Artificial Intelligence Newsletter

人工智能
前沿快报

2025 年第 07 期
总第 24 期

学会官方公众号

广东省图象图形学会

GDSIG

学术交流、科普宣传、科技成果鉴定

目 录

学 术 动 态

一. Causal-Copilot: An Autonomous Causal Analysis Agent……1

二. MegaHan97K: A Large-Scale Dataset for Mega-Category Chinese Character Recognition with over 97K Categories……4

三. Towards AI Search Paradigm……6

四. A Survey of AIOps in the Era of Large Language Models……8

五. Energy-Based Transformers are Scalable Learners and Thinkers……10

六. AI4Research: A Survey of Artificial Intelligence for Scientific Research……12

七. GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning…14

八. AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench……16

九. A Survey on Latent Reasoning……18

十. Mixture-of-Recursions: Learning Dynamic Recursive Depths for Adaptive Token-Level Computation……20

十一. RoboBrain 2.0 Technical Report……22

十二. Dynamic Chunking for End-to-End Hierarchical Sequence Modeling……24

十三. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities……26

十四. Learning without training: The implicit dynamics of in-context learning……28

十五. Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos……30

十六. A Survey of Context Engineering for Large Language Models……32

十七. Gemini 2.5 Pro Capable of Winning Gold at IMO 2025……35

十八. AlphaGo Moment for Model Architecture Discovery……37

目 录

技 术 动 态

一. Kimi K2: Open Agentic Intelligence……39
二. 《Nature》发文赞美 Kimi K2 是另一个 DeepSeek 时刻……41
三. OpenAI 神秘新模型斩获 IMO2025 金牌……42
四. 英伟达开源了 OpenReasoning-Nemotron……43
五. 全球首个 IMO 金牌 AI 诞生！谷歌 Gemini 拿下 35 分碾碎奥数神话……44
六. Qwen3-Coder: Agentic Coding in the World 阿里巴巴发布开源大模型 Qwen3-Coder，性能明显超越 Kimi-K2、DeepSeek V3、GPT-4.1……45
七. DeepRare 重磅发布：全球首个可循证智能体诊断系统，直击医学 Last Exam 难题……46
八. 千问发布多语言翻译模型 QWen-MT，支持 92 种语言……48
九. 前 Meta 员工正式推出其首款大型视觉记忆模型；上海 AI Lab 发布 241B 多模态大模型 Intern-S1，性能超越 InterVL3， QWen2.5-VL……50
十. 智谱发布大模型 GLM 4.5，综合性能超过 Claude 4、QWen3……52
十一. 阿里巴巴开源视频生成 MOE 大模型 Wan 2.2……54
十二. Qwen3-30B-A3B-Instruct-2507 发布，3B 激活参数 3090 可运行……55

资 源 分 享

一. 《Foundations of Large Language Models》……56

鸣 谢 支 持

一. 深圳软牛科技集团股份有限公司……57

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 18 篇最新学术动态、12 篇技术动态、以及 1 个资源推荐给广大读者。

ACADEMIC TRENDS
学术动态

● 论文一

Causal-Copilot: An Autonomous Causal Analysis Agent

日期

2025-04-21

Causal-Copilot: An Autonomous Causal Analysis Agent

Xinyue Wang* XIW159@UCSD.EDU
Kun Zhou* FRANCISKUNZHOU@GMAIL.COM
Wenyi Wu* WEW058@UCSD.EDU
Har Simrat Singh H6SINGH@UCSD.EDU
Fang Nan FNAN@UCSD.EDU
Songyao Jin SOJ007@UCSD.EDU
Aryan Philip ARPHILIP@UCSD.EDU
Saloni Patnaik SAPATNAIK@UCSD.EDU
Hou Zhu HOZ020@UCSD.EDU
Shivam Singh SHS046@UCSD.EDU
Parjanya Prashant PPRASHANT@UCSD.EDU
Qian Shen Q1SHEN@UCSD.EDU
Biwei Huang BIH007@UCSD.EDU
University of California San Diego, CA 92093, USA

内容摘要

Causal- Copilot 是一款 LLM 驱动的自主因果分析代理，面向表格与时序数据，能够自动完成因果发现、因果推断、算法筛选、超参数优化、结果解释与可操作洞见等全流程工作。系统将 20 余种先进因果分析技术封装在统一框架中，并允许用户通过自然语言进行交互式细化，从而大幅降低非专业人士使用门槛，同时为因果研究者提供真实场景反馈，实现方法与应用的良性循环。实验结果显示，该代理在多种复杂数据场景下整体性能显著优于传统算法与无上下文 LLM 基线，兼顾可靠性、可扩展性与易用性。

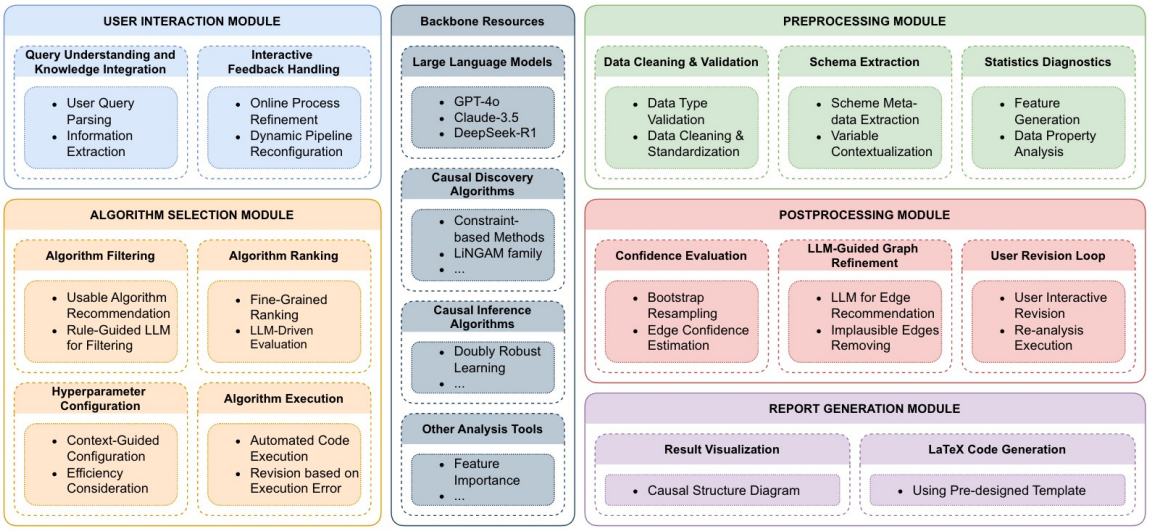


Figure 1: The overall architecture of our Causal-Copilot.

主要亮点

- **端到端自动化**：从数据预处理、算法选择到报告生成均由代理自主完成，并支持自然语言微调与检查，真正实现“一键式”因果分析。
- **算法库丰富且加速友好**：内置 >20 种因果发现、推断与辅助工具，涵盖线性/非线性、表格/时序场景，并提供 CPU/GPU/并行加速方案以保证大规模数据处理效率。
- **LLM- 驱动智能决策**：融合因果知识记忆与统计诊断，自动为每个数据集制定算法与参数策略，同时通过统计置信度 + LLM 推理 + 用户反馈三重机制校正结果；在合成与真实复杂场景中均显著超越 GPT- 4o 及传统方法。

相关链接

Paper: https://arxiv.org/pdf/2504.13263

Github: https://github.com/Lancelot39/Causal-Copilot

Demo： https://causalcopilot.com/

● 论文二

MegaHan97K: A Large-Scale Dataset for Mega-Category Chinese Character Recognition with over 97K Categories

日期

2025-06-05（PR 录用论文）

MegaHan97K: A Large-Scale Dataset for Mega-Category Chinese Character Recognition with over 97K Categories

Yuyi Zhang[†a,b], Yongxin Shi[†a], Peirong Zhang[a], Yixin Zhao[a], Zhenhua Yang[a], Lianwen Jin[*a,b]

[a]*School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China.*
[b]*SCUT-Zhuhai Institute of Modern Industrial Innovation, Zhuhai, China*

内容提要

MegaHan97K 是首个覆盖 GB18030- 2022 全部 97 455 类中文字符的超大规模数据集，包含手写、历史文献与合成三种子集，总计 461 万样本，其中约 331 万由 FontDiffuser 自动生成并经古籍风格增强，有效缓解稀有字与长尾分布问题。作者通过众包书写（94 位志愿者、900 k 手写样本）与网页爬取构建历史子集，并为每类字符提供平均 35 个多风格样本，用于模型训练与零样本评估。系统基准显示，现有 CCR 与 ZSL 方法在 MegaHan97K 场景下精度显著下降并面临存储膨胀、相似字混淆等新挑战；而将该数据集与 HWDB、AHCDB 等联合训练，可显著提升模型对稀有、复杂字符的泛化能力，凸显其作为“97K 级”评测平台与文化遗产数字化资源的价值。

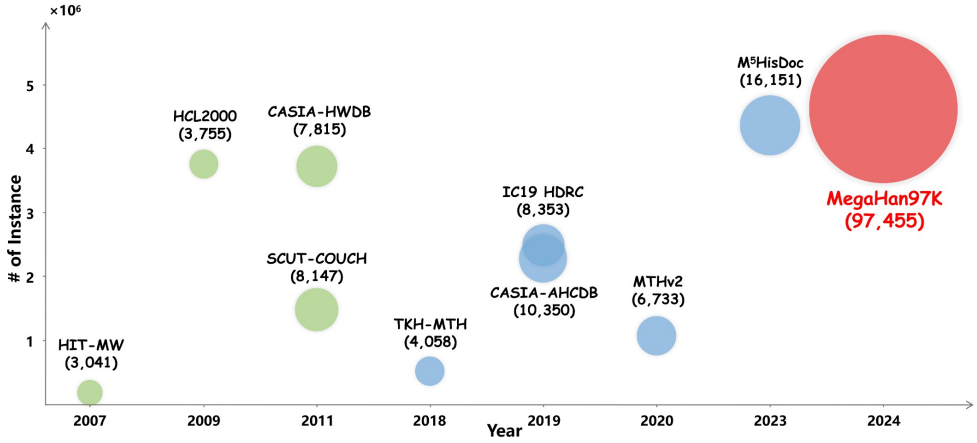


Figure 1: Comparing MegaHan97K with existing Chinese character datasets. The green and blue bubbles represent handwritten and historical datasets, respectively. The area of the bubble and the number in brackets denote categories in each dataset.

主要亮点

- **规模空前、标准完整：**首个全面支持 GB18030- 2022 的中文字符数据集，类别数较过去最大数据集扩大 6 倍，覆盖 97 455 类。
- **三子集 + 合成字体缓解长尾：**手写、历史、合成三源数据结合 FontDiffuser 一次生成 35 样本/类并做古籍风增强，大幅平衡类别分布并提升训练数据多样性。
- **基准与迁移实验揭示新挑战与价值：**在存储、零样本识别与相似字区分方面提出新难题，同时跨数据集联合训练显著提高模型对稀有字的鲁棒性，验证其研究与应用潜力。

相关链接

Paper: https://arxiv.org/pdf/2506.04807

Github: https://github.com/SCUT-DLVCLab/MegaHan97K

● 论文三

Towards AI Search Paradigm

日期

2025-06-20

Towards AI Search Paradigm

Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, Changle Qu, Keyi Kong, Wenwen Ye, Lixin Su, Xinyu Ma, Long Xia, Daiting Shi, Jiashu Zhao, Haoyi Xiong, Shuaiqiang Wang, Dawei Yin†

Baidu Search

{yuchenli13, yindawei}@acm.org

内容提要

本文系统性地提出并阐述了 AI Search 范式，即基于大语言模型（LLMs）进行信息检索与问题解决的新型范式。不同于传统搜索引擎仅提供信息片段，AI Search 聚焦于以智能体方式主动完成任务，整合搜索、理解、生成与行动能力。作者明确 AI Search 的定义、关键能力（包括指令跟随、多轮对话、Web 交互、信息记忆等）、评估维度与挑战，并建立了统一的任务框架与基准平台（AgentS）对现有方法进行全面评估。此外，论文还回顾了相关研究现状，总结了 AI Search 的主要发展方向与未来前景。

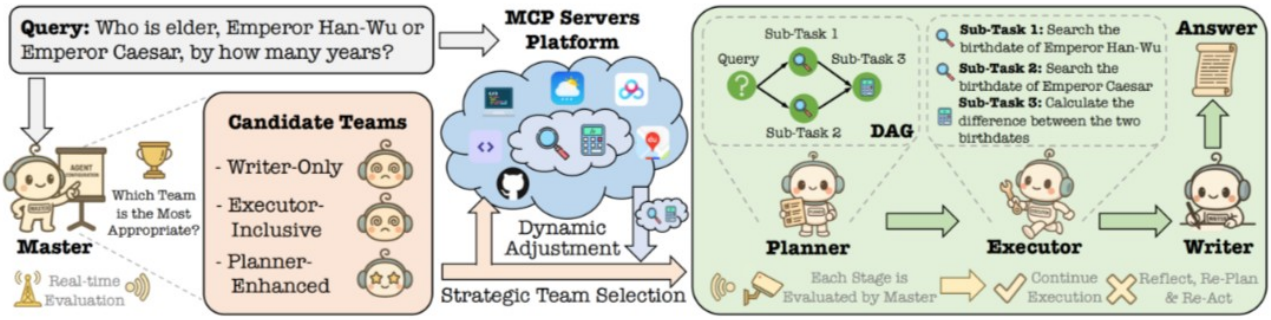


Figure 1 | The Overview of AI Search Paradigm.

