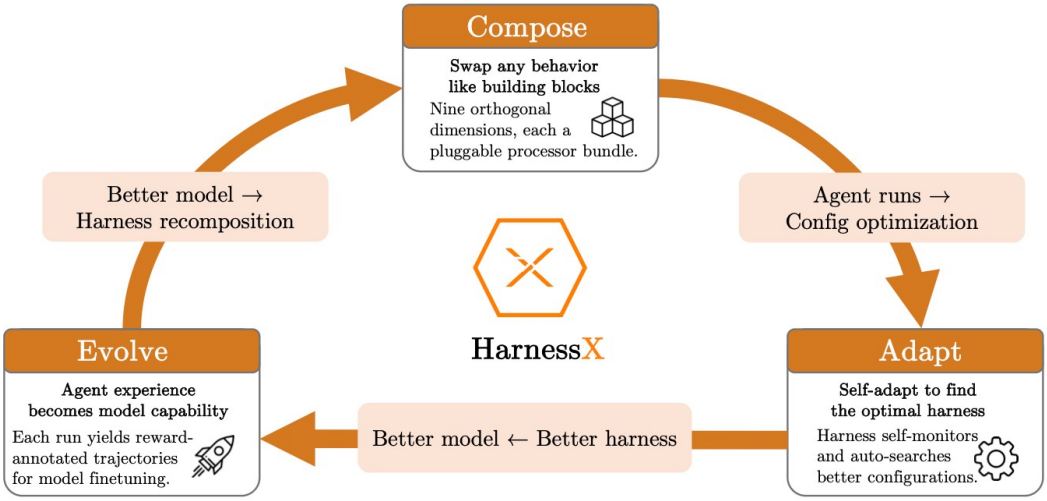


Figure 1 HarnessX overview.

主要亮点

- 把提示、工具、记忆、控制流和反馈机制抽象为可组合、可替换的 harness primitives。
- 通过 AEGIS 从失败轨迹中自动生成并验证 harness 修改，实现运行时结构进化。
- 展示不改模型参数也能提升 Agent 性能，补充了单纯模型 scaling 的改进路线。

相关链接

Paper： https://arxiv.org/abs/2606.14249

● 论文十五

ViT-Up: Faithful Feature Upsampling for Vision Transformers

日期

2026-06-12

ViT-Up: Faithful Feature Upsampling for Vision Transformers

Krispin Wandel, Jingchuan Wang, and Hesheng Wang

内容摘要

ViT-Up 研究如何为 Vision Transformer 生成忠实的高分辨率特征，以服务语义分割、语义对应和其他密集视觉任务。传统方法通常通过插值 patch token，或者借助浅层图像编码器补充高频纹理，但前者容易丢失边界与细节，后者则可能引入与主干特征空间不一致的纹理泄漏。ViT-Up 利用 ViT 中间层状态构造 layer-wise query，并在连续图像坐标上预测对应特征，使上采样结果尽量保持在原始 backbone 的表征空间之内。实验显示，该方法可提升 DINO 系列等 ViT backbone 在密集预测任务上的表现，并且随着 backbone 能力增强获得更明显收益。论文的核心贡献在于强调“忠实性”：上采样不应只是视觉上更平滑，而应保留 ViT 表征本身的语义结构。

