Artificial Intelligence Newsletter

人工智能前沿快报

2025 年第 12 期
总第 30 期

学会官方公众号

广东省图象图形学会
GDSIG
学术交流、技术咨询、科学普及

目 录

学 术 动 态

技 术 动 态

目 录

资 源 分 享

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 10 篇最新学术动态、10 篇技术动态、以及 4 个资源推荐给广大读者。

ACADEMIC TRENDS
学术动态

● 论文一

TIDE: Temporal-Aware Sparse Autoencoders for Interpretable Diffusion Transformers in Image Generation

日期

2025-08-12

TIDE : Temporal-Aware Sparse Autoencoders for Interpretable Diffusion Transformers in Image Generation

Victor Shea-Jay Huang[*1,2], Le Zhuo[*1,2], Yi Xin[2,3], Zhaokai Wang[2,4], Fu-Yun Wang[1], Yuchi Wang[1], Renrui Zhang[1], Peng Gao[2], Hongsheng Li[†1]

[1]CUHK, MMLab [2]Shanghai AI Laboratory [3]NJU [4]SJTU

jeix782@gmail.com, hsli@ee.cuhk.edu.hk

内容摘要

本文提出了一种针对扩散 Transformer（Diffusion Transformers, DiTs）的新型框架——TIDE（Temporal-aware sparse autoencoders for Interpretable Diffusion transformErs），旨在提升 DiT 模型在图像生成过程中的可解释性与特征表达能力。与传统 U-Net 结构的扩散模型相比，DiT 的中间激活特征具有潜在的解释性，但缺乏有效的分析机制。TIDE 通过引入时间感知的稀疏自编码器，在不同降噪步骤中抽取稀疏且具有解释性的激活表示，使得 DiT 在捕捉概念层次（例如类别、语义和几何信息）方面更具可解释性，从而增强模型对下游图像生成任务的理解深度与表现。

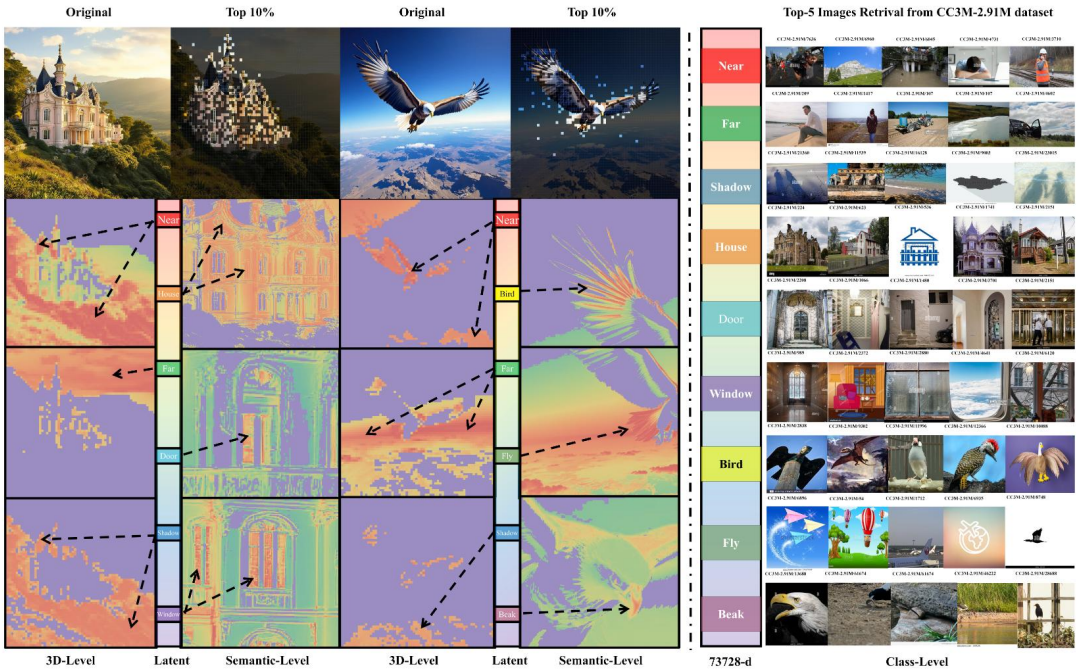


Figure 1: TIDE effectively encodes interpretable features across different levels (class, semantic, and 3D) in pre-trained diffusion transformer (Chen, Yu et al. 2023). This validates that the diffusion model inherently captures and organizes these multi-level concepts by large-scale generative pre-training, enabling it to perform various downstream diffusion tasks effectively(see Sec.Diffusion Really Learned Features for details).

主要亮点

- **时间感知的稀疏自编码框架：**引入 temporal-aware sparse autoencoders，有效捕捉扩散过程各阶段激活特征的时间动态，使模型在不同噪声级别下得到更稀疏且解释性强的表示。
- **增强 DiT 可解释性：**相比以往只关注生成质量的扩散模型设计，TIDE 能解构中间特征的语义与层次结构（如类别、语义和 3D 形状信息），为理解扩散 Transformer 的内部机制提供了直观视角。
- **提高下游生成表现与结构性理解：**通过对特征思想的结构化抽取，TIDE 帮助 DiT 在图像生成任务中不仅生成更高质量图像，也提升了模型对复杂概念的解析能力，有望促进可解释 AI 在生成领域的实际落地。

相关链接

Paper： https://arxiv.org/pdf/2503.07050

● 论文二

A Chinese AI model taught itself basic physics — what discoveries could it make?

日期

2025-11-14

NEWS | 14 November 2025

A Chinese AI model taught itself basic physics — what discoveries could it make?

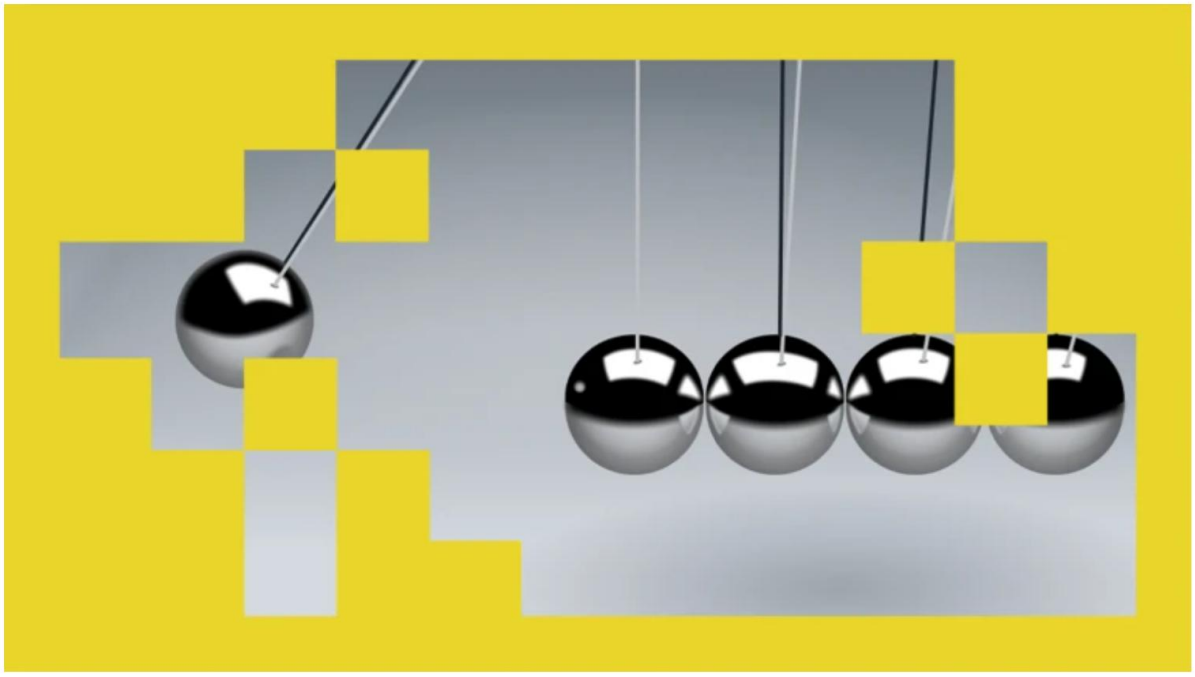
A tool called AI-Newton can derive scientific laws from raw data, but is some way from developing human-like reasoning.

By Mohana Basu

内容摘要

本文报道了一项来自中国的最新研究进展：研究团队提出了一种名为 AI-Newton 的人工智能系统，能够仅基于原始实验数据，自主推导出基础物理定律，例如牛顿第二定律。不同于主流深度学习模型侧重模式拟合和预测，AI-Newton 采用符号回归（symbolic regression）的方法，搜索最优的数学表达式来解释物理现象，并通过持续积累中间概念（如速度、质量）来逐步构建物理知识体系。

在 46 个涵盖自由运动、碰撞、振动与摆动系统的模拟物理实验中，AI-Newton 成功从噪声数据中恢复关键物理关系，并在后续任务中复用已学概念，表现出类似人类科学推理的“渐进式建模”特征。尽管该系统距离真正具备自主科学发现能力仍有差距，但研究展示了 AI 在“从数据到定律”这一核心科学能力上的重要一步，也凸显了当前大模型在概念抽象与物理归纳方面的局限性。



Researchers gave an AI data from physics experiments involving systems using pendulum-like motion to see if it could derive basic laws of physics. Credit: stefilyn/Getty

主要亮点

- **从“预测现象”走向“发现定律”的 AI 科学推理范式：**与仅能拟合或预测物理轨迹的主流深度模型不同，AI-Newton 的目标是从实验数据中直接推导可解释、可复用的物理定律。它不仅输出数值结果，还能生成符合物理直觉的数学公式（如速度、力与质量关系），标志着 AI 从“黑箱预测器”向“科学建模者”的重要转变。

- **基于符号回归的概念发现与知识累积机制：**AI-Newton 使用符号回归搜索简洁、可推广的数学表达式，并将已发现的概念（如速度）存入知识库，在后续任务中反复调用。这种概念递进式学习机制更接近人类科学发现过程，而非一次性端到端拟合，使模型具备一定程度的结构化知识内化能力。

- **揭示基础模型在物理概念抽象上的根本局限：**报道将 AI-Newton 与大型基础模型（如 GPT、Claude、LLaMA）进行对比，指出后者即便能准确预测物理系统行为，也往往无法归纳出简洁、合理的物理定律，甚至