主要亮点

- **系统定义 AI Search 范式：**首次明确提出 AI Search 的任务边界与核心能力，填补了领域内概念模糊的空白。
- **构建统一评估平台 AgentS：**涵盖 10 类任务、460+搜索目标，支持自动评测与用户交互，系统性衡量模型表现。
- **全景式回顾现有方法：**从 prompt-based 到 agent-based 系统，全面总结 AI Search 的模型演进与技术路径。
- **提出未来挑战与方向：**包括多模态检索、长期记忆、Web 智能交互等，为后续研究指明路线。

相关链接

Paper: https://arxiv.org/pdf/2506.17188

● 论文四

A Survey of AIOps in the Era of Large Language Models

日期

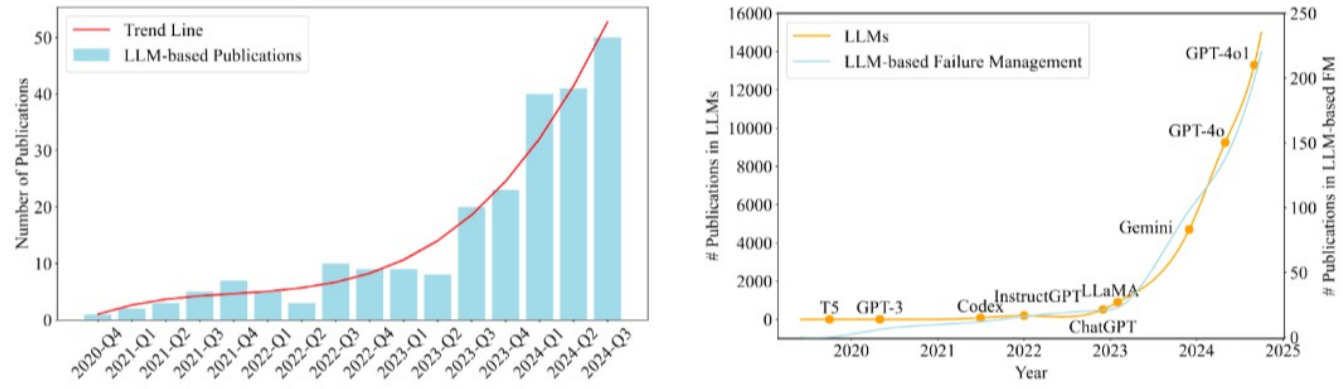
2025-06-23

A Survey of AIOps in the Era of Large Language Models

LINGZHE ZHANG, Peking University, China
TONG JIA*, Peking University, China
MENGXI JIA, Peking University, China
YIFAN WU, Peking University, China
AIWEI LIU, Tsinghua University, China
YONG YANG, Peking University, China
ZHONGHAI WU, Peking University, China
XUMING HU, The Hong Kong University of Science and Technology (Guangzhou), China
PHILIP S. YU, University of Illinois Chicago, United States
YING LI*, Peking University, China

内容提要

本文是首个针对“大模型时代 AIOps”全流程的系统综述，系统检索并分析了 2020 年 1 月-2024 年 12 月间发表的 183 篇研究，围绕四大研究问题（RQ1-RQ4）展开：① LLM 如何改造传统日志/指标等故障数据来源并引入软件文档、源代码、QA 等新数据；② AIOps 任务如何演进——从故障感知扩展到根因报告生成、脚本自动执行等新子任务；③ LLM 方法学的五大范式（基础模型、精调、嵌入、提示、知识增强）在各任务中的适用性与局限；④ 评估指标与数据集的更新趋势。作者还总结了当前研究的最新进展、空白与挑战，并给出未来发展方向。



(a) Number of LLM-based publications in the field of AIOps

(b) Number of publications in the field of LLM-based AIOps and LLMs (the data for "# Publications in LLMs" is extended based on [171])

Fig. 1. Analysis of Publication Trends in LLM-Based AIOps

主要亮点

- **覆盖 AIOps 全链路**：首次同时梳理“数据预处理→任务演进→方法体系→评测标准”四大维度，为学界和工业界提供一站式参考框架。
- **大规模定量元分析**：基于 183 篇论文的出版趋势与任务分布，提出包含 20 + 新任务/数据源的细粒度分类与方法图谱。
- **提出四大研究挑战**：指出算力-成本权衡、数据多样利用、模型泛化与软件演进适应、与现有工具链融合等关键难题，给出未来研究路线图。

相关链接

Paper: https://arxiv.org/abs/2507.12472

● 论文五

Energy-Based Transformers are Scalable Learners and Thinkers

日期

2025-07-02

Energy-Based Transformers are Scalable Learners and Thinkers

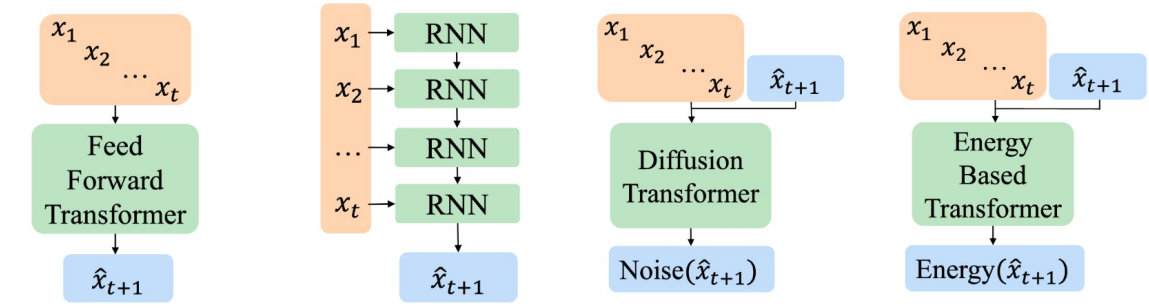
Alexi Gladstone[1,2], Ganesh Nanduru[1], Md Mofijul Islam[1,3], Peixuan Han[2],
Hyeonjeong Ha[2], Aman Chadha[3,4], Yilun Du[5], Heng Ji[3], Jundong Li[1], Tariq Iqbal[1]

[1]UVA [2]UIUC [3]Amazon GenAI† [4]Stanford University [5]Harvard University

energy-based-transformers.github.io github.com/alexiglad/EBT

内容提要

本 Energy- Based Transformers (EBT) 将每个“上下文- 候选预测”对映射为能量分数，并通过梯度下降不断最小化能量，从而在纯无监督预训练中让“System 2 Thinking（长推理 + 自验证）”自然涌现。在文本与视觉两种模态上，EBT 训练收敛速度均显著快于改进版 Transformer++：在数据量、参数量、FLOPs、深度等六条轴上最大可提升 35% 的 scaling rate。推理阶段若增加思考步数并执行能量最小值选择，语言任务可再获得 29% 的性能增益，而标准 Transformer++ 完全无收益。同时，EBT 对分布外数据提升更大，体现出更强泛化。在图像去噪等连续任务上，EBT 仅用 1% 前向次数即可超越 Diffusion Transformer，并取得更高 PSNR 与十倍以上的 ImageNet 线性探测准确率



(a) AR Transformer (b) RNN (c) Diffusion Transformer (d) EBT

Existing Autoregressive Approaches

Figure 1: **Autoregressive Architecture Comparison.** (a) Autoregressive (AR) Transformer is the most common, with (b) RNNs becoming more popular recently [22, 23]. (c) Diffusion Transformers [26] (which are often bidirectional but can also be autoregressive) are the most similar to EBT, being able to dynamically allocate computation during inference, but predict the noise rather than the energy [27, 28]. Consequently, diffusion models cannot give unnormalized likelihood estimates at each step of the thinking process, and are not trained as explicit verifiers, unlike EBTs.

主要亮点

- **无监督 System 2 推理机制**：利用显式能量函数 + 梯度优化，模型自行决定思考步数与自验证策略，无需额外监督或奖励设计。
- **全轴向可扩展性**：相较 Transformer++，在数据、批量、深度、参数、FLOPs 等多维度 scaling 率最高提升 35%，首度表现出更佳数据与算力效率。
- **推理与泛化双优**：思考更久 + 自验证带来语言任务 29% 额外收益，并对 OOD 数据改进更显著；在图像去噪仅用 1% 计算量即超过 DiT，显示出对连续模态的卓越泛化与表征能力。

相关链接

Paper: https://arxiv.org/abs/2507.02092

Github: https://github.com/alexiglad/EBT

● 论文六

AI4Research: A Survey of Artificial Intelligence for Scientific Research

日期

2025-07-02

AI4Research: A Survey of Artificial Intelligence for Scientific Research

Qiguang Chen[1*] Mingda Yang[1*] Libo Qin[2,✉] Jinhao Liu[1] Zheng Yan[1] Jiannan Guan[1]
Dengyun Peng[1] Yiyan Ji[1] Hanjing Li[1] Mengkang Hu[3] Yimeng Zhang[4] Yihao Liang[4]
Yuhang Zhou[5] Jiaqi Wang[6] Zhi Chen[7] Wanxiang Che[1,✉]

[1] LARG, Research Center for Social Computing and Interactive Robotics, Harbin Institute of Technology,
[2] School of Computer Science and Engineering, Central South University, [3] The University of Hong Kong,
[4] Independent Researcher, [5] Fudan University, [6] Chinese University of Hong Kong,
[7] ByteDance Seed (China)

内容提要

本文首次对“AI4Research”——人工智能赋能全流程科研——进行系统综述。作者基于近年 LLM 在逻辑推理、实验编程等方面的突破，提出一个覆盖**科学理解、学术综述、科学发现、论文写作、同行评议**五阶段的统一框架；并从功能组成、目标公式到评价指标对各子模块加以正式化描述，阐明 AI 在提升科研效率、创新能力与可重复性方面的作用。文章进一步梳理跨自然科学、工程技术与社会科学的典型应用，汇集数据集、工具链与开放平台等丰富资源，并对跨学科大模型、动态实验、伦理与安全、可解释性等前沿方向进行展望，旨在为研究者提供快速入口与未来研究指南。

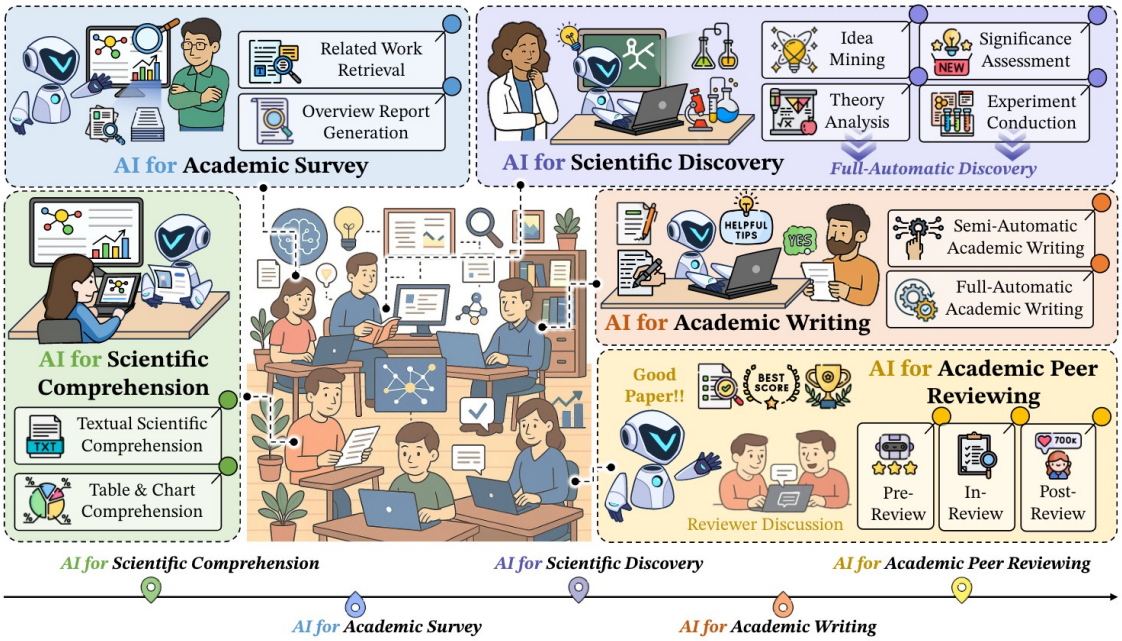


Figure 1: The mainstream processes and categories of AI4Research, which can be divided into five key areas: (1) AI for Scientific Comprehension, (2) AI for Academic Survey, (3) AI for Scientific Discovery, (4) AI for Academic Writing, and (5) AI for Academic Peer Review. Each of these areas contributes to improving the effectiveness and efficiency of AI-integrated research and publication.

主要亮点

- **提出五大任务系统分类**：将 AI4Research 明确划分为科学理解、学术综述、科学发现、论文写作与同行评议五个核心任务，并给出形式化定义。
- **前沿议题与研究缺口洞察**：重点强调跨学科基础模型、自动化大规模实验的严谨性与可扩展性以及伦理安全等未来研究方向。
- **资源汇编与社区贡献**：整理多学科应用场景、数据语料与工具平台，为学界与工业界构建一站式资源门户，促进后续创新。

相关链接

Paper: https://arxiv.org/abs/2507.01903

Github: https://github.com/alexiglad/EBT

● 论文七

GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning

日期

2025-07-02

GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning

GLM-V Team

Zhipu AI & Tsinghua University

(For the complete list of authors, please refer to the Contribution section)

内容提要

GLM- 4.1V- Thinking 是一款面向通用多模态推理的视觉- 语言模型（VLM）。作者先以海量图像- 文本、文档、视频与 GUI 数据进行大规模预训练，奠定强视觉与语言基础；随后提出“课程抽样强化学习”（RLCS），按难度自适应挑选跨领域样本，在 STEM 解题、视频理解、OCR、GUI 代理等任务上系统提升模型推理能力。9 B 参数版本在 28 项公开基准中整体超越 Qwen2.5- VL- 7B，并在 18 项任务上媲美或优于 72 B 的 Qwen2.5- VL- 72B 及闭源 GPT- 4o。团队已开源推理模型、基础模型与奖励系统，为研究者和开发者提供高效、可扩展的多模态推理平台。

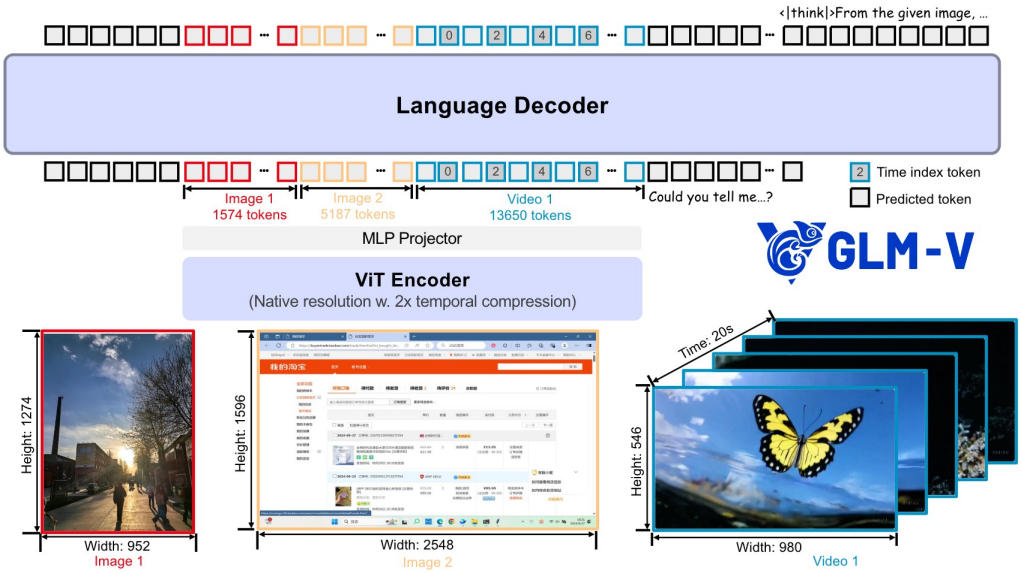


Figure 2: **The architecture of GLM-4.1V-Thinking.** The proposed model consists of three components: (1) a ViT Encoder to process and encode images and videos, (2) an MLP Projector to align visual features to tokens, (3) a Large Language Model as a Language Decoder to process multimodal tokens and yield token completions. GLM-4.1V-Thinking can perceive images and videos as their native resolutions and aspect ratios. For video inputs, additional time index tokens are inserted behind each frame to enhance the model's temporal understanding capability.

