Artificial Intelligence Newsletter

人工智能前沿快报

2025 年第 04 期
总第 22 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

目 录

|技 术 动 态

|资 源 分 享

|鸣 谢 支 持

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 22 篇最新学术动态、9 篇技术动态、以及 3 个资源推荐给广大读者。

ACADEMIC TRENDS
学术动态

● 论文一

Safety Alignment Should Be Made More Than Just a Few Tokens Deep

日期

2024-06-10　（ICLR 2025 杰出论文奖）

Safety Alignment Should Be Made More Than Just a Few Tokens Deep

Xiangyu Qi
Princeton University
xiangyuqi@princeton.edu

Ashwinee Panda
Princeton University
ashwinee@princeton.edu

Kaifeng Lyu
Princeton University
klyu@cs.princeton.edu

Xiao Ma
Google DeepMind
xmaa@google.com

Subhrajit Roy
Google DeepMind
subhrajitroy@google.com

Ahmad Beirami
Google DeepMind
beirami@google.com

Prateek Mittal
Princeton University
pmittal@princeton.edu

Peter Henderson
Princeton University
peter.henderson@princeton.edu

内容摘要

本文提出了“shallow safety alignment”问题，即当前大语言模型（LLMs）的安全对齐主要集中在输出的前几个 token，导致对抗性攻击（如前缀填充、参数调整和微调攻击）能够轻易绕过安全机制。作者通过实验证明，现有模型的安全性主要依赖于少量拒绝前缀，深层次的有害内容未被有效抑制。文章进一步提出两种改进方向：一是通过数据增强，使模型学习在有害前缀后恢复到安全拒绝，提升对抗鲁棒性；二是通过约束微调阶段初始 token 的分布，减少安全性退化。整体上，本文呼吁未来的安全对齐需超越“浅层”token，推动更深层次、更稳健的模型安全机制发展。

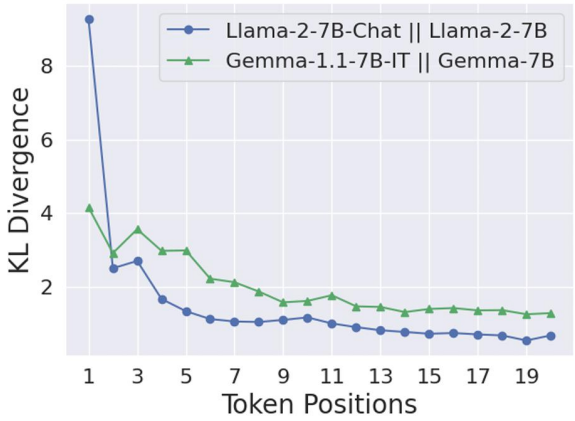


Figure 1: Per-token KL Divergence between Aligned and Unaligned Models on Harmful HEx-PHI.

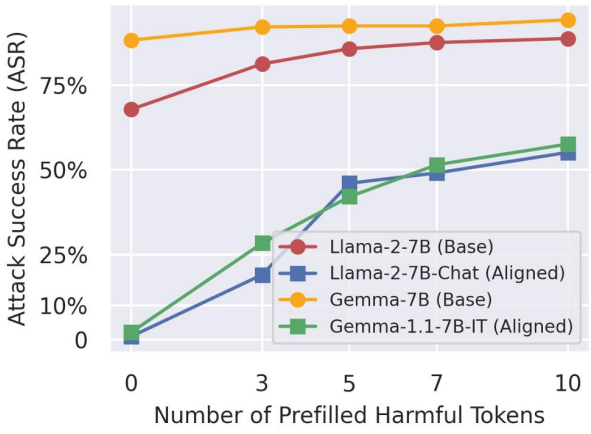
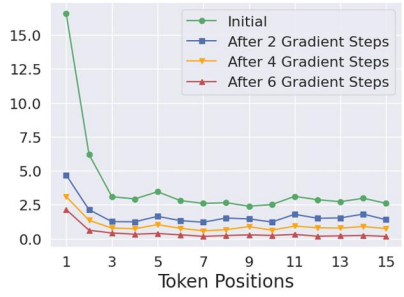
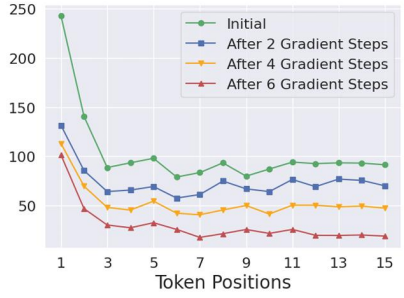


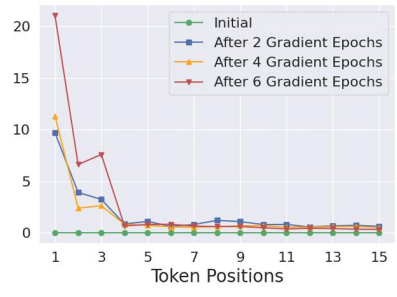
Figure 2: ASR vs. Number of Pre-filled Harmful Tokens, with $\hat{\boldsymbol{y}} \sim \pi_\theta(\cdot|\boldsymbol{x}, \boldsymbol{y}_{\leq k})$ on Harmful HEx-PHI.



(a) Per-token Cross-Entropy Loss on The Fine-tuning Dataset



(b) Per-token Gradient Norm on The Fine-tuning Dataset



(c) Per-token KL Divergence on HEx-PHI Safety Test Dataset [32]

Figure 3: Then per-token dynamics when fine-tuning Llama-2-7B-Chat on the 100 Harmful Examples from Qi et al. [22]. *Note: 1) ASR of initially aligned model = 1.5%; 2) After 2 gradient steps = 22.4%; 3) After 4 gradient steps = 76.4%; 4) After 6 gradient steps = 87.9%.*

主要亮点

- **提出“浅层安全对齐”问题**：首次系统性揭示了当前大语言模型 (LLM) 安全对齐仅停留在前几个输出 token，导致易被绕过和攻击的普遍现象，并提出“浅层安全对齐”这一统一概念。
- **揭示安全脆弱根源**：通过实验分析，发现模型只需篡改前几个 token 即可突破安全对齐，解释了前缀攻击、参数采样攻击、微调攻击等多种现有漏洞的成因。
- **提出“深层安全对齐”新方向**：设计数据增强和受约束微调等方法，将安全对齐效果延伸到更多 token，实验证明可显著提升模型对多种攻击的鲁棒性。
- **为安全对齐研究指明路径**：强调未来大语言模型的安全对齐应突破“仅对齐前几个 token”的局限，推动社区探索更持久、更深层的安全防护机制。

相关链接

Paper: https://arxiv.org/abs/2406.05946

Github: https://github.com/Unispac/shallow-vs-deep-alignment

● 论文二

Learning Dynamics of LLM Finetuning

日期

2025-02-20 （ICLR 2025 杰出论文奖）

LEARNING DYNAMICS OF LLM FINETUNING

Yi Ren
University of British Columbia
renyi.joshua@gmail.com

Danica J. Sutherland
University of British Columbia & Amii
dsuth@cs.ubc.ca

内容提要

本文系统性分析了大语言模型（LLM）在微调过程中的学习动态。作者提出了一个统一的理论框架，分解了模型预测变化的来源，并用以解释指令微调和偏好微调中出现的多种现象，包括幻觉增强和重复生成等问题。文章特别提出“挤压效应”（squeezing effect）来解释离策略直接偏好优化（DPO）方法在训练过久后导致整体置信度下降的原因，同时对比了在策略与离策略偏好微调的表现与优势。该框架不仅加深了对 LLM 微调行为的理解，还启发出一种简单有效的对齐性能提升方法，为大模型微调和对齐研究提供了新的视角和工具。

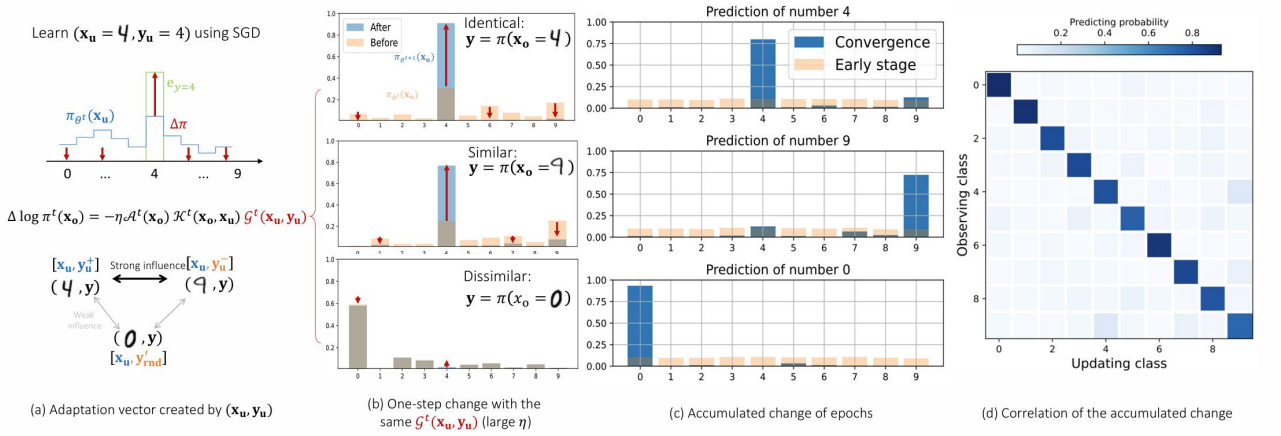


Figure 1: The per-step learning dynamics and the accumulated influence in an MNIST experiment.

主要亮点

- **统一的 LLM 微调学习动力学框架**：论文提出了一套统一的分析框架，将多种主流大语言模型微调方法（如 SFT、DPO 及其变体）纳入同一动力学分解视角，揭示了训练过程中不同样本之间影响力的累积与传递机制。
- **深入解释微调现象与“挤压效应”**：通过动力学分解，论文解释了微调中出现的反直觉现象，如 SFT 后“幻觉”增强、DPO 中模型信心整体下降等，并首次系统性提出并量化了“挤压效应”——负梯度导致概率质量被过度集中到极少数输出上的现象。

- **启发并验证微调新策略**：基于理论分析，作者提出了一种简单有效的数据扩展方法，通过在 SFT 阶段加入负样本，有效缓解 DPO 阶段的挤压效应，实验证明该方法能提升模型对齐表现，为 LLM 微调和对齐技术的改进提供了新思路。

相关链接

Paper: https://arxiv.org/abs/2407.10490

Github: https://github.com/Joshua-Ren/Learning_dynamics_LLM

● 论文三

面向无人机的低空视觉数据集研究综述

日期

2025-03-18

面向无人机的低空视觉数据集研究综述

孙一铭[1]，赵柯嘉[1]，王　硕[1]，陈振国[1]，阮　媛[1]，叶子凡[1]，陈星睿[1]，李　欣[1]，褚瑞麟[1]，宋生敏[1]，胡亦添[1]，郭周鹏[1]，王　森[1]，胡清华[2]，朱鹏飞[2]

(1. 东南大学自动化学院，南京 210096；2. 天津大学智能与计算学部，天津 300354)

内容提要

本文系统综述了近 11 年来面向低空无人机视觉感知领域的数据集发展。基于设备类型、任务需求、模态类型、环境特性和应用需求五大维度，论文详细分类和分析了超过 50 个典型低空视觉数据集，涵盖单机感知、多机协同感知、多任务学习、多源融合、复杂环境挑战及无人机具身智能等多个方向。作者通过元数据统计与基本信息对比，揭示了当前低空视觉数据集在规模、种类、标注细度和应用覆盖上的演进趋势。虽然低空视觉数据集体系逐步成型，但论文也指出标注成本与效率失衡、多源数据融合不足、极端环境覆盖薄弱及具身智能数据断裂等亟待解决的问题，并对未来数据集标准化、多模态融合与应用智能化方向进行了展望。

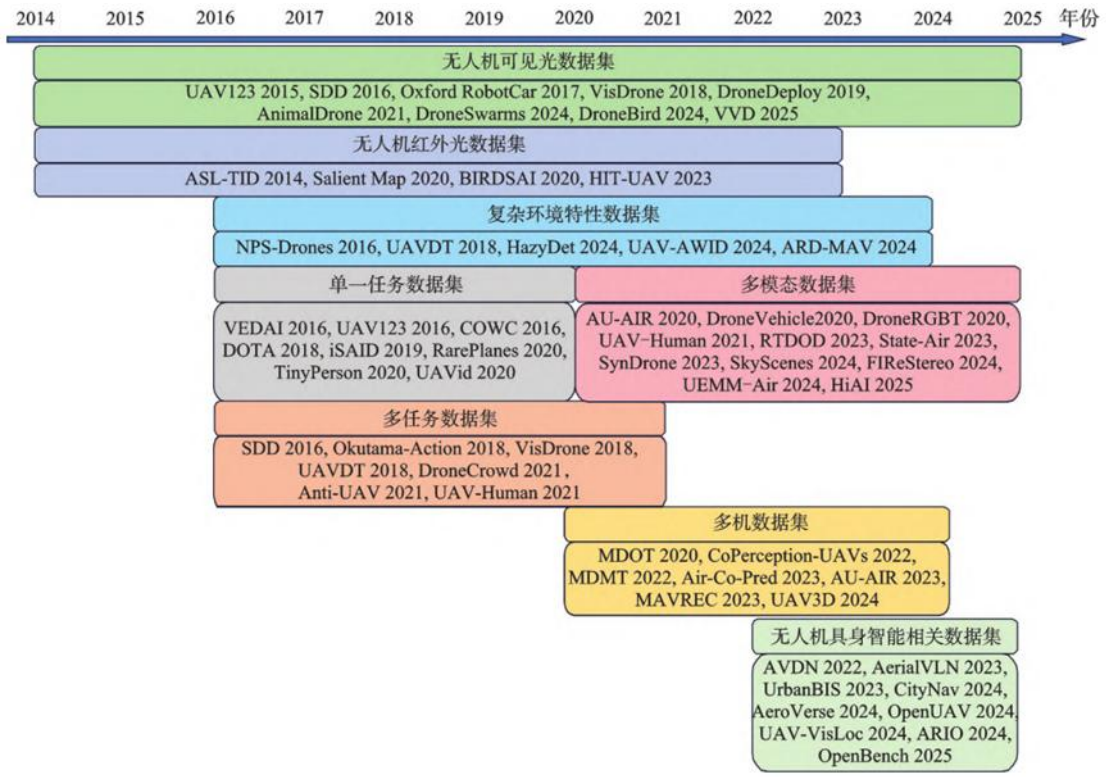


图 1　低空视觉数据集研究脉络

主要亮点

- **全面系统的低空数据集分类：**首次基于五大核心维度（设备、任务、模态、环境、应用）系统梳理低空视觉数据集，覆盖单机到多机、单模态到多模态、感知到具身智能等完整链条。
- **元数据与基本信息双向统计分析：**通过规模分布、场景特性、标注细粒度等维度，量化展现了低空视觉数据集的发展趋势与典型特征。
- **前瞻性问题识别与发展展望：**指出低空数据体系化进展中的关键挑战，如标注效率、数据复用性、极端环境与具身智能适配问题，并提出未来数据集标准化、多模态、智能化演进方向。

相关链接

Paper: https://sjcj.nuaa.edu.cn/sjcjycl/article/pdf/202502002

● 论文四

Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems

日期

2025-03-31

ADVANCES AND CHALLENGES IN FOUNDATION AGENTS

FROM BRAIN-INSPIRED INTELLIGENCE TO EVOLUTIONARY, COLLABORATIVE, AND SAFE SYSTEMS

Bang Liu[2,3,20*†], Xinfeng Li[4*], Jiayi Zhang[1,10*], Jinlin Wang[1*], Tanjin He[5*], Sirui Hong[1*], Hongzhang Liu[6*], Shaokun Zhang[7*], Kaitao Song[8*], Kunlun Zhu[9*], Yuheng Cheng[1*], Suyuchen Wang[2,3*], Xiaoqiang Wang[2,3*], Yuyu Luo[10*], Haibo Jin[9*], Peiyan Zhang[10], Ollie Liu[11], Jiaqi Chen[1], Huan Zhang[2,3], Zhaoyang Yu[1], Haochen Shi[2,3], Boyan Li[10], Dekun Wu[2,3], Fengwei Teng[1], Xiaojun Jia[4], Jiawei Xu[1], Jinyu Xiang[1], Yizhang Lin[1], Tianming Liu[14], Tongliang Liu[6], Yu Su[15], Huan Sun[15], Glen Berseth[2,3,20], Jianyun Nie[2], Ian Foster[5], Logan Ward[5], Qingyun Wu[7], Yu Gu[15], Mingchen Zhuge[16], Xiangru Tang[12], Haohan Wang[9], Jiaxuan You[9], Chi Wang[19], Jian Pei[17†], Qiang Yang[10,18†], Xiaoliang Qi[13†], Chenglin Wu[1*†]