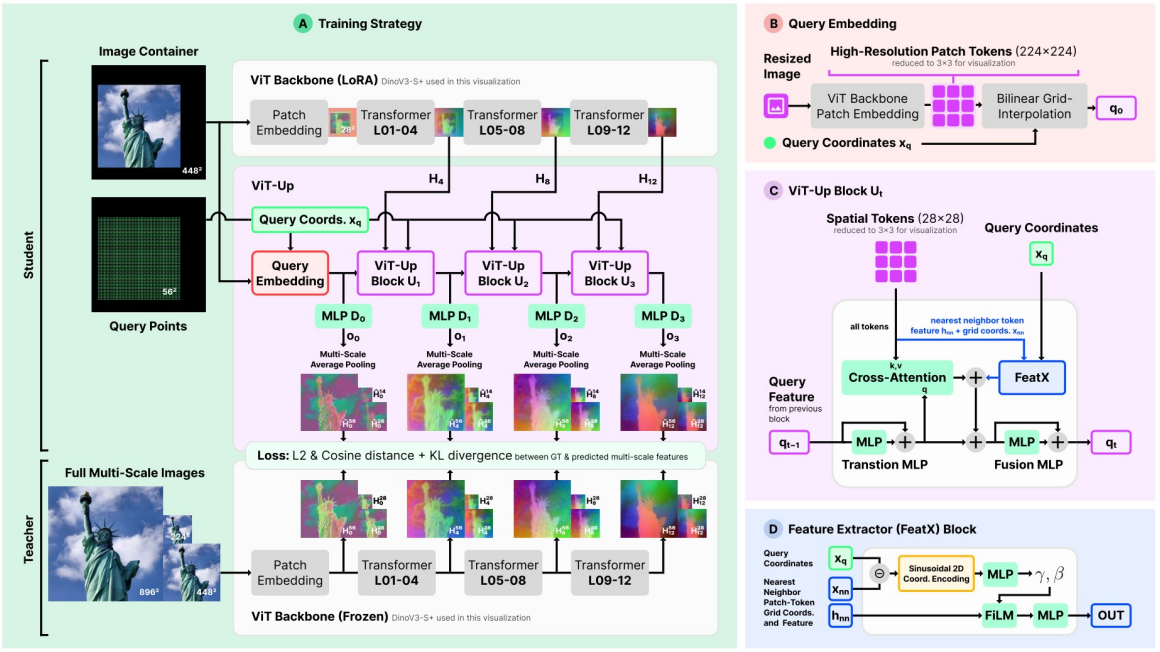


Fig. 2. Overview of ViT-Up. **(A)** Training strategy. A frozen teacher ViT is evaluated on the same image at resolutions 224×224, 448×448, and 896×896, providing feature targets at token-grid resolutions 14×14, 28×28, and 56×56. The student receives a downscaled version of the image pasted into a fixed-size image container, together with query coordinates sampled over the visible image region. A LoRA-adapted ViT backbone provides low-resolution intermediate hidden states to ViT-Up. For each sampled continuous coordinate, ViT-Up first constructs an initial query embedding and then refines it through T ViT-Up blocks that consume the corresponding backbone hidden states. For visualization, the figure shows three ViT-Up blocks consuming hidden states from layers 4, 8, and 12. In our DINOv3-S+/B main configurations, we use six ViT-Up blocks consuming hidden states from layers $\{2, 4, 6, 8, 10, 12\}$. The decoded query features are supervised at the finest teacher resolution and, after average pooling, at the corresponding coarser teacher token-grid resolutions using feature reconstruction losses and relational KL regularization. **(B)** Query embedding. The input image is processed with the ViT patch-embedding layer to obtain a 224×224 patch-token grid, which is bilinearly interpolated at the query coordinate x_q to initialize the query representation q_0. **(C)** ViT-Up block. Each block refines the previous query representation q_{t-1} by fusing token-level context from cross-attention with sub-token detail extracted from the nearest patch token using FeatX. **(D)** FeatX block. FeatX encodes the relative offset between the query coordinate and its nearest patch-token center, predicts FiLM parameters, and modulates the nearest-token feature to extract query-specific sub-token information.

主要亮点

- 提出忠实特征上采样目标，避免浅层图像引导造成纹理泄漏和边界碎片化。
- 利用 layer-wise query 与连续坐标建模，在任意图像位置预测 ViT 特征。
- 在分割和语义对应等密集视觉任务中提升 ViT backbone 表现，尤其适合强表征模型。

相关链接

Paper： https://arxiv.org/abs/2606.14024

Code： https://github.com/krispinwandel/vit-up

● 论文十六

VibeThinker-3B: Exploring the Frontier of Verifiable Reasoning in Small Language Models

日期

2026-06-15

Weibo AI

VibeThinker-3B: Exploring the Frontier of Verifiable Reasoning in Small Language Models

Sen Xu, Shixi Liu, Wei Wang, Jixin Min, Yingwei Dai, Zhibin Yin,
Yirong Chen, Xin Zhou, Junlin Zhang

Sina Weibo Inc.

内容摘要

VibeThinker-3B 是微博 AI 团队发布的 3B 级可验证推理小模型技术报告，关注小参数模型在数学、代码和可验证任务上的推理上限。报告提出 Spectrum-to-Signal 后训练流程，将课程化 SFT、多领域强化学习、离线自蒸馏和 claim-level 测试时扩展结合起来，使模型在可自动判分或可验证反馈的任务中逐步提升推理质量。与单纯扩大模型规模不同，该工作强调高质量后训练数据、可验证奖励和推理时多候选/声明级校验对小模型的重要性。实验显示，VibeThinker-3B 在 AIME、LiveCodeBench、LeetCode 等任务上达到接近甚至超过更大模型的结果，同时保持较好的指令遵循能力。该报告说明，在有足够强的后训练管线和验证信号时，小模型仍可能在特定推理领域达到很高能力密度。

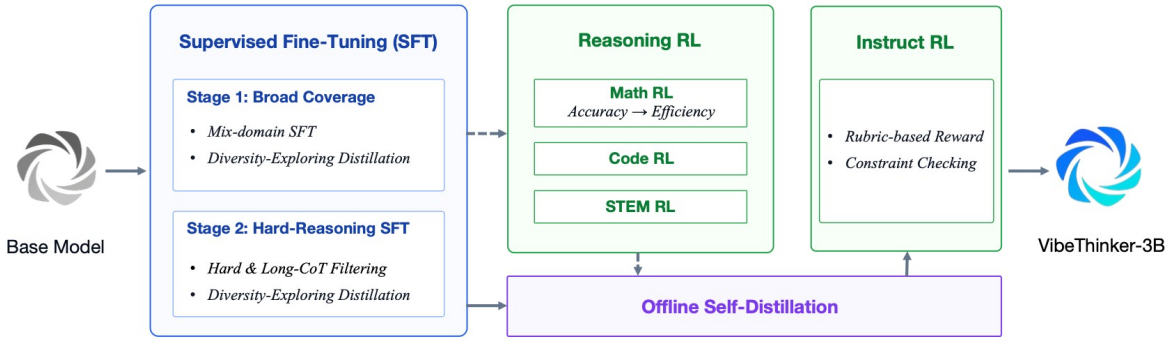


Figure 3: Overall training pipeline of VibeThinker-3B.