主要亮点

- **RLCS 驱动的跨领域强化学习：**提出 Reinforcement Learning with Curriculum Sampling，在训练中动态调整样本难度与任务配比，显著提升效率并保持稳定性。
- **小规模即 SOTA：**仅 9 B 参数的 GLM- 4.1V- Thinking 在 23/28 基准上斩获同级开源最佳，并在 18 项任务上追平或超越 72 B Qwen2.5- VL- 72B 与 GPT- 4o。
- **完整开源与易扩展架构：**公开推理模型、基础模型及领域奖励系统；采用 ViT Encoder + 3D- RoPE 语言解码器设计，可直接处理超宽高分辨率图像与长视频，为后续研究提供坚实基座。

相关链接

Paper： https://arxiv.org/abs/2507.01006

Project： https://github.com/THUDM/GLM-4.1V-Thinking

● 论文八

AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench

日期

2025-07-03

AI Research Agents for Machine Learning: Search, Exploration, and Generalization in MLE-bench

Edan Toledo[1,2,*], **Karen Hambardzumyan**[1,2,*], **Martin Josifoski**[1,*], **Rishi Hazra**[3,†], **Nicolas Baldwin**[1], **Alexis Audran-Reiss**[1], **Michael Kuchnik**[1], **Despoina Magka**[1], **Minqi Jiang**[1], **Alisia Maria Lupidi**[1], **Andrei Lupu**[1], **Roberta Raileanu**[1], **Kelvin Niu**[1], **Tatiana Shavrina**[1], **Jean-Christophe Gagnon-Audet**[1], **Michael Shvartsman**[1], **Shagun Sodhani**[1], **Alexander H. Miller**[1], **Abhishek Charnalia**[1], **Derek Dunfield**[1], **Carole-Jean Wu**[1], **Pontus Stenetorp**[2], **Nicola Cancedda**[1], **Jakob Nicolaus Foerster**[1], **Yoram Bachrach**[1]

[1]FAIR at Meta, [2]University College London, [3]Örebro University

*Equal contribution (author order determined by a game of UNO), †Work done while at Meta

内容提要

作者将 AI 研究代理形式化为“搜索策略 π + 算子集合 O”双轴框架，并推出可扩展实验平台 **AIRA-dojo**。在此平台上，他们设计了新算子集 **OAIRA**，并与 Greedy、MCTS、进化搜索等策略配合。仅替换算子即可将 MLE- bench Lite 奖牌成功率由 39.6 % 提升至 45.5 %，再结合 MCTS 可达 47.7 %，刷新公开记录。论文进一步量化“验证- 测试”泛化缺口：若最终节点按测试集而非验证集排序，奖牌率可再增 9- 13 个百分点。AIRA- dojo 采用 Apptainer 隔离与 Superimage 基础镜像，支持 H200 GPU + 24 核 CPU 的沙盒并发，验证了长时、千并发搜索的可靠性与可重复性。作者还讨论算子能力、环境配置与节点选择策略对自动科研规模化的关键作用。

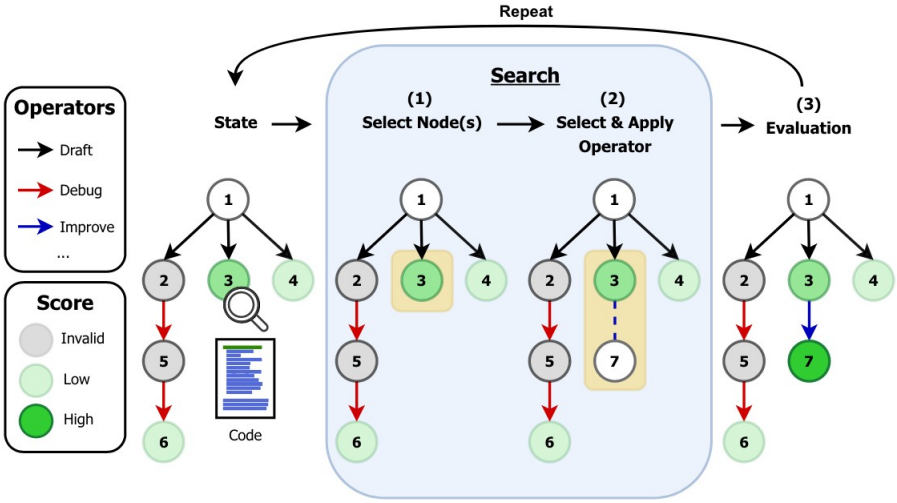


Figure 2 Overview of AIRA. Given a problem specification, AIRA maintains a search graph whose nodes are (partial) solutions. At each iteration, the agent (1) selects nodes via a *selection policy*, (2) picks an operator via an *operator policy* and applies this operator to the node, and (3) scores the resulting solution via a *fitness function*. Here, a `greedy` node-selection strategy applies the `improve` operator to the highest-scoring node.

主要亮点

- **框架解耦**：首次将研究代理拆解为可插拔的“搜索策略+算子”，实验证明算子是性能瓶颈，并以 OAIRA 显著缓解。
- **性能跃迁**：Greedy + OAIRA 将奖牌率提升至 45.5 %，MCTS + OAIRA 达 47.7 %，并揭示泛化缺口按测试集选节点可再提 9- 13 %。
- **可扩展平台**：AIRA- dojo 通过 Apptainer 隔离与预装 Superimage 镜像，支持 GPU/CPU 资源弹性调度，为大规模自动科研提供稳固基座。

相关链接

Paper：https://arxiv.org/abs/2507.02554

Project：https://github.com/facebookresearch/aira-dojo

● 论文九

A Survey on Latent Reasoning

日期

2025-07-10

A Survey on Latent Reasoning

Rui-Jie Zhu*,†, Tianhao Peng*, Tianhao Cheng*, Xingwei Qu*,
Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, Zhenhe Wu, Yongchi Zhao, Tianyu Liu, Jian Yang, Wangchunshu Zhou, Chujie Zheng, Chongxuan Li, Yuyin Zhou, Zhoujun Li, Zhaoxiang Zhang, Jiaheng Liu†, Ge Zhang†, Wenhao Huang, Jason Eshraghian†

UCSC, FDU, NJU, PKU, RUC, UoM, UW-Madison, PolyU, M-A-P

内容提要

本文系统性综述了 Latent Reasoning（潜在推理）这一新兴研究方向，旨在理解和构建具备抽象、压缩与可组合性的隐式中间推理结构。与显式链式推理不同，潜在推理强调构建潜在空间中的表征，通过引入中间变量（如思维草图、程序、图结构等）来辅助语言模型进行更复杂、更高效的推理。论文从建模形式、数据监督、训练策略和评估方法四个维度出发，对现有工作进行全面梳理，并指出当前面临的挑战与未来的研究趋势。

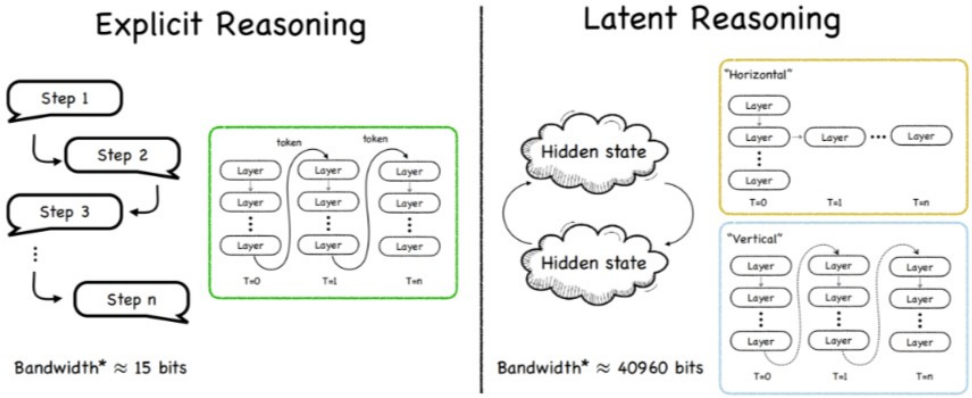


Figure 1. Explicit reasoning transmits discrete tokens (≈ 15 bits each), whereas latent reasoning exchanges full 2560-dimensional FP16 hidden states (≈ 40,960 bits each), revealing a $\sim 2.7 \times 10^3$-fold bandwidth gap between the two approaches.

主要亮点

- **系统化定义潜在推理**：明确潜在推理与传统显式推理的差异，提出统一的研究框架与建模目标。
- **全面综述四大维度**：从推理空间建模、监督信号、训练方法和评估机制四方面系统回顾已有工作。
- **分类总结代表性方法**：涵盖草图推理、程序生成、图结构建模、表格推理等不同潜在空间类型。
- **提出关键挑战与方向**：包括潜在空间设计、弱监督训练、泛化能力与多任务融合等，为未来研究提供路线图。

相关链接

Paper：https://arxiv.org/pdf/2507.06203

● 论文十

Mixture-of-Recursions: Learning Dynamic Recursive Depths for Adaptive Token-Level Computation

日期

2025-07-12

Mixture-of-Recursions: Learning Dynamic Recursive Depths for Adaptive Token-Level Computation

Sangmin Bae[1,*], Yujin Kim[1,*], Reza Bayat[2,*], Sungnyun Kim[1], Jiyoun Ha[3], Tal Schuster[4], Adam Fisch[4], Hrayr Harutyunyan[5], Ziwei Ji[4], Aaron Courville[2,6,†] and Se-Young Yun[1,†]

[1]KAIST AI, [2]Mila, [3]Google Cloud, [4]Google DeepMind, [5]Google Research, [6]Université de Montréal
*Equal Contribution, †Corresponding Authors

内容提要

本文提出了一种新型高效的 Transformer 架构——Mixture-of-Recursions (MoR)，旨在同时实现参数共享和自适应计算两种效率目标。MoR 以递归 Transformer 为基础，通过引入轻量级的路由器，实现了对每个 token 进行动态的递归深度分配，从而根据 token 的难度自适应地分配计算资源。该方法不仅通过重复使用参数块来降低模型参数量，还利用递归层级的 KV 缓存策略减少内存访问开销。此外，MoR 还引入 KV 共享机制，用于降低 prefill 阶段的延迟和显存使用。在多个模型规模（135M 至 1.7B）和相同 FLOPs 预算下，MoR 在验证损失、few-shot 准确率和推理吞吐率方面均优于传统 Transformer 和现有的递归 Transformer 方法。

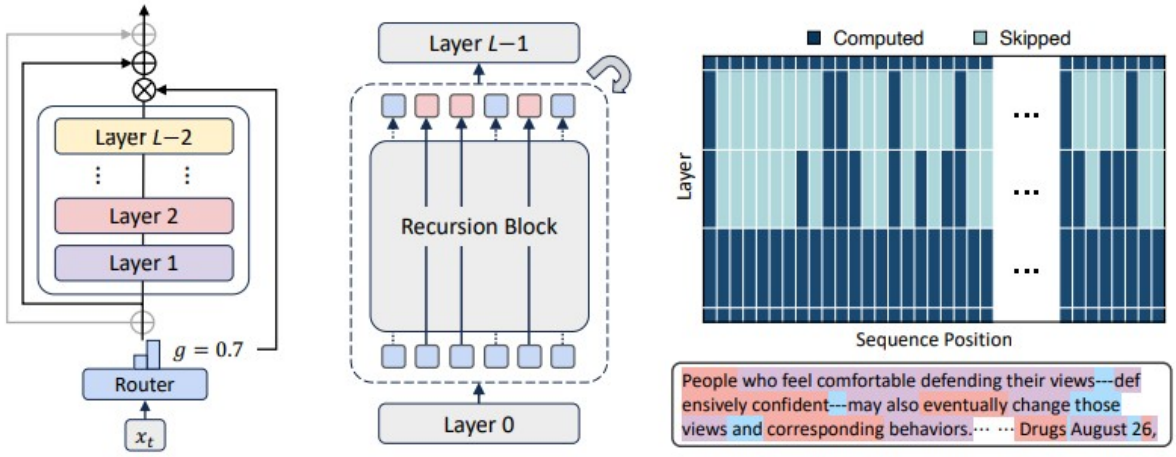


Figure 1: Overview of Mixture-of-Recursions (MoR). (*Left*) Each recursion step consists of a fixed stack of layers and a router that determines whether each token should pass through or exit. This recursion block corresponds to the gray box in the middle. (*Middle*) The full model structure, where the shared recursion step is applied up to N_r times for each token depending on the router decision. (*Right*) An example routing pattern showing token-wise recursion depth, where darker cells indicate active computation through the recursion block. Below shows the number of recursion steps of each text token, shown in colors: 1, 2, and 3.

主要亮点

- **统一高效建模框架：**MoR 首次将参数共享、自适应 token 级计算深度和内存高效的 KV 缓存机制统一在一个框架中，有效提升了训练和推理效率。
- **动态递归路由机制：**引入轻量级路由器，实现 token 级别的递归深度选择，使得模型能够根据 token 难度动态决定“思考”次数，提升模型计算的精细化分配能力。
- **创新 KV 缓存策略：**提出“递归级 KV 缓存”和“递归 KV 共享”两种机制，有效降低内存带宽和显存使用，同时在推理时显著提高吞吐率。
- **支持测试时递归深度可扩展：**MoR 允许在推理阶段灵活调整递归次数，在不改动模型参数的前提下提升生成质量。

相关链接

Blog：https://arxiv.org/pdf/2507.10524
GitHub：https://github.com/raymin0223/mixture_of_recursions

● 论文十一

RoboBrain 2.0 Technical Report

日期

2025-07-15

RoboBrain 2.0 Technical Report

BAAI RoboBrain Team

内容提要

RoboBrain 2.0 是 BAAI 发布的新一代具身视觉- 语言基础模型，提供 7B 与 32B 两个版本，采用轻量 ViT 编码器与 Decoder- only LLM 组成的异构架构。模型通过三阶段课程式训练——基础时空学习、具身时空增强与链式思考推理——结合大规模空间、时间及因果数据，统一实现环境感知、任务推理和长程规划能力。32B 版本在 BLINK- Spatial、Where2Place、Multi- Robot- Plan 等 12 项公开基准中有 6 项取得 SOTA，全面领先 GPT- 4o、Gemini- Pro 等闭源/开源对手。报告还开源权重、评测套件及 FlagScale 训练/推理框架，为通用具身智能研究与落地提供高效、可复现的基座。

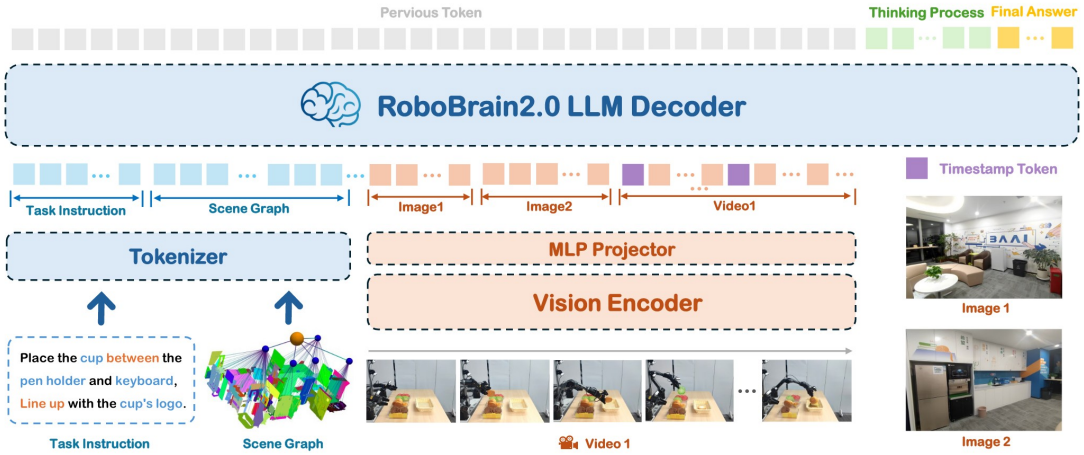


Figure 3 The Architecture of RoboBrain 2.0. The model supports multi-image, long video, and high-resolution visual inputs, along with complex task instructions and structured scene graphs on the language side. Visual inputs are processed via a vision encoder and an MLP projector, while textual inputs are tokenized into a unified token stream. All inputs are fed into an LLM decoder that performs long-chain-of-thought reasoning and generates a variety of outputs depending on the task, including structured plans, spatial relations, or relative and absolute coordinates.