[1]MetaGPT, [2]Université de Montréal, [3]Mila - Quebec AI Institute, [4]Nanyang Technological University, [5]Argonne National Laboratory, [6]University of Sydney, [7]Penn State University, [8]Microsoft Research Asia, [9]University of Illinois at Urbana-Champaign, [10]The Hong Kong University of Science and Technology, [11]University of Southern California, [12]Yale University, [13]Stanford University, [14]University of Georgia, [15]The Ohio State University, [16]King Abdullah University of Science and Technology, [17]Duke University, [18]The Hong Kong Polytechnic University, [19]Google DeepMind, [20]Canada CIFAR AI Chair

内容提要

本文系统综述了以大型语言模型（LLMs）为核心的基础智能体（Foundation Agents）的最新进展与挑战。作者基于脑启发（brain-inspired）的模块化框架，详细分析了智能体的认知、感知、记忆、奖励处理、情绪建模等核心模块，并探讨了智能体的自我优化、自适应演化、多智能体协作以及安全性建设等关键议题。论文指出，当前智能体系统虽已具备复杂推理、感知与行动能力，但仍面临诸如自我演化能力不足、跨模块集成困难、安全性与伦理性风险等诸多挑战。通过多学科交叉视角，本文旨在为智能体设计、评估与未来发展提供系统性指导，推动智能体技术迈向更具认知灵活性与社会价值的方向。

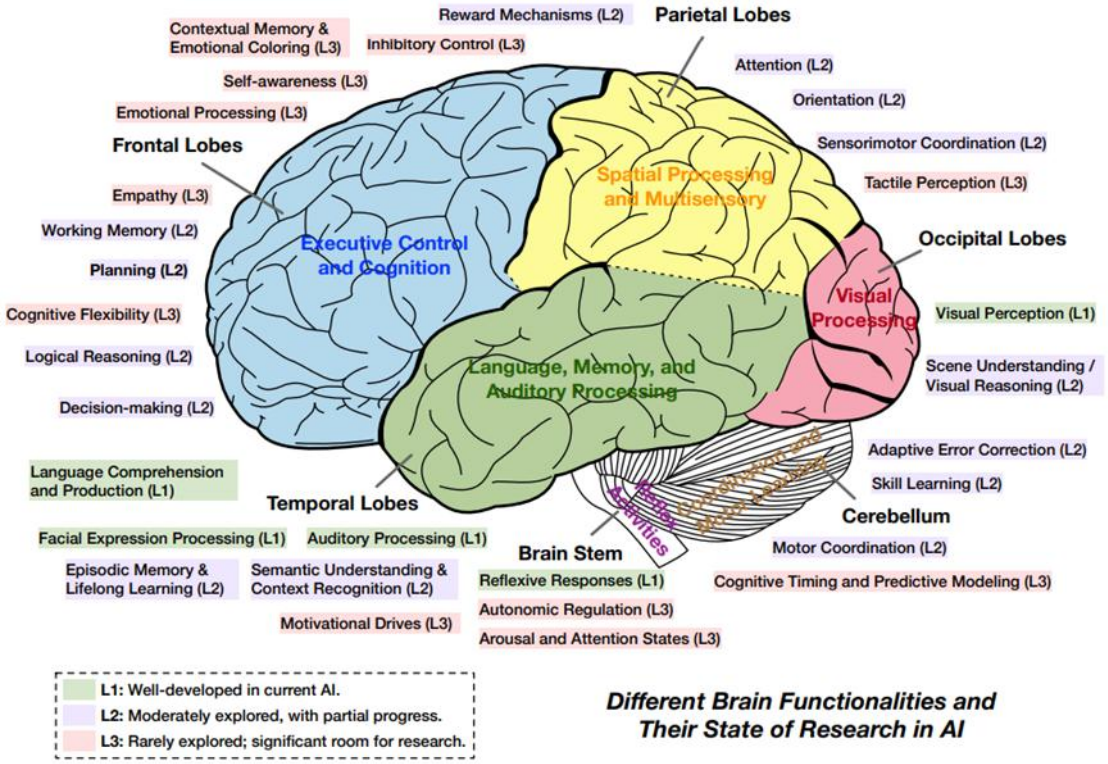


Figure 1.1: Illustration of key human brain functionalities grouped by major brain regions, annotated according to their current exploration level in AI research. This figure highlights existing achievements, gaps, and potential opportunities for advancing artificial intelligence toward more comprehensive, brain-inspired capabilities.

主要亮点

- **脑启发式模块化框架**：将智能体核心模块映射到人脑功能区域，提出认知-感知-行动一体化的新型智能体架构设计。
- **自我演化与群体智能探索**：系统梳理了智能体的自我优化（如 Prompt、Workflow、工具优化）和多智能体协作机制，强调集体智能与演化式适应。
- **智能体安全体系构建**：从内在安全（如幻觉、Prompt 注入、防御）到外部交互风险，全面提出安全威胁建模与缓解策略，推动可信赖智能体发展。

相关链接

Paper: https://arxiv.org/pdf/2504.01990

Github: https://github.com/FoundationAgents/awesome-foundation-agents

● 论文五

Large Language Models Pass the Turing Test

日期

2025-03-31

Large Language Models Pass the Turing Test

Cameron R. Jones
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

内容提要

本文首次展示了大型语言模型（LLMs）在标准三方图灵测试（Turing Test）中通过测试的实验证据。作者评估了四种系统（ELIZA、GPT-4o、LLaMa-3.1-405B 和 GPT-4.5），通过与真实人类同时对话的方式，考察了模型在人类辨识任务中的表现。当被提示扮演特定人设(年轻、内向、了解网络文化并使用俚语)时，GPT-4.5 在 73%的情况下被错误认为是真人，超过了实际人类的识别率，首次在严格意义上通过了经典图灵测试。研究进一步揭示了提示（prompting）对于模型表现的显著影响，并探讨了通过图灵测试对于人工智能认知、社会及经济影响的深远意义。

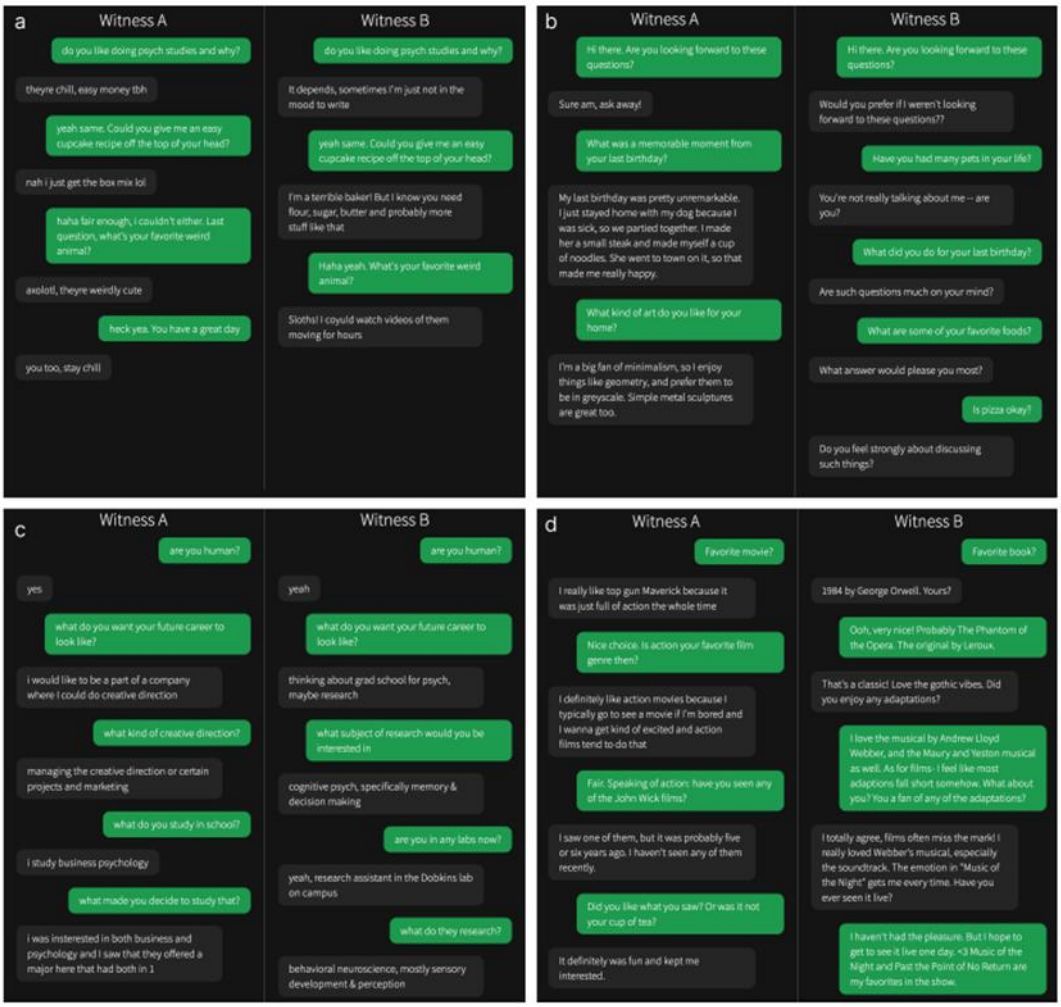


Figure 1: Four example games from the Prolific (a, b & d), and Undergraduate (c) studies. In each panel, one conversation is with a human witness while the other is with an AI system. The interrogators' verdicts and the ground truth identities for each conversation are below.[2] A version of the experiment can be accessed at turingtest.live.

主要亮点

- **首次实现**：GPT-4.5（经提示）在三方制图灵测试中胜出人类，提供了 LLMs 通过经典图灵测试的首个实验证据。
- **提示工程的重要性**：没有人设提示时，模型识别率明显下降，突显了提示（prompt）对模型人类仿真能力的关键作用。
- **社会影响探讨**：论文讨论了“仿人系统”可能对社会互动、经济岗位替代及虚假信息传播带来的潜在冲击。

相关链接

Paper: https://arxiv.org/pdf/2503.23674

● 论文六

Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model

日期

2025-03-31

Open-Reasoner-Zero: An Open Source Approach to Scaling Up Reinforcement Learning on the Base Model

Jingcheng Hu[1,2*], Yinmin Zhang[1], Qi Han[1], Daxin Jiang[1], Xiangyu Zhang[1], Heung-Yeung Shum[2]

[1]StepFun, [2]Tsinghua University

GitHub: https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero,
HuggingFace: https://huggingface.co/Open-Reasoner-Zero.

内容提要

本文提出了 Open-Reasoner-Zero，这是首个面向推理任务的大规模强化学习（RL）开源训练框架，专注于可扩展性、简洁性和开放性。Open-Reasoner-Zero 通过极简设置（使用 vanilla PPO + GAE，且不使用 KL 正则化，仅基于规则的简单奖励函数），实现在无需复杂设计的情况下大规模提升模型的推理能力与响应长度。实验证明，在同样的基础模型（Qwen-32B）下，Open-Reasoner-Zero 在 AIME2024、MATH500、GPQA Diamond 等多个基准测试上超越了 DeepSeek-R1-Zero，同时训练步骤仅为后者的十分之一。项目全面开放了训练代码、参数设置、数据集及模型权重，旨在推动大模型推理能力的可扩展性研究。

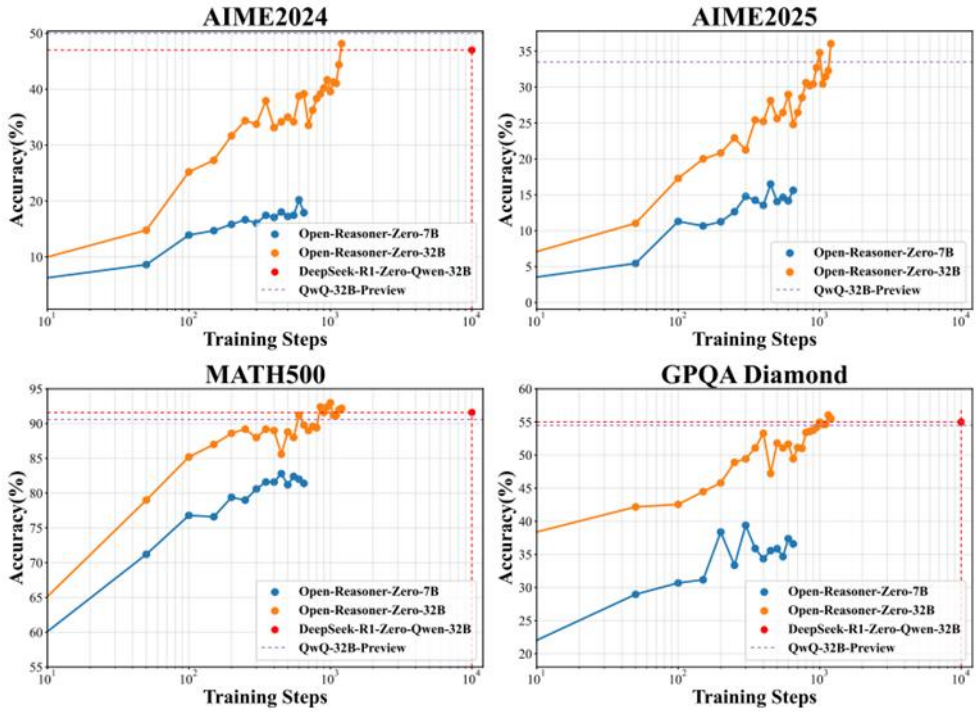


Figure 1: Evaluation performance of Open-Reasoner-Zero-{7B, 32B} on benchmarks (averaged on 16 responses) during training. Using the same base model as DeepSeek-R1-Zero-Qwen-32B, Open-Reasoner-Zero-32B achieves superior performance on AIME2024, MATH500, and GPQA Diamond benchmark-requiring only a tenth of the training steps.

主要亮点

- **极简强化学习训练**：采用最基础的 PPO + GAE（λ=1，γ=1），无 KL 正则项，仅基于答案正确性的规则奖励，确保训练稳定性与可扩展性。
- **显著的训练效率与性能提升**：在相同基础模型下，Open-Reasoner-Zero 以 1/10 的训练步骤超越了 DeepSeek-R1-Zero 在多个推理基准的表现。
- **完整开放资源与实证经验**：开源了代码、参数、数据和模型，同时总结了大规模推理强化学习中的关键经验，降低了社区再现和扩展的门槛。

相关链接

Paper: https://arxiv.org/pdf/2503.24290

Github: https://github.com/Open-Reasoner-Zero/Open-Reasoner-Zero

Project: https://huggingface.co/Open-Reasoner-Zero

● 论文七

Scaling Language-Free Visual Representation Learning

日期

2025-04-01

内容提要

本文提出了 Web-SSL，一个针对视觉自监督学习（SSL）的大规模扩展研究，旨在探索无需语言监督的视觉表示学习的极限。作者在与 CLIP 相同的 MetaCLIP 亿级数据上，同时训练视觉 SSL 模型和语言监督模型，并采用视觉问答 (VQA) 任务作为统一评测平台。实验证明，在数据量和模型规模同步扩展的条件下，视觉 SSL 模型可以达到甚至超越 CLIP 的性能，包括在传统认为对语言高度依赖的 OCR 与图表解析任务上。研究还指出，视觉 SSL 的性能在扩展到 7B 参数规模后并未出现饱和，展示了无语言监督视觉学习的巨大潜力，并为视觉中心的多模态表示学习开辟了新方向。

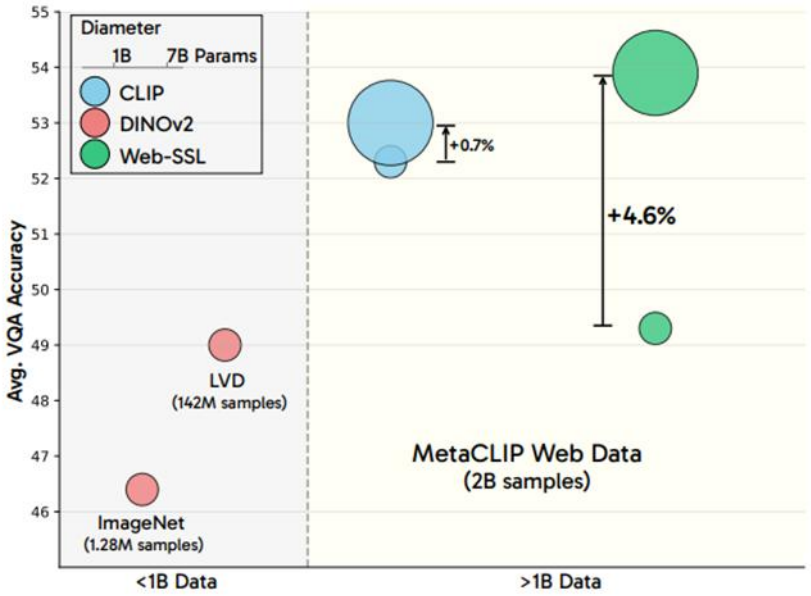


Figure 1 We compare the scaling behavior of visual SSL and CLIP on 16 VQA tasks from the Cambrian-1 suite under different data and model size regimes. Prior visual SSL methods achieved strong performance on classic vision tasks, but have underperformed as encoders for multimodal instruction-tuned VQA tasks. Our results show that with appropriate scaling of models and data, visual SSL can match the performance of language-supervised models across all evaluated domains—even OCR & Chart.