主要亮点

- 探索 3B 小模型在数学和代码等可验证推理任务上的能力边界。
- 结合课程化 SFT、RL、自蒸馏和 claim-level 测试时扩展，形成系统后训练流程。
- 在提升推理性能的同时保持指令遵循，展示小模型高能力密度的可能性。

相关链接

Paper： https://arxiv.org/pdf/2606.16140
Code: https://github.com/WeiboAI/VibeThinker
HuggingFace: https://huggingface.co/WeiboAI/VibeThinker-3B

● 论文十七

Moebius: 0.2B Lightweight Image Inpainting Framework with 10B-Level Performance

日期

2026-06-17 ECCV 2026 论文

Moebius: 0.2B Lightweight Image Inpainting Framework with 10B-Level Performance

Kangsheng Duan[1,*] Ziyang Xu[1,*,†] Wenyu Liu[1] Xiaohu Ruan[2]
Xiaoxin Chen[2] Xinggang Wang[1,✉]

[1] Huazhong University of Science and Technology
[2] VIVO AI Lab

内容摘要

Moebius 是一个约 0.22B 参数规模的轻量级图像修复框架，目标是在 inpainting 任务上以远小于 10B 级扩散模型的规模获得接近甚至更优的效果。论文提出 Local-λ Mix Interaction (LλMI) 模块，通过 Local-λ 和 Interactive-λ 在固定大小线性矩阵中汇聚局部与全局上下文，以增强小模型对遮挡区域和周围语义关系的建模能力；同时设计自适应多粒度潜空间蒸馏，将大模型的修复能力迁移到轻量模型中，避免直接在像素空间蒸馏带来的高成本。实验显示，Moebius 在参数量远少于 FLUX.1-Fill-Dev 等大模型的情况下获得相近或更好的视觉质量，并显著提升推理速度。该工作体现了图像生成任务中“专用轻量模型 + 蒸馏 + 高效交互模块”的路线。

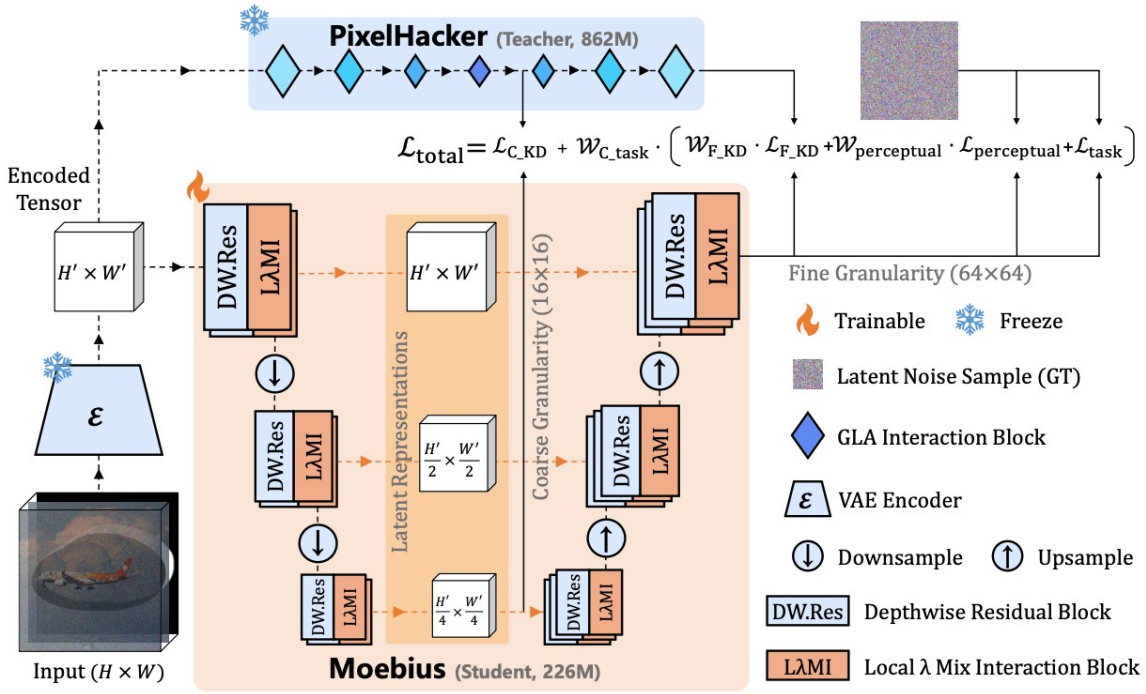


Fig. 1: Overall pipeline of Moebius. We adopt the Latent Diffusion Model (LDM) [32] framework equipped with Latent Categories Guidance (LCG) [54]. To achieve extreme architectural efficiency, the denoising U-Net is systematically restructured using our proposed $L\lambda MI$ blocks (detailed in Sec. 3.2). Furthermore, an adaptive multi-granularity distillation strategy (Sec. 3.3) is applied during training to align our lightweight specialist with the high-capacity teacher, successfully mitigating the capacity drop caused by extreme structural compression.

主要亮点

- 以约 0.2B 参数规模挑战 10B 级图像修复模型效果，突出轻量化专用生成模型价值。
- 提出 LλMI 高效交互模块，用固定大小矩阵建模局部与全局上下文。
- 采用自适应多粒度潜空间蒸馏，在降低训练/推理成本的同时保留大模型修复能力。

相关链接

Paper： https://arxiv.org/abs/2606.19195
Project: https://hustvl.github.io/Moebius

● 论文十八

Towards autonomous medical artificial intelligence agents

日期

2026-06-17 《Nature》论文

Article

Towards autonomous medical artificial intelligence agents

https://doi.org/10.1038/s41586-026-10675-5
Received: 29 April 2025
Accepted: 18 May 2026
Published online: 17 June 2026
Open access