主要亮点

- **三阶段课程式训练 + CoT**：采用“基础时空→具身增强→链式思考”递进策略，使模型同时具备空间理解、时序推理与因果链解释能力。
- **小规模即超越闭源**：仅 32B 参数即在多项空间与时间基准斩获 SOTA，整体性能超越 GPT-4o 与 Gemini-Pro 等大型闭源模型。
- **开源生态与高效基础设施**：公开模型权重、评测工具和 FlagScale-VeRL 训练/推理框架，支持混合并行、内存预分配与量化推理，大幅降低具身 VLM 研发成本。

相关链接

Paper：https://arxiv.org/abs/2507.02029v1
Blog：https://superrobobrain.github.io/

● 论文十二

Dynamic Chunking for End-to-End Hierarchical Sequence Modeling

日期

2025-07-15

Dynamic Chunking for End-to-End Hierarchical Sequence Modeling

Sukjun Hwang
Carnegie Mellon University
sukjunh@cs.cmu.edu

Brandon Wang
Cartesia AI
brandon.wang@cartesia.ai

Albert Gu
Carnegie Mellon University, Cartesia AI
agu@cs.cmu.edu, albert@cartesia.ai

内容提要

论文提出 **Dynamic Chunking (DC)** 与层次网络 **H- Net**，在训练过程中利用“路由判边界 + 平滑插值”机制自适应地将原始字节序列压缩成语义块，彻底替代手工 BPE 分词，实现真正的端到端序列建模。单层 H- Net 在与强 BPE- Transformer 算力、数据完全匹配的条件下就能取得更低困惑度；迭代为两层后，仅用 30 B 字节训练即追平参数翻倍的 Token 模型，并在随后的训练中保持更陡峭的性能增长曲线。实验进一步表明，H- Net 对中文、代码及 DNA 等弱分词场景具备 3–4 × 的数据效率和更强字符级鲁棒性。

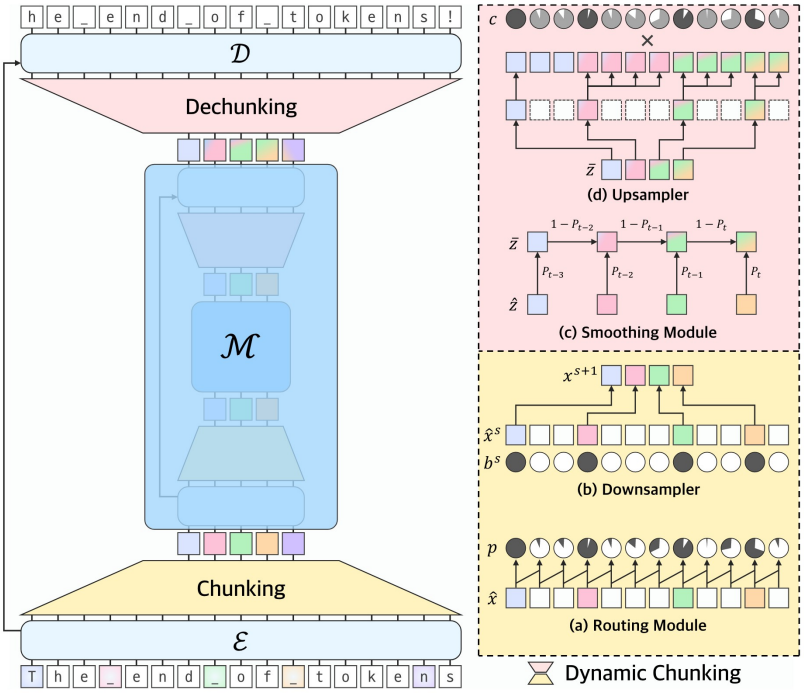


Figure 1: **(left)** Architectural overview of H-Net with a two-stage hierarchical design ($S = 2$). **(right)** Dynamic Chunking (DC). (**bottom-right**) Key components of a chunking layer: (a) a routing module for dynamically drawing chunk boundaries, and (b) a downsampler that selectively retains vectors based on boundary indicators, reducing sequence length while preserving semantically significant positions. (**top-right**) Key components of a dechunking layer: (c) a smoothing module for converting discrete chunks into interpolated representations, and (d) an upsampler that restores compressed vectors to their original resolution based on boundary indicators. Linear in equation (3) and STE in equation (9) are omitted in the illustration for brevity.

主要亮点

- **动态分块替代分词**：路由模块预测边界、平滑模块缓解离散梯度，模型自动学习内容与上下文相关的压缩策略，最终块大小≈ 5 bytes，与 BPE 粒度相当。
- **层次递归扩展性强**：两级 H- Net 仅训练 30 B 字节即可超越同算力 BPE- Transformer，后续优势随数据量持续扩大，验证了层次抽象提高 scaling 率的假设。
- **跨模态鲁棒与高效**：在中文、代码和 DNA 任务中，H- Net 降低 Bits- per- Byte 并在 XWinograd- zh、HellaSwag 等基准显著优于分词模型，DNA 序列建模数据效率提升近四倍。

相关链接

Paper：https://arxiv.org/abs/2507.07955
Blog：https://goombalab.github.io/blog/2025/hnet-past/

● 论文十三

Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities

日期

2025-07-17

Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.

Gemini Team, Google

内容提要

本文介绍了 Google 最新发布的多模态大模型系列 Gemini 2.5，包括 Gemini 2.5 Pro 与 Gemini 2.5 Flash 等模型。Gemini 2.5 Pro 是迄今为止最强大的 Gemini 模型，具备先进的推理能力、长上下文理解（超过 100 万 token）、多模态处理（文本、图像、音频、视频）以及工具使用能力。该系列模型在代码生成、数学推理、长文档检索、多语言支持、音视频理解等多个任务上达到了新的 SOTA 水平，并被集成到 Google 多个产品中，用于支持从教育、创作到智能代理等多种实际应用。

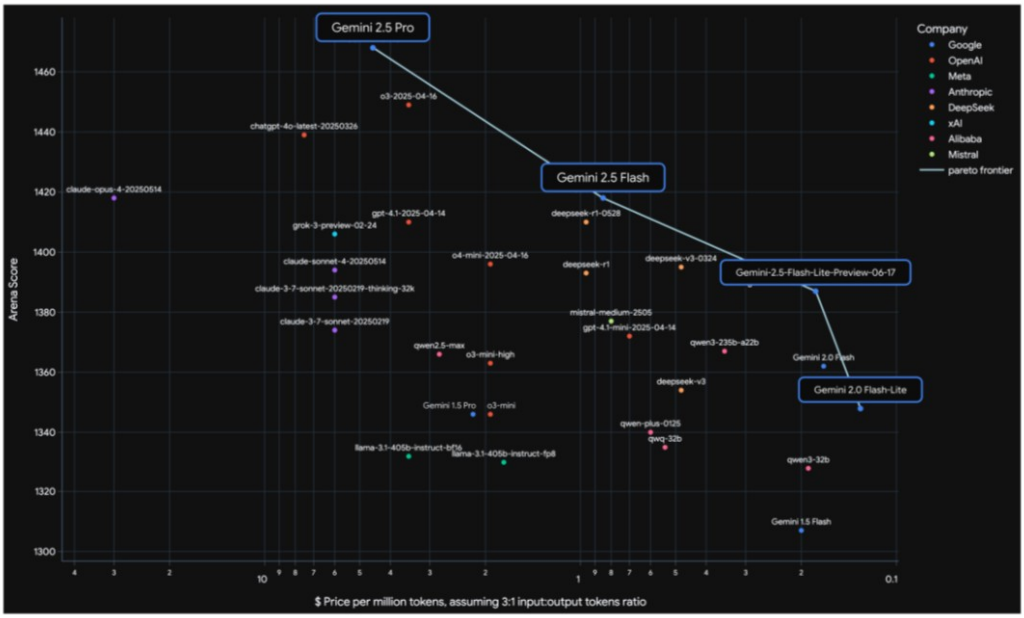


Figure 1 | Cost-performance plot. Gemini 2.5 Pro is a marked improvement over Gemini 1.5 Pro, and has an LMArena score that is over 120 points higher than Gemini 1.5 Pro. Cost is a weighted average of input and output tokens pricing per million tokens. Source: LMArena, imported on 2025-06-16.

主要亮点

- **动态推理机制（Thinking）显著提升模型推理能力：**引入“推理预算”机制，支持在推理阶段动态分配计算资源，使模型在多个基准上性能大幅跃升（如 AIME 2025、LiveCodeBench）。
- **强大的多模态与长上下文能力：**支持文本、图像、音频、视频等多模态输入，最大上下文窗口超百万 tokens，可处理 3 小时视频，并在多模态基准测试（如 VideoMME、MMMU）中实现 SOTA。
- **编码与工具使用能力大幅提升：**在多个代码基准如 LiveCodeBench、Aider Polyglot 上大幅超越旧模型，可操作完整代码库、构建交互式网页应用，具备自动搜索与多步工具调用能力。
- **覆盖全面的模型谱系与成本-能力权衡：**从轻量级的 Flash-Lite 到高性能的 Pro 模型，覆盖从超低延迟到最高能力的多种部署需求，推动模型能力-成本帕累托前沿。

相关链接

Paper： https://arxiv.org/pdf/2507.06261

● 论文十四

Learning without training: The implicit dynamics of in-context learning

日期

2025-07-21

Learning without training: The implicit dynamics of in-context learning

内容提要

本文提出了一种新颖的范式——无训练学习（Learning without Training, LwT），该方法完全跳过了传统的模型训练过程，直接在预训练大语言模型（LLMs）中构建任务特定模块，实现端到端推理。作者通过自动构建“模块图（Module Graph）”，将任务拆解为可执行的模块组合，使得 LLMs 能够在零训练的条件下完成文本分类、结构预测等任务，在多个基准测试中取得接近甚至超过微调模型的效果。

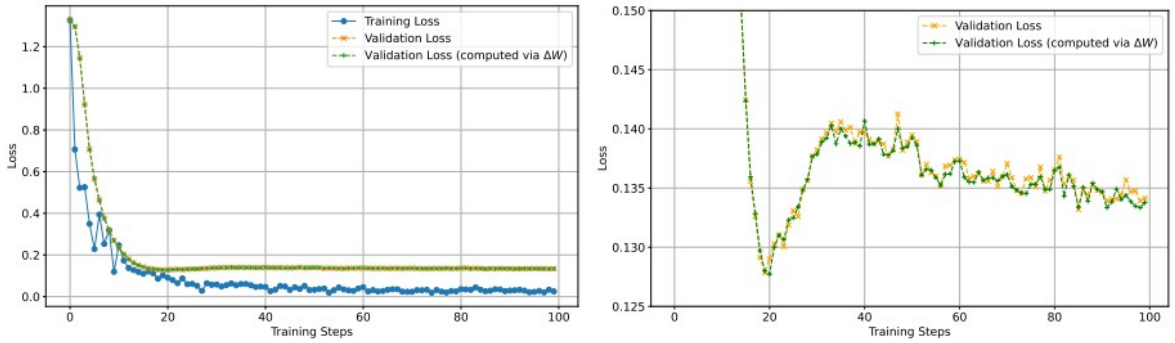


Figure 1: Train and Validation loss curves. Here, the "Validation loss (computed via ΔW)" refers the loss computed using $T_{W+\Delta W}(x)$; i.e., the trained model prediction given only x_{query} but with MLP weights modified by ΔW as defined in Equation (8). **Left:** Training loss and both validation Loss curves. **Right:** Close-up of validation loss computed both ways; i.e., using $T_W(C, x)$ vs. $T_{W+\Delta W}(x)$.

主要亮点

- **提出无训练学习新范式**：跳过微调与训练过程，仅基于模块组合与 LLM 推理完成任务。
- **设计模块图机制**：通过自动任务解析和模块装配生成图结构，实现可解释、可组合的推理流程。
- **广泛适用于多种任务类型**：在文本分类、情感分析、NER 等任务中均取得与训练模型相当的效果。
- **显著节省计算与时间成本**：完全摒弃参数更新过程，提升效率与可移植性。

相关链接

Paper：https://arxiv.org/pdf/2507.16003

● 论文十五

Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos

日期

2025-07-21

Being-H0: Vision-Language-Action Pretraining from Large-Scale Human Videos

Hao Luo[1,3,*] Yicheng Feng[1,3,*] Wanpeng Zhang[1,3,*] Sipeng Zheng[3,*]
Ye Wang[2,3] Haoqi Yuan[1,3] Jiazheng Liu[1] Chaoyi Xu[3] Qin Jin[2]
Zongqing Lu[1,3,†]

[1]Peking University [2]Renmin University of China [3]BeingBeyond

https://beingbeyond.github.io/Being-H0

内容提要

本文提出了一个统一的多模态多智能体对齐框架 Being-H0，旨在赋予大模型“身为人类”的能力，从而提升其作为智能体与人类交互的表现。Being-H0 通过构建多模态、多智能体、任务驱动的交互式对齐数据集 Being-H0 Bench，模拟现实中的人类互动场景，并采用增强对比学习等方法对大语言模型进行训练。实验结果表明，经过 Being-H0 对齐的模型在多项基准测试上显著优于未对齐模型，尤其在人类偏好评估中具备更强的共情能力与响应一致性。

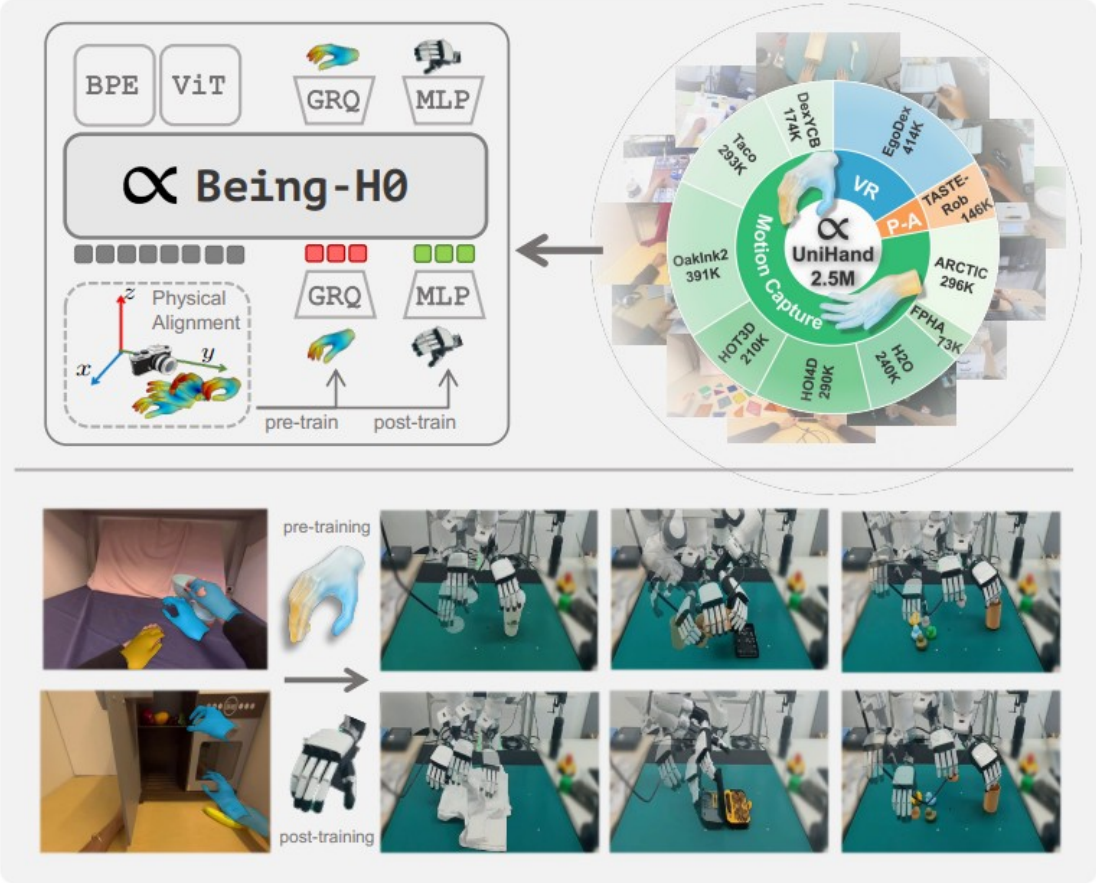


Figure 1: **Being-H0** acquires dexterous manipulation skills by learning from large-scale human videos in the UniHand dataset via **physical instruction tuning**. By explicitly modeling hand motion patterns, the resulting foundation model seamlessly transfers from human hand demonstrations to robotic manipulation.

