顶会观察

ICLR 2025

中国科学院计算技术研究所 徐逸峰 王中琦 何振梁

国际表征学习大会（International Conference on Learning Representations，ICLR）是机器学习领域的顶级会议之一，每年举办一次。ICLR 2025于4月24日至4月28日在新加坡举办，这也是该会议首次在亚洲举办。

一、 会议概况

ICLR 2025 共收到 11,500 份投稿，最终接收论文 3,705 篇，整体录用率为 32%。其中，213 篇论文被评为 Oral，380 篇被评为 Spotlight。本届会议的论文主题高度集中于大语言模型（LLMs）。以论文的第一个关键词进行统计，在排名前十的关键词中，有一半与大模型相关，包括“Large Language Models”及其各种大小写、缩写变体。此外，强化学习主题在前十中占据两个位置。前十关键词中还包括联邦学习、图神经网络以及扩散模型等热门方向。ICLR 2025 共组织了 40 个 Workshops，涵盖人工智能多个前沿领域。主会议的所有入选论文，包括获得口头报告机会的论文，均以墙报的形式展示，以便与会者深入交流讨论。

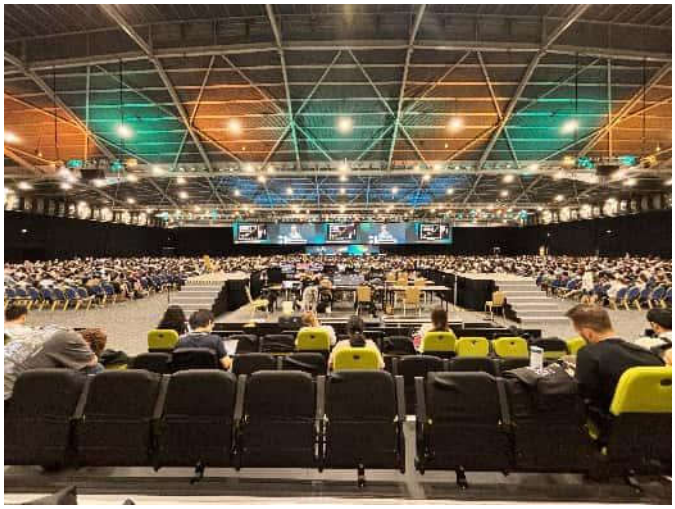

图 1　ICLR 2025 主会场一隅

二、 获奖论文

本次会议共选出了 3 篇杰出论文（Outstanding Papers）[1][2][3]、3 篇荣誉提名（Honorable Mentions）[4][5][6]、1 篇时间检验奖（Test of Time Awards）[7]和 1 篇时间检验候选（Runner Up）[8]。下面简要介绍 3 篇杰出论文和 1 篇时间检验奖。

2.1 杰出论文 1: Safety Alignment Should be Made More Than Just a Few Tokens Deep[1]

此文作者来自于普林斯顿大学和谷歌，主要内容为研究大语言模型（LLM）的对齐缺陷发现与修复。近年来，大语言模型的安全性主要依赖于安全对齐方法，涉及有监督微调(SFT)与基于偏好的优化方法(如 RLHF)等。然而，最近的研究表明这类对齐方法的脆弱性。对齐模型在对抗性输入或有限步数微调后仍会输出不安全内容。在本文中，作者深入分析了造成当前对齐脆弱性的原因，发现当前安全对齐仅作用于回答中的前几个 token，这种“安全对齐捷径”被作者称为大语言模型的“浅层安全对齐（Shallow Safety Alignment）”。具体的，作者计算了对齐模型和未对齐模型中，不同 token 长度处二者的特征 KL 散度，作者发现 KL 散度在前几个 token 中显著高于其他 token，证明安全对齐主要作用于前几个 token（如拒绝前缀“I cannot ...”）。此外，作者还通过实验证明了这种浅层对齐是许多脆弱性的来源：在推理阶段中，对齐模型可能会遭受前缀攻击、基于优化的后缀攻击和纯随机采样的越狱攻击。作者表明如果模型仅通过引导一个简短的“拒绝响应”的前缀 token 来屏蔽有害内容模型，攻击者可以轻易替换或优化肯定性前缀（如“Sure，here is...”）完成