Dyke Ferber[1,2], Lars Hilgers[2,3], Christiane Höper[2], Benedict Kinny-Köster[4,5], Jan-Niklas Eckardt[2,3], Katharina Egger-Heidrich[3], Marius Bill[3], Martin M. K. Schneider[3], Jan Clusmann[2,3,6,7], Lejla Kadric[8], Marcel Oehme[9], Maximilian Mayrhofer-Schmid[10,14], Alexander Oeser[11], Georg Wölflein[2], Isabella C. Wiest[2,6], Jan Moritz Middeke[2,3], A. John Iafrate[12], Daniel Truhn[13], Dirk Jäger[1] & Jakob Nikolas Kather[1,2,3]✉

内容摘要

这篇 Nature 论文提出 MIRA，一个面向医疗场景的自主 AI Agent，目标是从医学问答走向更接近真实临床流程的行动型系统。与只根据病例文本给出诊断建议的模型不同，MIRA 被放入沙盒化电子病历环境中，可以浏览患者病史，申请并解释实验室检查、影像和微生物结果，逐步形成诊断，并触发治疗、住院等临床动作。作者基于真实患者病例构建仿真评估，比较模型与医生在诊断准确性、指南一致性、用药安全、检查选择和收治决策等方面的表现。结果显示 MIRA 在若干指标上优于医生基线，但论文也强调，自动医疗 Agent 涉及高风险决策，实际临床部署仍需要前瞻性验证、责任边界、监管审查和严格安全治理。该工作标志着医疗 AI 评估从静态问答向闭环临床行动模拟迈进一步。

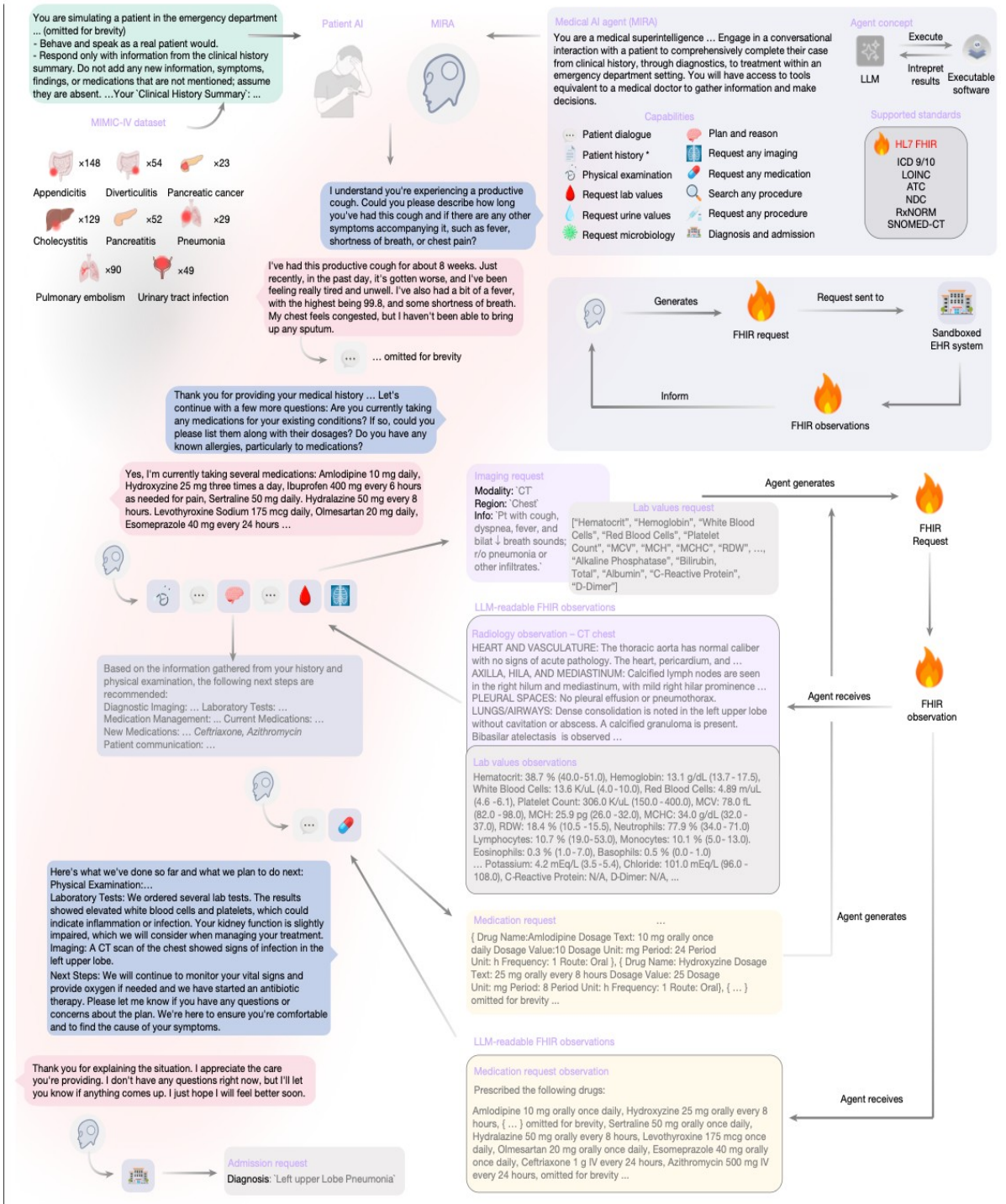


Fig. 1 | MIRA workflow. MIRA is an autonomous medical AI agent that operates within an EHR sandbox, using a suite of tools to simulate clinical workflows: it can order tests, synthesize results and produce diagnoses and treatment plans while interacting through chat with a patient AI agent that is grounded in the documented HPI extracted from retrospective notes from real cases. Left, exemplary conversation between patient and MIRA with interleaved tool calls. Right, the FHIR-based architecture that executes tool calls and records medical outputs. Note: data shown here are shortened and slightly modified to adhere to privacy restrictions of the dataset.