主要亮点

- **提出统一框架“Being-H0”以实现多模态多智能体对齐，**提升大模型在真实人类交互场景中的表现。
- **构建高质量的 Being-H0 Bench 数据集，**涵盖多模态输入、多智能体互动及多任务形式，贴近现实对齐需求。
- **引入增强对比对齐方法，**结合上下文响应对比与角色感知对比，有效提升响应质量与一致性。
- **在多个主观与客观评测中取得优异表现，**包括 GPT-4 多维评估、人类偏好、MT Bench、Arena-H 等。

相关链接

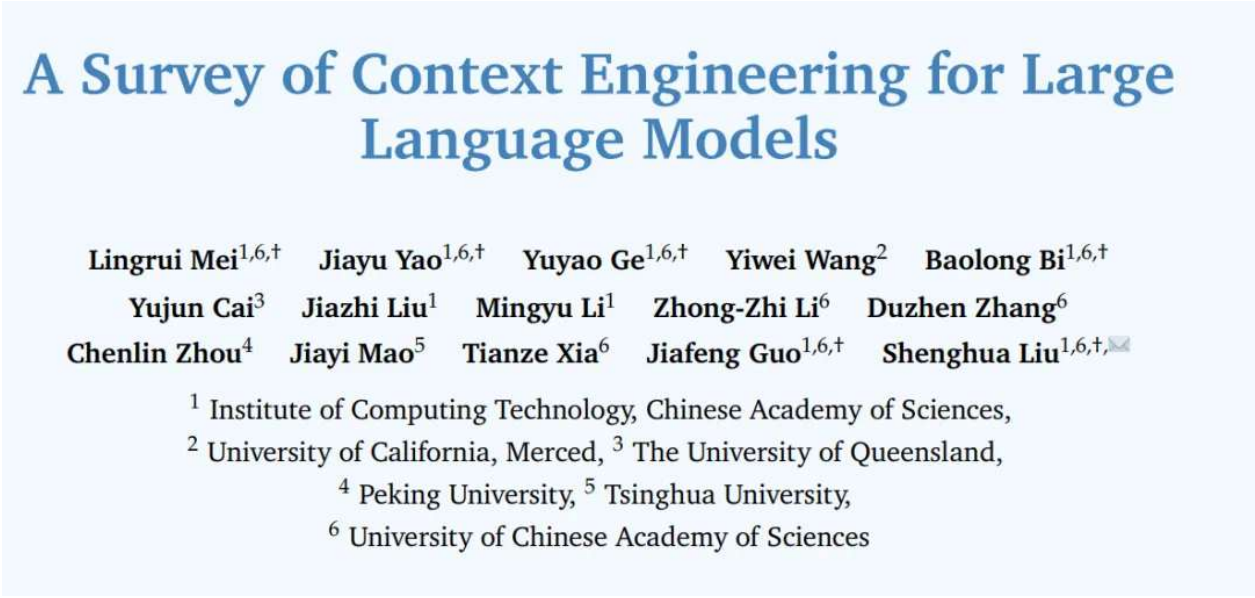
Paper： https://arxiv.org/pdf/2507.15597
GitHub： https://beingbeyond.github.io/Being-H0/

● 论文十六

A Survey of Context Engineering for Large Language Models

日期

2025-07-21


A Survey of Context Engineering for Large Language Models

Lingrui Mei[1,6,†] Jiayu Yao[1,6,†] Yuyao Ge[1,6,†] Yiwei Wang[2] Baolong Bi[1,6,†]
Yujun Cai[3] Jiazhi Liu[1] Mingyu Li[1] Zhong-Zhi Li[6] Duzhen Zhang[6]
Chenlin Zhou[4] Jiayi Mao[5] Tianze Xia[6] Jiafeng Guo[1,6,†] Shenghua Liu[1,6,†,✉]

[1] Institute of Computing Technology, Chinese Academy of Sciences,
[2] University of California, Merced, [3] The University of Queensland,
[4] Peking University, [5] Tsinghua University,
[6] University of Chinese Academy of Sciences

内容提要

本文首次把 **Context Engineering（上下文工程）** 确立为独立学科，突破传统“Prompt Engineering”视角，将“为 LLM 优化信息载荷”系统化为可分解、可度量、可迭代的工程流程。作者检索并分析 2020-2025 间 1400 余篇相关工作，提出“双层 3 + 4”总览框架：基础组件层涵盖 ① 上下文检索与生成、② 上下文处理、③ 上下文管理；系统实现层涵盖 ① RAG、② Memory Systems、③ Tool-Integrated Reasoning、④ Multi-Agent Systems，并辅以评测方法与未来方向。调查指出，现有研究在“理解复杂上下文”上已趋成熟，但在“生成同等复杂、长篇输出”方面仍显薄弱，形成显著能力不对称；同时，从算力-效率权衡、跨模态表示、长程推理到安全伦理，仍存多项关键挑战。文章最终给出理论基础、技术创新与应用驱动三大研究路线图，为学界与工业界提供统一术语、评价指标与资源索引。

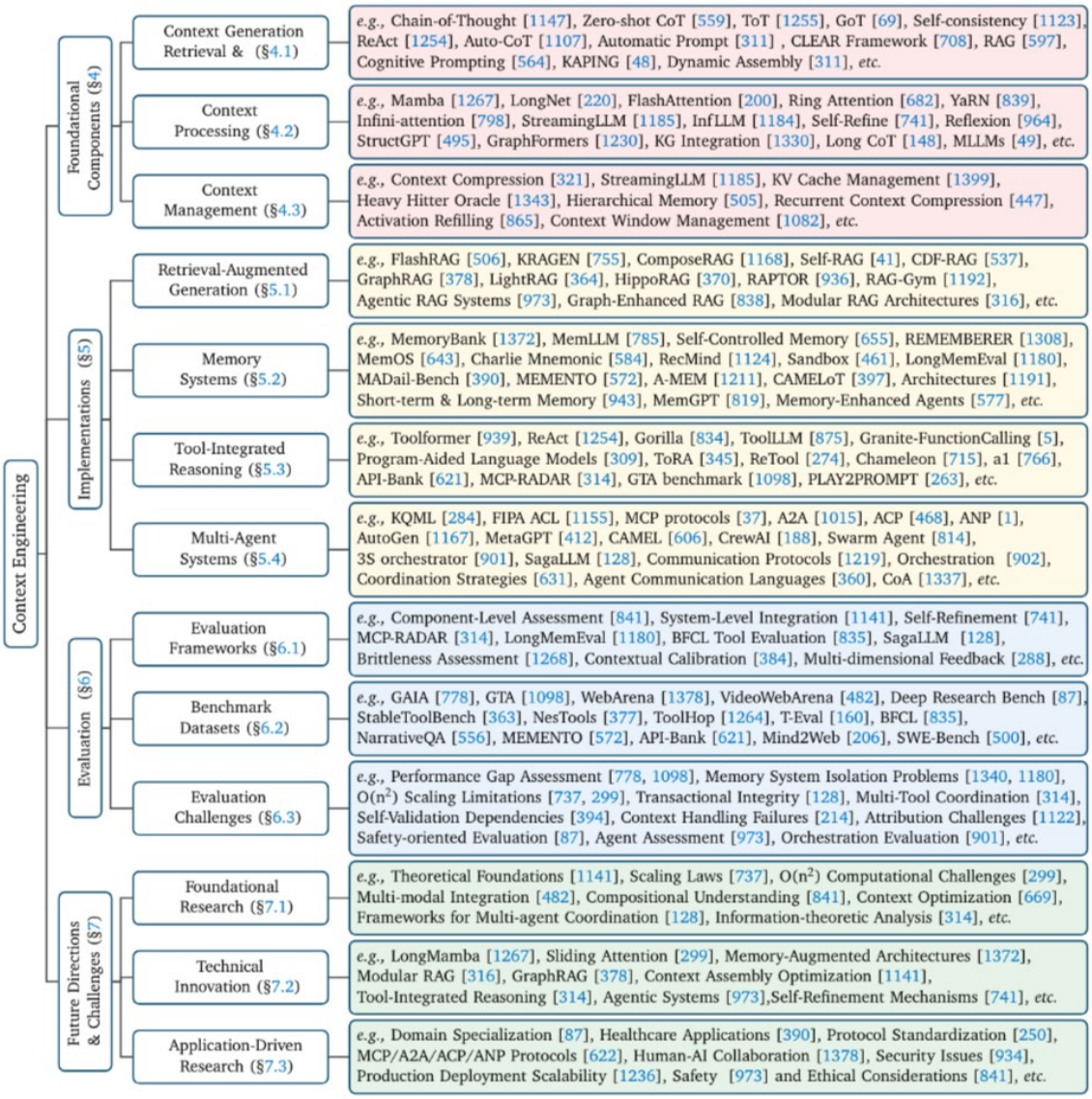


Figure 1: The taxonomy of Context Engineering in Large Language Models is categorized into foundational components, system implementations, evaluation methodologies, and future directions. Each area encompasses specific techniques and frameworks that collectively advance the systematic optimization of information payloads for LLMs.

主要亮点

- **概念奠基与公式化定义：**首次将 Context Engineering 从经验型“提示技巧”提升为可数学建模的优化问题，给出上下文组件集合 F 与期望奖励函数的正式表述。
- **“3 基础 + 4 实现”全景分类：**构建覆盖检索/生成-处理-管理到 RAG-记忆-工具-多智能体的统一知识图谱，厘清子领域间内在依赖，方便横向对比与纵向延展。
- **揭示能力不对称与未来路线图：**指出 LLM 对复杂上下文“理解强、生成弱”的本质落差，并从理论、算法、评测、部署、伦理五维提出后续研究与产业落地关键议题。

相关链接

Paper： https://arxiv.org/pdf/2506.04788

Project: https://github.com/Meirtz/Awesome-Context-Engineering

● 论文十七

Gemini 2.5 Pro Capable of Winning Gold at IMO 2025

日期

2025-07-22

Gemini 2.5 Pro Capable of Winning Gold at IMO 2025*

Yichen Huang (黄溢辰)† Lin F. Yang (杨林)‡

July 28, 2025

内容提要

本文介绍了 Google DeepMind 最新发布的多模态大模型 Gemini 2.5 Pro 的能力与进展，强调其已经达到可在 International Mathematical Olympiad (IMO) 中夺得金牌的水平。该模型展示了在数学推理、代码生成、视觉理解和跨模态任务中的卓越性能，且具备超过 1 百万 token 的长上下文处理能力。此外，Gemini 2.5 Pro 通过引入“思考（Think）”模式实现了计算资源的动态分配，从而在推理复杂任务中实现质的飞跃。

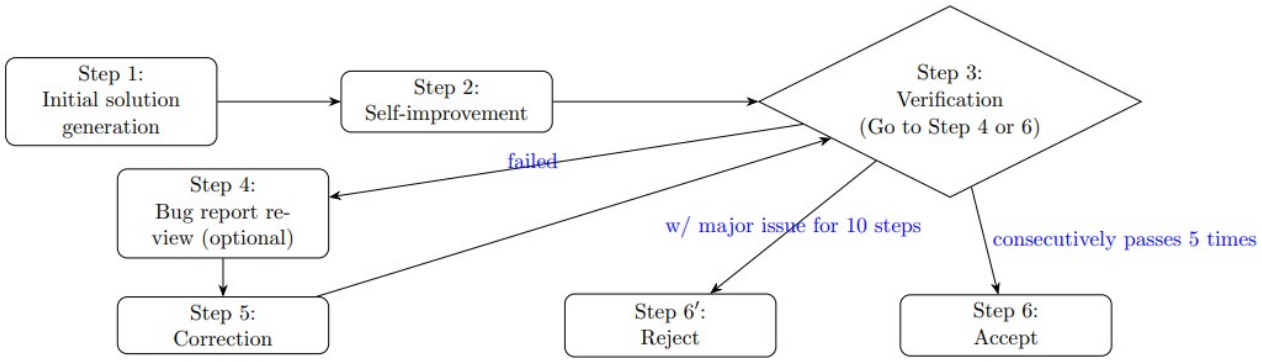


Figure 1: Flow diagram of our pipeline. See the main text for detailed explanations of each step.