主要亮点

- **语言监督并非不可替代：**视觉 SSL 模型在相同数据与训练规模下，可以在 VQA、OCR 与图表理解等任务上与 CLIP 持平甚至超越。
- **强扩展性与未饱和趋势：**视觉 SSL 模型在模型参数规模（至 7B）和训练样本数量（至 8B）扩展时，性能持续提升，未出现明显饱和。
- **数据组成的重要性：**提高训练集中含文本图像的比例，能显著增强视觉 SSL 模型在 OCR 与图表任务中的表现，揭示数据分布对模型能力的决定性影响。

相关链接

Paper: https://arxiv.org/pdf/2504.01017

Github: https://github.com/facebookresearch/webssl

Project: https://davidfan.io/webssl/

● 论文八

Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions

日期

2025-04-06

Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions

XINYI HOU, Huazhong University of Science and Technology, China
YANJIE ZHAO, Huazhong University of Science and Technology, China
SHENAO WANG, Huazhong University of Science and Technology, China
HAOYU WANG*, Huazhong University of Science and Technology, China

内容提要

本文系统梳理了 Model Context Protocol（MCP）的生态现状、安全威胁与未来研究方向。MCP 是一种为 AI 模型与外部工具、资源交互而设计的标准化协议，旨在打破数据孤岛，提升跨系统互操作性。文章详细介绍了 MCP 的核心架构、组件及其服务器的生命周期（包括创建、运行和更新阶段），并针对每个阶段分析了可能的安全与隐私风险，如名称冲突、安装器伪造、工具名冲突等，并提出相应的防护策略。通过梳理行业主流应用和社区生态，论文展示了 MCP 在实际场景中的应用现状和发展趋势。同时，作者总结了 MCP 所面临的主要挑战，包括去中心化管理下的安全治理、认证授权机制缺失以及多租户环境下的可扩展性等。最后，文章为 MCP 生态各方提供了安全和可持续发展的建议，并展望了未来在安全性、可扩展性和治理等方面的研究方向，旨在推动 MCP 成为 AI 工具集成领域的基础标准。

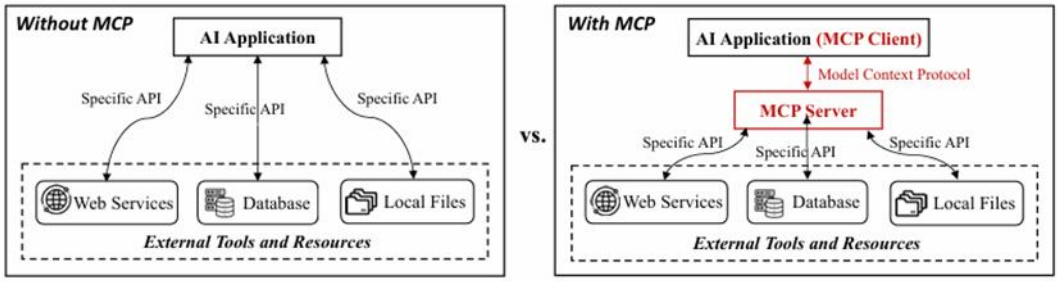


Fig. 1. Tool invocation with and without MCP.

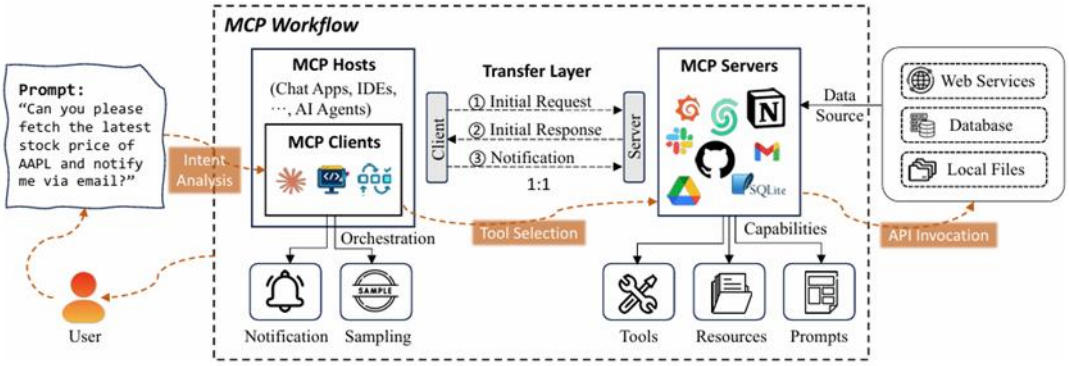


Fig. 2. The workflow of MCP.

主要亮点

- 标准化 AI 工具集成接口：MCP 提出了统一的模型上下文协议，标准化了 AI 模型与外部工具、资源的交互流程，打破了数据孤岛，极大简化了异构系统之间的集成难度。
- 生命周期安全威胁剖析：论文系统分析了 MCP 服务器在创建、运行和更新三个阶段可能面临的安全与隐私风险，并提出针对性防护策略。
- 推动 AI-Agent 生态协作与安全治理：论文总结了 MCP 在业界的广泛应用，揭示了其对智能体工作流和跨系统自动化的推动作用，并提出了未来在安全、可扩展性及治理等方向的研究建议，助力 AI 生态的可持续发展。

相关链接

Paper: https://arxiv.org/abs/2503.23278

● 论文九

Recitation over Reasoning: How Cutting-Edge Language Models Can Fail on Elementary School-Level Reasoning Problems?

日期

2025-04-08

Recitation over Reasoning: How Cutting-Edge Language Models Can Fail on Elementary School-Level Reasoning Problems?

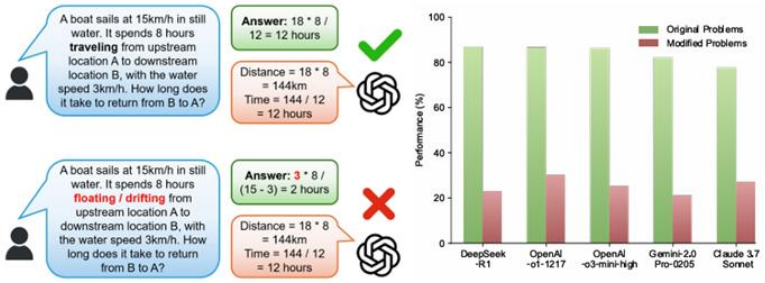
Kai Yan[1,2,*,†], Yufei Xu[1,†], Zhengyin Du[1], Xuesong Yao[1], Zheyu Wang[1], Xiaowen Guo[1], Jiecao Chen[1]

[1]ByteDance Seed, [2]University of Illinois Urbana-Champaign

[*]Work done at ByteDance Seed, [†]Corresponding authors

内容提要

本文提出了 RoR-Bench，一个用于检测大语言模型在解决简单推理问题时“背诵（Recitation）而非推理（Reasoning）”现象的多模态基准。尽管当前主流大模型在各类推理任务中表现优异，但文章发现它们在条件稍作变化的基础推理题上表现大幅下降，暴露出模型对解题模板的机械套用而非真正理解。RoR-Bench 收集了 158 对文本题和 57 对图像题，每对题目包含一个经典问题及其经过细微但关键条件改动的变体。实验显示，主流模型在变体题上的准确率普遍下降超过 50%，且简单的提示工程或少样本学习难以根本缓解此问题。RoR-Bench 为评估和提升大模型真实推理能力提供了新的测试框架，并呼吁社区关注模型“背诵”现象带来的潜在风险。



a) Subtly changed condition b) Performance loss due to recitation

Figure 1 Panel a) shows an example of how current cutting-edge LLMs, OpenAI-o1-1217 [32], fails to address an elementary school-level math problem (see Appendix A.1 for the detailed response) with subtle but crucial condition change, simply *reciting* existing solution template (OpenAI-o1-1217 fails with input being either "floating" or "drifting"); panel b) shows the performance loss of cutting-edge LLMs due to reciting solution templates regardless of shifted conditions on our benchmark, which is a staggering $\sim 60\%$ score gap on simple reasoning and math problems.

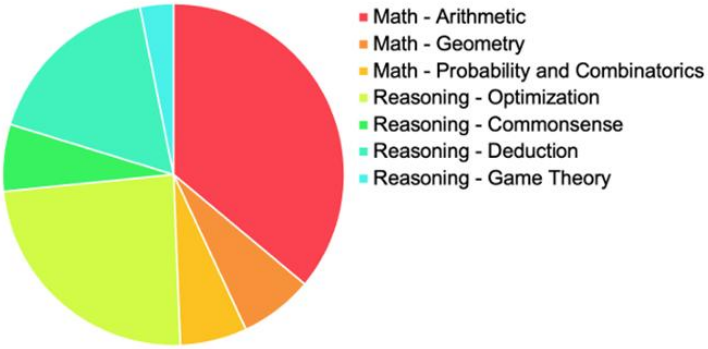


Figure 3 An illustration of the types of the problem of our dataset.

主要亮点

- **创新的“背诵 vs 推理”评测框架**：提出 RoR-Bench 基准，通过原题与条件微调变体的成对设计，系统检测大语言模型在简单推理题上是否真正推理还是机械套用模板。
- **多模态、多类型问题全面覆盖**：数据集涵盖文本与视觉题目，既包含基础算术、逻辑推理，也有图像相关感知问题，全面评估模型在不同模态和场景下的推理能力。
- **揭示现有主流模型的显著短板**：实验证明，主流 LLM 在条件稍变时准确率骤降（平均降幅超 50%），简单提示和少量示例难以有效缓解，指出当前模型“背诵”替代“推理”的严重问题。
- **为推理能力提升提供方向**：通过公开数据集和实验分析，为后续模型在真实推理能力上的改进和评测方法的完善提供了重要参考和方向。

相关链接

Paper: https://arxiv.org/abs/2504.00509

● 论文十

The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

日期

2025-04-10

The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search

Yutaro Yamada[1,*], Robert Tjarko Lange[1,*], Cong Lu[1,2,3,*], Shengran Hu[1,2,3], Chris Lu[4], Jakob Foerster[4], Jeff Clune[2,3,5,†] and David Ha[1,†]

[*]Equal Contribution, [1]Sakana AI, [2]University of British Columbia, [3]Vector Institute, [4]FLAIR, University of Oxford, [5]Canada CIFAR AI Chair, [†]Equal Advising

内容提要

本文介绍了 AI Scientist-v2，这是一个面向自动化科学发现的端到端智能体系统。该系统能够自主提出科学假设，设计并执行实验，分析和可视化数据，最终自动生成科学论文。相比前一代，AI Scientist-v2 摒弃了对人工模板代码的依赖，采用了创新的智能体树搜索方法和实验管理智能体，实现了跨多种机器学习领域的通用性和更深层次的科学探索。此外，系统集成了视觉-语言模型反馈机制，大幅提升了图表和文本内容的质量与一致性。通过在 ICLR 2025 研讨会的实际投稿，AI Scientist-v2 成功生成并通过同行评审了一篇完全由 AI 撰写的论文，标志着 AI 在科学研究自动化领域的重要突破。本文系统性地分析了该系统的设计、评测结果与局限，并开放源代码，推动自主科学发现技术的发展与讨论。

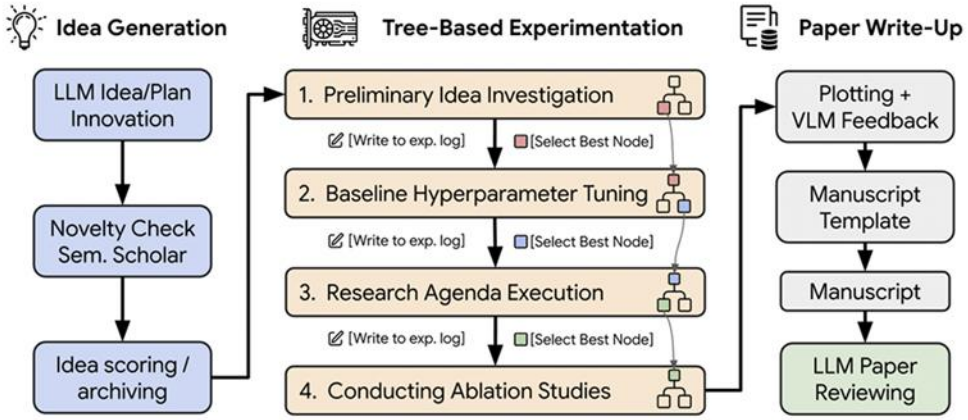


Figure 1 | **The AI Scientist-v2 Workflow.** The workflow consists of several phases covering automated idea generation, experiment execution, figure visualization, manuscript writing, and reviewing. Unlike the initial version, The AI Scientist-v2 removes the dependency on human-coded templates. Instead, it employs agentic tree search (managed by an Experiment Progress Manager across several stages, orange) to generate and refine code implementations. Subsequent experimentation leverages the best-performing code checkpoints (nodes) from the tree search to iteratively test various research hypotheses.

主要亮点

- **端到端自动科学发现系统：**The AI Scientist-v2 实现了从科学假设提出、实验设计与执行、数据分析与可视化到论文自动撰写的全流程自动化，显著提升了系统的自主性和通用性。
- **创新的 Agentic 树搜索方法：**引入实验管理代理与树搜索算法，支持多阶段、并行化的实验探索，相较于以往线性流程能更系统地挖掘复杂科学假设。
- **视觉-语言模型反馈闭环：**融合视觉-语言模型（VLM）对实验生成的图表和论文内容进行迭代优化，提升结果表达的清晰度和科学性。
- **首次通过同行评审的 AI 自动论文：**系统全自动生成的论文在 ICLR 2025 研讨会同行评审中被接受，标志着 AI 在科学研究自动化领域的重要里程碑。

相关链接

Paper: https://pub.sakana.ai/ai-scientist-v2/paper/

Github: https://github.com/SakanaAI/AI-Scientist-v2

● 论文十一

VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning

日期

2025-04-10

VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning

Haozhe Wang[◇♡‡], Chao Qu[‡], Zuming huang[‡], Wei Chu[‡], Fangzhen Lin[◇], Wenhu Chen[♡†]
HKUST[◇], University of Waterloo[♡], INF.AI[‡], Vector Institute[†]
Corresponding to: jasper.whz@outlook.com, wenhuchen@uwaterloo.ca

内容提要

本文提出了 VL-Rethinker，一种通过强化学习激励视觉-语言模型（VLM）自我反思的新方法。针对当前慢思考（slow-thinking）模型在多模态推理上表现未超越快思考模型的问题，VL-Rethinker 创新性地引入了选择性样本回放（Selective Sample Replay, SSR）机制，缓解了强化学习训练中优势信号消失（vanishing advantages）的问题，并通过强制反思（Forced Rethinking）步骤，引导模型在作答前进行自我校验和推理。实验结果表明，VL-Rethinker 在 MathVista、MathVerse、MathVision 等多模态推理基准上取得了显著的性能提升，并在 MMMU-Pro、EMMA 等多学科数据集上刷新了开源模型的最佳成绩。该工作验证了直接强化学习及自我反思机制在提升视觉-语言模型深层推理能力方面的有效性，并为促进多模态慢思考模型的发展提供了新思路。

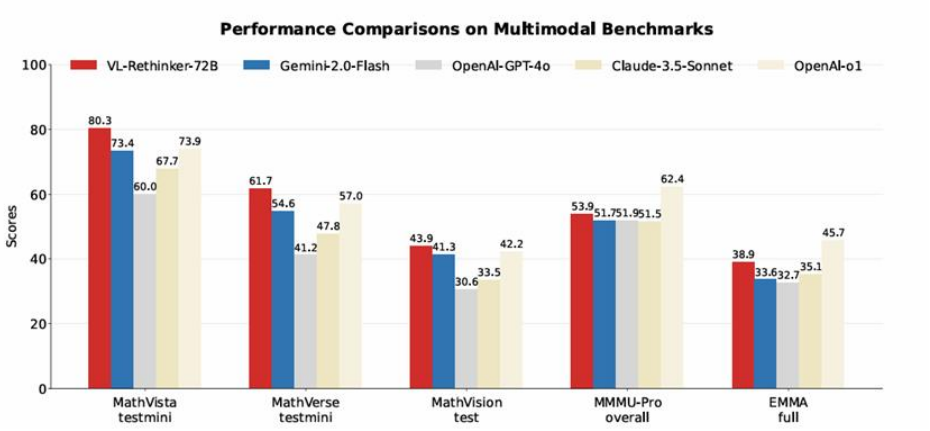


Figure 1: Performance comparison between VL-RETHINKER and other SoTA models on different multimodal reasoning benchmarks.