越狱攻击。在下游微调阶段中，作者也通过实验证明，基于微调的攻击主要扰动生成分布的前几个 token。为了实现“深层安全对齐（Deep Safety Alignment）”，作者设计了一种数据增强方法，其核心思想是训练模型在“开始被恶意诱导”的情况下依旧能做出拒绝回应。实验表明，数据增强的方法使模型得到了更深的对齐结果，并且原始性能也得到保留。此外，作者为了增强模型抗微调攻击，提出了一种约束性微调优化目标，通过限制初始 token 的分布变化，使安全对齐在下游微调时保持得更为持久。总结而言，该论文系统性验证了 LLM 对齐中“浅层安全对齐”现象，从根本上分析了 LLM 安全对齐的薄弱之处，并提出加强对齐深度的策略，对防御大模型越狱具有重要意义。

2.2 杰出论文 2: Learning Dynamics of LLM Finetuning[2]

此文作者来自不列颠哥伦比亚大学，主要内容为研究微调大语言模型（LLM）时模型的学习动态。近年来，大语言模型的强大表现使其成为自然语言处理研究的焦点。为了使这些模型更好地遵循人类指令、符合人类偏好，微调成为关键步骤，通常包括两个阶段：指令微调和偏好微调。尽管已有大量研究尝试从训练目标、最终模型表现或与强化学习的关系角度来理解微调机制，但此论文提出了一种全新的视角——从学习动态出发，对 LLM 微调过程进行深入解析。学习动态描述的是：在通过梯度下降更新模型参数的过程中，训练样本如何影响模型对其他样本的预测。这种动态视角不仅揭示了训练过程中一些有趣现象，如“Z 字形”学习路径和组合性概念空间的形成，也为设计更高效的训练算法提供了理论基础。在此文中，作者通过将模型预测变化分解为三个不同的组成部分，建立了一个通用的学习动态框架。该框架可适用于多种主流微调算法，包括：有监督微调、直接偏好优化、强化学习方法（如 PPO）。这一框架不仅统一了对不同微调算法训练过程的理解，还解释了若干在实践中观测到的、看似违反直觉的现象。例如：复读机现象，即偏好微调后，模型倾向于重复简单、模板化的短语；幻觉增强，即模型可能在回答一个问题时引用另一个问题的错误信息或无关事实；输出置信度下降，即在离线策略 DPO 训练过程中，几乎所有候选回答的置信度都降低。为了进一步解释这些现象，作者提出了一个关键机制——压缩效应。该效应源于交叉熵损失函数与 softmax 层结合下的梯度上升。具体来说，当模型对一个不太可能的标签进行负向梯度更新时，这种更新会压缩其对所有标签的预测概率，仅将概率推向少数最可能的标签上。这种压缩可能导致原本合理的输出也变得置信度较低，从而损害模型与人类偏好的一致性。该发现也解释了为何在线策略 DPO 通常优于离线策略方法。作者通过深入的梯度分析，揭示了不同优化策略在动态层面上的根本差异。在理论分析的基础上，此文还提出了一种简单而有效的对齐提升方法，尽管这种方法在直觉上可能与常规策略相反，却在实验证明中显示出显著提升的效果。总结而言，此论文通过构建统一的学习动态框架，为理解大语言模型的微调过程提供了全新视角。该框架不仅帮助解释多个实证现象，还揭示了不同微调方法的本质差异，并进一步启发出实用的优化策略。对研究者和工程师而言，这项工作不仅拓展了理论边界，也为提升模型性能和对齐质量指明了方向。

2.3 杰出论文 3: AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models[3]