主要亮点

- **数学能力：**在 AIME 模拟测试中达到金牌标准，数学推理能力显著提升。
- **长上下文处理：**支持超过 100 万 tokens 的上下文长度，可处理极长文档和多轮交互。
- **多模态能力：**具备对图像、文本、音频、视频等多模态输入的理解与生成能力。
- **动态推理机制：** “Think”模式可根据任务复杂度动态调整计算预算，提高复杂任务准确率。

相关链接

Paper：https://arxiv.org/pdf/2507.15855

● 论文十八

AlphaGo Moment for Model Architecture Discovery

日期

2025-07-24

AlphaGo Moment for Model Architecture Discovery

Yixiu Liu[1,2,4*] Yang Nan[2,4 *] Weixian Xu[1,2,4 *] Xiangkun Hu[2,4]
Lyumanshan Ye[1,2,4] Zhen Qin[3] Pengfei Liu[1,2,4†]

[1]Shanghai Jiao Tong University [2]SII [3]Taptap [4]GAIR

SII-GAIR/ASI-Arch Model Gallery

内容提要

本文提出将模型架构发现（Model Architecture Discovery）视为下一个类 "AlphaGo Moment" 的突破方向，主张构建一个通用框架，以系统性地探索、评估和进化神经网络架构。作者指出，当前模型架构主要依赖人类设计和渐进式优化，而未来的发展需要借助自动化搜索、代理评估器与演化机制，形成类似 AlphaGo 中强化学习与树搜索结合的“闭环”。本文也强调模型架构发现不仅是工程问题，更蕴含科学探索价值。

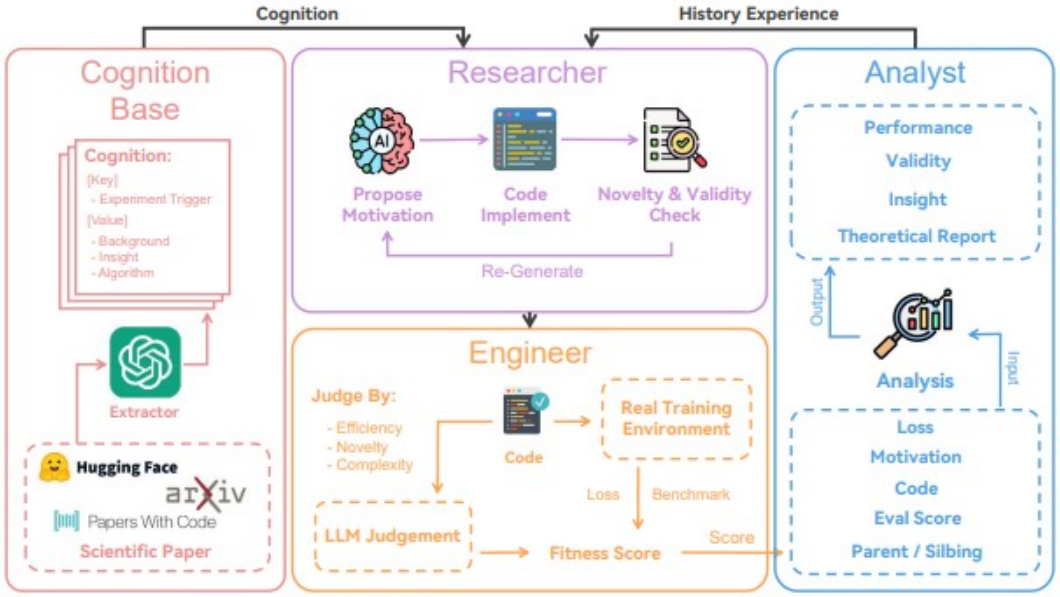


Figure 4: An overview of our four-module ASI-ARCH framework, which operates in a closed evolutionary loop. The cycle begins with the Researcher (purple) proposing a new architecture based on historical data. The Engineer (orange-yellow) handles the subsequent training and evaluation. Finally, the Analyst (blue) synthesizes the experimental results, enriching its findings with knowledge from the Cognition module (red). The output of this analysis informs the next evolutionary step, enabling the system to continuously improve.

主要亮点

- **提出突破性的架构发现范式**：将模型结构设计从“人类主导”过渡到“自动发现+评估+优化”的闭环系统。
- **引入代理评估器机制**：以代理模型代替真实训练，极大提升架构搜索的效率和可扩展性。
- **强调架构发现的科学潜力**：从科学发现视角看待模型设计，主张其不应局限于性能优化，而是通向理解智能本质的路径。
- **分析未来关键挑战**：包括评估器可靠性、搜索空间设计、通用性权衡与算法收敛等问题。

相关链接

Paper: https://arxiv.org/pdf/2507.18074

Github： https://gair-nlp.github.io/ASI-Arch/

TECHNICAL UPDATES
技术动态

技术动态一

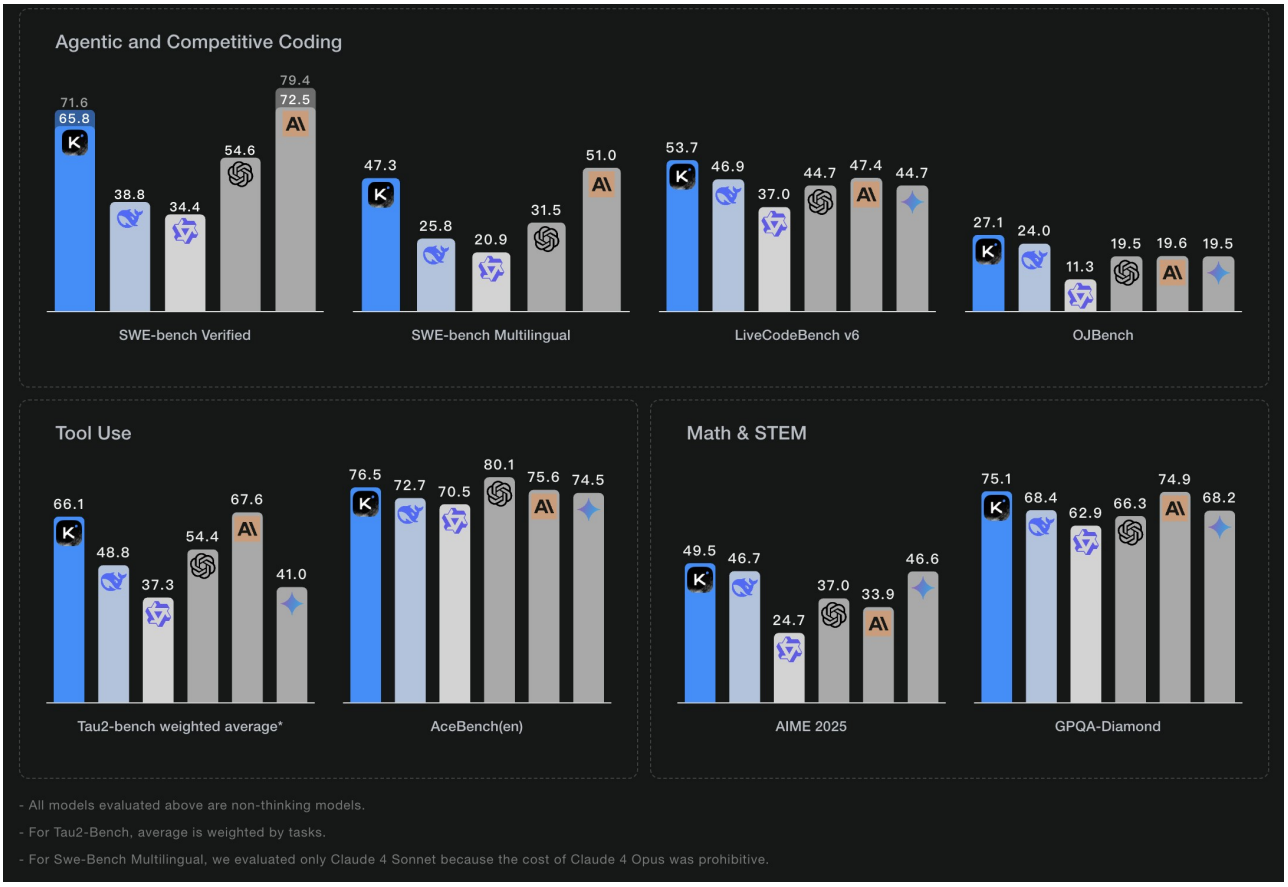
Kimi K2: Open Agentic Intelligence

日期

2025-07-12

内容提要

7 月 12 日，月之暗面深夜开源首个万亿参数混合专家模型 **Kimi K2**（激活 320 亿、总参数 1 万亿），API 价格仅 16 元/百万 token。官方同步释出可商用的 **Kimi- K2- Base** 与 **Kimi- K2- Instruct**，上线 20 分钟 Hugging Face 下载量即破 1.2 万。模型在 15.5 T token 预训练中引入 **MuonClip + qk- clip** 优化器以抑制 logits 爆炸，并保持全程稳定。借助大规模 Agentic 数据合成与通用 RL，K2 学会自动工具调用与环境感知。在 LiveCode Bench、AIME 2025、GPQA- Diamond 等 28 项公开基准全面超越 DeepSeek- V3- 0324、Qwen3- 235B- A22B，并在 SWE- bench Verified 取得 65.8 % pass@1，逼近闭源 GPT- 4.1 的水平。由此，Kimi K2 被视作开源领域的新 SOTA，对 OpenAI 等闭源模型形成直接压力。



相关链接

Blog： https://moonshotai.github.io/Kimi-K2/

GitHub： https://github.com/MoonshotAI/Kimi-K2

● 技术动态二

《Nature》发文赞美 Kimi K2 是另一个 DeepSeek 时刻

日期

2025-07-16

Moonshot AI's Kimi K2 model has been praised for its writing and coding abilities.

内容提要

这篇发表在 Nature 上的文章主要对中国初创公司 Moonshot AI 于 2025 年 7 月推出的大语言模型 **Kimi K2** 进行了介绍。Kimi K2 是一个开源权重的 Agent 型语言模型，具备执行多步骤任务、网页浏览、调用数学工具等能力，尽管它并非一个专注于推理的 Reasoner 模型。它在代码生成和写作方面表现尤为突出，在多个基准测试（如 LiveCodeBench、Creative Writing v3）上取得优异成绩。该模型使用“专家混合”（Mixture of Experts）架构，虽总参数达 1 万亿，但实际激活仅 32B，兼顾性能与计算效率。其发布被誉为“中国 AI 发展中的又一 DeepSeek 时刻”，对全球开源 AI 生态造成深远影响。

相关链接

Paper： https://www.nature.com/articles/d41586-025-02275-6

• 技术动态三

OpenAI 神秘新模型斩获 IMO 2025 金牌

日期

2025-07-19

内容提要

OpenAI 近日披露一款尚未命名的通用推理模型，在 IMO 2025 官方赛制下以 35/42 分解出 6 题中的 5 题，斩获金牌，成绩远超此前最佳的 Gemini 2.5 Pro （13 分）。比赛采用与人类相同的双 4.5 小时闭卷规则，禁止一切网络与外部工具，模型仅凭自然语言写出完整证明，由三位前金牌得主评分。OpenAI 联合创始人 Greg Brockman 等人在社交媒体宣布这一里程碑，引发硅谷热议；研究团队强调该模型并非 GPT- 5，而是一项仍处实验阶段、不会公开的新型推理技术，被视为可能摆脱传统 CoT 范式、展现数小时持续创造性思维与跨领域泛化能力的重大突破。

相关链接

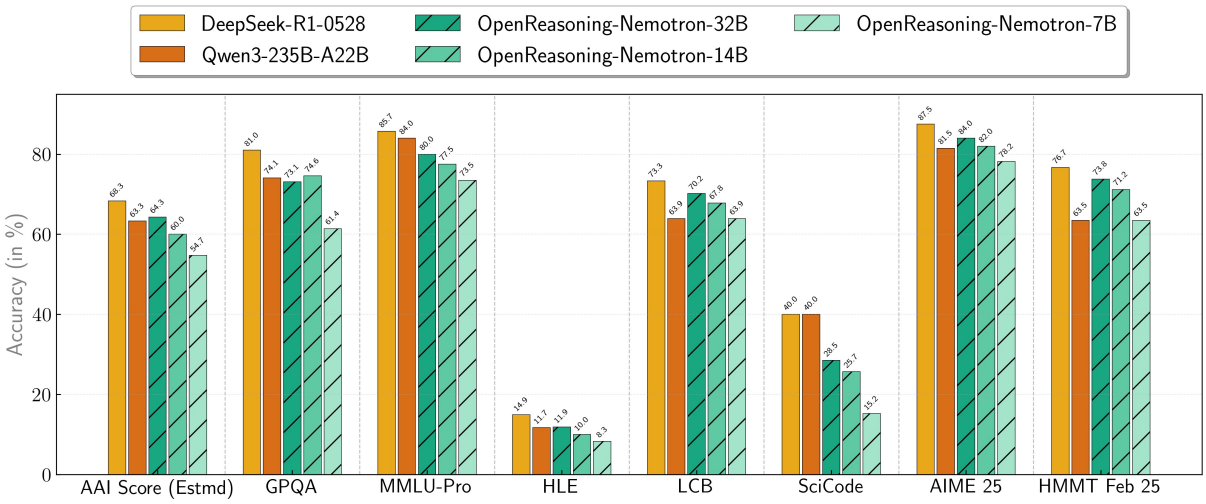
Blog： https://x.com/alexwei_/status/1946477742855532918

● 技术动态四

英伟达开源了 OpenReasoning-Nemotron

日期

2025-07-20



内容提要

OpenReasoning- Nemotron 是英伟达基于 Qwen 2.5 架构、蒸馏 DeepSeek- R1- 0528 数据推出的开源推理模型家族，覆盖 1.5B、7B、14B、32B 四档规模，可在本地 CPU/GPU 运行，用于数学、科学与代码任务。模型在 5 M 条推理轨迹上仅经监督微调便于数学综合基准整体超越 OpenAI o3，并结合 AIMO - 2 提出的 GenSelect@64 采样再次扩大优势。32B 版本在代码 LCB 基准 pass@1 由 70.2 升至 75.3，7B 版本较旧模型数学成绩提升近 20%，展现规模与跨域泛化收益。作者还讨论 TIR 与 CoT 行为切换及未来引入在线 RL 提升一致性的可行性。

相关链接

Blog： https://x.com/NVIDIAAIDev/status/1946281437935567011

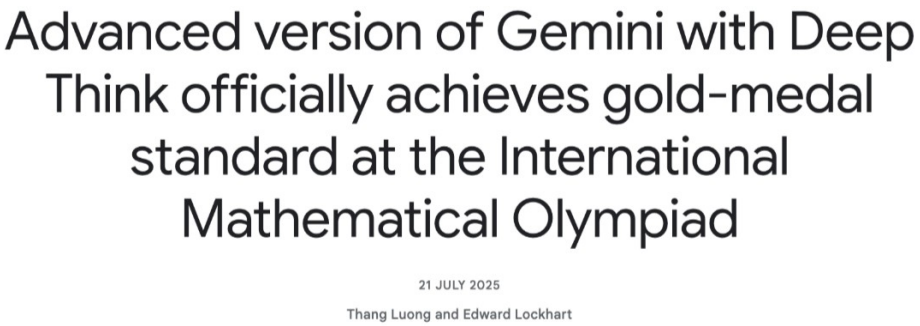
Hugging face： https://huggingface.co/nvidia/OpenReasoning-Nemotron-7B

• 技术动态五

全球首个 IMO 金牌 AI 诞生！谷歌 Gemini 拿下 35 分碾碎奥数神话

日期

2025-07-21


Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

Thang Luong and Edward Lockhart

内容提要

Google DeepMind 最新推出的 Gemini Deep Think 在 IMO 2025 正式赛制中仅用两段各 4.5 小时闭卷时间、纯自然语言写出完整证明，攻克 6 题中的 5 题，以 35/42 分获金牌——这一成绩经 IMO 组委会官方认证，成为全球首个达成此成就的 AI 系统。相较 AlphaProof 等依赖形式化语言的先例，Gemini 直接端到端处理英文题面，在限定时长内生成严谨证明。其“Deep Think”高级模式结合并行思考与自适应多步推理训练，突破传统线性 CoT 限制，并通过强化学习与高质量解题库进一步增强推理深度和稳定性。该里程碑标志通用大模型在高难度数学推理上的跨越式进展，也引发与 OpenAI“自评金牌”成绩的激烈对比与业界热议。