主要亮点

- 从医学问答扩展到行动型 Agent，能够在 EHR 环境中执行检查、诊断和治疗相关动作。
- 评估更贴近临床流程，不只看最终答案，还考察检查选择、用药安全和入院决策。
- 在展示自主医疗 Agent 潜力的同时，强调前瞻性验证、监管和责任边界的重要性。

相关链接

Paper： https://doi.org/10.1038/s41586-026-10675-5
Code: https://github.com/Dyke-F/MIRA

● 论文十九

DSpark: Confidence-Scheduled Speculative Decoding with Semi-Autoregressive Generation

日期

2026-06-27

DSpark: Confidence-Scheduled Speculative Decoding with Semi-Autoregressive Generation

Xin Cheng[1,2,*], Xingkai Yu[2,*], Chenze Shao[2,*], Jiashi Li[2,*], Yunfan Xiong[2,*]
Yi Qian[2], Jiaqi Zhu[2], Shirong Ma[2], Xiaokang Zhang[2], Jiasheng Ye[2], Qinyu Chen[2],
Chengqi Deng[2], Jiping Yu[2], Damai Dai[2], Zhengyan Zhang[2], Yixuan Wei[2], Yixuan Tan[2],
Wenkai Yang[2], Runxin Xu[2], Yu Wu[2], Zhean Xu[2], Xuanyu Wang[2], Muyang Chen[2],
Rui Tian[2], Xiao Bi[2], Zhewen Hao[2], Shaoyuan Chen[2], Huanqi Cao[2], Wentao Zhang[2],
Anyi Xu[2], Huishuai Zhang[1], Dongyan Zhao[1], Wenfeng Liang[2]
[1]Peking University [2]DeepSeek-AI
{chengxin, xingkai, shaochenze, js.li, yunfanxiong}@deepseek.com

内容摘要

DSpark 是 DeepSeek 与北京大学等团队提出的面向大模型推理加速的 speculative decoding 框架，核心目标是在不改变目标模型输出分布的前提下提升高并发服务中的生成速度。论文指出，现有并行 drafter 能一次提出长 token 块，但由于块内 token 缺乏依赖，后缀接受率快速衰减；同时，在高负载系统中无差别验证长草稿会浪费批处理容量。DSpark 用半自回归架构解决草稿质量问题：保留并行 backbone 以保持高吞吐，再加入轻量顺序模块建模块内依赖；同时引入置信度调度验证，根据每个请求的 prefix survival probability 和引擎吞吐曲线动态决定验证长度。离线实验中，DSpark 在多种目标模型和任务上提升平均接受长度；在线部署到 DeepSeek-V4 Flash/Pro 后，相比 MTP-1 生产基线在匹配吞吐下显著提升用户侧生成速度，并扩展了严格交互 SLA 下的服务 Pareto frontier。

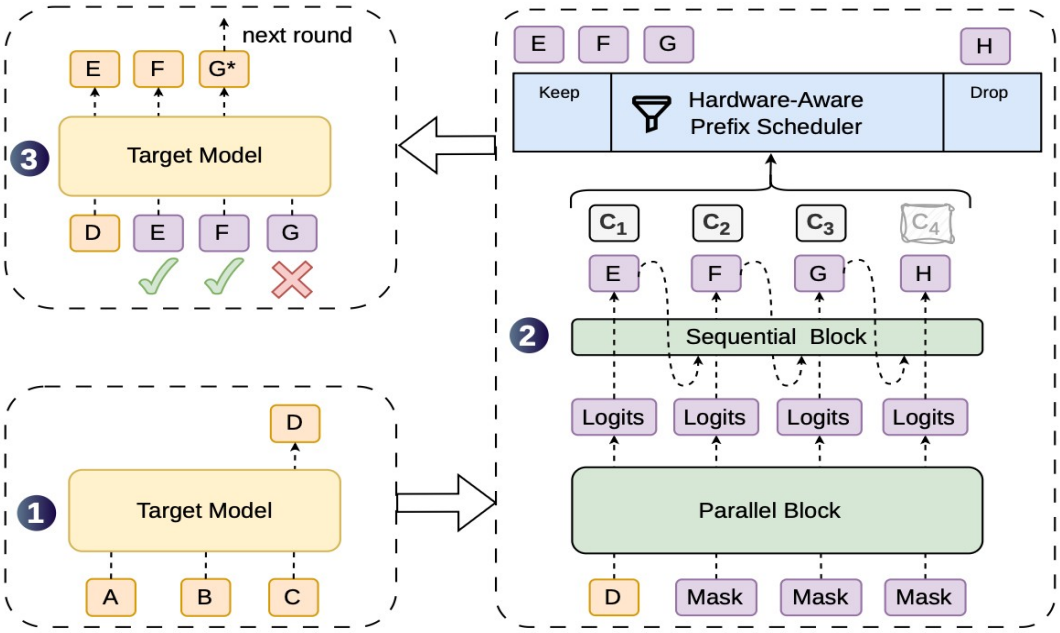


Figure 1 | **The DSpark architecture and decoding cycle.** Given prompt tokens ABC, the target model executes one step to generate the next token D, which serves as the anchor for the drafting phase. Using D as the input, DSpark employs a heavy parallel backbone and a lightweight sequential head to generate draft tokens EFGH along with their corresponding confidence scores c_1–c_4. The Hardware-Aware Prefix Scheduler then evaluates these scores to retain the prefix EFG and drop the low-confidence token H. Finally, the target model verifies the scheduled prefix in parallel. As illustrated, E and F are accepted while G is rejected, prompting the model to generate a corrected token G^* to complete the current round.