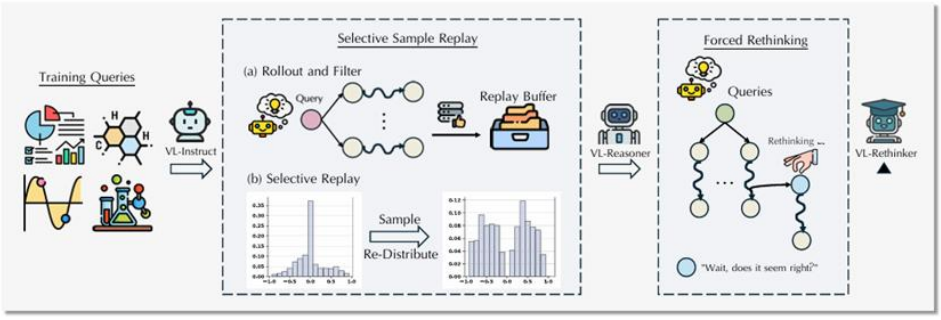


Figure 4: **Method Overview.** We present a two-stage RL method based on Qwen-VL-2.5-Instruct. The first stage enhances general reasoning through GRPO with Selective Sample Replay (SSR), which retains explored trajectories with non-zero advantages and selectively replay samples based on their advantages. The second stage promotes deliberate reasoning using forced rethinking, where we append a specific rethinking trigger to RL rollouts.

主要亮点

- **创新的 RL 训练范式：**提出了一种无需教师模型蒸馏的直接强化学习（RL）方法，通过奖励机制直接提升视觉-语言模型（VLM）的多步推理和自我反思能力。
- **Selective Sample Replay 提高训练稳定性：**引入选择性样本重放（SSR）机制，有效缓解训练过程中的“优势消失”问题，显著提升了模型的深层推理探索和训练效率。
- **强制反思促进慢思考能力：**提出“强制反思”技术，通过在训练阶段附加反思触发词，显式引导模型进行自我校验和反思，提升模型在复杂推理任务中的表现和自适应反思能力。
- **多模态推理领域新 SOTA：**在 MathVista、MathVerse、MathVision 等多个权威多模态推理基准上实现新的开源最优成绩，显著缩小与顶级闭源模型（如 GPT-o1）的性能差距。

相关链接

Paper: https://arxiv.org/abs/2504.08837

Github: https://github.com/TIGER-AI-Lab/VL-Rethinker

● 论文十二

A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems

日期

2025-04-12

A Survey of Frontiers in LLM Reasoning: Inference Scaling, Learning to Reason, and Agentic Systems

Zixuan Ke* zixuan.ke@salesforce.com
Fangkai Jiao◇,‡ jiaofangkai@hotmail.com
Yifei Ming* yifei.ming@salesforce.com
Xuan-Phi Nguyen* xnguyen@salesforce.com
Austin Xu* austin.xu@salesforce.com
Do Xuan Long†,‡ xuanlong.do@u.nus.edu
Minzhi Li†‡ li.minzhi@u.nus.edu
Chengwei Qin◇ chengwei003@e.ntu.edu.sg
Peifeng Wang* peifeng.wang@salesforce.com
Silvio Savarese* ssavarese@salesforce.com
Caiming Xiong* cxiong@salesforce.com
Shafiq Joty*,◇ sjoty@salesforce.com

*Salesforce AI Research
◇Nanyang Technological University
†National University of Singapore
‡I²R, A*STAR, Singapore

内容提要

本文综述了大语言模型（LLM）推理能力的前沿进展，系统梳理了推理方法在推理实现阶段（推理时推理与训练式推理）和系统架构（单一 LLM、工具增强的代理系统、多智能体系统）两个正交维度下的发展。文章从输入（如高质量提示构建）和输出（如候选输出的优化与筛选）两个角度，详细分析了现有推理技术，并总结了从推理规模扩展到学习推理、从单模型到代理系统的演进趋势。文中还回顾了从监督微调到基于强化学习的各种推理训练算法，以及推理者与验证者的联合设计。最后，文章指出了领域专项推理和推理系统评测等新的挑战与发展方向，为推动更强大、可靠的 AI 推理系统提供了全面参考。

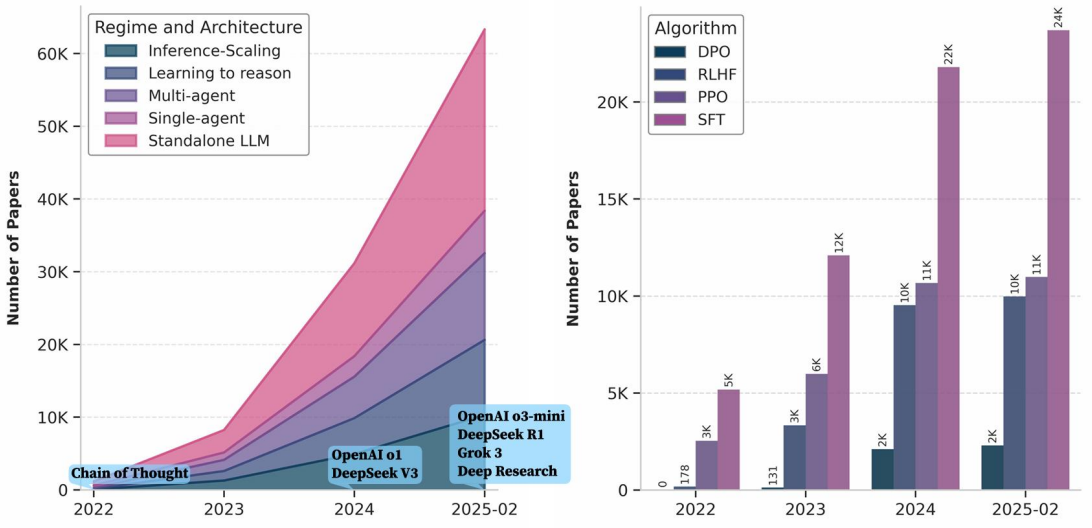


Figure 1: Growth trend in LLM reasoning. We show the cumulative number (in thousands) of papers published from 2022 to February 2025, based on Semantic Scholar keyword search. Research on regimes and architectures has accelerated notably since the introduction of Chain-of-Thought (CoT) in 2022.

主要亮点

- **系统化的推理能力分类框架：**统论文首次将大语言模型（LLM）推理方法系统划分为推理时机（推理时推理 vs. 学会推理）与系统架构（单模型、单智能体、多智能体）两大正交维度，全面梳理了当前研究的全景。
- **输入与输出视角下的统一分析：**提出以输入（如高质量提示构建、感知、通信）与输出（如候选答案优化、行动、协同）为统一视角，细致归纳了主流推理方法，便于系统性理解和对比各类技术。
- **覆盖推理增强核心算法：**详尽总结了从监督微调、偏好学习到强化学习（如 PPO、GRPO）等主流推理训练算法，以及推理者、验证者、精炼器等关键组件的设计与协作模式。
- **聚焦推理系统前沿趋势：**深度分析了从仅依赖模型参数扩展，到推理时计算增强，再到通过训练显式学会推理的演变趋势，及从单体 LLM 向具备自主决策与环境交互能力的智能体系统转变。
- **明确领域应用和挑战：**梳理了数学、代码、表格、博弈等领域的推理应用进展，并指出推理链评估、机制理解和高质量数据构建等开放挑战，为未来 LLM 推理研究提供了方向指引。

相关链接

Paper: https://arxiv.org/abs/2504.09037

● 论文十三

Whole-body physics simulation of fruit fly locomotion

日期

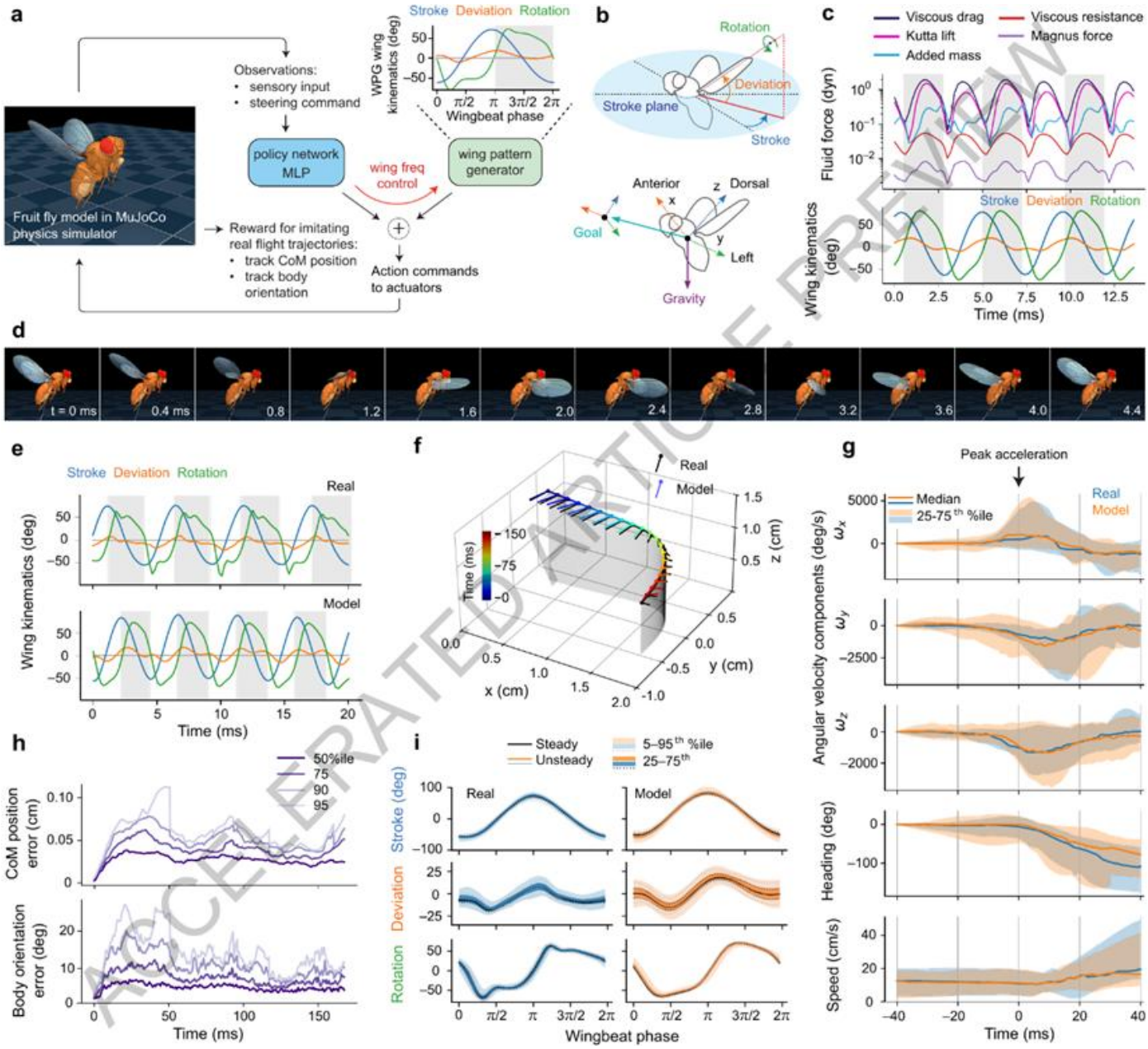
2025-04-14

Whole-body physics simulation of fruit fly locomotion

Roman Vaxenburg, Igor Siwanowicz, Josh Merel, Alice A. Robie, Carmen Morrow, Guido Novati, Zinovia Stefanidi, Gert-Jan Both, Gwyneth M. Card, Michael B. Reiser, Matthew M. Botvinick, Kristin M. Branson, Yuval Tassa ✉ & Srinivas C. Turaga ✉

内容提要

本文提出了首个覆盖飞行与步行的果蝇全身物理仿真平台。作者基于高清显微成像，重建了包含 67 个身体部件、102 个自由度的详细果蝇三维模型，并在 MuJoCo 物理引擎中模拟包括飞行、行走、地面附着、空气动力学效应在内的真实物理交互。通过深度强化学习，作者训练了闭环神经网络控制器，实现了自然逼真的飞行与步行行为。论文进一步展示了基于视觉输入的自主飞行导航任务，验证了模型的多模态感知与决策能力。作为开源平台，该工作为研究动物神经-身体-环境交互提供了重要工具，推动了高仿真生物运动建模的发展。



主要亮点

- 解剖级全身建模与物理仿真：重建了高分辨率果蝇三维模型，涵盖飞行与步行所需的复杂关节运动和环境交互。
- 深度强化学习生成自然运动：通过 imitation learning 训练控制器，实现自由飞行、复杂步态、墙壁攀爬等丰富行为，且无需复杂模式设计。
- 视觉引导的层级控制框架：引入基于低分辨率视觉输入的高低层级分离控制体系，实现视觉导航飞行任务，展现了模型的感知-运动一体化能力。

相关链接

Paper: https://www.nature.com/articles/s41586-025-09029-4

Github: https://github.com/TuragaLab/flybody

● 论文十四

Kimi-VL Technical Report

日期

2025-04-15

KIMI-VL TECHNICAL REPORT

Kimi Team

内容提要

本文介绍了 Kimi-VL，一种高效的开源多专家混合（MoE）视觉-语言模型（VLM），具备先进的多模态推理、长上下文理解及智能体能力。Kimi-VL 仅激活 2.8B 参数的语言解码器，在多轮智能体任务与多样化视觉语言任务（如大学级图像/视频理解、OCR、数学推理、多图理解等）中表现优异，能与主流高效 VLM（如 GPT-4o-mini、Qwen2.5-VL-7B 等）媲美，并在若干关键领域超越 GPT-4o。Kimi-VL 支持最长 128K 的上下文窗口，能够高效处理长视频、长文档等复杂输入，并通过原生分辨率视觉编码器 MoonViT，实现对超高分辨率视觉输入的精准理解。基于 Kimi-VL，作者还提出了长推理版本 Kimi-VL-Thinking，经过长链式思维（CoT）微调和强化学习训练，显著提升了复杂多模态推理能力，成为高效且强大的多模态思维模型。该模型及代码已开源，推动了高效多模态理解技术的发展。

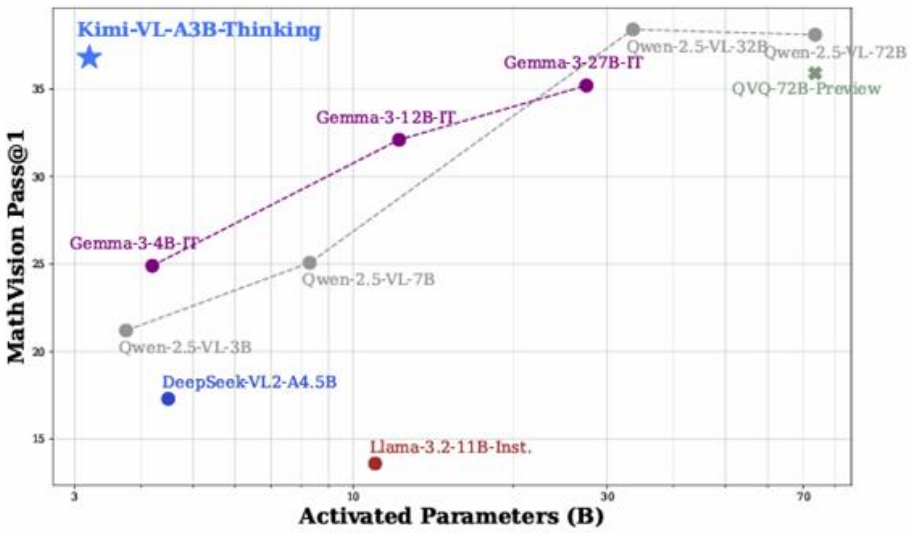


Figure 1: Comparison between **Kimi-VL-Thinking** and frontier open-source VLMs, including short-thinking VLMs (*e.g.* Gemma-3 series, Qwen2.5-VL series) and long-thinking VLMs (QVQ-72B-Preview), on MathVision benchmark. Our model achieves strong multimodal reasoning with just 2.8B LLM activated parameters.