相关链接

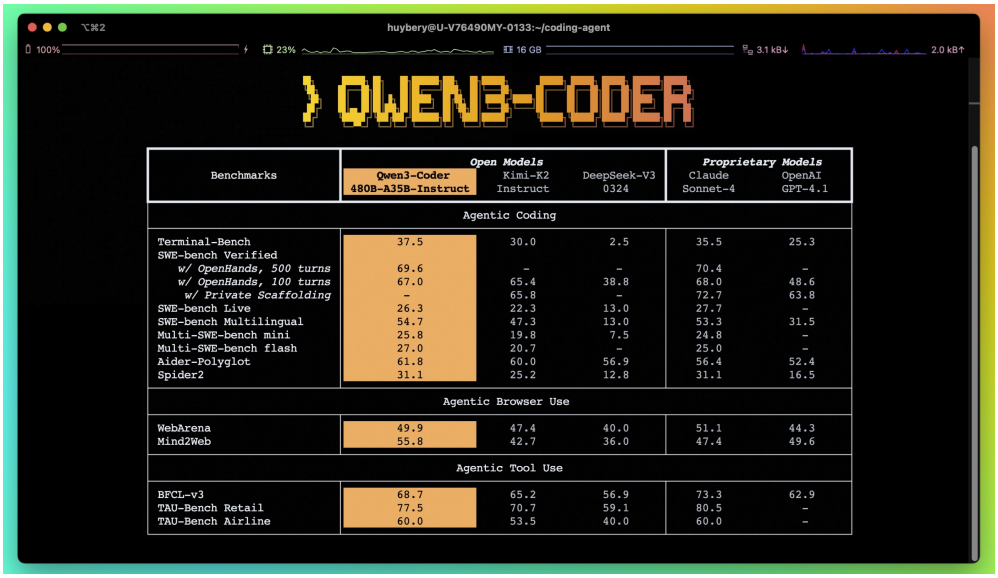
Blog:https://deepmind.google/discover/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

• 技术动态六

Qwen3-Coder: Agentic Coding in the World 阿里巴巴发布开源大模型 Qwen3-Coder，性能明显超越 Kimi-K2、DeepSeek v3、GPT-4.1

日期

2025-07-22



内容提要

Qwen3-Coder 是 Qwen 团队最新发布的代理式代码大模型系列，主力版本 Qwen3-Coder-480B-A35B-Instruct 采用 1 T 总参数、35 B 激活的 MoE 架构，原生支持 256 K token 上下文并可借助 YaRN 扩展至 1 M，针对仓库级与动态 PR 场景优化代码理解与生成。预训练阶段从数据、上下文与合成三个维度扩展至 7.5 T token（代码占 70%），并使用 Qwen2.5-Coder 清洗重写低质样本。后训练阶段引入 Scaling Code RL 与 Agent RL：前者在“大量可验证、易判断”真实任务上释放执行驱动强化学习潜力；后者通过 20 k 并行环境训练长程交互策略，在 SWE-bench Verified 夺得开源 SOTA。此外，官方推出 CLI 工具 Qwen Code，适配 prompt 与工具协议，进一步激活模型的 Agentic Coding 能力。

相关链接

Code： https://github.com/QwenLM/qwen-code

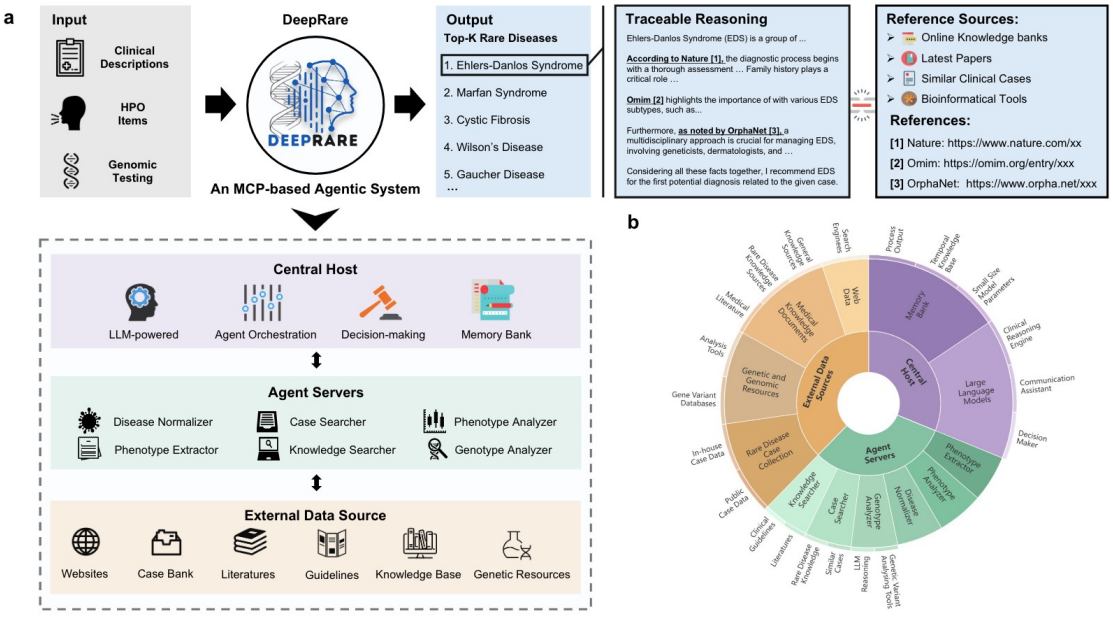
Model： https://hf.co/Qwen/Qwen3-Coder-480B-A35B-Instruct

● 技术动态七

DeepRare 重磅发布：全球首个可循证智能体诊断系统，直击医学 Last Exam 难题

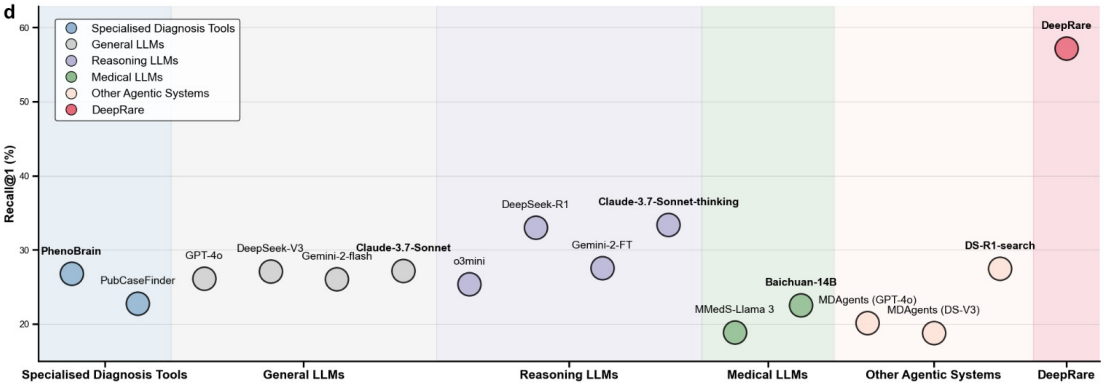
日期

2025-07-24



c

	RareBench (MME)	RareBench (HMS)	RareBench (LIRICAL)	RareBench (RAMEDIS)	MyGene2	DDD	MIMIC-IV-Rare	Xinhua Hosp.
Cases	40	88	370	624	146	2283	1875	975
Avg Info Content	52.7	103.7	59.5	46.1	29.4	70.9	50.1	16.4
Avg HPO Ids	12.2	19.4	14.3	10.1	7.9	18.0	10.1	4.0
Rare Diseases	17	39	252	74	58	2150	355	314
Source	Literature	**Clinical Center** (Germany)	Literature	Scientist Uploaded	Patients Uploaded	Literature	**Clinical Center** (USA)	**Clinical Center** (China)
Public	✓	✓	✓	✓	✓	✓	✓	✗
Gene	○	○	○	○	◐	◐	○	●



内容提要

DeepRare 是首个面向罕见病诊断的 LLM 多智能体系统，可同时处理病历自由文本、HPO 表型项与 VCF 基因变异，输出排序后的候选疾病及可溯源推理链。其三层架构由长记忆中央主机、40 余个专业代理服务器及多源医学知识库组成，实现信息收集- 自反思双阶段闭环推理。在来自欧美亚 8 个数据集、共 6401 例患者、2919 种疾病的评测中，DeepRare 对 1013 种疾病达到 100 % 准确率；平均 Recall@1 为 57.18 %，较次优方法高 23.79 个百分点，多模态输入时 Recall@1 达 70.6 %（Exomiser 仅 53.2 %）。对 180 例病例的人工核查表明，其推理链与医学证据一致率 95.4 %，彰显临床可信度。

相关链接

Paper： https://arxiv.org/abs/2506.20430

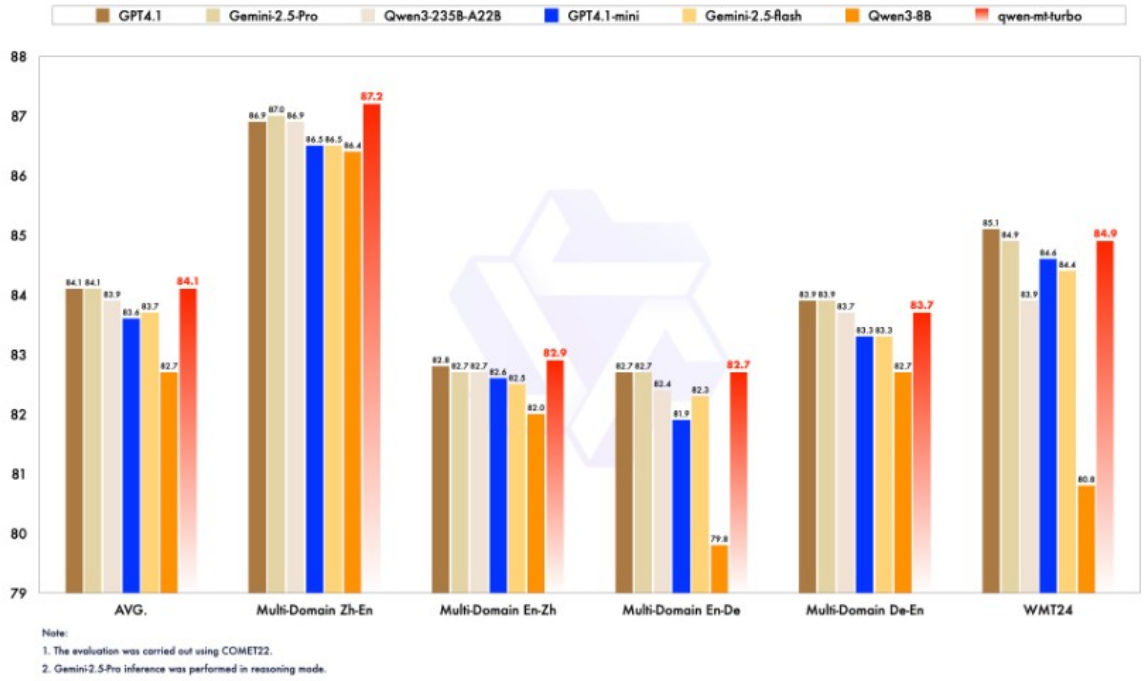
Demo： http://raredx.cn

• 技术动态八

千问发布多语言翻译模型 QWen-MT，支持 92 种语言

日期

2025-07-24



内容提要

阿里巴巴 Qwen 团队发布最新翻译模型 Qwen- MT（Turbo），基于 Qwen3 架构构建，训练数据涵盖数万亿多语种翻译样本，支持机器翻译任务的强化学习训练，显著提升翻译的准确性与流畅性。Qwen- MT 覆盖 92 种语言与方言，可适用于全球大部分语言用户。在多项机器和人工评测中均超越业界主流模型，如 GPT- 4.1- mini、Gemini 2.5- Flash，甚至在某些情况下不亚于 GPT- 4.1 和 Gemini 2.5- Pro。同时采用轻量化 MoE（专家混合）架构，兼顾延迟低、成本低。语言定制能力强，支持术语干预（terminology intervention）、领域提示（domain prompts）和翻译记忆（translation memory）等特性。

相关链接

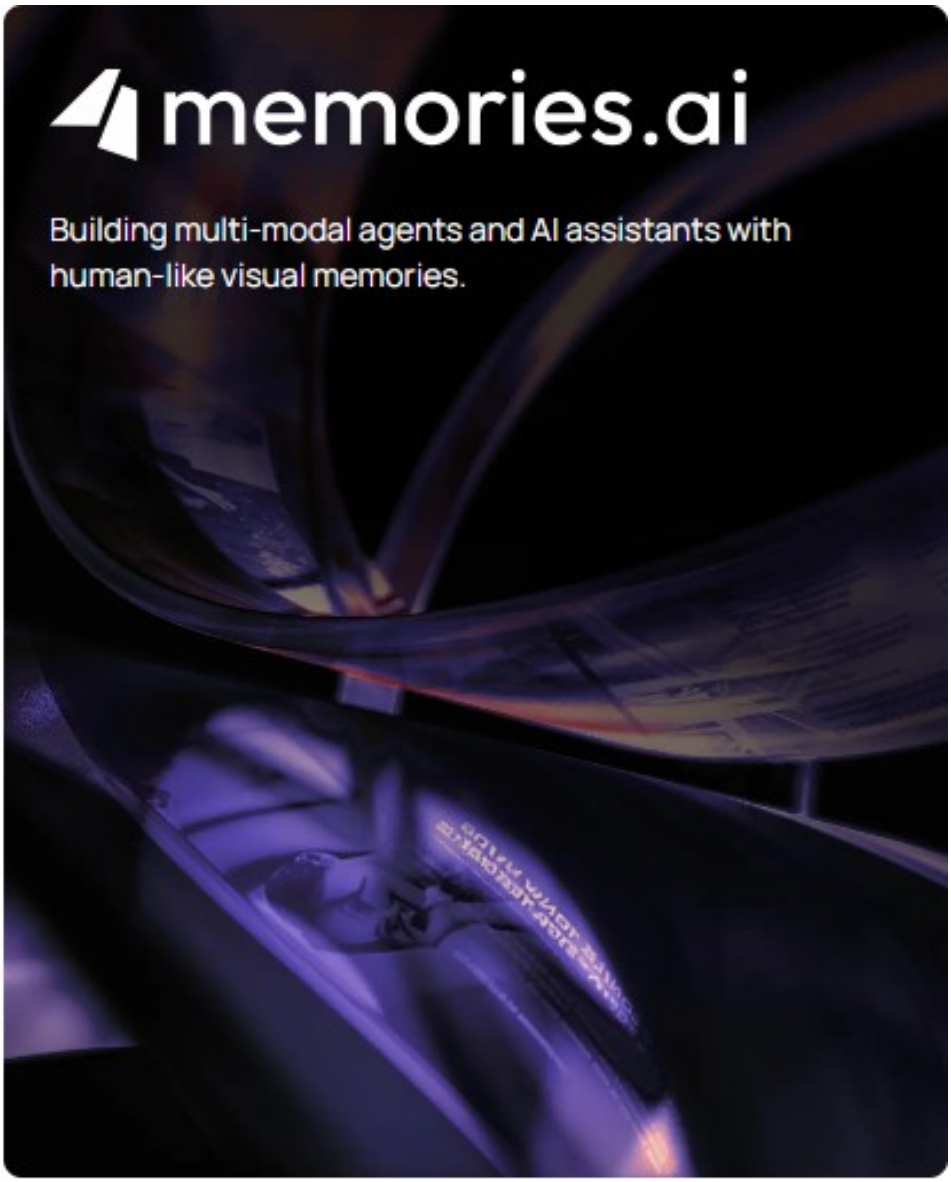
Hugging Face Demo： https://huggingface.co/spaces/Qwen/Qwen3-MT-Demo
ModelScope Demo： https://modelscope.cn/studios/Qwen/Qwen3-MT-demo
API Doc： https://alibabacloud.com/help/en/model-studio/translation-abilities
Blog： https://qwenlm.github.io/blog/qwen-mt/