产生“非人类式”的错误引力规律。这一对比强调：仅靠大规模数据与预测目标，难以自然涌现科学概念，结构化归纳机制仍不可或缺。

相关链接

Paper： https://www.nature.com/articles/d41586-025-03659-4

● 论文三

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning

日期

2025-11-27

DeepSeekMath-V2: Towards Self-Verifiable Mathematical Reasoning

Zhihong Shao*, Yuxiang Luo*, Chengda Lu*†, Z.Z. Ren*
Jiewen Hu, Tian Ye, Zhibin Gou, Shirong Ma, Xiaokang Zhang
DeepSeek-AI
zhihongshao@deepseek.com

https://github.com/deepseek-ai/DeepSeek-Math-V2

内容摘要

DeepSeekMath-V2 是一套专注于自我验证数学推理能力的模型与训练体系，在大型语言模型（LLM）基础上设计了一个面向数学定理证明的生成–验证循环结构。传统方法通常仅对最终答案准确性给予奖励，而忽略推理过程的完整性与逻辑严谨性，这在定理证明等任务中尤为不足。作者提出先训练一个可靠的证明验证器（verifier），再以验证器作为奖励模型指导证明生成器不断改进自身逻辑推理过程。此外引入扩展验证计算来标注难以自动验证的证明样本，从而形成一个针对性更强的训练集。DeepSeekMath-V2 在国际数学奥林匹克

（IMO）、中国数学奥林匹克（CMO）和 Putnam 大赛等严苛测试中展现了强大的数学推理与自我验证能力。

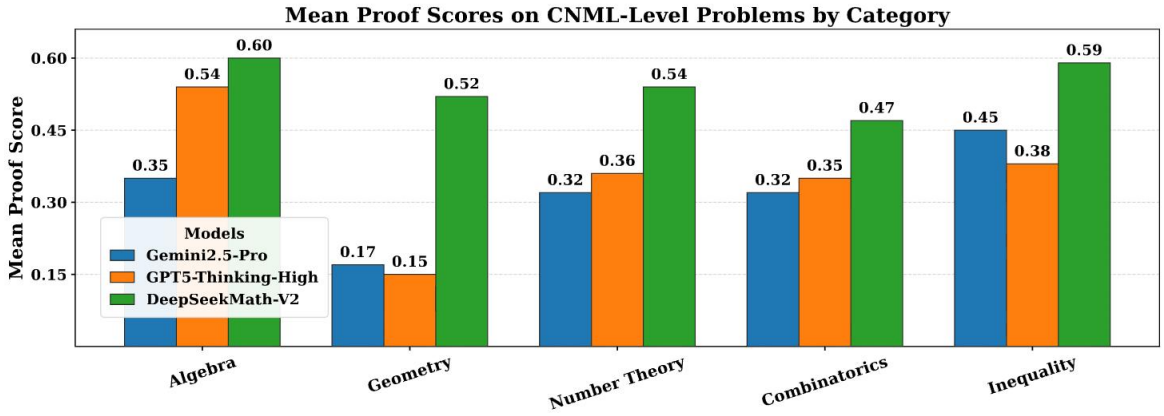


Figure 1 | Average proof scores on CNML-level problems by category and model, as evaluated by our verifier.

主要亮点

- **自我验证机制**：不仅训练生成模型去提出数学证明，更训练一个 verifier 去发现证明中的逻辑错误，使模型具备检查自身推理合理性的能力，这种双向学习机制显著提升了逻辑严谨性。
- **迭代生成–验证优化循环**：利用 verifier 作为奖励模型反哺 proof generator，使生成过程不断修正和完善自己的证明，这种内循环机制突破了只看最终答案的传统奖励设计。
- **高性能实证结果**：在高难度数学竞赛（IMO 2025、CMO 2024、Putnam 2024 等）上的优异表现，展示了模型不仅能解题，还能在没有已知标准答案的情况下评估和提升自身输出的合理性。

相关链接

Paper：https://arxiv.org/abs/2511.22570

Github：https://github.com/deepseek-ai/DeepSeek-Math-V2

● 论文四

How Far Are We from Genuinely Useful Deep Research Agents?

日期

2025-12-01

oppo

How Far Are We from Genuinely Useful Deep Research Agents?

OPPO AI Agent Team

内容摘要

本文系统性评估了当前所谓的深度研究代理（Deep Research Agents, DRAs）在执行真实世界科研任务（如撰写分析性报告、综合文献与实验洞察）方面的能力差距。作者指出现有 DRA 多在问答式基准任务上验证，而对于综合性、结构化的分析报告生成缺乏合适的评估标准。为此提出了 FINDER：一个包含 100 个由专家细化的研究任务和 419 条检查清单项的精细评估基准，以更严格衡量 DRAs 在结构完整性、事实可信度、分析深度等维度的表现。同时构建了 Deep rEsearch Failure Taxonomy (DEFT)，系统归纳 DRA 在实践中常见的失败类型。这一研究为衡量与推进 DRA 真正有用性提供了基础工具与诊断视角。

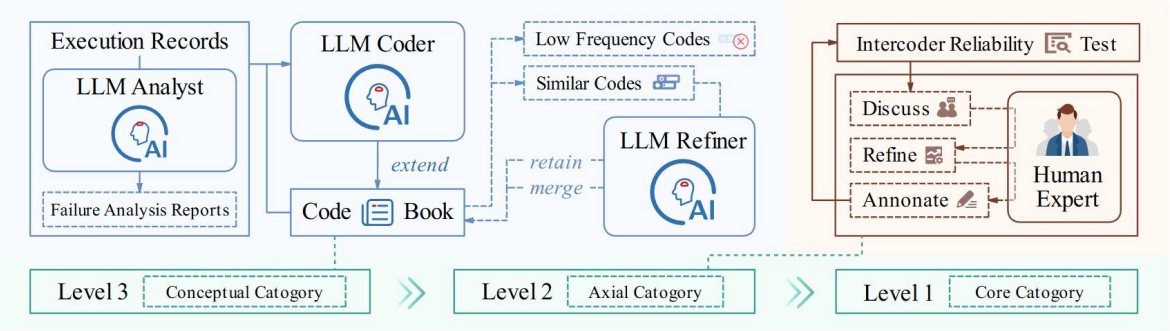


Figure 2 Overview of the DEFT Construction.

主要亮点

- **FINDER 精细基准**：提出了一个比传统 QA 基准更具真实科研报告特性的数据集与任务集合，从结构严谨度、分析深度到事实引用等方面评估 DRA 生成质量。
- **失败行为分类（DEFT）**：首次梳理出了 DRA 在复杂研究任务中可能出错的 14 类典型模式，为后续改进提供了可操作的诊断框架。
- **强调科研代理实际效用差距**：指出现存方法在合成报告、深入综述和逻辑严谨推理方面仍明显不足，这种高度结构化的任务比简单问答更能揭示代理能力瓶颈，推动更接近真实科研场景的 agent 设计。

相关链接

Paper：https://arxiv.org/abs/2512.01948

Github：https://github.com/OPPO-PersonalAI/FINDER_DEFT

● 论文五

Stabilizing Reinforcement Learning with LLMs: Formulation and Practices

日期

2025-12-03

Stabilizing Reinforcement Learning with LLMs: Formulation and Practices

Chujie Zheng* Kai Dang Bowen Yu* Mingze Li Huiqiang Jiang
Junrong Lin Yuqiong Liu Hao Lin Chencan Wu Feng Hu
An Yang Jingren Zhou Junyang Lin
Qwen Team, Alibaba Inc.

内容摘要

本文系统讨论了为什么 LLM 的强化学习（RL）训练容易不稳定，并给出一个更“可解释”的统一视角：在很多 LLM-RL 实践中，我们用的是序列级奖励（整段回答给一个分数），但优化时却常用 token 级目标（如 REINFORCE/GRPO 的 token-level 形式），这在理论上存在“奖励-优化粒度不匹配”的疑问。作者提出一种新的公式化解释：token-level 目标可以被视为真实序列级目标的一阶近似，但这一近似要“站得住”，需要同时满足两个条件——训练-推理引擎差异（training–inference discrepancy）足够小，以及策略陈旧性（policy staleness）足够小。基于该洞见，论文进一步解释了为什么若干业内常用稳定化技巧（重要性采样修正、clipping，以及 MoE 模型中特别关键的 Routing Replay）能显著改善稳定性，并在一个 30B MoE 模型上进行了大规模实证，给出 on-policy / off-policy 场景下更可靠的训练“配方”。作者还观察到：一旦训练过程稳定，长时间优化后不同冷启动（SFT/初始化）差异会显著缩小。

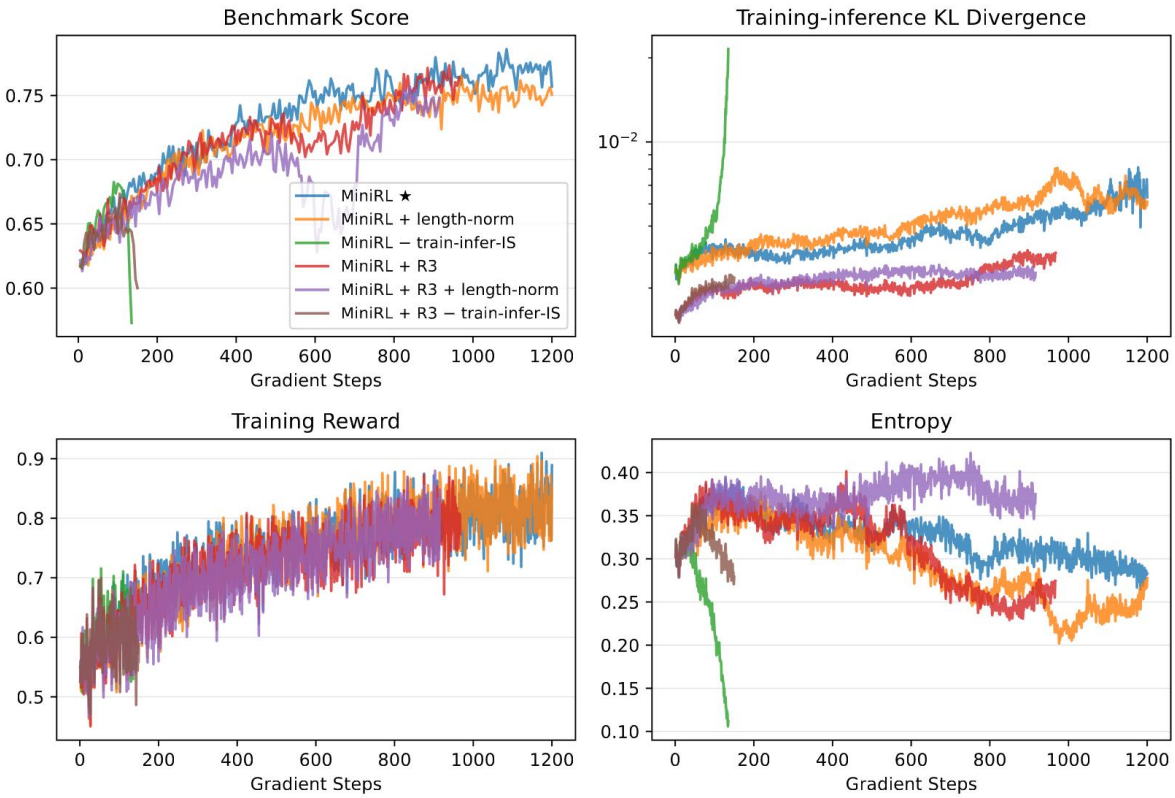


Figure 1: Results of on-policy training with gbs (global batch size) = mbs (mini-batch size) = 1,024.

