

顶会观察

ICLR 2026

南京大学助理教授 傅朝友

国际表征学习会议（International Conference on Learning Representations, ICLR）作为全球人工智能领域的顶级盛会，备受学术界和产业界的关注。第 14 届 ICLR 大会于 2026 年 4 月底在巴西里约热内卢成功举办。本届会议在经历了激增的投稿量、严苛的 LLM 审查机制以及开放评审安全突发事件考验后，依然保持了学术前瞻性与严谨性。本届会议不仅组织了涵盖核心机器学习与交叉学科的特邀报告，还开展了 40 场聚焦智能体、AI for Science、基础模型对齐等热门议题的专题研讨会，并开展 35 场 Oral 报告，系统性地展现了该领域的最新研究进展。

一、会议亮点

评审机制革新：规范 LLM 使用与“幻觉文献”零容忍。面对 LLM 在论文写作与审稿中的全面渗透，本届会议出台了严格的管控与检测政策。组委会在审稿端部署了双重 LLM 内容检测器，以辅助 AC 甄别低质量的 AI 自动生成审稿意见，同时在投稿端重点排查大模型生成的幻觉参考文献。通过自动化提取比对与多级人工核查（每篇涉事论文至少由 3 名人类复核）相结合，所有被证实编造虚假文献的论文均被作 Desk Reject 处理，切实维护了学术严谨性。

突发危机响应：采取分数重置与人员重分配机制保障评审公正。在本届会议的 Rebuttal 阶段，遭遇了 OpenReview API 恶意抓取事件，导致部分作者、审稿人与 AC 身份泄露，并引发了学术骚扰与串通打分威胁。面对该违规事件，组委会紧急采取了冻结讨论、将评分重置回 Rebuttal 前状态、并为受波及论文重新分配 AC

等应对措施。此外，官方还适当延长了 Meta-Review 周期，最终有效保障了整体录用标准。

学术生态拓展：“开放科学”与 Workshop 赛道规模显著增长。本届大会注重开放科学与跨学科交流，Workshop 赛道提案数量创下新高，共收到 151 份提案（较去年增长约 24%），最终精选出 40 场。议题呈现出明显的跨学科前沿导向，诸如智能体系统、AI for Science（涵盖偏微分方程、基因组学、材料科学）、基础模型 Test-Time Updates 等成为主要探讨方向，进一步拓宽了表征学习的研究边界。

二、录用情况

投稿与录用：大会共收到了 19,525 篇有效投稿，相较于 2025 年的 11,672 篇出现了爆发式增长。经过多轮严格评审，最终接收 5,355 篇论文，录用率约为 27.4%，整体延续了 ICLR 近年来将录用率稳定在 30% 上下的趋势。在录用的论文中，有 223 篇被选为 Oral，占比仅为 1.13%。其余 5,120 篇录用论文以 Poster 形式展示。此外，受本届严格的合规性与 LLM 使用审查影响，Desk Reject (779 篇) 与 Withdraw (5,042 篇) 的数量也创下新高。

评审团队：为了应对海量且激增的投稿，大会组织了由 18,054 名审稿人构成的评审团，共计提供了逾 7.6 万份评审意见。在遭遇突发事件的压力下，大会紧急优化了 AC 的接管与校准流程，通过延长 Meta-Review 周期，在维持一贯录用标准的前提下，确保了评审质量。

区域与机构表现：本届 ICLR 的论文区域分布呈

现出显著的结构变化。据第三方开源项目 TradingView 提取的论文机构署名数据显示,在接收的论文中,中国大陆机构参与的论文占比为 43.7%,超过了美国的 31.9%。高校方面,清华大学以 332 篇录用论文位列全球第一;上海交大 (240 篇)、浙江大学 (232 篇) 紧随其后。企业及新型机构方面,阿里巴巴和上海 AI Lab 以 135 篇录用位居国内企业前列、华为 (111 篇)、字节跳动 (107 篇) 等机构亦有显著产出。不过值得注意的是,在代表更高学术评价的 Oral 论文中,美国机构占比约为 40% (中国约 30%), 表明其在顶尖研究上仍保持一定的引领优势。

三、 热门研究方向

根据主题演讲、Workshop 议题及录用论文分布, ICLR 2026 呈现出以下核心技术趋势:

智能体系统与自主可靠性全面深化: 纯文本大模型的单点研究趋于收敛, 智能体在复杂环境中的应用成为重点探讨方向。本届会议设置了多个相关的专题 Workshop, 重点涵盖大模型记忆机制 (Memory)、终身学习智能体 (Lifelong Agents) 以及真实世界可靠自主性 (Agentic AI in the Wild) 等议题, 探讨模型在动态环境下的规划与执行能力。

AI for Science 领域的跨学科融合: AI 与自然科学的结合正向底层机制探索演进。本届大会的特邀报告与多个 Workshop 集中探讨了这一交叉领域。具体方向包括偏微分方程求解 (AI&PDE)、加速材料设计 (AI4MAT)、生成式基因组学 (Gen²) 等。研究重点涉及如何将物理先验知识、空间对称性及热力学原理整合至表征学习模型中, 以辅助科学问题的计算与发现。

开放模型开发与推理时计算 (Test-Time Compute): 针对当前前沿基础模型训练细节不够透明的现状, 大会特邀报告探讨了借助 Marin 等开源平台, 实现从数据配比、优化器设计到 Scaling Laws 的公开实验与开发。此外, 在推理架构层面, 相关 Workshop 集中探讨了超越传统 CoT 的 “Latent & Implicit Thinking” 以及 “Test-Time Updates” 等技术, 旨在研究模型在推理阶段动态分配计算资源的有效机制。

四、 热点论文

本届 ICLR 官方评审委员会最终在浩如烟海的录用论文中, 评选出 **2 篇杰出论文** (Outstanding Papers) 和 **1 篇荣誉提名** (Honorable Mention):

(Outstanding Paper) Transformers are Inherently Succinct^[1]: 随着 LLM 的广泛应用, 理解 Transformer 架构的理论表达能力成为核心议题。该论文从形式语言理论的角度, 提出 “简洁性 (Succinctness)” 作为衡量 Transformer 表达能力的新标准。研究在理论上证明了: 相较于有限状态自动机 (Finite Automata)、线性时序逻辑 (LTL) 以及传统 RNN (含近期热门的 SSMs), 固定精度的 Transformer 能够以 “指数级” 甚至 “双指数级” 的简洁性来编码复杂的概念模式。这项纯理论工作不仅为 “Transformer 架构为何如此强大” 提供了深刻的数学解答, 同时也更加严谨证明了对 Transformer 内部属性进行形式化验证是一个困难 (EXPSPACE-Complete) 的计算问题。

(Outstanding Paper) LLMs Get Lost In Multi-Turn Conversation^[2]: 这篇研究揭示了当今 LLM 在真实交互场景中的关键局限。现有的 LLM 评测多集中于 “单轮、指令明确” 的理想环境, 而现实中用户与模型的交互往往是多轮且指令不明。该研究团队设计了一种可扩展的多轮对话评估方法, 对 15 款主流大模型进行了逾 20 万次模拟测试。结果表明: 当面临多轮、信息不全的交互时, 无论是开源小模型还是顶尖闭源模型 (如 GPT-4o, Gemini 2.5 Pro), 其整体性能均出现显著下滑 (平均降幅达 39%), 且输出可靠性大幅降低。论文指出, 这种 “在对话中迷失” 的根源在于模型倾向于在早期做出错误假设, 并过度依赖先前的错误尝试而无法完成自我纠错。这项工作为未来评测基准的演进提供了新的方向。

(Honorable Mention) The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm^[3]: 随着深度学习规模的扩大, 高效的底层优化器至关重要。这篇研究聚焦近期在训练大模型时具备竞争力的 Muon 优化

器，提出了一种名为“Polar Express”的全新矩阵极分解算法。该方法利用近似理论设计最佳多项式，适配 GPU 的高吞吐量需求并规避了低精度计算陷阱。实验证明，将其集成至 Muon 优化器后，能显著且稳定地降低 GPT-2 等语言模型的验证损失。

五、 时间检验奖

时间检验奖 (Test of Time Award) 是 ICLR 的最高荣誉之一，旨在回溯并表彰 10 年前 (即 2016 年) 发表，且经长期实践证明对人工智能领域产生深远、基石性影响的经典工作。本年度获得该殊荣的是两篇在生成式人工智能与深度强化学习领域具有里程碑意义的奠基性工作：