• 技术动态九

前 Meta 员工正式推出其首款大型视觉记忆模型

日期

2025-07-25



内容提要

前 Meta 员工沈俊潇（Shawn Shen）近日在海外社交平台 X 上宣布，他与 Enmin Zhou 共同创办的 Memories.ai 正式发布其首个大型视觉记忆模型（Large Visual Memory Model）。该技术旨在为多模态大语言模型（multi-modal LLMs）赋予视觉记忆回溯与推理能力。

该模型具备强大的视觉记忆检索功能：可理解用户意图、检索相关视觉片段、整合关联信息，并基于记忆内容与用户查询进行深入推理。其核心创新在于引入智能决策机制，能够自主判断何时、如何、调取哪些视觉记忆。在完成记忆整合后，模型可生成引用并以标准格式输出回答，使多模态大模型具备处理“无限长”视觉上下文的能力。现在用户可以享受首月免费体验。

相关链接

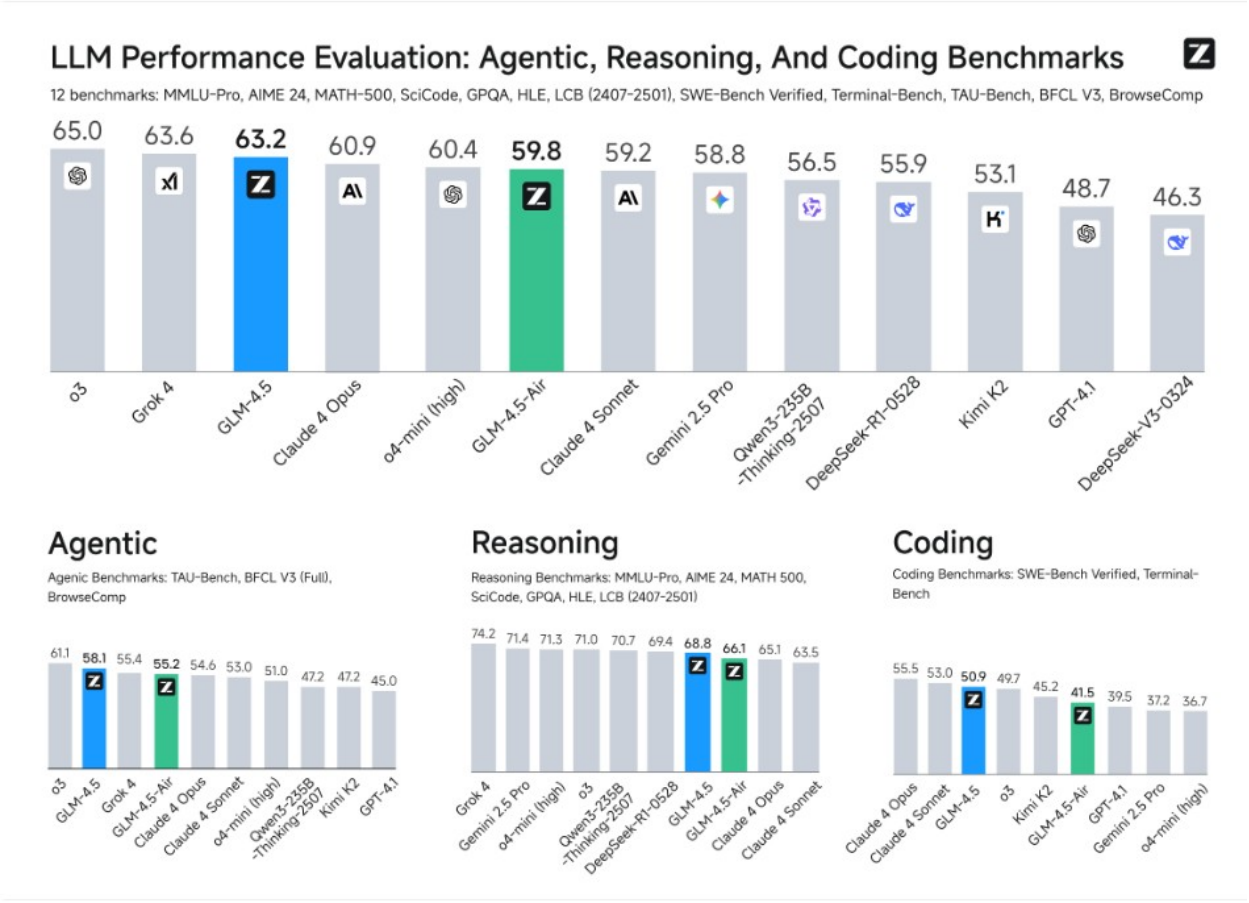
WebSite: https://memories.ai/app

• 技术动态十

智谱发布大模型 GLM 4.5，综合性能超过 Claude 4、QWen 3

日期

2025-07-28



内容提要

Zhipu AI 近期发布了旗舰大模型系列 GLM- 4.5 和 GLM- 4.5 Air。GLM- 4.5 拥有 3550 亿总参数(其中 320 亿为激活参数），Air 版本则轻量许多，仅为 1060 亿总参、激活参数 120 亿。两款模型支持 混合推理模式：包括用于复杂推理与工具调用的“思考模式”（thinking mode），以及用于快速响应的“非思考模式”（non- thinking mode）。它们具备超长上下文窗口（达 128K tokens）、原生函数调用能力，开放权重并支持 HuggingFace、ModelScope、本地部署与 Z.ai 平台调用。

在涵盖 agentic（代理决策）、逻辑推理与代码生成的 12 项主流 Benchmark 上，GLM- 4.5 表现位列整体第三，GLM- 4.5 Air 排名第六。尤其在 TAU-bench、BFCL-v3、BrowseComp 等代理任务中对标顶尖模型;在数学和逻辑推理(如 AIME24、MATH500、GPQA)以及 SWE- bench Verified 和 Terminal Bench 编码任务中均表现优异。同时，它提供极具性价比的推理成本，并且模型权重开源、许可宽松，便于定制本地化部署。

相关链接

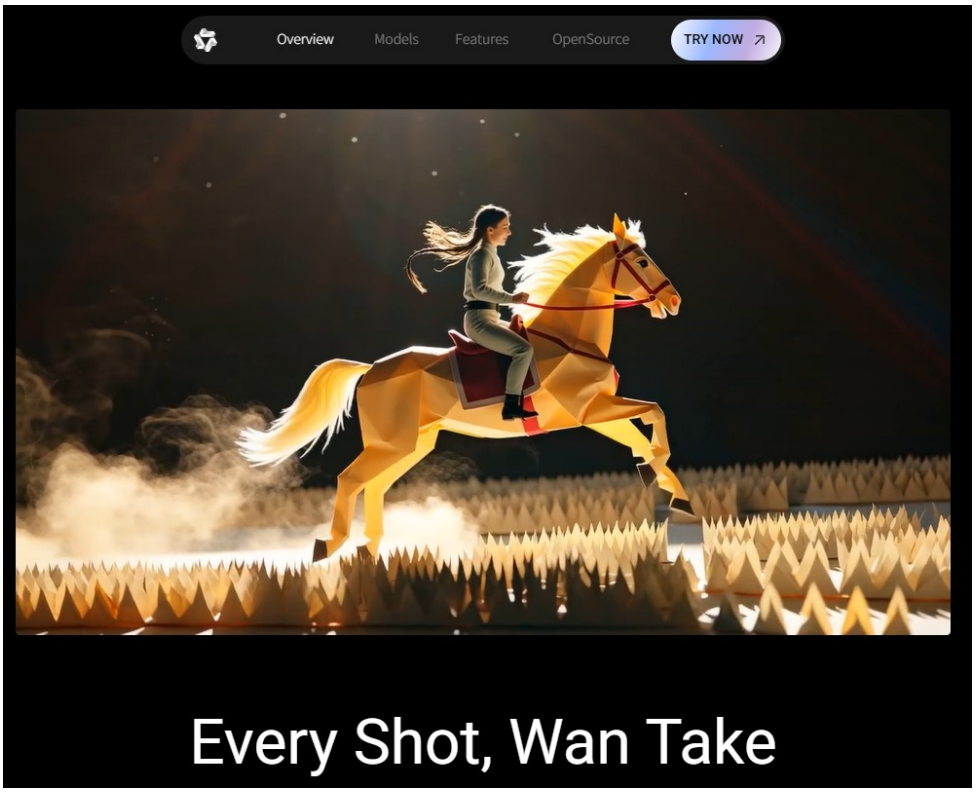
Hugging Face： https://huggingface.co/collections/zai-org/glm-45-687c621d34bda8c9e4bf503b

• 技术动态十一

阿里巴巴开源视频生成 MOE 大模型 Wan 2.2

日期

2025-07-28



内容提要

Wan.Video 是由阿里巴巴 Tongyi Lab 推出的开放式 AI 视频创作平台与大型视频生成模型套件，核心产品为 Wan 系列模型（如 Wan 2.1、Wan 2.2）。Wan 提供从文本或静态图像生成高质量视频的能力，支持多任务场景（文本转视频、图像转视频、视频编辑等）。Wan 模型具有开源授权（Apache- 2.0），支持在消费级 GPU（如 RTX 4090）下运行，兼顾高性能与资源可及性。

相关链接

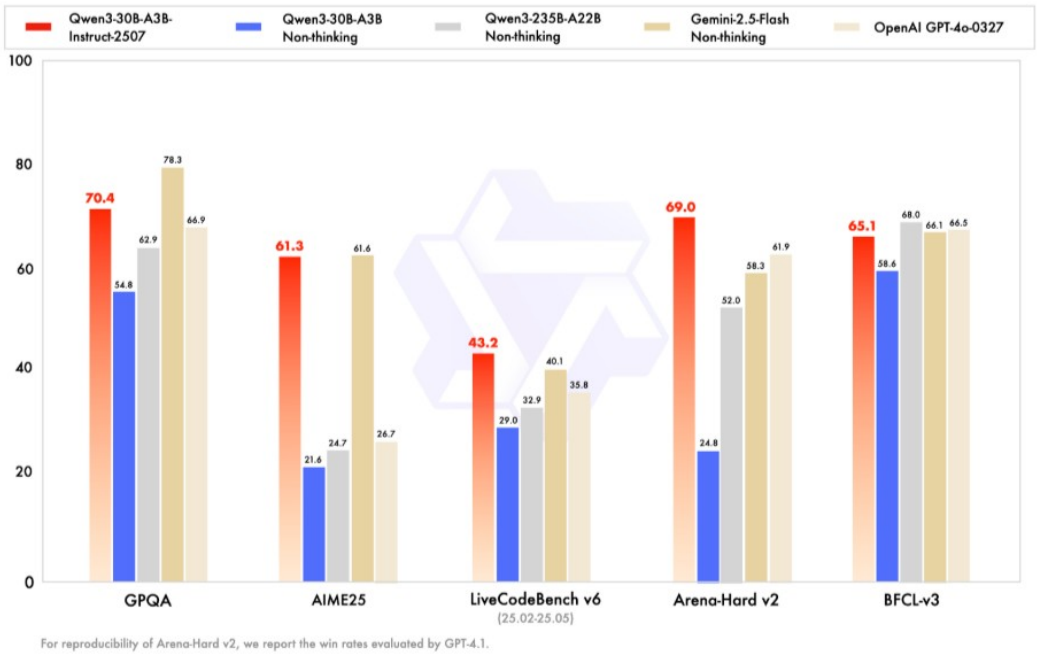
Wan Web: https://wan.video/welcome
GitHub: https://github.com/Wan-Video/Wan2.2
Hugging Face: https://huggingface.co/Wan-AI
ModelScope: https://modelscope.cn/organization/Wan-AI

• 技术动态十二

Qwen3-30B-A3B-Instruct-2507 发布，3B 激活参数 3090 可运行

日期

2025-07-28



内容提要

Qwen3-30B-A3B-Instruct-250 为一款非思考模式 (non-thinking mode) 模型，亮点在于仅激活 30 亿 (3B) 参数，便展现出与业界领先的闭源模型——如谷歌的 Gemini 2.5-Flash (非思考模式) 和 OpenAI 的 GPT-4o——相媲美的强大性能。这标志着模型在效率与性能优化方面取得了重要突破。与更新前的 Qwen3-30B-A3B 相比，新版本在多项测试中都实现了跨越式提升，比如 AIME25 从之前的 21.6 提升到了 61.3，Arena-Hard v2 成绩从 24.8 提升到了 69.0。

相关链接

Hugging Face: https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507

RESOURCE SHARING

资源分享

资源分享一

《Foundations of Large Language Models》

类型：电子书

内容提要

本书系统性地回顾和总结了大语言模型（LLM）的基础理论与技术发展脉络，涵盖其结构设计、训练机制、能力表现、对齐方法、多模态扩展以及面临的挑战等多个方面。作者从底层原理出发，梳理了 Transformer 架构演化、预训练与微调策略、多语言与多任务能力的涌现规律，同时探讨了大模型的对齐、安全、推理效率和未来发展方向，是一本关于 LLM 的具有指导性的综述类书籍。

资源链接：

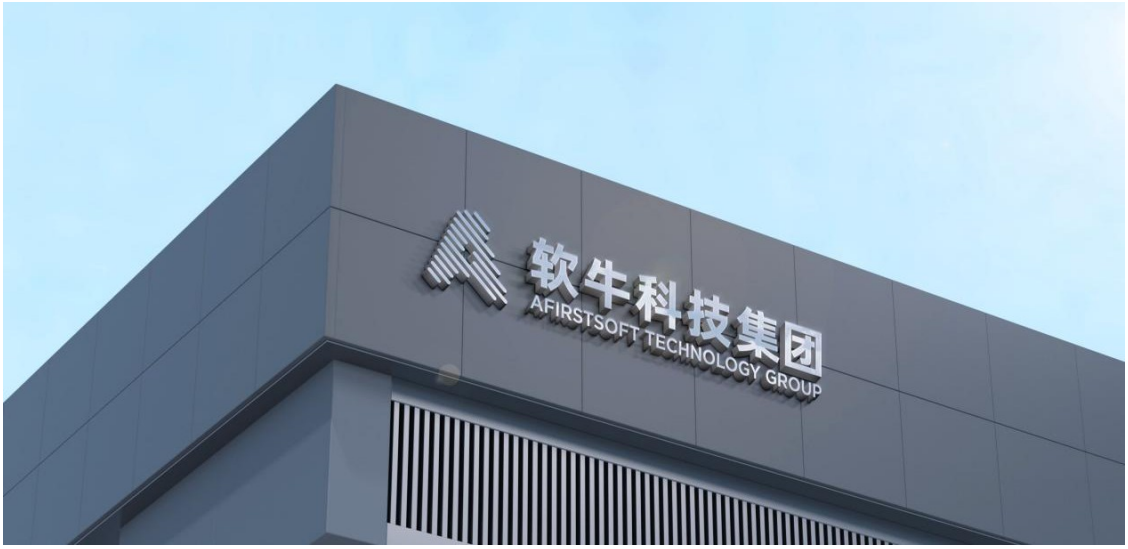
Paper: https://arxiv.org/abs/2501.09223v2

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

2025 年第 07 期

总第 24 期

广东省图象图形学会

主办 **广东省图象图形学会**

本期编委

人工智能
前沿快报

2025 年第 07 期
总第 24 期
2025.08.04

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，文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、科普宣传、科技成果鉴定