主要亮点

- **token 级优化合理：一阶近似 + 两个稳定性条件：**论文最核心的贡献是给出一个很“工程可落地”的理论解释：token-level surrogate 并非天然正确，而是序列级目标的一阶近似；要让近似有效，必须尽量同时压低 training–inference discrepancy 与 policy staleness。这为 LLM-RL 训练中常见的崩溃/震荡现象提供了可诊断的原因框架。

- **统一解释了“常用稳定化 trick”为何有效，并把它们对准不同的误差源：**文章将重要性采样修正、clipping、以及 MoE 训练里常被提及但缺乏统一理论解释的 Routing Replay，纳入同一推导视角：这些方法并不是零散的工程技巧，而是在不同层面分别抑制训练-推理不一致与策略陈旧性带来的偏差积累；尤其对 MoE 而言，

动态路由会额外放大这种偏差，Routing Replay 的作用因此变得“必要且可解释”，这对大规模 MoE 的 RL 训练具有直接指导意义。

- **给出了可复用的大规模实践结论：**稳定后，长期优化会弱化冷启动差异，训练配方比初始化更重要：作者通过 30B MoE 的大规模实验展示：on-policy 往往更稳，而在 off-policy 或多轮更新场景中，策略陈旧性会快速成为主要不稳定源，需要相应的稳定化机制兜底；更重要的是，论文观察到只要训练过程足够稳定并持续优化，不同 SFT/初始化之间的差距会显著收敛，说明决定最终上限的关键不在“起跑线”，而在 RL 过程中的稳定动力学控制与工程配方选择。

相关链接

Paper： https://arxiv.org/abs/2512.01374

● 论文六

From Code Foundation Models to Agents and Applications: A Comprehensive Survey and Practical Guide to Code Intelligence

日期

2025-12-03

From Code Foundation Models to Agents and Applications: A Comprehensive Survey and Practical Guide to Code Intelligence

BUAA-SKLCCSE, Alibaba, ByteDance, M-A-P, BJTU, OPPO, HKUST (GZ), BUPT, TeleAI, Shanghai AI Lab, Manchester, StepFun, UoS, SCU, CASIA, NJU, Kuaishou, HIT, Huawei Cloud, Tencent, Monash/CSIRO, NTU, ZJU, BIT, Ubiquant, NUS, HNU, PKU, CSU

内容摘要

本文对代码大型语言模型（Code LLMs）进行了全面概述，涵盖了其生命周期、技术细节和应用领域。文章首先介绍了代码 LLMs 的兴起和发展，并分析了通用和专用模型的特点和优缺点。随后，文章详细探讨了数

据准备、指令微调、多模态和推理能力等关键技术，并强调了安全性和可解释性的重要性。最后，文章展望了代码 LLMs 的未来发展趋势，例如软件工程代理和自主学习。

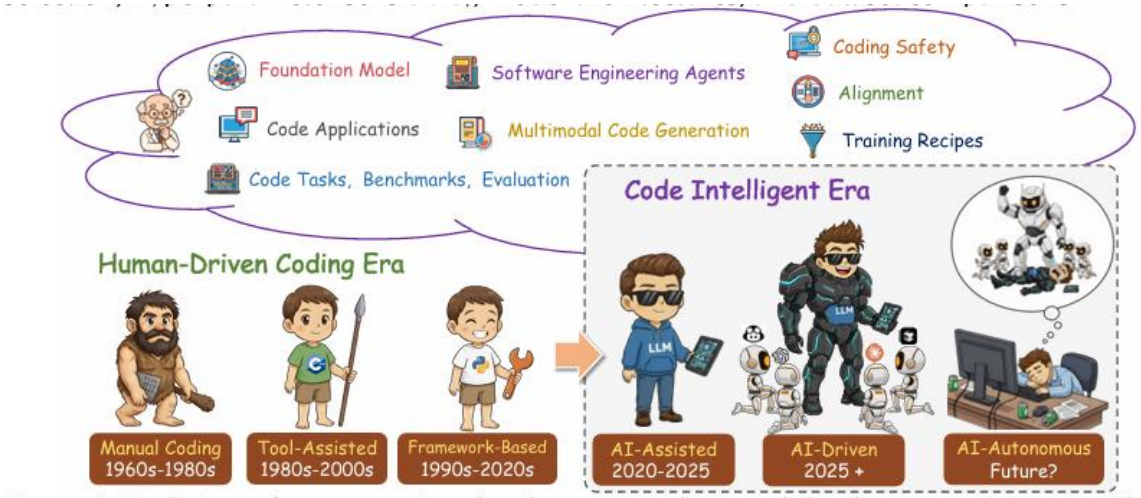


Figure 1. Evolution of programming development and research landscapes in AI-powered code generation. The upper section highlights the key research areas covered in this work. The timeline below illustrates the six-stage evolution from the human-driven coding era to the emerging code intelligence era.

主要亮点

- **代码 LLMs 的全面概述**：从模型架构、数据准备、指令微调到应用领域，文章提供了一个系统性的分析。
- **通用与专用模型的对比**：分析了通用模型在代码能力方面的优势和局限性，以及专用模型在特定领域的优越性。
- **关键技术探讨：**深入探讨了数据准备、指令微调、多模态和推理能力等关键技术，并强调了它们对模型性能的影响。

相关链接

Paper： https://arxiv.org/abs/2511.18538

● 论文七

State of AI: An Empirical 100 Trillion Token Study with OpenRouter

日期

2025-12-04

SIMA 2: A Generalist Embodied Agent for Virtual Worlds

SIMA Team, Google DeepMind[1]

内容摘要

Figure 1 | **SIMA 2** is a Gemini-based agent that reasons, acts, and engages in dialogue across diverse embodied 3D virtual worlds. In the top left panel, we see an example of the agent responding to the user in No Man's Sky. As compared with SIMA 1, SIMA 2 is a step-change improvement in embodied performance, and it is even capable of self-improving in previously unseen environments.

SIMA 2，一种通用的具身代理，能够理解并在各种三维虚拟世界中行动。SIMA 2 建立在双子座基础模型之上，代表了在具身环境中实现主动、目标导向互动的重要一步。与之前仅限于简单语言命令的 SIMA 1 不同，SIMA 2 作为一个交互伙伴，能够推理高层目标、与用户对话，并处理通过语言和图像传递的复杂指令。在多

样化的游戏组合中，SIMA 2 大幅缩小了与人类表现的差距，并展示了对前所未见环境的强有力推广，同时保留了基础模型的核心推理能力。此外，我们展示了开放式自我提升的能力：通过利用 Gemini 生成任务和提供奖励，SIMA 2 能够在新环境中自主从零学习新技能。这项工作验证了为虚拟世界乃至最终物理世界创造多功能且持续学习的代理的路径。

主要亮点

- **多模型生态系统的兴起：**不同于单一模型主导的局面，LLM 使用呈现多元化趋势，开源和闭源模型都占据重要份额。这表明未来 LLM 使用将是模型无关的，开发者需要灵活选择最适合的任务的模型，而模型提供商则需要不断创新和差异化。
- **使用多样性超越生产力：**除了编程、写作等生产力任务，LLM 在角色扮演和娱乐方面的使用量也很大。这表明 LLM 在陪伴和探索方面具有巨大潜力，并为模型评估指标带来新的思考。
- **代理推理的兴起：**LLM 使用正在从单次交互转向代理推理，模型能够进行规划、推理和执行多步骤操作。这意味着模型评估的重点将转向任务完成和效率，而不仅仅是语言质量。

相关链接

Paper：https://arxiv.org/abs/2512.04797

● 论文八

Implicit Reasoning in Large Language Models: A Comprehensive Survey

日期

2025-12-04

State of AI:
An Empirical 100 Trillion Token Study with OpenRouter

Malika Aubakirova[*†], **Alex Atallah**[‡], **Chris Clark**[‡], **Justin Summerville**[‡], **and Anjney Midha**[†]

[‡]OpenRouter Inc. [†]a16z (Andreessen Horowitz)