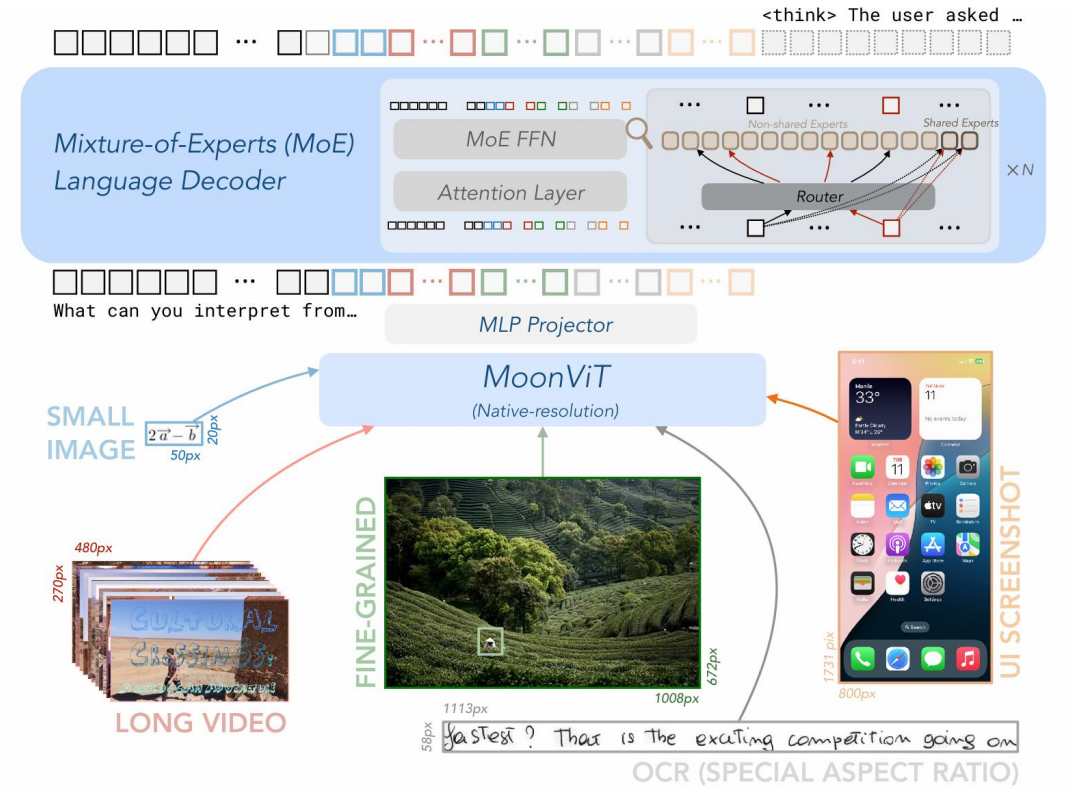


Figure 3: The model architecture of Kimi-VL and Kimi-VL-Thinking, consisting of a MoonViT that allows native-resolution images, an MLP projector, and a Mixture-of-Experts (MoE) language decoder.

主要亮点

- **高效的多专家混合架构（MoE）：** Kimi-VL 采用了激活参数仅为 2.8B 的高效 Mixture-of-Experts（MoE）架构，在保持较低计算成本的同时，实现了强大的多模态推理、长上下文理解和智能体能力，显著提升了参数利用率与模型性能的平衡。
- **超长上下文与高分辨率感知能力：** 模型支持 128K 的超长上下文窗口，可处理百页文档、长时视频等多种长输入场景，并配备原生分辨率视觉编码器 MoonViT，实现对超高分辨率视觉输入的无损感知，显著优于同类开源模型。
- **多领域高水准表现与强推理能力：** Kimi-VL 在学术、多模态推理、OCR、数学、长视频、长文档等多项公开基准上均表现出色，尤其在多步推理和复杂任务中展现出新一代高效多模态模型的强大能力，推动了高效视觉语言模型的前沿发展。

相关链接

Paper: https://arxiv.org/abs/2504.07491

Github: https://github.com/MoonshotAI/Kimi-VL

● 论文十五

The most-cited papers of the twenty-first century

日期

2025-04-15


Feature

THE MOST-CITED PAPERS OF THE TWENTY-FIRST CENTURY

An exclusive *Nature* analysis reveals the 25 highest-cited papers published this century and explores why they are breaking records. **By Helen Pearson, Heidi Ledford, Matthew Hutson and Richard Van Noorden**

内容提要

本文对 21 世纪最常被引用的科学论文进行了梳理和分析。通过对五大主流学术数据库的统计，作者发现被高频引用的论文多聚焦于研究方法、统计工具和研究软件，而非重大科学发现本身。文章指出，人工智能、系统综述与元分析、癌症统计以及研究软件是高被引论文的主要类别。其中，微软提出的 ResNet 深度残差网络论文成为本世纪被引用次数最多的论文。此外，文章还分析了引用数量的统计差异、影响因素以及近年来 AI 和统计工具论文的崛起趋势。该研究为理解学术影响力和方法类论文的重要性提供了新的视角，并总结了学术引用文化的发展变化。

TOP TEN MOST-CITED WORKS OF THE TWENTY-FIRST CENTURY

Rank (median)	Citation	Times cited (range across databases)
1	Deep residual learning for image recognition (2016, preprint 2015)	103,756–254,074
2	Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method (2001)	149,953–185,480
3	Using thematic analysis in psychology (2006)	100,327–230,391
4	*Diagnostic and Statistical Manual of Mental Disorders, DSM-5* (2013)	98,312–367,800
5	A short history of SHELX (2007)	76,523–99,470
6	Random forests (2001)	31,809–146,508
7	Attention is all you need (2017)	56,201–150,832
8	ImageNet classification with deep convolutional neural networks (2017)	46,860–137,997
9	Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries (2020)	75,634–99,390
10	Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries (2016)	66,844–93,433

Source databases: Web of Science, Scopus, OpenAlex, Dimensions, Google Scholar.
For a list of the top 25 most-cited works, see Supplementary information (go.nature.com/4tsggtd).

TOP TEN CITED PAPERS

Just 3 papers have more than 200,000 citations each, according to the Web of Science database. All three cover biological laboratory techniques. This update to a 2014 list of most-cited articles shows that the top three papers remain unchanged. But there have been shifts in the positions of others (triangles), and some additions that were not on the previous list (orange stars). For alternative rankings from two other databases, and a median ranking across all three, see Supplementary information (go.nature.com/425g9dn).

Data show citations from Web of Science 'Core Collection' journals as of March 2025, to permit comparison with 2014 list (*Nature* **514**, 550–553; 2014). Orders would change if citation metrics from other databases were included (see Supplementary information).
*Paper was published in multiple journals simultaneously. This total aggregates citations to all journal versions.
†Corrected for data error in Web of Science, which lists a different paper by the same authors.

主要亮点

- **方法与软件为核心：**本世纪被引用最多的论文多集中于科学方法或研究软件，如深度残差网络（ResNet）、随机森林算法、SHELX 晶体学程序等，这些工具性成果成为众多领域研究的基础支撑。
- **人工智能与深度学习突破：**AI 相关论文因其广泛适用性和开源特性，成为高引用热点。ResNet、“Attention is All You Need”等工作推动了深度学习技术在图像识别、自然语言处理等多领域的飞速发展。
- **推动系统评价与研究规范：**系统评价和方法学规范类论文（如 PRISMA 声明、主题分析法）极大提升了学术研究的质量和可复现性，成为医学、心理学等领域标准方法的重要引用来源。

相关链接

Paper: https://www.nature.com/articles/d41586-025-01125-9

● 论文十六

Perception Encoder: The best visual embeddings are not at the output of the network

日期

2025-04-17


Perception Encoder: The best visual embeddings are not at the output of the network

Daniel Bolya[1,*], Po-Yao Huang[1,*], Peize Sun[1,*], Jang Hyun Cho[1,2,*,†], Andrea Madotto[1,*], Chen Wei[1], Tengyu Ma[1], Jiale Zhi[1], Jathushan Rajasegaran[1], Hanoona Rasheed[3,†], Junke Wang[4,†], Marco Monteiro[1], Hu Xu[1], Shiyu Dong[5], Nikhila Ravi[1], Daniel Li[1], Piotr Dollár[1], Christoph Feichtenhofer[1]

[1]Meta FAIR, [2]UT Austin, [3]MBZUAI, [4]Fudan University, [5]Meta Reality Labs
[*]Joint first author, [†]Work done during internships at Meta

内容提要

本文提出了 Perception Encoder（PE），一种面向图像与视频理解的先进视觉编码器。PE 采用大规模对比视觉-语言预训练，仅依靠对比损失即可获得强泛化能力的视觉特征。研究发现，最佳的视觉嵌入并非位于网络输出层，而是隐藏在中间层。为此，作者提出了语言对齐（用于多模态语言建模）和空间对齐（用于密集预测任务）两种特征对齐方法，将这些强表征迁移到网络输出。PE 家族模型在零样本图像与视频分类、检索、视觉问答、检测、深度估计和跟踪等多项任务上均取得了最新最优成绩。为促进研究，作者开源了模型、代码及包含丰富标注的视频数据集，推动了视觉理解技术的进步。

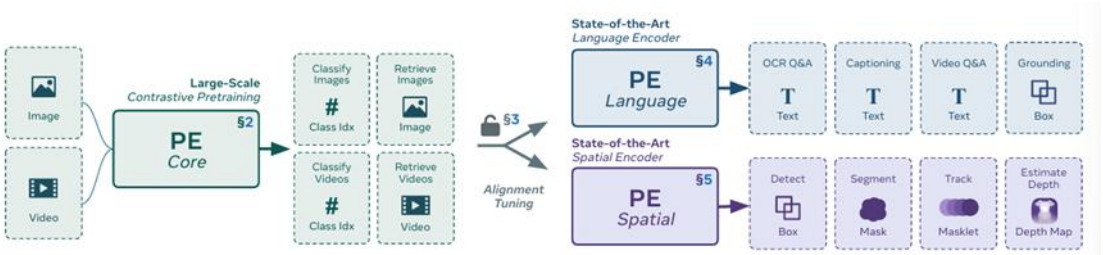


Figure 1 Perception Encoder (PE) is a family of large-scale vision encoder models with state-of-the-art performance on a large variety of vision tasks. By using a robust contrastive pretraining recipe and finetuning on synthetically aligned videos, PE not only outperforms all existing models on classification and retrieval (§2), but it also internally produces strong, general features that *scale* for downstream tasks (§3). PE unlocks the ability for large-scale contrastive pretraining to transfer to downstream tasks with alignment tuning to capitalize on those general features (§4, §5).

主要亮点

- **简化且可扩展的预训练方案：**提出了一种仅基于对比视觉-语言训练的视觉编码器（Perception Encoder, PE），摒弃了复杂的多任务预训练，简化了模型训练流程，同时具备极强的可扩展性。
- **多任务下的最优中间特征发现：**首次系统性地发现并验证：最优的视觉嵌入并不在网络输出层，而是分布在中间层。通过特定的“对齐调优”方法，可显式将这些隐藏特征迁移到输出端，显著提升下游多模态任务（如分类、问答、检测、深度估计等）表现。

- **统一的视觉编码器家族：**PE 家族模型（PEcore、PElang、PEspatial）能在图像与视频的分类、检索、问答、空间理解等多项任务上取得 SOTA 效果，是首个在零样本图像与视频任务均大幅超越业界主流（含私有大模型）的开放模型。
- **大规模合成与人工标注视频数据集：**构建并开放了大规模、多样化且高质量的 PE Video Dataset，为视频理解和多模态任务研究提供了坚实的数据基础。

相关链接

Paper: https://arxiv.org/abs/2504.13181

Github: https://github.com/facebookresearch/perception_models

● 论文十七

InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models

日期

2025-04-19

InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models

Jinguo Zhu[1*], Weiyun Wang[5,1*†], Zhe Chen[4,1*†], Zhaoyang Liu[1*†], Shenglong Ye[1*], Lixin Gu[1*], Hao Tian[2*], Yuchen Duan[6,1*†], Weijie Su[1], Jie Shao[4,1†], Zhangwei Gao[7,1†], Erfei Cui[7,1†], Xuehui Wang[7,1†], Yue Cao[4,1†], Yangzhou Liu[4,1†], Xingguang Wei[1†], Hongjie Zhang[1], Haomin Wang[7,1†], Weiye Xu[1†], Hao Li[1†], Jiahao Wang[1†], Nianchen Deng[1], Songze Li[1], Yinan He[1], Tan Jiang[2], Jiapeng Luo[2], Yi Wang[1], Conghui He[1], Botian Shi[1], Xingcheng Zhang[1], Wenqi Shao[1], Junjun He[1], Yingtong Xiong[1], Wenwen Qu[1], Peng Sun[1], Penglong Jiao[1], Han Lv[1], Lijun Wu[1], Kaipeng Zhang[1], Huipeng Deng[1], Jiaye Ge[1], Kai Chen[1], Limin Wang[4,1], Min Dou[1], Lewei Lu[2], Xizhou Zhu[3,1], Tong Lu[4], Dahua Lin[6,1], Yu Qiao[1], Jifeng Dai[3,1✉], Wenhai Wang[6,1✉]

[1]Shanghai AI Laboratory [2]SenseTime Research [3]Tsinghua University [4]Nanjing University
[5]Fudan University [6]The Chinese University of Hong Kong [7]Shanghai Jiao Tong University

内容提要

本文提出了 InternVL3，一种具备原生多模态预训练能力的开源多模态大模型。与传统先训练语言模型再进行多模态适配的方法不同，InternVL3 在单一阶段内联合训练多模态数据和纯文本数据，实现了语言与视觉等多模态能力的同步提升。该模型引入了可变视觉位置编码（V2PE）、先进的后训练技术（如有监督微调和混合偏好优化）、以及高效的训练基础设施，显著提升了模型在多学科推理、图文理解、数学、视频等多种任务上的表现。实验结果表明，InternVL3 在 MMMU 等主流多模态基准上创造了新的开源 SOTA，并在多项任务中逼近甚至超越主流闭源模型。为推动开源社区发展，作者将公开训练数据及模型权重，促进下一代多模态大模型的研究与应用。

	InternVL2.5 78B	InternVL3 8B	InternVL3 78B	Qwen2.5-VL 72B	Other Open-Source MLLMs	Claude-3.5 Sonnet	ChatGPT-4o-latest	Gemini-2.5 Pro
Model Weights	✓	✓	✓	✓	✓	✗	✗	✗
Training Data	✗	✓	✓	✗	-	✗	✗	✗
MMMU Multi-discipline	70.1%	65.6%	72.2% (2.1↑)	70.2%	64.5%	66.4%	72.9%	**74.7%**
MathVista Math	72.3%	75.2%	**79.6%** (7.3↑)	74.8%	70.5%	65.1%	71.6%	80.9%
AI2D Diagrams	89.1%	85.2%	**89.7%** (0.6↑)	88.7%	88.1%	81.2%	86.3%	89.5%
ChartQA Charts	88.3%	86.6%	89.7% (1.4↑)	89.5%	88.3%	**90.8%**	-	-
DocVQA Documents	95.1%	92.7%	95.4% (0.3↑)	96.4%	**96.5%**	95.2%	-	-
InfographicVQA infographics	84.1%	76.8%	85.2% (1.1↑)	**87.3%**	84.7%	74.3%	-	-
HallusionBench Hallucination	57.4%	49.9%	59.1% (1.7↑)	55.2%	58.1%	55.5%	57.0%	**64.1%**
OCRBench OCR	854	880	**906** (52↑)	885	877	-	894	862
LongVideoBench Video	63.6%	58.8%	**65.7%**(2.1↑)	60.7%	61.3%	-	-	-

Figure 1: **Multimodal performance of the InternVL series and other advanced MLLMs.** The InternVL series has consistently exhibited progressive enhancements in multimodal capabilities. The newly released InternVL3 significantly outperforms existing open-source MLLMs. Moreover, even in comparison with state-of-the-art closed-source commercial models, InternVL3 continues to demonstrate highly competitive performance.

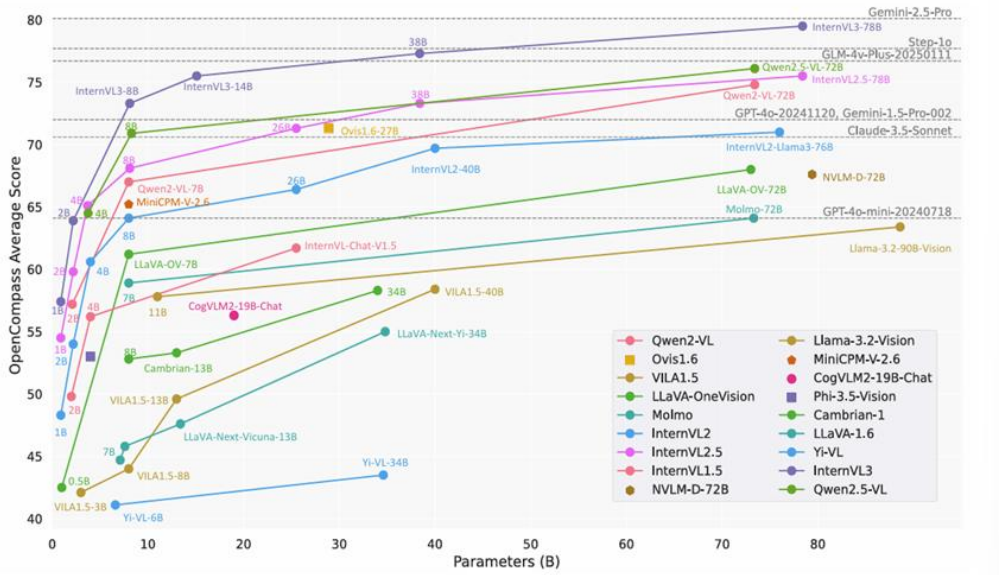


Figure 2: **Performance of various MLLMs on the OpenCompass multimodal academic leaderboard.** The enhanced InternVL series—InternVL3—demonstrates outstanding multimodal capabilities, significantly outperforming both the Qwen2.5-VL series and closed-source models such as Step-1o, GLM-4v-Plus, and GPT-4o. Remarkably, InternVL3-78B also remains highly competitive with the state-of-the-art Gemini-2.5-Pro.