此文作者来自于中国科学技术大学和新加坡国立大学，主要研究内容为面向大型语言模型（LLM）的知识编辑问题。近年来，为解决 LLM 中出现的幻觉问题，许多工作从知识编辑角度进行实验。目前主流的编辑方式是“先定位-再编辑”，然而有研究表明，现有编辑方法会不可避免地会破坏模型中原有知识。作者认为这是由于编辑方法导致模型隐空间表示发生分布偏移。为了解决这个问题，作者引入了零空间（Null Space）的概念。零空间指的是那些不会改变原始知识表现的模型参数扰动方向所构成的空间。具体来说，作者先识别出用于保持原有知识的重要表示，然后将编辑过程中的参数扰动严格投影到这一零空间中，从而实现在编辑目标知识的同时，保护原有知识不被破坏。特别地，由于这一思想十分简洁，仅需在现有的大多数模型编辑方法中增加一行代码，即可将参数扰动投影到零空间，从而显著提升编辑后的性能与原有知识的保留效果。在实验中，作者表明该工作在编辑有效性和泛化性能上要优于基线 12.54%以上。此外，在连续编辑任务中，当编辑 3000

个样本规模时，编辑模型仍能保持原始性能。总体而言，该工作开创性地引入以零空间投影为核心的语言模型编辑新范式，为实现可持续、可靠的大模型知识更新提供了强有力的方法论支持。

2.4 时间检验奖：Adam: A Method for Stochastic Optimization[7]

Adam 是由 Kingma 和 Ba 提出的一种用于优化随机目标函数的一阶梯度优化算法。它能够自适应地估计梯度的一阶和二阶矩来调整每个参数的学习率，从而提升优化效率与稳定性。该方法实现简单、计算高效，能够很好地处理目标函数非平稳、梯度稀疏或噪声较大的情况，而且通常无需复杂的调参，因此特别适合用于大规模数据或参数量庞大的问题。如今，Adam 已成为深度学习领域应用最广泛的优化算法之一，彻底改变了神经网络训练方式，其在实际应用中的卓越表现使之成为从计算机视觉、自然语言处理到强化学习等众多领域最先进模型的首选优化器，展现了卓越的通用性。在颁奖仪式上，两位作者用生动有趣的语言分享了开发 Adam 优化器背后不为人知的故事，讲述了论文先被 ICLR 拒绝、到写邮件申诉、最终被接收的曲折经历。这启示我们，有价值的研究成果可能不会立即获得认可，但随着时间的推移，真正优秀的工作最终能够经受时间的考验，被学界发现并广泛应用。

三、 大会演讲

3.1 Zico Kolter: Building Safe and Robust AI Systems

演讲者 Jeremy Zico Kolter 是卡内基梅隆大学的教授，并担任该校机器学习系主任。他的研究主要聚焦于模型优化与鲁棒性。在此次演讲中，演讲者系统回顾了过去十年他和他的团队在神经网络优化、可验证对抗鲁棒性、深度学习实践以及 AI 安全等方向的阶段性研究进展。演讲者针对各个研究阶段都分享了自己选择该研究方向的“动机”以及“方法”。其中，在关于 AI 安全的部分，演讲者提出了几个颇具启发性的观点：1. 当前被广泛关注的安全问题，可能会随着更强大模型的出现而被自然解决；2. 我们可能只能识别安全问题，却无法真正修复它们；3. AI 安全研究可能是当前学术界对人工智能发展影响最大的切入点。在演讲的最后，演讲者针对不同阶段的研究人员提出了具体建议：1. 对于学生和初级研究者：不必担心你的研究路径看起来不够“战略性”。研究的真正进展往往来源于偶然的交流、对问题的好奇心以及自然的探索过程；2. 对更资深的教职员工/研究人员来说，请慷慨支持那些尝试走出常规路线、探索非传统方向的教职员工。