December, 2025

内容摘要

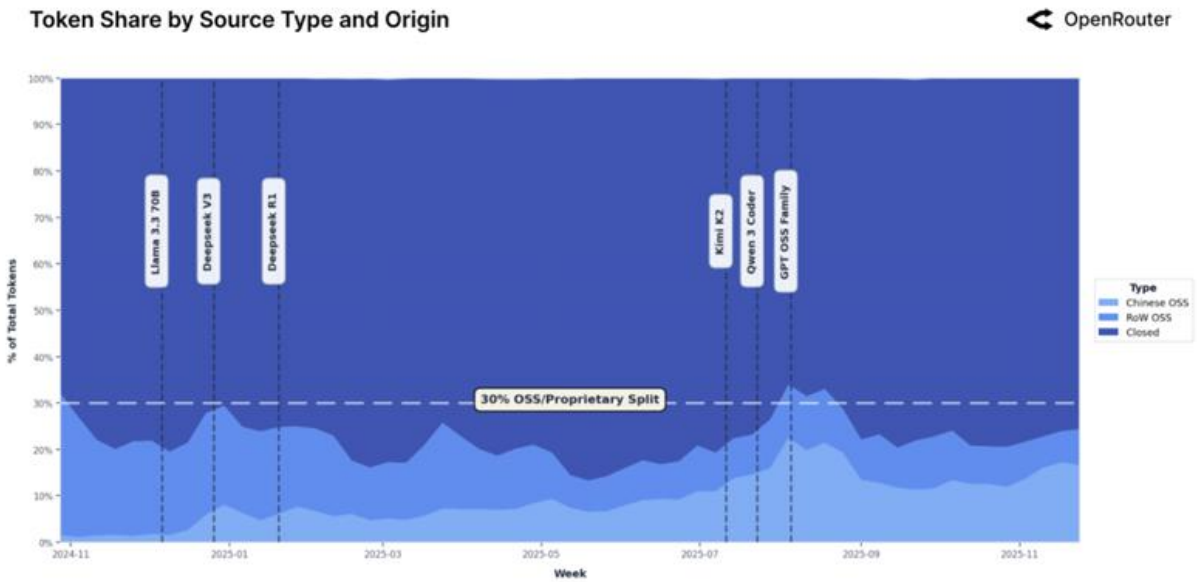


Figure 1: **Open vs closed source models split.** Weekly share of total token volume by source type. Lighter blue shades represent open-weight models (China vs Rest-of-World), while dark blue corresponds to proprietary (closed) offerings. Vertical dashed lines mark the release of key open weight models including Llama 3.3 70B, DeepSeek V3, DeepSeek R1, Kimi K2, GPT OSS family, and Qwen 3 Coder.

本文基于 OpenRouter 平台上的 100 万亿个 LLM 交互数据，分析了 LLM 在现实世界中的应用情况。研究发现了开源模型和闭源模型并存的趋势，开源模型在角色扮演和编程领域占据主导地位。此外，研究还发现了代理推理的兴起，用户开始将 LLM 集成到更大的自动化系统中，并进行多步骤推理。研究还分析了 LLM 使用的地域差异和用户留存模式，并提出了“灰姑娘水晶鞋”现象，即早期用户对模型的忠诚度更高。最后，研究讨论了 LLM 使用的成本和效益动态，以及 LLM 市场的竞争格局。

主要亮点

- **多模型生态系统的兴起：**不同于单一模型主导的局面，LLM 使用呈现多元化趋势，开源和闭源模型都占据重要份额。这表明未来 LLM 使用将是模型无关的，开发者需要灵活选择最适合的任务的模型，而模型提供商则需要不断创新和差异化。
- **使用多样性超越生产力：**除了编程、写作等生产力任务，LLM 在角色扮演和娱乐方面的使用量也很大。这表明 LLM 在陪伴和探索方面具有巨大潜力，并为模型评估指标带来新的思考。
- **代理推理的兴起：**LLM 使用正在从单次交互转向代理推理，模型能够进行规划、推理和执行多步骤操作。这意味着模型评估的重点将转向任务完成和效率，而不仅仅是语言质量。

相关链接

Paper：https://openrouter.ai/state-of-ai

● 论文九

Evaluating Large Language Models in Scientific Discovery

日期

2025-12-17

Evaluating Large Language Models in Scientific Discovery

Zhangde Song[1, †, ‡], Jieyu Lu[1, †], Yuanqi Du[2, †], Botao Yu[3, †], Thomas M. Pruyn[4, †], Yue Huang[5, †], Kehan Guo[5, †], Xiuzhe Luo[6, †], Yuanhao Qu[7, †], Yi Qu[8, ‡], Yinkai Wang[9, ‡], Haorui Wang[10, ‡], Jeff Guo[11, ‡], Jingru Gan[12, ‡], Parshin Shojaee[13, ‡], Di Luo[14, 15, ‡], Andres M Bran[11], Gen Li[16], Qiyuan Zhao[1], Shao-Xiong Lennon Luo[17], Yuxuan Zhang[18, 33, 34], Xiang Zou[4], Wanru Zhao[19], Yifan F. Zhang[21], Wucheng Zhang[22], Shunan Zheng[23], Saiyang Zhang[23], Sartaaj Takrim Khan[4], Mahyar Rajabi-Kochi[4], Samantha Paradi-Maropakis[4], Tony Baltoiu[24], Fengyu Xie[25], Tianyang Chen[26], Kexin Huang[7], Weiliang Luo[27, 28], Meijing Fang[29], Xin Yang[27], Lixue Cheng[30], Jiajun He[20], Soha Hassoun[9], Xiangliang Zhang[5], Wei Wang[12], Chandan K. Reddy[13], Chao Zhang[10], Zhiling Zheng[31], Mengdi Wang[21], Le Cong[7], Carla P. Gomes[2], Chang-Yu Hsieh[29], Aditya Nandy[32], Philippe Schwaller[11], Heather J. Kulik[27, 28], Haojun Jia[1, *], Huan Sun[3, *], Seyed Mohamad Moosavi[4, 18, *], and Chenru Duan[1, ‡, *]

[1]Deep Principle, Hangzhou, China
[2]Department of Computer Science, Cornell University, Ithaca, NY, USA
[3]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA
[4]Department of Chemical Engineering & Applied Chemistry, University of Toronto, Toronto, ON, Canada
[5]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA
[6]QuEra Computing Inc., Boston, MA, USA
[7]Department of Pathology, Department of Genetics, Cancer Biology Program, Stanford University School of Medicine, Stanford, CA, USA
[8]Harvard Law School, Cambridge, MA, USA
[9]Department of Computer Science, Tufts University, Medford, MA, USA
[10]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[11]Laboratory of Artificial Chemical Intelligence, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
[12]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, USA
[13]Department of Computer Science, Virginia Tech, Arlington, VA, USA
[14]Department of Physics, Tsinghua University, Beijing, China
[15]Institute for Advanced Study, Tsinghua University, Beijing, China
[16]Department of Chemistry, Princeton University, Princeton, NJ, USA
[17]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[18]Vector Institute for Artificial Intelligence, Toronto, ON, Canada
[19]Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom
[20]Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[21]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA
[22]Department of Physics, Princeton University, Princeton, NJ, USA
[23]Department of Physics, The University of Texas at Austin, Austin, TX, USA
[24]Department of Mechanical Engineering, McGill University, Montreal, QC, Canada
[25]College of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei, Anhui, China
[26]Department of Chemical Engineering, Stanford University, Stanford, CA, USA
[27]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA
[28]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[29]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang, China
[30]Department of Chemistry, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China
[31]Department of Chemistry, Washington University in St. Louis, St. Louis, MO, USA
[32]Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, Los Angeles, CA, USA
[33]Department of Chemical Engineering & Applied Chemistry, University of Toronto, Toronto, ON, Canada
[34]Institute of Physics, Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
[†]These authors contribute equally
[‡]Project contributor
[*]Correspondence to: haojunjia@deepprinciple.com, sun.397@osu.edu, mohamad.moosavi@utoronto.ca, duanchenru@gmail.com

内容摘要

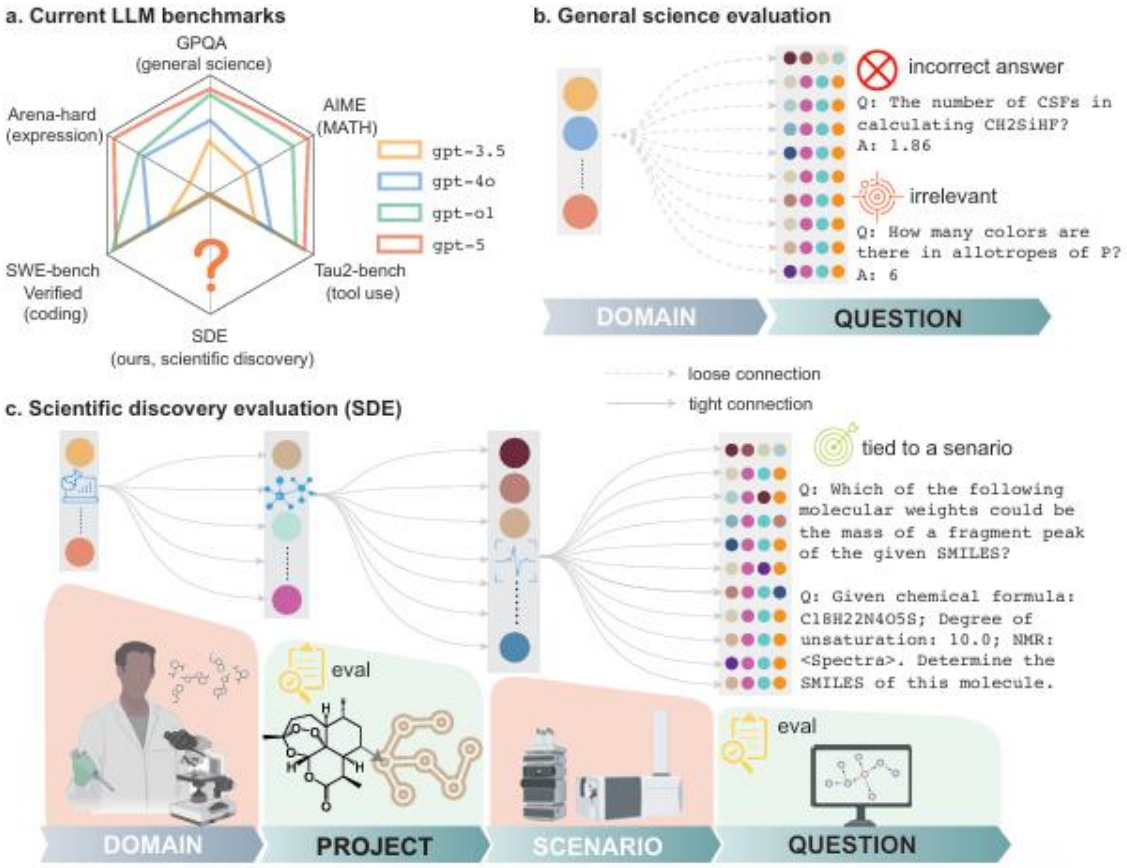


Fig. 1 | From evaluating LLMs on general-science quizzes to scenario-grounded scientific discovery. a. Schematic comparison of representative LLM benchmarks. GPQA, AIME, Arena-hard, SWE-bench verified, Tau2-bench, alongside our scientific discovery evaluation (SDE) are shown. Shaded polygons indicate relative performance of four models (`gpt-3.5`, `gpt-4o`, `gpt-o1`, `gpt-5`) across benchmarks. Only GPT series are shown as representatives to show their performance improve with time. **b.** Limitations of general-science Q&A. Existing benchmarks often contain questions that are less relevant to scientific discovery or incorrect answers as ground-truth. **c.** The SDE framework anchors assessment to projects and realistic research scenarios within each scientific domain, producing tightly coupled questions, enabling more faithful evaluation of LLMs for scientific discovery. LLMs are evaluated on both question and project levels. A project of discovering new pathways for artemisinin synthesis is shown as an example, which comprises multiple scenarios, such as forward reaction prediction and structure elucidation from nuclear magnetic resonance (NMR) spectra, where the question sets are finally collected.