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks^[4]。在 GAN 被提出的初期，其训练很不稳定且难以生成高分辨率图像。在这篇论文中，研究团队首次成功将深度卷积神经网络与 GAN 结合，提出了一系列网络架构设计规范 (如使用步长卷积代替池化、引入 Batch Normalization、取消全连接层等)，有效缓解了早期 GAN 的训练崩溃难题。DCGAN 不仅率先实现了对复杂自然图像的稳定合成，其在隐空间中展示的“向量算术”特性也揭示了模型深层的语义组合能力，为现代计算机视觉中生成模型 (及后续演进架构) 的底层设计确立了范式。

Continuous Control with Deep Reinforcement Learning^[5]。该工作大幅拓展了深度强化学习在物理控制领域的应用边界。在此之前，经典的 DQN 算法主要处理离散动作空间，面对物理机器人等“连续且高维”的动作空间时存在显著局限。为了解决“维度灾难”与动作离散化带来的精度折损，该论文创新性地将确定性策略梯度 (DPG) 与 Actor-Critic 架构、深度神经网络及经验回放机制相融合，提出了 DDPG 算法。该算法率先实现了端到端 (直接从高维像素输入到连续动作输出) 的运动控制，极大推动了深度强化学习在复杂机器人与连续控制任务中的广泛应用。

六、 总结展望

从 ICLR 2026 创纪录的近两万篇投稿，以及仅 1.13% 的 Oral 录用率可以看出，全球深度学习领域的研究热度仍在持续攀升，而顶级学术成果的筛选标准依然保持着极高门槛。在这一背景下，中国大陆研究机构的论文录用占比首次在顶级人工智能会议中超过美国 (达到 43.7%)，不仅反映出中国人工智能研究生态的快速成熟，也体现了其在基础研究领域长期积累后的集中释放。

纵观本届会议的研究图景，一个值得关注的变化是：学术界的关注重点正逐步从单纯追求模型规模扩张 (Scaling Laws)，转向对模型底层机制、系统可靠性以及高效优化方法的深入探索。研究者开始更加关注大模型在真实复杂环境中的行为表现与能力边界。一方面，越来越多工作系统性揭示并尝试解决大模型在复杂多轮交互、长程推理以及信息不完备场景下出现的脆弱性与性能退化问题；另一方面，研究社区也在从更加严谨的数学视角重新审视神经网络的表达能力与理论边界，并探索在算力约束条件下，通过最佳多项式近似等数学工具实现训练与推理效率提升的新路径。

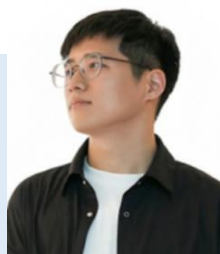
这一趋势表明，人工智能研究正在逐渐告别以静态基准测试为核心的能力评估范式，转而更加关注模型在开放环境中的实际表现，以及在有限算力、高可靠性和可持续部署条件下的综合能力。如何构建真正稳定、可信且具备长期泛化能力的智能系统，正成为学界与产业界共同关注的核心议题。

展望未来，随着多模态原生架构、推理时计算以及 AI for Science 等方向的持续融合，人工智能领域有望迎来新的技术突破。如何通过系统架构创新赋予模型动态规划、反思与自我纠错能力，如何将物理规律、领域知识与先验约束有效融入神经网络的学习过程，从而提升模型的推理效率、可靠性与泛化能力，都有望成为推动下一代人工智能向更高水平通用智能演进的重要研究命题。

责任编辑 魏秀参

参考文献

- [1] Pascal Bergstraßer, et al. Transformers are Inherently Succinct. ICLR 2026.
- [2] Philippe Laban, et al. LLMs Get Lost In Multi-Turn Conversation. ICLR 2026.
- [3] Noah Amsel, et al. The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm. ICLR 2026.
- [4] Alec Radford, et al. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. ICLR 2016.
- [5] Timothy P. Lillicrap, et al. Continuous Control with Deep Reinforcement Learning. ICLR 2016.



傅朝友

南京大学智能科学与技术学院研究员、助理教授、博导，入选中国科协“青年人才托举工程”。2022 年博士毕业于中科院自动化所模式识别实验室。主要研究方向为多模态内容分析，谷歌学术引用 9800 余次，开源项目累计获 2 万余次 GitHub Stars。担任 Pattern Recognition/IEEE T-BIOM 期刊编委、ICLR/ICML 会议领域主席。

Email: cyfu@nju.edu.cn