主要亮点

- 提出半自回归 drafter，在并行生成速度和 token 间依赖建模之间取得折中，缓解后缀接受率衰减。
- 引入置信度调度验证，根据请求级存活概率和系统吞吐状态动态裁剪低价值验证 token。
- 在 DeepSeek-V4 生产服务中验证收益，并开源 DSpark checkpoints 与 DeepSpec 训练工具链。

相关链接

Paper： https://github.com/deepseek-ai/DeepSpec/blob/main/DSpark_paper.pdf
Code: https://github.com/deepseek-ai/DeepSpec
Models: https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro-DSpark

TECHNICAL UPDATES
技术动态

技术动态一

OpenAI 推进 ChatGPT 史上最大改版 从单一聊天机器人迈向 AI 超级应用

日期

2026-06-07



内容摘要

OpenAI 即将对 ChatGPT 进行自 2022 年上线以来规模最大的一次升级，推动其从单一问答式聊天机器人转型为集成编程工具、人工智能体和任务执行能力的“超级应用”。此次改版将重点加码 Codex 编程产品线，强化代码生成、企业协同、SDK 调用和管理工具等能力，使 ChatGPT 不仅能回答问题，也能自主完成更复杂的开发、办公和业务流程任务。在 AI 模型训练和推理成本持续高企、用户增长逐渐放缓的背景下，OpenAI 正通过 Codex、企业客户销售和广告测试等方式拓展营收渠道，为未来上市和盈利路径打基础。整体来看，此次改版标志着 OpenAI 的产品重心正在从“更强模型”转向“更高频使用、更强任务闭环和更清晰商业变现”，也反映出 AI 行业竞争正进入智能体应用与商业化落地的新阶段。

相关链接

https://finance.sina.com.cn/jjxw/2026-06-07/doc-iniaqpen9606317.shtml

• 技术动态二

Anthropic 发布 Claude Fable 5 与 Mythos 5 强化长程智能体与高风险能力安全释放

日期

2026-06-09



内容摘要

Anthropic 正式发布 Claude Fable 5 与 Claude Mythos 5，进一步提升 Claude 系列模型在长程任务执行、软件工程、知识工作、视觉理解、科研推理和生命科学研究等场景中的能力。其中，Fable 5 面向通用用户开放，强调复杂任务规划、生产级代码处理、长上下文理解和持久化记忆能力；Mythos 5 则面向网络安全防御者、关键基础设施提供方及部分可信科研机构受限开放，用于支持网络安全、药物设计、蛋白设计、分子生物学假设生成等高价值场景。此次发布的重点不只是模型能力提升，也在于 Anthropic 对高风险能力采取分级释放与安全分类机制，体现出前沿 AI 模型正在从单纯性能竞争转向“能力释放、安全治理与专业场景落地并重”的新阶段。

相关链接

https://www.anthropic.com/news/claude-fable-5-mythos-5

• 技术动态三

美国升级前沿 AI 出口管制 Anthropic Fable 5 与 Mythos 5 陷入监管风波

日期

2026-06-13



内容摘要

Anthropic 最新模型 Claude Fable 5 与 Claude Mythos 5 发布后迅速引发监管风波。美国政府以国家安全为由，对两款前沿 AI 模型实施出口管制，要求 Anthropic 暂停外国政府、企业、个人以及美国境内外国人的访问权限，使其被纳入类似高端芯片和敏感软件的管制框架。此次事件的直接背景，是监管部门对 Mythos 5 潜在越狱风险和高能力模型外溢风险的担忧，也反映出美国正在将最先进 AI 系统视为重要战略资产进行管理。对 Anthropic 而言，其一方面因模型能力过强受到政府审查，另一方面又需要在商业化、用户访问和合规要求之间寻找平衡。整体来看，Fable 5 与 Mythos 5 风波标志着前沿 AI 竞争正在从模型能力竞赛，进一步延伸到国家安全、出口管制和全球技术治理层面。

相关链接

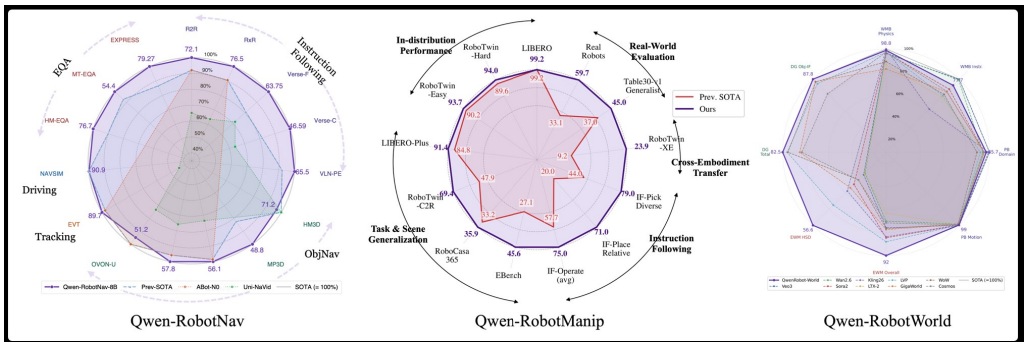
https://www.reuters.com/technology/us-blocks-foreign-access-anthropics-most-advanced-ai-models-axios-reports-2026-06-13/