大型语言模型（LLMs）越来越多地应用于科学研究，然而主流科学基准却探讨了脱离语境化的知识，忽视了推动科学发现的迭代推理、假设生成和观察解释。我们引入了一个基于场景的基准测试，评估生物学、化学、材料和物理领域的大型语言模型，领域专家定义真正有趣的研究项目，并将其分解为模块化的研究场景，从中抽取经过筛选的问题。该框架在两个层面评估模型：（i）情景相关题目的问题级准确性，以及（ii）项目级表现，模型需提出可检验的假设，设计模拟或实验，并解释结果。将这一两阶段科学发现评估（SDE）框架应用

于最先进的大型语言模型，显示出相对于一般科学基准测试存在持续的性能差距、模型规模和推理扩展的收益递减，以及不同提供者顶级模型间系统性存在的弱点。研究场景中性能差异很大，导致在评估的科学发现项目中，选择最佳模型的选择发生变化，表明所有现有 LLM 都远离一般科学“超级智能”。尽管如此，LLMs 已经在各种科学发现项目中展现出潜力，包括成分情景评分较低的案例，凸显了引导探索和偶然性在发现中的作用。该 SDE 框架为 LLM 的发现相关评估提供了可重复的基准，并规划了推动其科学发现发展的实用路径。

主要亮点

- **SDE 评估框架:** 提出了一个名为“科学发现评估”（SDE）的新框架，该框架通过将问题与实际研究场景和项目联系起来，评估大型语言模型（LLMs）在科学发现中的能力。SDE 评估框架超越了传统的科学问答基准，更贴近科学发现的实际情况。
- **LLMs 的局限性:** 通过 SDE 框架对最先进的 LLMs 进行评估，发现它们在科学发现方面存在明显的局限性，包括推理能力的提升逐渐停滞、模型之间存在共性的失败模式以及项目层面的性能波动。
- **未来发展方向:** 文章提出了几个改进 LLMs 在科学发现中应用的方向，包括针对问题制定和假设生成进行针对性训练、多样化预训练数据来源、将工具使用融入训练过程以及为科学推理开发强化学习策略。

相关链接

Paper： https://arxiv.org/abs/2512.15567

● 论文十

LLaDA2.0: Scaling Up Diffusion Language Models to 100B

日期

2025-12-24

LLaDA2.0: Scaling Up Diffusion Language Models to 100B

Tiwei Bie[1], Maosong Cao[1], Kun Chen[1], Lun Du[1], Mingliang Gong[1], Zhuochen Gong[1], Yanmei Gu[1], Jiaqi Hu[1,3], Zenan Huang[1], Zhenzhong Lan[1,4,†], Chengxi Li[1], Chongxuan Li[2], Jianguo Li[1,†], Zehuan Li[1], Huabin Liu[1], Lin Liu[1], Guoshan Lu[1], Xiaocheng Lu[1,5], Yuxin Ma[1], Jianfeng Tan[1], Lanning Wei[1], Ji-Rong Wen[2], Yipeng Xing[1], Xiaolu Zhang[1], Junbo Zhao[1,3,†], Da Zheng[1,†], Jun Zhou[1], Junlin Zhou[1], Zhanchao Zhou[1,4], Liwang Zhu[1], Yihong Zhuang[1]

[1]Ant Group, [2]Renmin University of China, [3]Zhejiang University,
[4]Westlake University, [5]HongKong University of Science and Technology

内容摘要

LLaDA2.0——一组离散扩散大型语言模型（dLLM），通过系统转换从自回归（AR）模型扩展至 100B 总参数，建立了前沿规模部署的新范式。LLaDA2.0 没有从零开始进行昂贵的训练，而是坚持知识继承、渐进适应和效率意识设计原则，并无缝将预训练的增强现实模型无缝转换为基于 WSD 的三阶段模块级 WSD 训练方案：模块扩散（预热）中逐步增加块大小，稳定实现大规模全序列扩散，并恢复到紧凑大小的模块扩散（衰减）。在与 SFT 和 DPO 的培训后对齐外，我们还获得了 LLaDA2.0-mini (16B) 和 LLaDA2.0-flash (100B)，这两种经过指令调优的专家混合（Mixture-of-Experts，MoE）变体，专为实际部署优化。通过保留并行解码的优势，这些模型在前沿尺度上提供了卓越的性能和效率。这两种模型都是开源的。

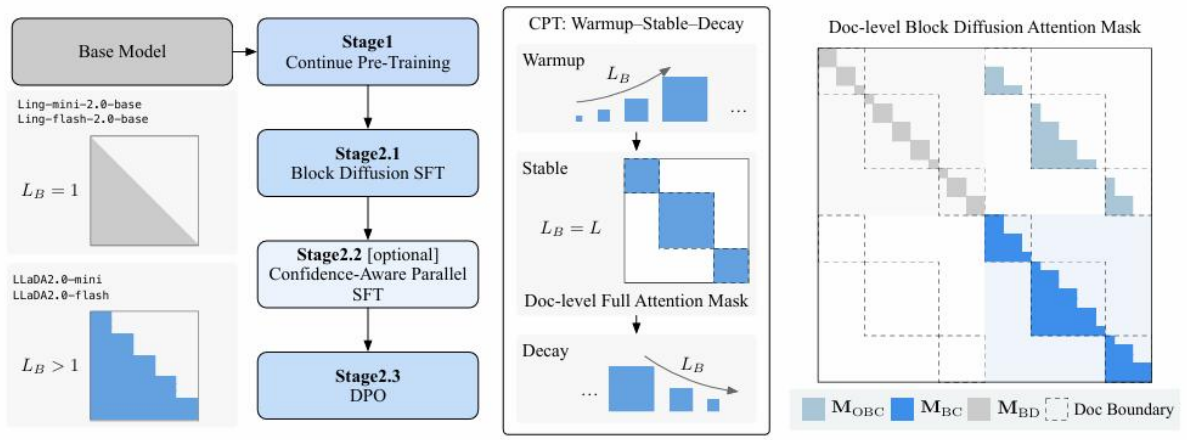


Figure 2: **A schematic of the progressive training framework for transforming an AR model into a MDLM.** Continual Pre-Training Stage facilitates the **Warmup-Stable-Decay** strategies by scheduling block size L_B enables smooth, stable, and effective attention mask adaptation. Post-training Stage facilitates the same block diffusion configuration conducting the instruction SFT, Confidence-Aware Parallel SFT, and DPO. The right panel illustrates the document-level block diffusion attention mask,which enables an efficient, vectorized forward pass by constructing a single input sequence from multiple noisy and clean examples, such as $[\boldsymbol{x}_{\text{noisy1}}, \ldots, \boldsymbol{x}_{\text{clean1}}, \ldots]$. The forward pass then employs a combination of block-diagonal ($\mathbf{M}_{\text{BD}}$), offset block-causal ($\mathbf{M}_{\text{OBC}}$), and block-causal ($\mathbf{M}_{\text{BC}}$) masks.

主要亮点

- **WSD 转换策略：**LLaDA2.0 提出了 WSD 策略，通过渐进式增加块大小、大规模全序列扩散和恢复到紧凑块大小的扩散，实现了 AR 到 dLLM 的平滑转换，有效解决了直接转换导致的优化不稳定和知识退化问题。
- **文档级注意力掩码：**LLaDA2.0 引入文档级注意力掩码，确保了语义边界内的双向建模一致性，防止了跨文档干扰，从而提高了学习稳定性和语义理解能力。
- **置信度预测损失：**LLaDA2.0 通过引入置信度预测损失，使模型能够更准确地预测每个 token 的概率分布，从而实现了高效的并行解码，提高了推理速度。

相关链接

Paper： https://arxiv.org/abs/2512.15745

GitHub： https://github.com/inclusionAI/LLaDA2.0

设计了以目标为中心的奖励函数，能够在视角变化时提供准确的反馈；同时，采用基于课程学习的训练策略，逐步提升智能体在复杂场景中的跟踪性能。在仿真器和真实图像数据集上的实验结果表明，该方法在复杂场景无人机视觉主动跟踪任务上显著优于现有最优算法。

主要亮点

- **提出了统一的无人机主动跟踪基准 DAT**：能够充分验证智能体在跨场景和跨域条件下的迁移能力。
- **提出了以目标为中心的主动跟踪算法 GC-VAT**：利用以目标为中心的奖励函数和基于课程学习的训练策略，显著提升复杂场景下无人机的跟踪性能。
- **全面实验说明基准和方法的有效性**：在仿真器和真实图像上的实验充分验证了 DAT 基准和 GC-VAT 方法的有效性。