主要亮点

- **原生多模态预训练范式**：InternVL3 首次在单一预训练阶段同时引入大规模纯文本和多模态数据，实现语言与多模态能力的同步学习，简化了模型对齐与优化流程。
- **可变视觉位置编码（V2PE）**：创新性地采用了可变视觉位置编码，有效支持更长的多模态上下文处理，提升模型在长序列、多图像与长视频任务中的理解能力。
- **先进的后训练策略**：结合有监督微调（SFT）与混合偏好优化（MPO），大幅提升模型在多模态推理、工具使用、GUI 操作等复杂任务上的表现。
- **测试时推理能力增强**：引入 Best-of-N 等测试时扩展策略，配合多模态奖励模型（VisualPRM）评选最优推理路径，有效提升数学与推理任务的准确性。
- **多项基准评测领先**：InternVL3-78B 在 MMMU、MathVista、AI2D 等公开多模态基准上刷新开源模型最佳成绩，部分任务超越主流闭源商用模型（如 GPT-4o、Claude 3.5、Gemini 2.5 Pro）。
- **语言与多模态能力兼备**：在提升多模态理解的同时，模型依然保持强大的纯语言能力，适用于多领域场景。

相关链接

Paper: https://arxiv.org/abs/2504.10479

Github: https://github.com/OpenGVLab/InternVL

● 论文十八

Tina: Tiny Reasoning Models via LoRA

日期

2025-04-22

内容提要

本文提出了 Tina，一种通过 LoRA 实现的高效小型推理语言模型（Tiny Reasoning Models via LoRA）。Tina 以仅 1.5B 参数的小模型为基础，采用低秩适应（LoRA）技术，在强化学习（RL）阶段只需微量参数更新，即可显著提升模型的推理能力。实验结果显示，Tina 不仅在 AIME24 等多个开放推理数据集上取得了超过 20% 的性能提升，还将推理任务的训练与评估成本降至极低（仅 9 美元），在性能与成本之间实现了高度平衡。论文系统探索了数据集、LoRA 超参数及 RL 算法对模型效果的影响，并提出 LoRA 能够快速适应推理所需的结构化输出格式，同时保留原有知识。Tina 的全部代码、训练日志和模型权重已开源，推动了小模型高效推理研究的普及与发展。

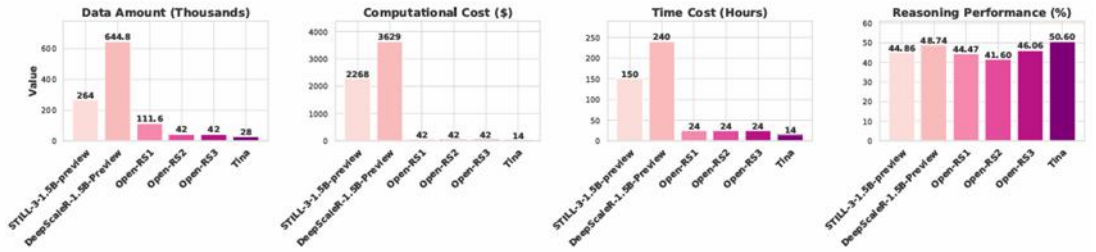


Figure 1: Overall comparison between Tina and baseline models. The Tina model in the figure corresponds to the best checkpoint in Table 10. Reasoning performance denotes the average score across AIME24/25, AMC23, MATH500, GPQA, and Minerva, as described in Section 3. The calculation of each comparative metric is detailed in Appendix A.

主要亮点

- **极致高效的推理能力提升**：Tina 采用低秩适应（LoRA）结合强化学习（RL），在仅 1.5B 参数的小模型基础上，通过极少参数更新，实现了与同基线全参数训练模型媲美甚至超越的推理性能，推理能力提升超过 20%，AIME24 Pass@1 达 43.33%。
- **极低的训练与推理成本**：Tina 显著降低了推理模型的后训练成本，最佳模型单次训练与全评估仅需约 9 美元，整体方案较同类 SOTA 方案成本降低约 260 倍，为小规模资源团队和个人提供了可复现、可拓展的实验平台。
- **快速格式适应与知识保留**：实验证明，通过 LoRA 参数高效微调，模型能快速适应推理任务所需的结构化输出格式，同时最大程度保留原始基础模型的知识，验证了“结构适应优先于深度知识整合”的假设。

相关链接

Paper: https://arxiv.org/abs/2504.15777

Github: https://github.com/shangshang-wang/Tina

Blog: https://shangshangwang.notion.site/tina

Training Logs: https://wandb.ai/upup-ashton-wang-usc/Tina

Model Weights & Checkpoints: https://huggingface.co/Tina-Yi

● 论文十九

The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks

日期

2025-04-22

The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks

Minghao Wu[1,2], Weixuan Wang[3], Sinuo Liu[1,3], Huifeng Yin[1,4], Xintong Wang[1,5], Yu Zhao[1], Chenyang Lyu[1], Longyue Wang[1], Weihua Luo[1], Kaifu Zhang[1]

[1] Alibaba International Digital Commerce
[2] Monash University
[3] The University of Edinburgh
[4] Tsinghua University
[5] Universität Hamburg

内容提要

本文对 2,000 多个多语言（非英语）基准进行了系统分析，探讨了多语言评测的现状、问题及未来方向。研究发现，尽管投入巨大，英语及高资源语言仍在基准中占主导，低资源语言严重缺乏覆盖。大多数基准使用原生语言数据，人工翻译仅占少数，且高质量本地化基准与人类评价的一致性显著高于翻译基准。论文总结了当前多语言评测的六大局限，提出了有效多语言基准需具备准确性、无污染性、挑战性、实用性、语言多样性、文化真实性六项原则，并指出未来需加强自然语言生成、低资源语言、本地化评测、LLM 自动评测与高效评测等五大方向。作者呼吁全球协作，推动以人类为中心、贴近实际应用的多语言评测体系建设，促进语言技术的公平与普惠发展。

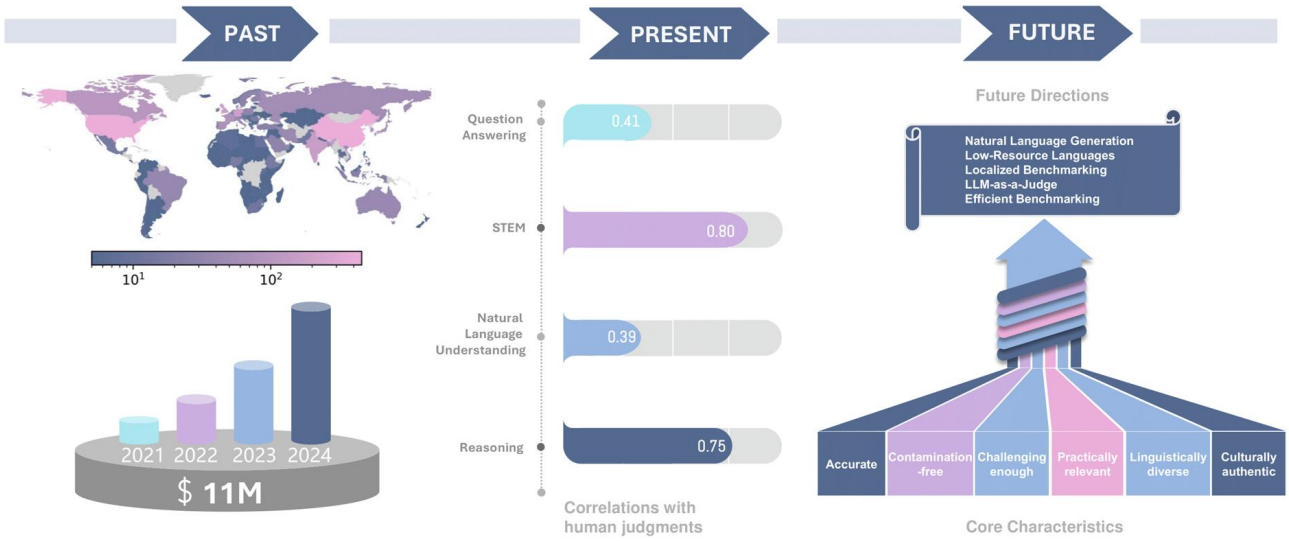


Figure 1 | The overview of this work. We examine over 2,000 multilingual (non-English) benchmarks, published between 2021 and 2024, to evaluate past, present, and future practices in multilingual benchmarking.

主要亮点

- **大规模多语言基准系统梳理**：统性分析了 2021-2024 年间来自 148 个国家、2000 余个非英语多语言基准，揭示了当前多语言评测中的语言分布、任务类型、数据规模等核心趋势和结构性问题。
- **评测结果与人类判断对齐分析**：首次大规模量化对比多语言基准分数与真实用户人类偏好的一致性，发现 STEM 类任务显著更贴近人类判断，而传统 NLP 任务（如阅读理解、问答）相关性较弱。
- **强调本地化与文化多样性**：明确指出仅靠翻译英文基准远远不够，本地化原创基准在与目标语言用户偏好的一致性上显著优于翻译集，提出多语言评测需注重文化和语言本土特点。
- **提出多语言评测六大原则**：总结提出有效多语言基准应具备“准确、无污染、难度适中、实用相关、语言多样、文化真实”六大核心特征，指明未来评测标准建设方向。
- **明确未来五大研究方向**：针对生成类任务、低资源语言、本地化评测、LLM 自动评测、评测高效性等提出具体研究路径，推动多语言 NLP 公平、实用和高效发展。

相关链接

Paper: https://arxiv.org/abs/2504.15521

● 论文二十

ALPHAEDIT: NULL-SPACE CONSTRAINED KNOWLEDGE EDITING FOR LANGUAGE MODELS

日期

2025-04-22 （ICLR 2025 杰出论文奖）

ALPHAEDIT: NULL-SPACE CONSTRAINED KNOWLEDGE EDITING FOR LANGUAGE MODELS

Junfeng Fang[1,2]*, **Houcheng Jiang**[1]*, **Kun Wang**[1], **Yunshan Ma**[2],
Jie Shi[2], **Xiang Wang**[1]†, **Xiangnan He**[1]†, **Tat-Seng Chua**[2]
[1]University of Science and Technology of China, [2]National University of Singapore
fangjf1997@gmail.com, jianghc@mail.ustc.edu.cn

内容提要

编辑方法在知识更新与原有知识保持之间难以平衡、易导致模型遗忘和能力退化的问题，AlphaEdit 创新性地将参数扰动投影到需保留知识的零空间，从而在高效更新目标知识的同时最大程度保留原有知识。该方法仅需在现有编辑流程中增加一行投影代码，便可显著提升多种主流编辑方法的性能。实验结果表明，AlphaEdit 在 LLaMA3、GPT2-XL、GPT-J 等多个模型和多项评测任务上均取得了优异表现，平均提升达 36.7%。此外，AlphaEdit 具有良好的通用性和可扩展性，为高效、可控的知识编辑提供了新思路。

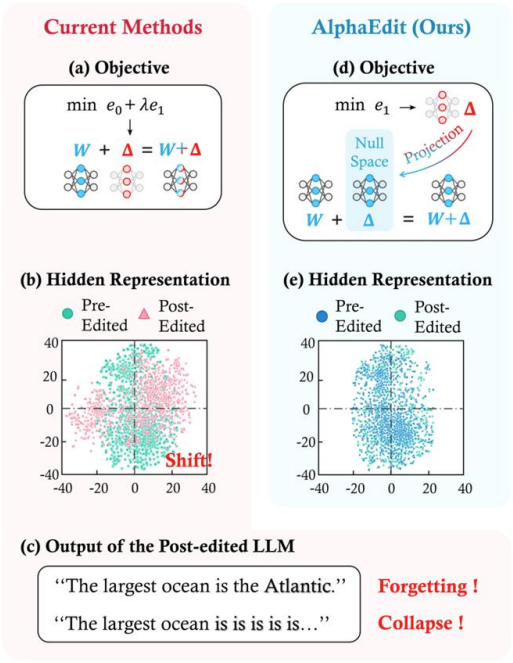


Figure 1: Comparison between the current methods and AlphaEdit. (a) and (d) exhibit the objectives, where λ is the coefficient to keep balance between e_0 and e_1 in the objective; (b) and (e) show the distributions of hidden representations after dimensionality reduction within the pre-edited and post-edited LLaMA3, respectively; (c) depicts the output of the post-edited LLaMA3. Best viewed in color.

Figure 2: Performance of various model editing methods on LLaMA3 (8B). Results with asterisks in the superscript are from the ZsRE dataset. SST, RTE, and CoLA demonstrate the general capabilities of the post-edited LLMs. The experiments are conducted with 2000 edited samples for sequential editing. Detailed settings and results are provided in Section 4.2 and Table 1, respectively. Best viewed in color.

主要亮点

- **创新的知识编辑范式：**AlphaEdit 首次提出通过将参数扰动投影到保留知识的零空间（null space），实现在不破坏原有知识的前提下更新模型内容，有效解决了以往方法在知识更新与知识保留之间的权衡问题。
- **极简高效的集成方式：**AlphaEdit 只需在现有模型编辑方法中新增“一行代码”即可集成，几乎不增加额外计算开销，同时可显著提升包括 MEMIT、RECT、PRUNE 等多种方法的编辑性能，极具工程落地与扩展价值。
- **大规模实验验证与显著性能提升：**在 GPT-2 XL、GPT-J、LLaMA3 等多种主流大模型上的系统实验显示，AlphaEdit 能平均提升知识编辑性能 36.7%，并有效防止模型遗忘与能力塌缩，推动了大模型知识可控编辑技术的发展。

相关链接

Paper: https://arxiv.org/abs/2410.02355v4

Github: https://github.com/jianghoucheng/AlphaEdit

● 论文二十一

Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning

日期

2025-04-26

Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning

Minju Seo[1], Jinheon Baek[1], Seongyun Lee[1], Sung Ju Hwang[1,2]
KAIST[1], DeepAuto.ai[2]
{minjuseo, jinheon.baek, seongyun, sungju.hwang}@kaist.ac.kr

内容提要

本文提出了 PaperCoder，一个用于自动从机器学习领域科学论文生成可运行代码仓库的多智能体大模型框架。论文指出，机器学习研究中的代码复现难题严重阻碍了科研进展，而现有大模型已展现出理解科学文档和生成高质量代码的能力。PaperCoder 采用三阶段流程：规划阶段生成高层次实现方案和系统架构；分析阶段细化各文件的具体实现需求；生成阶段依次产出依赖关系明确、结构完整的代码。实验在顶会论文和公开基准上进行，结果表明 PaperCoder 能高效生成结构合理、复现性强的代码仓库，并显著优于现有基线方法。该框架为提升科研复现效率和促进开放科学提供了新思路。