• 技术动态四

Qwen-Robot 发布 打通大模型走向物理世界的行动闭环

日期

2026-06-15



内容摘要

通义千问团队正式发布 Qwen-Robot 具身智能基础模型系列，面向机器人从“看懂世界”到“真实行动”的核心鸿沟，构建了覆盖导航、操作和世界建模的三类专业模型：Qwen-RobotNav 负责指令跟随、目标搜索、目标追踪和自动驾驶等导航任务，Qwen-RobotManip 通过统一状态—动作表示和跨本体对齐机制支持机械臂、双臂、灵巧手等多类机器人操作，Qwen-RobotWorld 则以自然语言作为统一动作接口，预测符合物理规律的未来场景。该系列模型通过模块化协同，将多模态理解、物理行动和世界动力学预测连接起来，并可进一步作为智能体系统中的物理世界工具调用。整体来看，Qwen-Robot 的发布标志着大模型能力正在从语言、视觉和代码场景进一步延伸到真实物理世界，也为具身智能在导航、抓取、环境探索和开放式任务执行中的规模化落地提供了新的技术路径。

相关链接

https://qwen.ai/blog?id=qwen-robotsuite

技术动态五

微博 AI 发布 VibeThinker-3B 探索小模型可验证推理能力边界

日期

2026-06-17

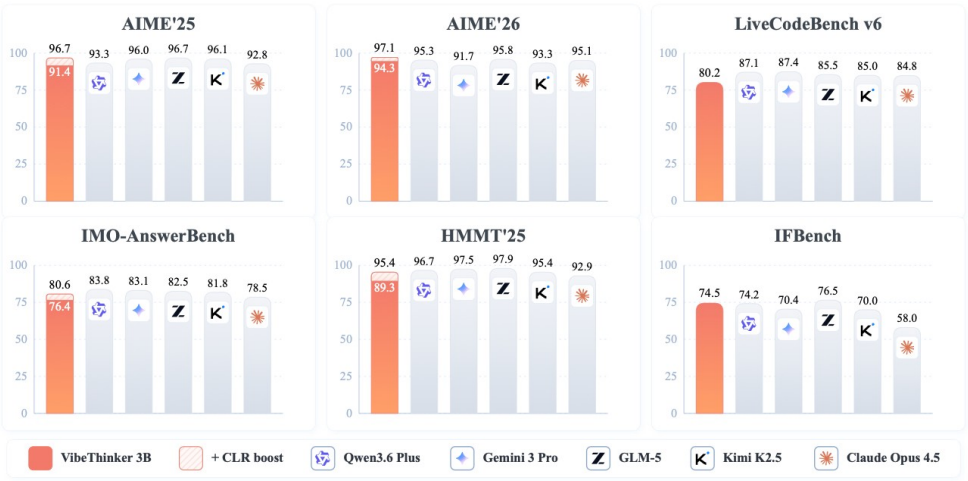


Figure 1: VibeThinker-3B reaches frontier reasoning performance at 3B scale. CLR denotes Claim-Level Reliability Assessment, a claim-level test-time scaling strategy.

内容摘要

微博 AI 发布 VibeThinker-3B 技术报告与开源模型，尝试在仅 3B 参数规模下逼近前沿推理模型能力边界。该模型基于 Qwen2.5-Coder-3B，通过课程式监督微调、多领域强化学习、离线自蒸馏和指令对齐等后训练流程，重点强化数学、代码和 STEM 等可验证任务中的复杂推理能力。报告显示，VibeThinker-3B 在 AIME26、LiveCodeBench v6、HMMT25 和 LeetCode 近期竞赛等任务上取得较强表现，并在测试时引入 Claim-Level Reliability Assessment 进一步提升答案可靠性。其核心意义不在于证明小模型可以全面替代大模型，而在于提出“可验证推理能力可以被压缩进紧凑推理核心”的思路，表明小模型在结构化推理、约束满足和可验证任务中仍具备重要研究价值，也为低成本部署、学术研究和垂直场景推理模型提供了新的技术方向。

相关链接

https://arxiv.org/pdf/2606.16140

• 技术动态六

OpenAI 发布全新网络安全模型 GPT-5.5-Cyber 表现超越 Anthropic Mythos

日期

2026-06-22



内容摘要

OpenAI 正式发布全新网络安全专用模型 GPT-5.5-Cyber，面向漏洞发现、漏洞验证、补丁生成、攻击路径分析和威胁建模等高价值防御场景提供更强的智能化安全能力。作为 Daybreak 网络安全计划的核心模型，GPT-5.5-Cyber 在多项网络安全评测中取得领先表现，并被 OpenAI 称为在关键能力上超越 Anthropic Mythos，显示出前沿 AI 模型正在加速进入专业安全基础设施领域。相比通用大模型，GPT-5.5-Cyber 更强调授权环境下的可控安全能力释放，配合 Codex Security、Daybreak Cyber Partner Program 和 Patch the Planet 等机制，将模型能力嵌入代码审计、漏洞修复和开源生态防护流程。整体来看，GPT-5.5-Cyber 的发布不仅意味着 OpenAI 正式加码网络安全垂直模型，也反映出 AI 巨头竞争正从通用智能、编程助手进一步延伸到高风险、高价值的网络攻防能力体系。