相关链接

Paper：https://arxiv.org/pdf/2412.00744

GitHub：https://github.com/SHWplus/DAT_Benchmark

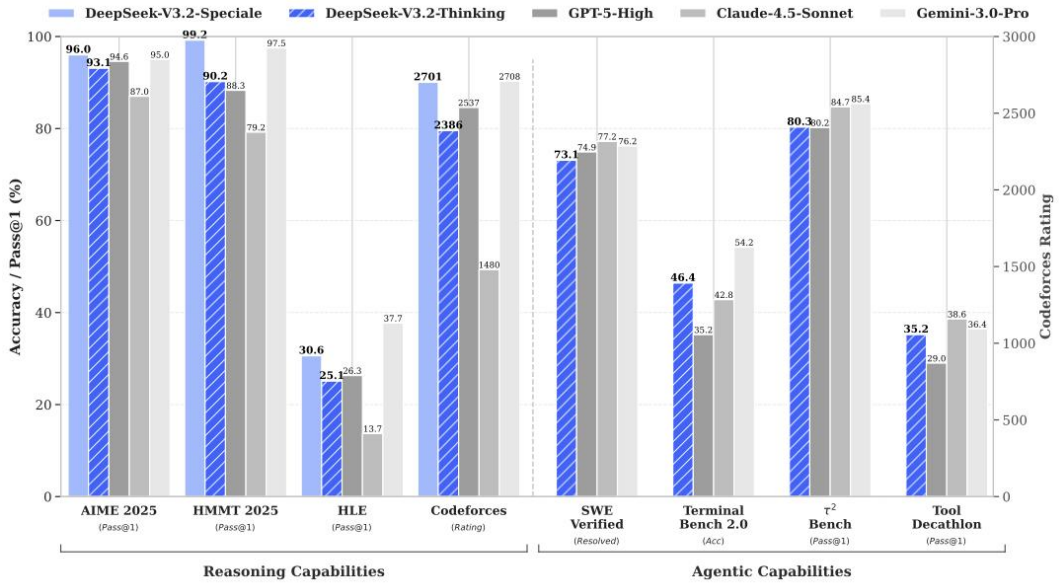
TECHNICAL UPDATES
技术动态

• 技术动态一

DeepSeek-AI 推出高效推理与 Agent 模型 DeepSeek-V3.2

日期

2025-12-01



内容摘要

DeepSeek-AI 正式发布 DeepSeek-V3.2，这是一款旨在协调高计算效率与卓越推理及智能体（Agent）性能的开源大语言模型。该模型在多项基准测试中展现了与 GPT-5 及 Kimi-k2-thinking 相当的实力，其高算力变体 DeepSeek-V3.2-Speciale 更是超越了 GPT-5，在推理能力上与谷歌的 Gemini-3.0-Pro 持平，并在 2025

年国际数学奥林匹克（IMO）和国际信息学奥林匹克（IOI）中均取得了金牌级表现。DeepSeek-V3.2 的核心技术亮点包括：(1) DeepSeek 稀疏注意力（DSA）：一种高效的注意力机制，通过细粒度的 Token 选择大幅降低了长上下文场景下的计算复杂度，同时保持模型性能；(2) 可扩展的强化学习框架：利用 GRPO 算法并大幅增加后训练阶段的算力投入（超过预训练成本的 10%），有效提升了模型的推理能力；(3) 大规模 Agent 任务合成流水线：开发了一种全新的合成管道，生成了超过 1800 个环境和 8.5 万个复杂提示词，将推理能力深度整合到工具使用场景中。DeepSeek-V3.2 的推出显著缩小了开源模型与顶尖闭源专有模型在复杂推理和 Agent 任务上的差距，并已开源相关实现，为开发低成本、高通用性的 AI 智能体奠定了重要基础。

相关链接

技术博客：https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.2/resolve/master/assets/paper.pdf

DeepSeek-V3.2：

Hugging Face： https://huggingface.co/deepseek-ai/DeepSeek-V3.2

ModelScope： https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.2

DeepSeek-V3.2-Speciale：

HuggingFace： https://huggingface.co/deepseek-ai/DeepSeek-V3.2-Speciale

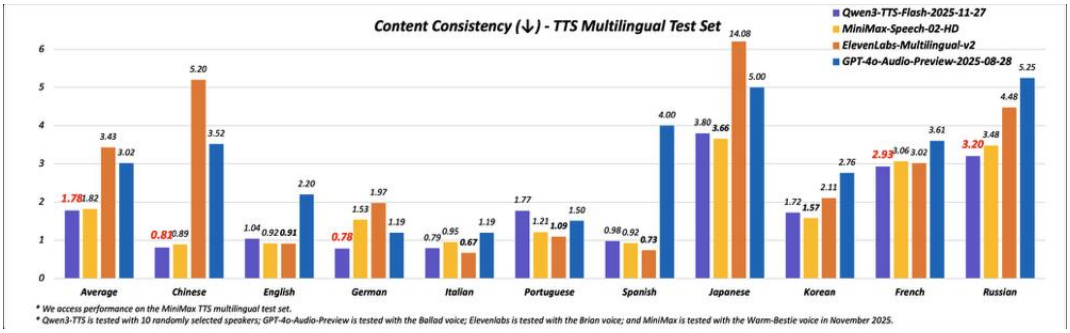
ModelScope： https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.2-Speciale

• 技术动态二

通义千问 Qwen3-TTS 发布：零样本语音克隆与全能语音设计引擎

日期

2025-12-05



内容摘要

阿里云通义千问（Qwen）团队正式发布新一代语音生成大模型 Qwen3-TTS，该模型在多语种合成与情感表达上取得了突破性进展。Qwen3-TTS 的核心亮点在于其强大的 Voice Design（语音设计） 与 Voice Cloning（语音克隆） 能力：用户既可以通过自然语言提示词（如“一个略带紧张、语速稍快的年轻男性”）精准定制语音风格，也能仅凭 3 秒的参考音频实现高保真的零样本声音克隆。该模型内置了 49 种具有鲜明个性的角色声线（覆盖旁白、游戏 NPC、甚至特定情感状态），并原生支持包括中、英、德、法在内的 10 种主流语言以及四川话、粤语等 10 种中国方言。在性能评测方面，Qwen3-TTS 在 MiniMax TTS 多语言测试集上的单词错误率（WER）显著优于 MiniMax 和 ElevenLabs，其生成的语音自然度（MOS）高达 4.6 分，极度接近真人水平（4.8 分）。此外，模型支持首包延迟低于 300ms 的实时流式输出，并已通过 API 向开发者开放，旨在为有声读物、智能客服及实时交互应用提供低成本、高表现力的语音解决方案。

相关链接

技术博客：https://qwen.ai/blog?id=qwen3-tts-1128

Hugging Face：http://hf.co/spaces/Qwen/Qwen3-TTS-Demo

ModelScope：http://modelscope.cn/studios/Qwen/Qwen3-TTS-Demo

技术动态三

OpenRouter 发布 2025 年 AI 现状报告：推理模型与 Agent 主导的百万亿 Token 洞察

日期

2025-12-06


State of AI

An Empirical 100 Trillion Token Study with OpenRouter

Malika Aubakirova * Alex Atallah † Chris Clark † Justin Summerville † Anjney Midha *

* a16z (Andreessen Horowitz) • † OpenRouter Inc.

* Lead contributors. Please see Contributions section for details.

December, 2025

内容摘要

OpenRouter 联合 a16z 正式发布《2025 年 AI 现状报告》，基于平台积累的 100 万亿 Token 真实交互数据，深入剖析了全球大语言模型（LLM）产业的格局演变。报告核心发现指出，AI 行业正经历从“单次文本生成”向“多步推理与 Agent 智能体（Agentic Inference）”的根本性范式转变，推理优化模型（如 o1、DeepSeek-R1 等）的 Token 使用占比已从 2025 年初的微不足道飙升至年底的 50%以上，成为处理复杂任务的默认选择。在模型生态方面，开源模型（OSS）已稳固占据约三分之一的市场份额，且呈现出显著的“多元化”与“去中心化”趋势：市场不再由单一厂商垄断，而是形成了包括 DeepSeek、Qwen、Moonshot (Kimi)、Minimax 及 GPT-OSS 系列在内的多强并存格局。报告还提出了“中型模型即新小型模型”（Medium is the New Small）的论断，指出 15B-70B 参数模型正因其在性能与效率上的最佳平衡，逐渐取代小模型成为主流。在应用场景上，编程（Programming）已超越通用问答，成为全平台增长最快且占比最高的（>50%）专业用例，而角色扮演 (Roleplay) 则继续主导开源模型的流量。此外，报告通过“水晶鞋效应”（Glass Slipper Effect）

揭示了用户留存规律，指出早期在特定模型上成功跑通工作流的用户群体表现出极高的粘性，证明了“模型-任务匹配度”是构建竞争壁垒的关键。

相关链接

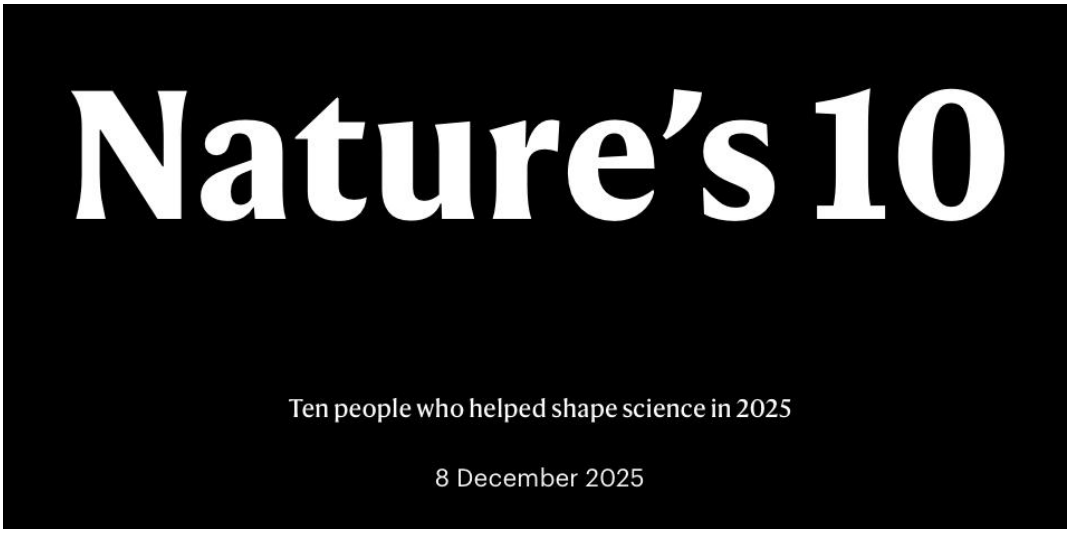
技术博客：https://openrouter.ai/state-of-ai

• 技术动态四

《Nature》评选 2025 年度十大人物：DeepSeek 创始人梁文峰入选

日期

2025-12-08



内容摘要

顶级科学期刊《Nature》正式发布 2025 年度十大人物榜单，DeepSeek（深度求索）创始人梁文峰因其

在人工智能领域的革命性影响被评为“技术颠覆者”。《Nature》高度评价了梁文峰及其团队在这一年推出的 DeepSeek-R1 与 V3 模型，指出其凭借极高的计算效率和开源策略，在性能上成功匹敌 OpenAI 等顶尖闭源模型，同时大幅降低了训练成本。这一成就不仅打破了美国科技巨头在 AI 领域的绝对垄断，引发了全球科技股（如 Nvidia）的价值重估，更实质性地推动了 AGI 技术的民主化与普及。同期入选的还有另一位中国科学家——来自中国科学院深海科学与工程研究所的杜梦然，她因在千岛-堪察加海沟发现地球已知最深处的动物生态系统而获誉“深海潜行者”。榜单其他入选者涵盖了公共卫生、天文学及生物医学等领域的关键人物，包括：在政治压力下坚守疫苗科学立场的前美国 CDC 主任 Susan Monarez、推动全球首个大流行病条约签署的 Precious Matsoso、Vera Rubin 天文台的奠基人 Tony Tyson，以及首位接受个性化 CRISPR 基因编辑疗法的婴儿 KJ Muldoon 等。这份榜单深刻折射了 2025 年全球科学界在技术突破、科研伦理及地缘政治博弈中的演变图景。