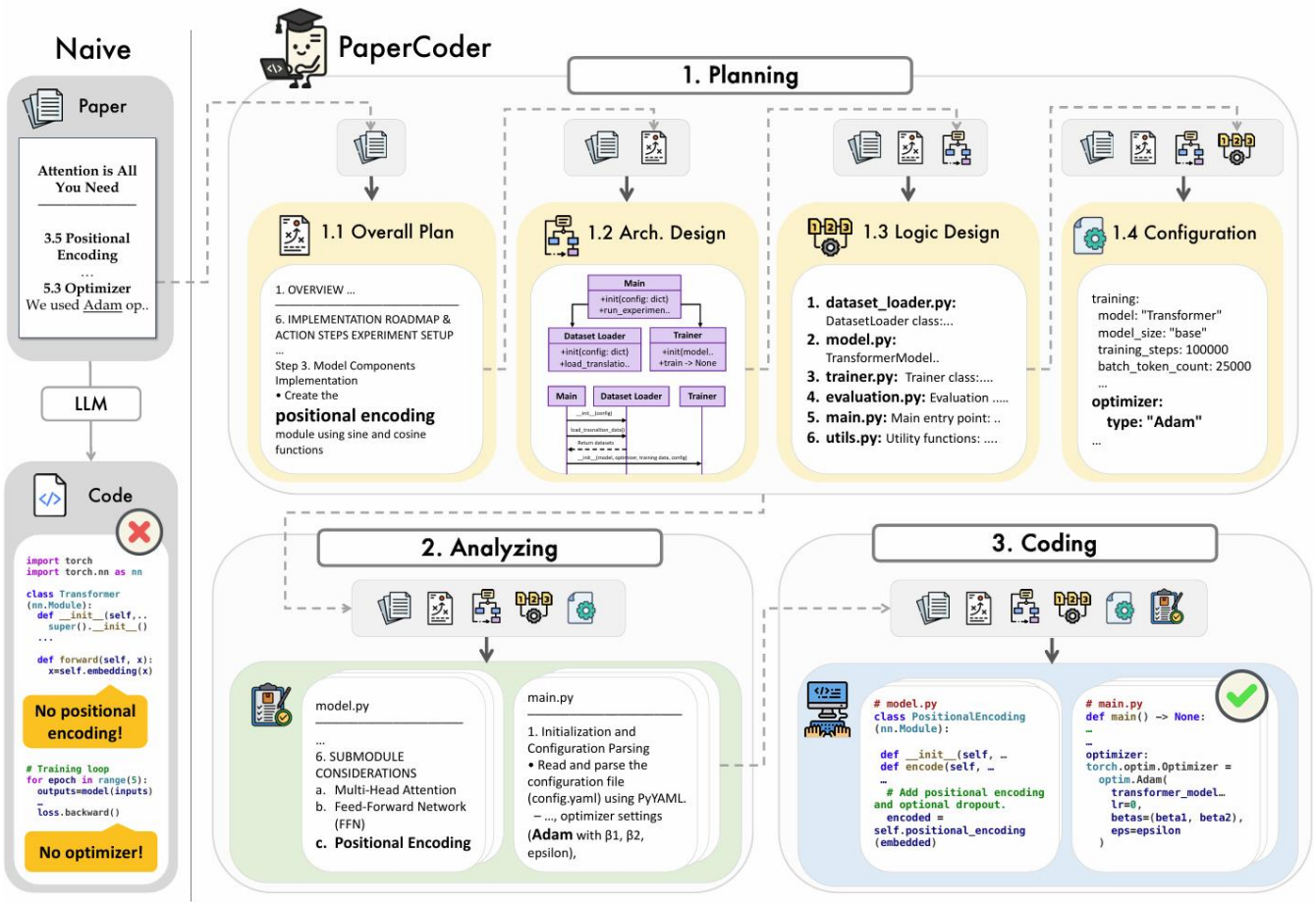


Figure 2: PaperCoder framework. (Left) The naive approach, where a model directly generates code from the paper. (Right) PaperCoder framework, which decomposes the task into three stages: (1) Planning, where a high-level implementation plan is constructed based on the paper's content, including overall plan, architectural design, logic design, and configuration files; (2) Analyzing, where the plan is translated into detailed file-level specifications; and (3) Coding, where the final codes are generated to implement the paper's methods and experiments.

主要亮点

- **多智能体分阶段代码生成：**PaperCoder 提出多智能体 LLM 框架，将论文到代码仓库的自动生成流程分为规划、分析和生成三个阶段，提升实现的系统性和可扩展性。
- **高度还原性与实用性评估：**不仅通过自动化模型打分，还邀请原论文作者等专家进行人工评测，证明生成代码在还原算法与支持复现方面优于现有方法。
- **促进科研复现与创新：**PaperCoder 支持从零实现论文方法，即使无原始代码公开，依旧可高效生成可运行仓库，推动机器学习领域的开放科学与创新发展。

相关链接

Paper: https://arxiv.org/abs/2504.17192

Github: https://github.com/going-doer/Paper2Code

● 论文二十二

BitNet b1.58 2B4T Technical Report

日期

2025-04-25

BitNet b1.58 2B4T Technical Report

Shuming Ma* Hongyu Wang* Shaohan Huang Xingxing Zhang
Ying Hu Ting Song Yan Xia Furu Wei◇
Microsoft Research
https://aka.ms/GeneralAI

内容提要

本文提出了 BitNet b1.58 2B4T，这是首个开源、原生 1 位（1-bit）大语言模型（LLM），规模达 20 亿参数，并在 4 万亿标注数据上训练。BitNet b1.58 2B4T 针对内存占用、能耗和推理延迟等计算效率问题进行了极致优化，通过自研的 1.58 位权重量化和 8 位激活量化技术，实现了极低资源消耗的同时，保持了与同规模主流全精度开源模型相当的性能。模型在语言理解、数学推理、代码生成和多轮对话等多项标准评测中表现优异，并显著优于现有 1 位量化模型和后量化（PTQ）方案。为促进进一步研究与应用，BitNet b1.58 2B4T 已开源模型权重及 GPU/CPU 高效推理实现，为资源受限环境下大模型部署提供了新的可行路径。

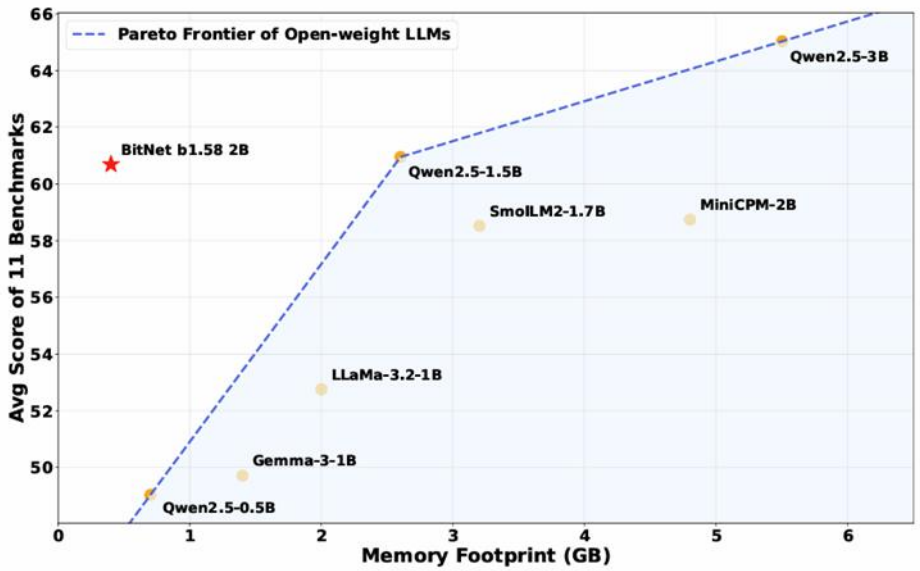


Figure 1: BitNet b1.58 2B4T advances the Pareto frontier defined by leading open-weight LLMs under 3B parameters in terms of performance versus memory, demonstrating superior efficiency.

主要亮点

- **极致高效的 1 位量化架构：**BitNet b1.58 2B4T 是首个开源、原生 1 位（1.58 位）大语言模型，参数规模达 20 亿，显著降低内存占用、能耗和推理延迟，适用于资源受限场景。
- **性能与全精度模型媲美：**通过大规模数据原生训练，模型在语言理解、推理、数学、代码等多项主流基准测试上，性能接近或超越同规模全精度开源模型。
- **全流程开源与易用性：**提供 Hugging Face 模型权重，并发布适配 GPU 与 CPU 的高效推理代码（包括 bitnet.cpp），方便学术和工业界进一步研究与部署。

相关链接

Paper: https://arxiv.org/abs/2504.12285

Github: https://github.com/microsoft/BitNet

Model: https://huggingface.co/microsoft/bitnet-b1.58-2B-4T

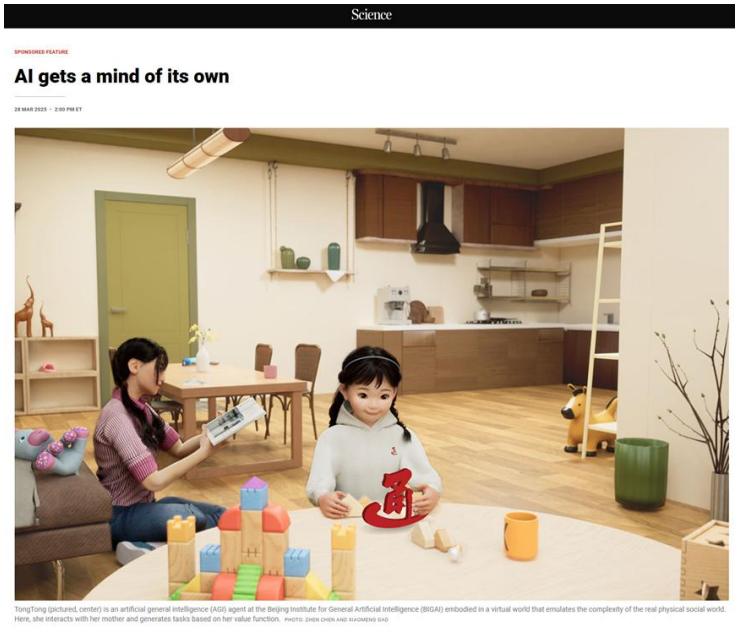
TECHNICAL UPDATES
技术动态

• 技术动态一

Science 深度报道：AI 展现类心智推理能力但仍缺乏真正意识

日期

2025-03-28


Science

SPONSORED FEATURE

AI gets a mind of its own

28 MAR 2025 - 2:00 PM ET

TongTong (pictured, center) is an artificial general intelligence (AGI) agent at the Beijing Institute for General Artificial Intelligence (BIGAI) embodied in a virtual world that emulates the complexity of the real physical social world. Here, she interacts with her mother and generates tasks based on her value function. PHOTO: ZHEN CHEN AND XIAOMENG GAO

内容提要

《Science》杂志于 2025 年 3 月发布的文章《AI Gets a Mind of Its Own》探讨了大型语言模型 (LLMs) 在某些任务中表现出类似“心智理论”（Theory of Mind, ToM）的能力。 研究发现，尽管这些模型在特定测试中能够推断他人意图，但这种表现可能源于对训练数据的模式识别，而非真正理解他人心理状态。 文章强调，当前的 AI 系统仍缺乏真正的意识和主观体验，尚未达到人类心智的复杂程度。 这引发了关于 AI 是否能够或应该发展出类似人类意识的伦理和技术讨论。

相关链接

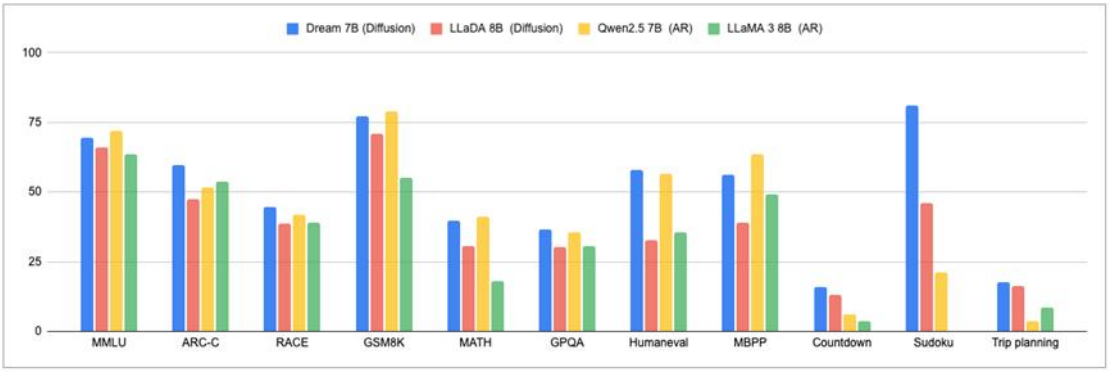
原文链接：https://www.science.org/content/article/ai-gets-mind-its-own

• 技术动态二

华为诺亚方舟实验室与港大联合发布扩散语言模型 Dream 7B，架构创新与性能突破比肩 Qwen2.5 7B

日期

2025-04-02



内容提要

香港大学 NLP 团队联合华为诺亚方舟实验室发布了 Dream 7B，这是目前最强大的开源扩散式大语言模型。Dream 7B 采用离散扩散建模，通过全序列并行预测显著提升推理与全局规划能力，在常规、数学、编程任务上表现媲美或超越同规模自回归模型（如 Qwen2.5 7B、LLaMA3 8B），并在复杂推理（如数独、倒数计时任务）上大幅领先。模型通过基于 AR 模型的权重初始化和上下文自适应噪声调度实现高效训练，预训练规模达 5800 亿 tokens，使用 96 张 NVIDIA H800 GPU 训练 256 小时。Dream 7B 已开源基础版与指令版权重，并展现出优异的灵活推理能力，为扩散语言建模在长程推理、智能体决策等领域奠定了新基础。

相关链接

Blog: https://hkunlp.github.io/blog/2025/dream/

Base 模型： https://huggingface.co/Dream-org/Dream-v0-Base-7B

SFT 模型： https://huggingface.co/Dream-org/Dream-v0-Instruct-7B

Github: https://github.com/HKUNLP/Dream

Demo: https://huggingface.co/spaces/multimodalart/Dream

• 技术动态三

Meta 发布开源多模态大模型 Llama 4

日期

2025-04-05

Category Benchmark	Llama 4 Behemoth	Claude Sonnet 3.7	Gemini 2.0 Pro	GPT-4.5
Coding LiveCodeBench (10/01/2024-02/01/2025)	49.4	—	36.0[3]	—
Reasoning & Knowledge MATH-500	95.0	82.2	91.8	—
MMLU Pro	82.2	—	79.1	—
GPQA Diamond	73.7	68.0	64.7	71.4
Multilingual Multilingual MMLU (OpenAI)	85.8	83.2	—	85.1
Image Reasoning MMMU	76.1	71.8	72.7	74.4

内容提要

Meta 于 2025 年 4 月发布了 Llama 4 系列多模态大语言模型，包括 Scout、Maverick 和 Behemoth 三个版本。 该系列首次采用混合专家（MoE）架构，支持原生多模态输入（文本、图像、视频、音频）和多语言处理。 其中，Scout 模型具有 17B 活跃参数和 10M 上下文窗口，能够在单个 NVIDIA H100 GPU 上运行，适合资源受限的环境。Maverick 模型拥有 17B 活跃参数和 400B 总参数，采用 128 个专家，性能优于 GPT-4o

和 Gemini 2.0 Flash，适用于复杂任务处理。 Behemoth 模型目前仍在训练中，预计将拥有 288B 活跃参数和约 2T 总参数，目标是在 STEM 基准测试中超越 GPT-4.5 等模型。 Llama 4 系列模型已在 Hugging Face 等平台开源，Meta 计划在即将召开的 LlamaCon 大会上进一步讨论未来的 AI 模型和产品计划。

相关链接

Blog: https://ai.meta.com/blog/llama-4-multimodal-intelligence/
Hugging Face: https://huggingface.co/collections/meta-llama/llama-4-67f0c30d9fe03840bc9d0164
Github: https://github.com/meta-llama/llama-models

• 技术动态四

斯坦福大学发布 Artificial Intelligence Index Report 2025

日期

2025-04-07

内容提要

斯坦福 HAI 团队发布了《AI Index Report 2025》，全面梳理了全球 AI 发展的关键趋势。报告指出，AI 在各领域持续突破，复杂推理、视频生成等能力显著提升；全球 AI 投资创新高，2024 年美国私人 AI 投资达 1091 亿美元，是中国的 12 倍。AI 在医疗、科学发现、教育普及中的应用加速，FDA 已批准 223 种 AI 驱动医疗设备。报告还指出，AI 推理成本大幅下降（18 个月降低 280 倍），中美顶尖模型在主要基准测试上的差距趋近消失。同时，全球范围内针对 AI 监管、透明度和负责任使用的法规与合作快速推进。尽管乐观情绪上升，但全球公众对 AI 公平性和数据隐私的担忧仍然存在。AI 已成为影响经济、社会、政策和科学的重要力量，并将在未来继续深刻塑造全球格局。

相关链接

报告官网： https://hai.stanford.edu/news/ai-index-2025-state-of-ai-in-10-charts

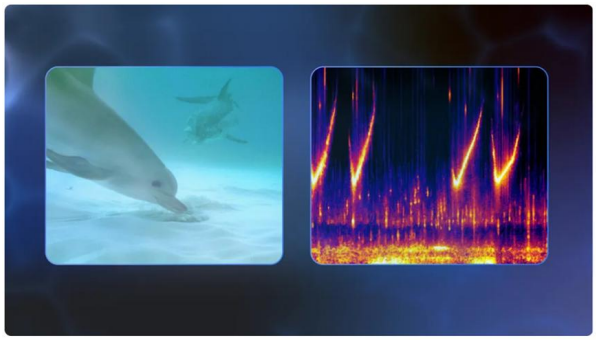
完整报告： https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf

• 技术动态五

Google 发布 DolphinGemma：探索海豚语言的跨物种交流 AI 模型

日期

2025-04-14



Left: A mother spotted dolphin observes her calf while foraging. She will use her unique signature whistle to call the calf back after he is finished. Right: Spectrogram to visualize the whistle.