相关链接

https://arxiv.org/pdf/2606.16140

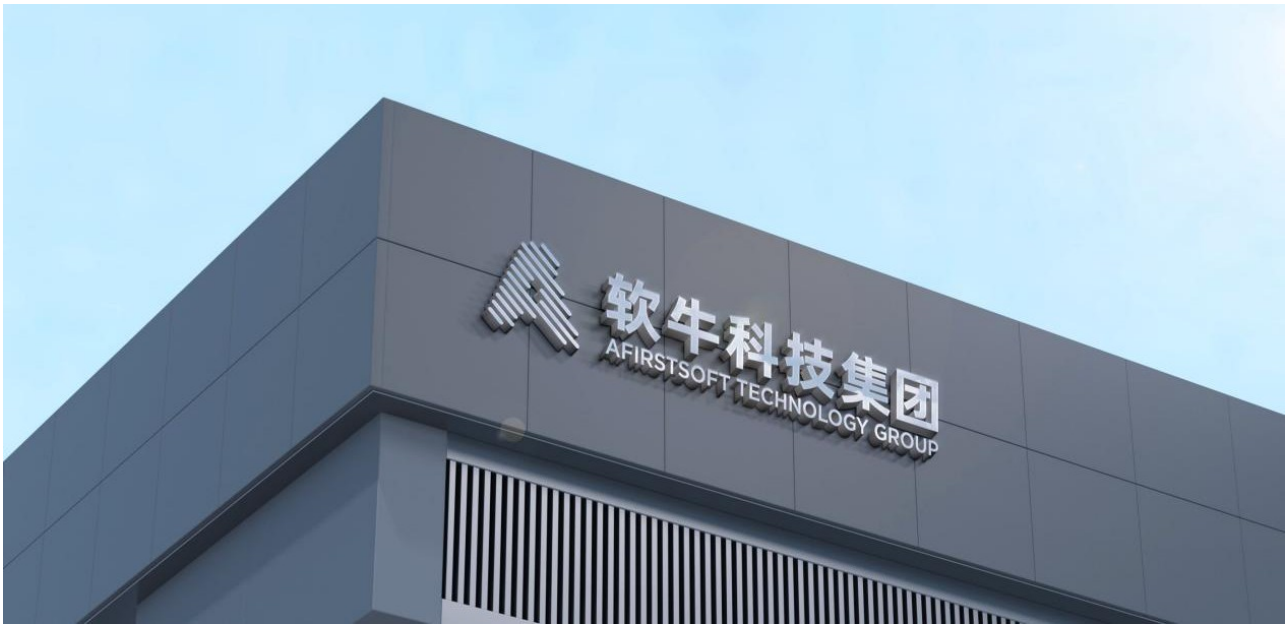


深圳软牛科技集团股份有限公司成立于 2014 年 6 月，作为国家级专精特新“小巨人”企业，始终以“让创意成为现实”为使命，专注于通过人工智能视觉大模型技术驱动全球数字创意生态。

公司深度聚焦图像修复、视频增强、多模态内容生成等视觉大模型核心技术，联合西安电子科技大学、香港中文大学（深圳）等顶尖高校，攻克了视频抠像、画质修复等难题，并与深圳大学共建“人工智能技术联合创新中心”，推动 AI 赋能的图像增强技术规模化落地。

目前，软牛科技能够为广电机构、影视制作、工业检测、农业测报等多个行业提供专业解决方案。从微米级工业精密测量，到食品异物智能检测；从农业害虫 AI 识别防治系统，到 AR 设备巡检实时画面 AI 增强与场景识别；从 VR 全息空间建模助力文博展陈数字化升级，到跨终端智能影像协作软件的全链路画质优化引擎，软牛科技正以技术革新重新定义新质生产力。

在消费级市场，旗下牛小影（https://www.niuxuezhang.cn/video-enhancer.html）、HitPaw 等海内外知名品牌，以“一键式 AI 视觉创意”为核心，为全球超 1 亿用户提供视频修复、老电影 AI 上色等场景化服务。例如，牛小影支持 4K 或 8K 超高清修复技术，帮助家庭用户重现祖辈黑白影像的生动细节；HitPaw Edimakor 通过 AI 语音合成、多语言实时翻译等功能，让自媒体创作者轻松实现跨语言内容生产。



这些创新成果的背后，是公司每年近亿元的研发投入和占比超 50%的顶尖技术团队，以及 CMMI、ISO56005 等国际权威认证体系支撑的硬核实力。立足深圳总部，我们通过北京、新加坡、香港等战略支点构建起辐射全球 200 多个国家和地区的智慧服务网络。作为高新技术企业、深圳市潜在科技独角兽，软牛科技始终以“技术无界”为理念，持续拓展 AIGC+产业应用的边界。

未来，软牛科技将加速推进物理世界与 AI 视觉的无缝衔接，让每个人都能通过指尖的艺术级创作，开启属于自己的数字时代。

人工智能前沿快报

Artificial Intelligence Newslettler

2026 年第 6 期

总第 32 期

广东省图象图形学会

主办 广东省图像图形学会

本期编委

金连文	华南理工大学
谭明奎	华南理工大学
钟亦武	北京大学
林上港	华南农业大学
李　斌	深圳大学
谢晓华	中山大学
王泽宇	香港科技大学（广州）
李冠彬	中山大学
熊龙飞	金山办公
段纪伟	金山办公
孙　成	金山办公
胡晋武	华南理工大学
王宇丰	华南理工大学
王骞玥	华南理工大学
李成昊	华南理工大学
许毅平	华南农业大学
陈嘉杜	华南农业大学



人工智能前沿快报

2026 年第 6 期
总第 32 期
2026.07.01

声明：

（1）本快报内容提要、文章亮点等部分内容由 AI 生成，不一定准确及全面，章完整内容及思想应以原文为准。

（2）本文大部分插图来源于相关论文或文章，版权归原作者或原出版社所有。

（3）本快报摘录文章观点不代表编委会及主办方立场。

学会官方公众号

广东省图象图形学会
GDSIG
学术交流、技术咨询、科学普及