相关链接

技术博客：https://www.nature.com/immersive/d41586-025-03848-1/index.html

技术动态五

OpenAI 发布 GPT Image 1.5：四倍速生成与精准修图，ChatGPT 新增独立图像创作中心

日期

2025-12-16

内容摘要

OpenAI 正式发布新一代旗舰图像生成模型 GPT Image 1.5，并同步推出了深度集成的 ChatGPT Images 创作体验。此次更新主要解决了 AI 绘图中“修改难”的痛点，引入了强大的“意图匹配”编辑功能。用户现在可以上传图片或针对已生成的图像发出自然语言指令，模型能够在完美保留原图光影、构图、人物面部特征及背景细节的前提下，进行高精度的局部调整（如更改服装材质、添加特定物体或调整环境氛围）。在性能指标上，GPT Image 1.5 的生成速度较前代提升了 4 倍，并显著增强了对图像中文本渲染的准确性。交互方面，

ChatGPT 侧边栏新增了独立的“Images”入口，提供丰富的预设艺术风格、滤镜及灵感提示词库，使用户无需精通提示工程即可创作专业级图像。对于开发者，OpenAI 已开放 gpt-image-1.5 API，不仅性能更强，其调用成本也相比上一代模型降低了 20%，旨在将 ChatGPT 打造为兼具效率与精度的全能创意工作室。

相关链接

技术博客：https://openai.com/zh-Hans-CN/index/new-chatgpt-images-is-here/

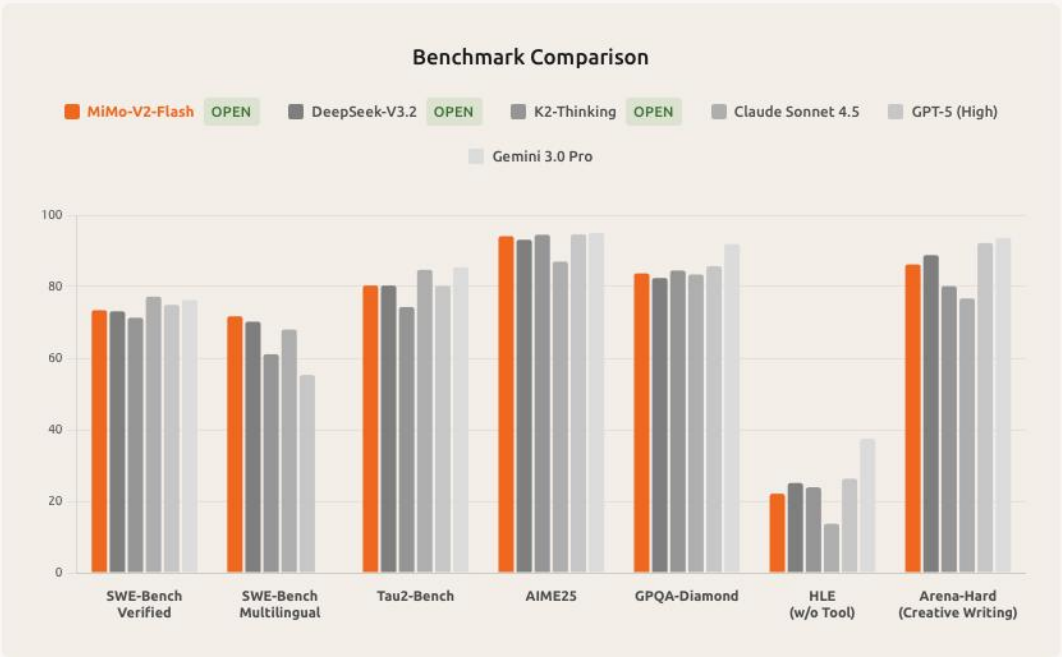
Link： https://chatgpt.com/images

• 技术动态六

小米发布 MiMo-V2-Flash：开源 MoE 性能新标杆，编程与长窗口推理双优

日期

2025-12-16



内容摘要

小米正式推出并开源 MiMo-V2-Flash，这是一款兼顾高计算效率与卓越智能体（Agent）性能的 MoE 架构大模型（总参数 309B/激活 15B）。该模型凭借 5:1 混合注意力机制 与 256k 超长上下文，在多项基准测试中表现亮眼：在软件工程榜单 SWE-bench Verified 中登顶开源第一，在 AIME 2025 数学竞赛与 GPQA 科学评测中稳居开源前二，综合实力对标 DeepSeek-V3.2 与 Kimi-k2。核心技术突破包括：(1) Multi-Token Prediction (MTP)：引入原生多 Token 预测训练与并行解码技术，实现 2.0-2.6 倍的推理加速；(2) MOPD 后训练范式：利用“多教师在线策略蒸馏”大幅降低强化学习算力成本。MiMo-V2-Flash 深度优化了编程与工具调用能力，支持“思考模式”切换及一键生成交互式前端代码，完美适配 Vibe-coding 工作流。目前，模型权重已通过 Hugging Face 全球开源（MIT 协议），并适配 SGLang 推理框架。

相关链接

技术博客：https://mimo.xiaomi.com/mimo-v2-flash

Hugging Face： https://huggingface.co/XiaomiMiMo/MiMo-V2-Flash

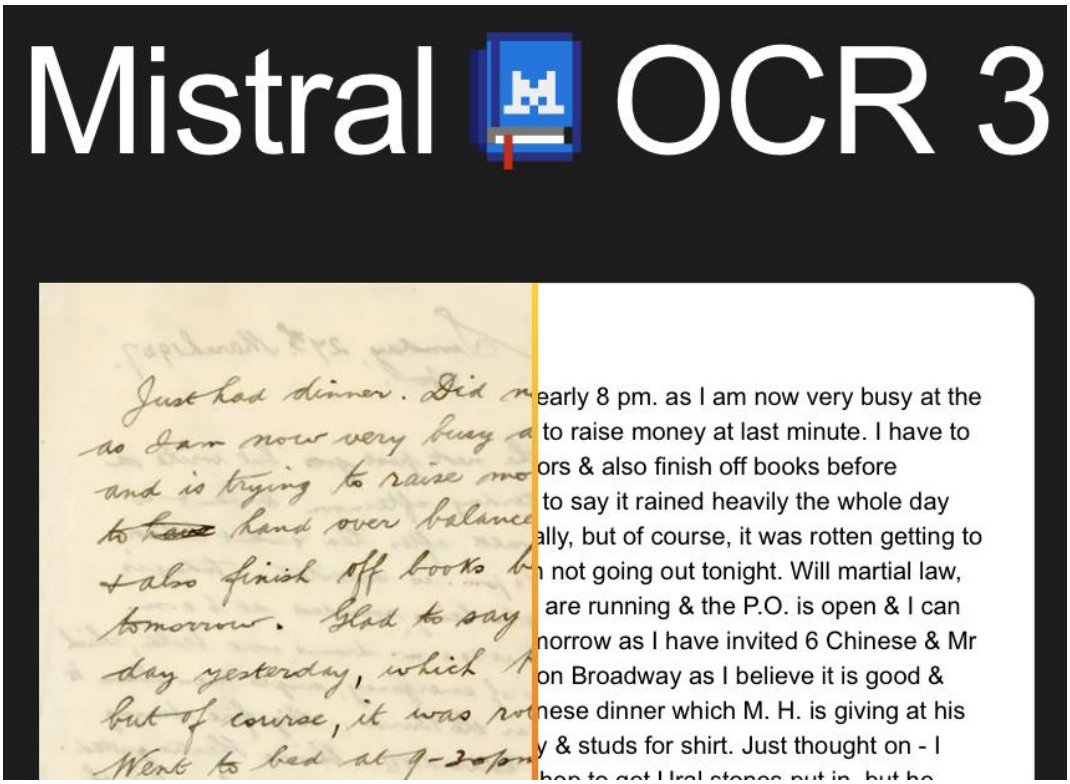
Link： https://platform.xiaomimimo.com/#/docs/welcome

● 技术动态七

Mistral AI 发布 Mistral OCR 3：全能文档解析引擎，重塑多模态 RAG 数据基座

日期

2025-12-17



内容摘要

Mistral AI 宣布推出第三代文档解析模型 Mistral OCR 3，旨在解决企业级 AI 应用中“非结构化数据处理”这一核心痛点。Mistral OCR 3 不再仅仅是一个字符识别工具，而是一个端到端的视觉-语义转换引擎。该模型能够以接近人类的理解力处理复杂的 PDF 文档、扫描件及手写笔记，完美还原多栏排版、嵌套表格、数学公式及内嵌图表，并将其直接输出为结构化极强的 Markdown 或 JSON 格式。技术亮点方面，Mistral OCR 3 引入了“全局文档注意力机制”，使其能够跨页理解上下文逻辑（例如跨页表格的合并或脚注的语义关联），从而显著提升了长文档检索增强生成的准确率。在基准测试中，Mistral OCR 3 在 DocVQA 和 ChartQA 数据集上的表现超越了 GPT-5 Vision 及传统的 AWS Textract 服务，同时处理速度提升了 4 倍，API 调用成本却降低了 60%。Mistral AI 表示，该工具已集成至 Le Platform，将成为构建下一代高精度知识库和智能档案系统的标准基础设施。