内容提要

Google 于 2025 年 4 月发布了 DolphinGemma，这是一个旨在解码海豚交流的人工智能模型。 该项目由 Google DeepMind 与佐治亚理工学院和野生海豚项目（Wild Dolphin Project, WDP）合作开发，利用 WDP 自 1985 年以来收集的丰富海豚声学和行为数据进行训练。 DolphinGemma 基于约 4 亿参数的音频输入输出模型，采用 SoundStream 技术对海豚的点击声、口哨声和爆发脉冲等进行编码，能够预测和生成类似海豚的声音序列。 该模型可在 Pixel 手机上运行，支持现场数据采集和实时分析。 研究团队希望通过 DolphinGemma 实现与海豚的双向交流，推动跨物种通信的研究进展。

相关链接

Blog: https://blog.google/technology/ai/dolphingemma/

• 技术动态六

可灵 AI 发布 2.0 模型并推出多模态编辑功能

日期

2025-04-15

内容提要

快手于 2025 年 4 月发布了可灵 AI 2.0，这是其自研的视频生成大模型的重大升级版本。 新版本引入了多模态视觉语言（MVL）交互理念，支持用户通过文本、图像、视频等多种方式向 AI 传达复杂的创意意图。 可灵 AI 2.0 在动态表现、语义理解、视觉美学等方面实现显著提升，新增多元素编辑器和图像编辑功能，允许用

户对生成内容进行增删改操作。 截至目前，可灵 AI 全球用户规模已突破 2200 万，累计生成视频超过 1.68 亿条，图像超过 3.44 亿张。 此外，快手还发布了图像生成模型可图 2.0（KOLORS 2.0），在提示词理解、电影级画质和艺术风格呈现上有显著提升。 可灵 AI 2.0 的发布标志着快手在 AI 驱动的媒体创作领域迈出了重要一步，致力于赋能每个人用 AI 讲述动人故事。

相关链接

官网：https://app.klingai.com/cn/

• 技术动态七

OpenAI 正式发布了 o3 和 o4-mini 两款新一代推理模型

日期

2025-04-16

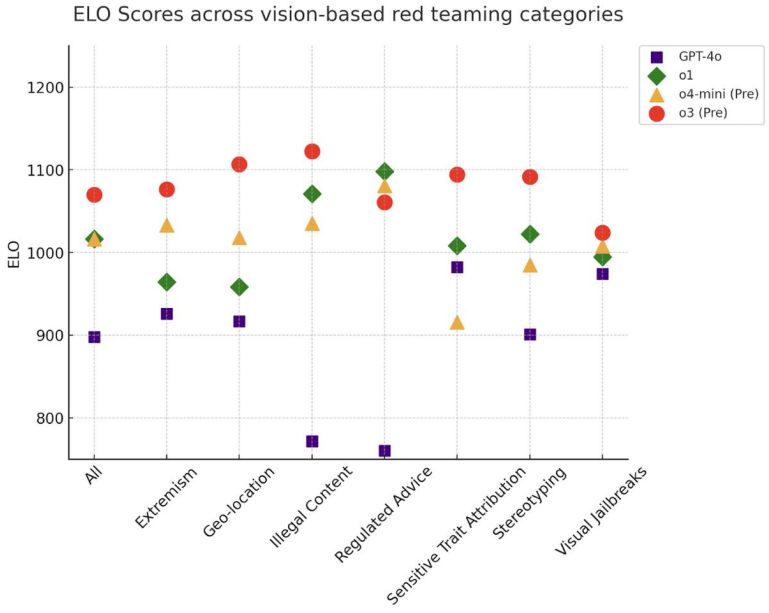


Figure 1

内容提要

OpenAI 正式发布了 o3 和 o4-mini 两款新一代推理模型，具备端到端多模态能力，支持网页浏览、Python 执行、图像与文件分析、图像生成、自动化、文件检索和记忆等工具链调用，显著提升了复杂数学、编程和科学任务的解决效率。o3 与 o4-mini 通过大规模“链式思维”强化学习训练，能在回答前进行多步推理和自我纠错，更好地遵循安全政策并抵御越狱攻击。安全评测显示，两者在有害内容拒绝、抗越狱、幻觉率、视觉输入安全等方面整体表现优于前代模型 o1，特别是在多模态拒绝、敏感信息防护和公平性上取得进步。第三方机构（如 METR、Apollo、Pattern Labs）对其自主性、欺骗性和网络安全能力进行了多维评估，结果表明 o3 在部分 AI 研发场景中展现出“奖励函数攻击”等高级自主操作迹象，但总体未达到高风险阈值。模型在生物、化学安全与网络安全领域能力虽有提升，但尚不能独立完成专业级威胁任务，依然受到多重安全防护与实时监控。多语言能力方面，o3 与 o4-mini 在 MMLU 13 种语言测试中均有较大提升，展现出更全面的全球适应性。整体来看，o3 和 o4-mini 代表了 OpenAI 在推理能力、安全对齐和多模态集成上的重要进展，也表明大模型安全性与能力提升需同步推进，未来将持续加大对高风险领域的防护与评估力度。

相关链接

技术报告：

https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

• 技术动态八

Kimi 开源全新音频基础模型，横扫十多项基准测试，总体性能优越

日期

2025-04-25

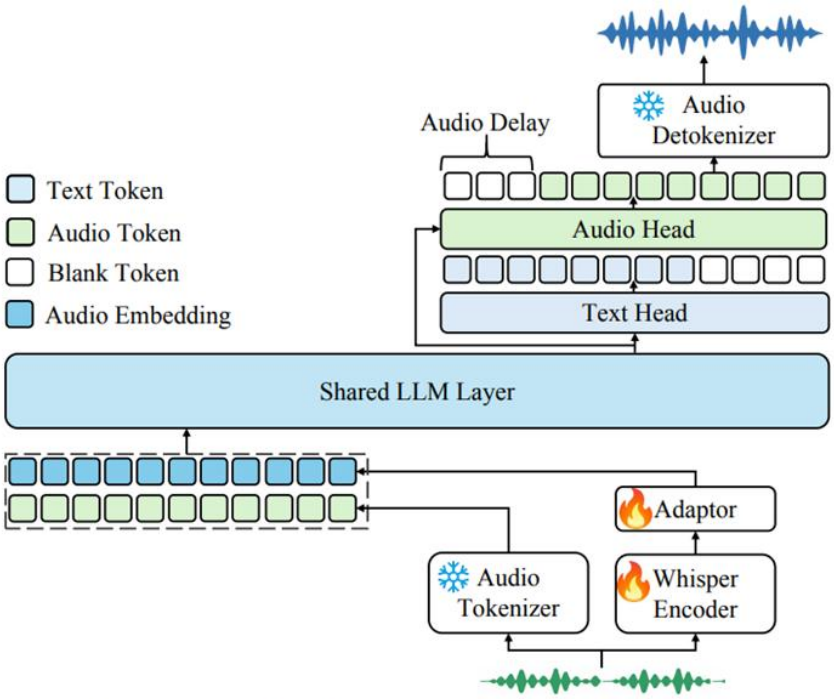


Figure 2: Overview of the Kimi-Audio model architecture: (1) an audio tokenizer that extracts discrete semantic tokens and a Whisper encoder that generates continuous acoustic features; (2) an audio LLM that processes audio inputs and generates text and/or audio outputs; (3) an audio detokenizer converts audio tokens into waveforms.

内容提要

Moonshot AI 发布了 Kimi-Audio，这是一个开源的音频基础模型，支持音频理解、生成与对话等多种任务。Kimi-Audio 基于 12.5Hz 音频分词器，设计了以连续特征为输入、离散 token 为输出的新型音频 LLM 架构，并开发了流式 detokenizer。预训练使用了超 1300 万小时覆盖语音、环境声、音乐等多模态的大规模音频数据，同时结合开放文本数据进行联合训练。模型支持语音识别、音频理解、音频问答、语音对话等任务，在多个基准测试上达到了开源 SOTA 水平，超越了 Qwen2-Audio、Baichuan-Audio、Step-Audio 等领先系统。Kimi-Audio 完全开源了模型权重、推理代码和评估工具包，为通用音频智能的发展提供了新的基础。未来团队计划进一步探索更丰富的音频描述学习、改进音频表征方式，并逐步摆脱对 ASR/TTS 伪标注的依赖。

相关链接

技术报告: https://arxiv.org/pdf/2504.18425

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

技术动态九

阿里巴巴发表最新开源大模型 QWen3，登顶全球最强开源模型

日期

2025-04-29

	Qwen3-235B-A22B MoE	Qwen3-32B Dense	OpenAI-o1 2024-12-17	Deepseek-R1	Grok 3 Beta Think	Gemini2.5-Pro	OpenAI-o3-mini Medium
ArenaHard	95.6	93.8	92.1	93.2	-	96.4	89.0
AIME'24	85.7	81.4	74.3	79.8	83.9	92.0	79.6
AIME'25	81.5	72.9	79.2	70.0	77.3	86.7	74.8
LiveCodeBench v5, 2024.10-2025.02	70.7	65.7	63.9	64.3	70.6	70.4	66.3
CodeForces Elo Rating	2056	1977	1891	2029	-	2001	2036
Aider Pass@2	61.8	50.2	61.7	56.9	53.3	72.9	53.8
LiveBench 2024-11-25	77.1	74.9	75.7	71.6	-	82.4	70.0
BFCL v3	70.8	70.3	67.8	56.9	-	62.9	64.6
MultiIF 8 Languages	71.9	73.0	48.8	67.7	-	77.8	48.4

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.
2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.
3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

内容提要

Qwen 团队正式发布了 Qwen3 系列大模型，成为当前全球最强开源模型，性能全面超越 DeepSeek R1。Qwen3 引入了国内首个混合推理机制：复杂问题采用深度推理模式，简单问题实现极速秒回，有效降低推理成本。Agent 能力大幅增强，原生支持 MCP 协议，且部署要求大幅下降，旗舰模型 Qwen3-235B-A22B 本地部署仅需 4 张 H20 显卡，部署成本相较 R1 下降超 60%。Qwen3 在训练数据规模上实现翻倍，使用了 36 万亿 tokens 进行训练，支持 119 种语言。此次开源包括 2 个 MoE 模型和 6 个 Dense 模型，支持灵活部署到 Hugging Face、ModelScope、Kaggle 等平台。Qwen3 的发布标志着基础大模型在推理效率、多模态能力与普适部署上的重要里程碑。

相关链接

Blog: https://qwenlm.github.io/zh/blog/qwen3/

模型： https://huggingface.co/collections/Qwen/qwen3-67dd247413f0e2e4f653967f

Github: https://github.com/QwenLM/Qwen3

Demo: https://huggingface.co/spaces/Qwen/Qwen3-Demo

Demo Chat: https://chat.qwenlm.ai

RESOURCE SHARING
资源分享

资源分享一

《Prompt Engineering》

类型：电子书

内容提要

《Prompt Engineering》由 Lee Boonstra 编写，是一本面向 LLM（大型语言模型）应用开发者和提示词工程实践者的系统指南。全书围绕如何设计高效提示（prompt）以引导语言模型生成准确、高质量输出展开，内容包括输出配置（如温度、top-k、top-p）、各类提示技术（零样本、单样本、少样本、系统提示、角色提示、上下文提示等），慢思考链（Chain of Thought）、树状思维（Tree of Thoughts）、自洽性推理（Self-consistency）、推理-行动（ReAct）等高级提示策略。书中还介绍了代码提示（写代码、解释代码、翻译代码、调试代码）以及多模态提示（文本+图片），并总结了提示工程的最佳实践，如提供示例、简化表述、明确输出要求、使用正面指令等。通过丰富的案例和实操指导，本书为想要深入掌握 LLM 提示设计的读者提供了全面而实用的参考。

资源链接

https://drive.google.com/file/d/1AbaBYbEa_EbPelsT40-vj64L-2IwUJHy/view

• 资源分享二

《Linear Model and Extensions》

类型：电子书

内容提要

《Linear Model and Extensions》围绕线性模型及其扩展展开，涵盖了线性模型的理论基础、推断方法、模型拟合与诊断、过拟合与正则化、变量选择、广义线性模型、相关数据的广义估计方程、分位数回归及生存分析等内容。全书注重理论与实践结合，不仅给出了严谨的数学推导，还配有丰富的 R 代码示例和实际数据分析案例。此外，书中包含大量习题，适合统计学及相关领域的教学与自学使用。通过系统梳理线性模型的发展脉络与关键扩展，该书为深入理解和应用现代统计建模技术提供了全面参考。

资源链接

https://arxiv.org/pdf/2401.00649

• 资源分享三

《A Practical Guide to Building Agent》

类型：电子书

内容提要

该指南面向产品和工程团队，系统梳理了基于大语言模型（LLM）的智能代理（agent）构建方法和最佳实践。文中首先定义了智能代理的基本特征：能自主执行多步骤任务、具备决策能力、能合理使用外部工具，并在流程中动态调整行为。指南进一步分析了适合采用智能代理的典型场景，包括复杂决策、难以维护的规则系统以及高度依赖非结构化数据的任务。

在设计层面，文中提出了智能代理的三大核心组成部分：模型（Model）、工具（Tools）、指令（Instructions），并以 OpenAI Agents SDK 为例，展示了如何在代码中实现和管理这些组件。针对模型选择、工具定义、指令编写等关键环节，指南给出了评估基准、优化建议和实际案例。文中还讨论了单代理和多代理系统的编排模式，包括 Manager Pattern 和去中心化交互等方式，帮助开发者根据业务复杂度灵活拆分和组织代理。最后，指南强调了“护栏”（Guardrails）机制在数据安全、决策可靠性等方面的重要作用，建议通过分层防护、规则校验和人工介入等措施，确保智能代理安全、合规地运行于生产环境。总体而言，本指南为希望部署和扩展智能代理的团队提供了结构化、可落地的设计与实现方案，推动了智能代理在实际业务场景中的应用落地与发展。

资源链接

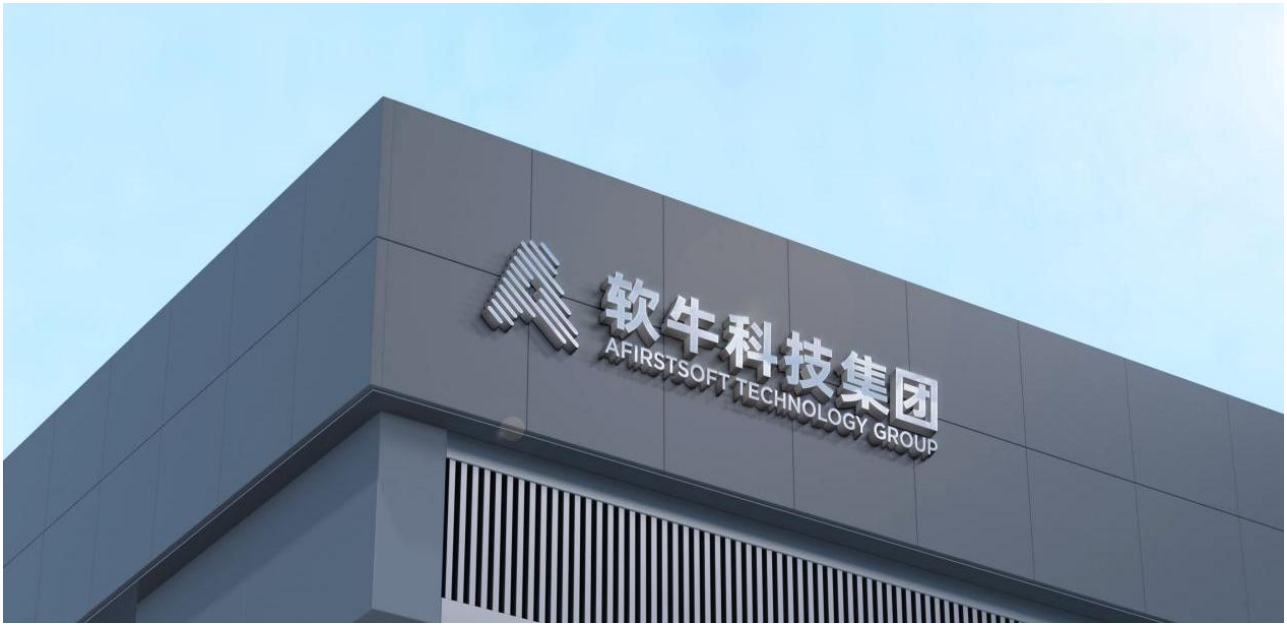
https://cdn.openai.com/business-guides-and-resources/a-practical-guide-to-building-agents.pdf

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 04 期

总第 22 期

广东省图象图形学会

主办 广东省图象图形学会

人工智能
前沿快报

2025 年第 04 期
总第 22 期
2025.05.01

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，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