相关链接

技术博客：https://mistral.ai/news/mistral-ocr-3

Demo：https://console.mistral.ai/build/document-ai/ocr-playground

● 技术动态八

字节跳动发布 Seed-1.8：通用全模态 Agent 模型，原生视觉感知重构人机交互

日期

2025-12-18

内容摘要

字节跳 Seed 团队正式发布 Seed-1.8，这是一款专为“通用现实世界代理”设计的全模态大模型。与传统的视觉语言模型不同，Seed-1.8 不再局限于被动的信息理解，而是将感知、推理与行动深度集成于单一架构中。该模型具备原生视觉感知能力，能够直接解读 GUI 界面（如软件截图、文档、图表及视频流），并在无 API 接口的情况下像人类一样操作软件环境。技术特性方面，Seed-1.8 引入了可配置的“思考模式”，允许用户在推理深度与响应延迟之间灵活平衡，并内置了统一的 Agent 接口，支持复杂的联网搜索、代码生成及即时执行。在 BrowseComp-en 等权威基准测试中，Seed-1.8 取得了 67.6 的高分，超越了当前市场上的顶尖

模型。该模型的发布标志着多模态 AI 从“读图对话”向“自主执行”迈出了关键一步，为自动化软件工程、复杂工作流编排及下一代人机交互界面提供了强大的底层驱动力。

相关链接

技术博客：https://github.com/ByteDance-Seed/Seed-1.8/blob/main/Seed-1.8-Modelcard.pdf

Github： https://github.com/ByteDance-Seed/Seed-1.8/tree/main

● 技术动态九

快手发布 Kling-Omni 技术报告：统一视频生成与物理推理的端到端世界模拟器

日期

2025-12-18

内容摘要

快手可灵团队（Kling Team）正式发布 Kling-Omni 技术报告，推出一款集视频生成、编辑与智能推理于一体的端到端通用框架。该模型摒弃了传统碎片化的流水线方案，创新性地引入 MVL 多模态视觉语言 交互范式，将文本指令、参考图像及视频上下文统一编码，从而赋予模型对物理规律和复杂用户意图的深度理解力。Kling-Omni 不仅在生成质量上超越了 Veo 与 Runway 等竞品，更通过包含指令预训练、监督微调及 DPO 强

化学习在内的多阶段训练策略，实现了从简单的像素合成向逻辑推理的质变。此外，凭借高效的两阶段蒸馏技术，模型成功将推理步数从 150 步大幅压缩至 10 步，为构建能实时交互的多模态世界模拟器奠定了算力基础。

相关链接

技术博客：https://arxiv.org/abs/2512.16776

Hugging Face：https://huggingface.co/papers/2512.16776

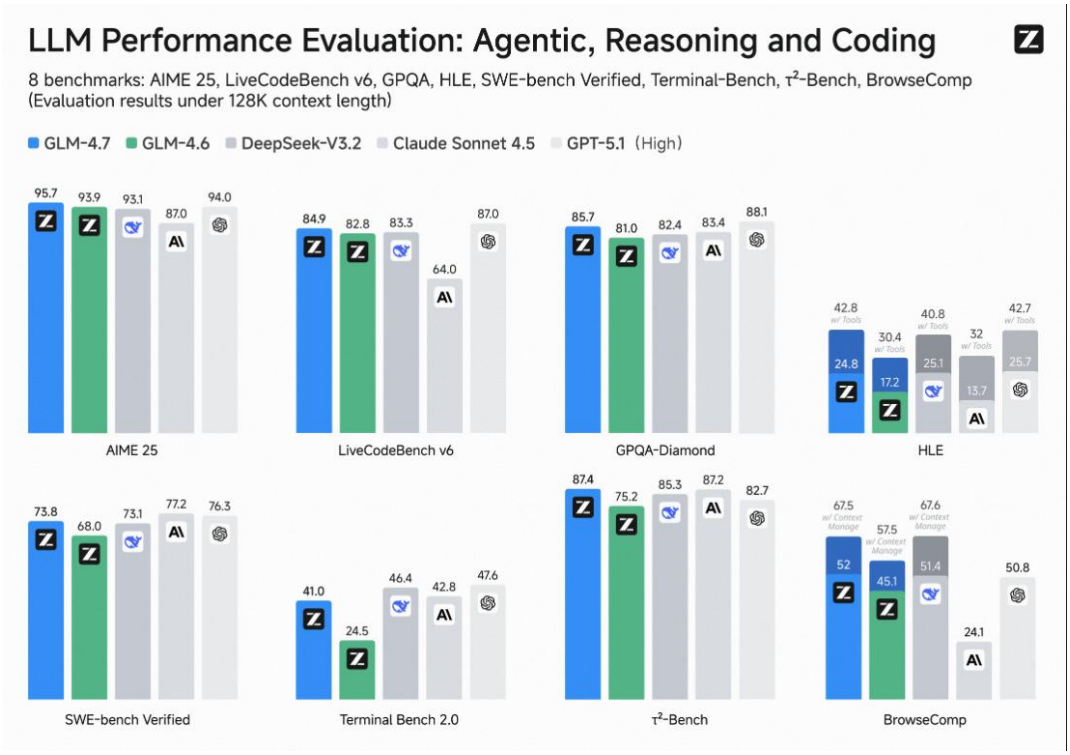
ModelScope：https://modelscope.cn/models/deepseek-ai/DeepSeek-V3.2

● 技术动态十

智谱 AI 发布 GLM-4.7：开源最强编程与思维模型，定义 Agent 交互新范式

日期

2025-12-22



内容摘要

智谱 AI 正式发布年度旗舰模型 GLM-4.7，这是一款专为复杂编程与 Agent 智能体任务设计的“思考型”开源大模型。GLM-4.7 的核心突破在于其独创的 “混合思维模式”（Interleaved & Preserved Thinking），该机制允许模型在执行工具调用或编写代码前进行显式的多步推理，并能在多轮对话中完整保留思考链路，从而有效解决了长程任务中的逻辑断层问题。在性能表现上，GLM-4.7 刷新了开源模型的多项纪录：它在 SWE-bench Verified 软件工程榜单上取得了 73.8% 的高分，超越了 DeepSeek-V3.2 及 MiMo-V2-Flash，并在高难度的 HLE（Human Last Exam）基准中达到了 42.8% 的准确率，击败了闭源模型 GPT-5.1。此外，模型特别针对 UI/UX 生成 进行了视觉对齐训练，能够直接生成极具美感的现代网页与演示文稿代码，完美契合当下流行的“Vibe Coding”工作流。GLM-4.7 支持 200k 上下文与 128k 超长输出，并已在 Hugging Face 开放权重，成为当前构建自主编程 Agent 与复杂工作流的首选基座。

相关链接

技术博客：https://z.ai/blog/glm-4.7

Hugging Face：https://huggingface.co/zai-org/GLM-4.7

Github：https://github.com/zai-org/GLM-4.5

RESOURCE SHARING
资源分享

● 资源分享一

Stanford Agentic Reviewer

类型：网站

内容提要

这是一款基于代理架构的革命性论文预审工具。它通过检索 arXiv 上的海量文献进行知识增强，彻底解决了通用大模型“幻觉”严重的问题，使得审稿反馈具备极高的学术专业性。该工具不仅能从原创性、实验可靠性等七个维度对论文进行 1-10 分的深度诊断，更能提供如“补充特定基准测试”等极具操作性的修改建议。研究者只需上传 PDF 即可在短时间内获得详尽的评审报告，这极大地缩短了学术迭代周期，帮助作者在正式提交会议前针对性地弥补学术漏洞，是将科研草稿打磨为顶会精品的必备利器。

资源链接

https://paperreview.ai

● 资源分享二

25 天 AI Agent 课程

类型：网站

内容提要

谷歌最新推出的 Advent of Agents 是一场为期 25 天的“硬核” AI Agent 实战课程，带开发者从零跨越到企业级生产部署。虽然披着圣诞季趣味外壳，其内容却极具深度。课程依托 Gemini 系列模型与 Google ADK 套件，涵盖了从基础的函数调用、多 Agent 协作，到前沿的 A2A 跨生态协议等全栈技术。最令人惊艳的是其对实时双向流技术的拆解，使 AI 具备类似真人的即时打断与响应能力，解决了复杂场景下的交互痛点。通过每日一个主题的系统路径，学员能掌握状态管理、生产级监测等高阶实战技能。

资源链接

https://adventofagents.com/

• 资源分享三

AI 原生应用架构白皮书

类型：书籍

内容提要

这是 阿里云开发者社区 发布的一本系统性技术白皮书，由 15 位行业专家联名推荐、40 多位一线工程师撰写、全书超 20 万字，首次从 AI 原生应用的 DevOps 全生命周期 角度全面拆解了构建智能化应用的架构和实践难点。白皮书围绕大模型、提示词、RAG、记忆、工具与网关、运行时、可观测性、评估和安全等 11 个关键要素 给出清晰定义和解题思路，结合业界最新趋势与工程实践，为企业和开发者从设计到落地部署 AI 原生应用提供了完整的参考路线图。

资源链接

https://developer.aliyun.com/ebook/8479

• 资源分享四

State of Agent Engineering

类型：网站

内容提要

由 LangChain 发布的 State of Agent Engineering 系统性梳理了当前 AI Agent 工程化的整体图景与实践共识，从 Agent 架构设计、状态管理、工具调用、记忆机制到多 Agent 协作与评测方法，深入总结了真实落地过程中面临的关键挑战与可行解法。该报告并非停留在概念层面，而是基于大量一线经验，提炼出“哪些设计真的有效、哪些常见做法会失败”，为构建可扩展、可维护、可观测的 Agent 系统提供了清晰的工程路线图，对从研究走向生产级 Agent 应用具有很强的指导意义。

资源链接

https://www.langchain.com/state-of-agent-engineering

人工智能前沿快报

Artificial Intelligence Newslettler

2025 年第 12 期

总第 30 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 12 期
总第 30 期
2026.01.04

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、技术咨询、科学普及