3.2 Song-Chun Zhu: Framework, Prototype, Definition and Benchmark

演讲者朱松纯教授是北京通用人工智能研究院院长、北京大学讲席教授、清华大学基础科学讲席教授。此演讲系统介绍了北京通用人工智能研究院与北京大学在通用人工智能（AGI）方面的研究成果。有趣的是，朱教授指出汉字“通”不仅有通用人工智能的含义，也正好包含了 A、G、I 三个英文字母，因此也将通用人工智能称作“TongAI”。演讲者首先展示了一个数字智能体原型——一个名为“通通”的生活在高度真实物理模拟环境中的虚拟小女孩，她能够在多物理属性和社会互动交织的情境中持续学习和成长。通通具备自我驱动的价值系统，拥有欲望与目标，能够据此生成计划并采取行动。她所体现的智能体行为不是外部指令控制的产物，而是基于内在动机和价值的自主行为。围绕这一智能体原型，演讲者进一步阐述了其背后的理论框架，即由三个核心组成部分构成的系统：认知架构（C）、潜能函数（U）和价值函数（V）。各种 AGI 系统可以被视为三维空间（C,U,V）中的一个点。这个框架代表了一种范式转变：从以大数据驱动小任务的统计范式转向以价值驱动的小数据、大任务的范式。演讲还介绍了团队提出的 TongTest——一个用于评估通用智能的全新测试标准与平台。TongTest 不仅远超图灵测试的复杂性，还融合了心理学与人类学的理论成果，评估智能体在复杂多模态、具身任务中的表现。目前，TongTest 认为通通的智能水平可类比人类 3 至 4 岁的儿童。此外，演讲还展示了一些近期在类人机器人及其实际应用方面的研究进展，并探讨了东方哲学视角下关于人与智能的本质，以及道德与社会规范如何在 CUV 框架下自然涌现，从而为 AGI 的安全性提供理论基础。总体而言，此次演讲不仅提出了一个具备理论完备性与实践原型的 AGI 研究框架，也结合认知科学、哲学与技术系统展示了一

种跨学科的、面向长期通用智能目标的研究路径。

3.3 Yi Ma: Pursuing the Nature of Intelligence

演讲者马毅教授是香港大学穆斯克特基金会数据科学研究所（HKU IDS）及计算机科学系的讲席教授，其研究方向涵盖计算机视觉、高维数据分析与智能系统。在此次演讲中，马教授从历史、科学、数学和计算等多个维度系统阐述了智能的多层次机制与本质。演讲伊始，马教授回顾了智能在自然界中的演化历程——从系统发育、个体发育再到社会层面的演进，以及人工智能的发展轨迹。在此基础上，他提出了智能学习的三个核心问题，并逐一给出解答：1. “学什么？”：智能应从外部世界中学习那些具有可预测性的数据。所有可预测的数据都可被编码于分布 p(x) 中，而该分布实际存在于一个低维的“子空间”里。尽管我们所观察到的数据是高维的，但其中有用和有规律的信息往往隐藏在一个更低维度的空间中。因此，学习应聚焦于最终的低维表征。2. “如何学？”演讲者强调，学习低维结构的一个基本且统一的机制是：通过压缩来降低观察到的（噪声）数据分布的熵。3. “如何判断正确性？”演讲者指出，低维结构学习的正确性可以通过双向的“编解码”过程来验证——即既能压缩信息，又能解码还原。这种双向验证机制为学习结果的可靠性提供了坚实基础。在探讨当前智能发展的基础上，演讲者展望了“下一代智能”的发展方向，即自动智能（Autonomous Intelligence）。他强调，我们应思考如何构建能够持续且自动学习连续表征的人工智能系统。对此，自然智能提供了丰富的启发。例如，视觉皮质已具备某种形式的离散编码机制，这为人工系统的构建提供了生物学依据。在演讲结尾，演讲者提出了关于智能的核心观点：“智能的核心在于如何编码和改进信息，以更好地预测世界。”总体而言，该演讲提出了一个基于“压缩编码—解码”原理的数学框架，作为理解与解释智能本质的理论基础。

3.4 Dawn Song: Towards Building Safe and Secure AI: Lessons and Open Challenges

演讲者 Dawn Song 教授是加州大学伯克利分校“负责任去中心化智能研究中心（RDI）”的教员联合主任，同时也是伯克利人工智能研究实验室（BAIR Lab）的一员。她的研究工作主要聚焦于深度学习、安全性以及区块链技术。在此次演讲中，Dawn Song 教授深入探讨了构建与部署人工智能和大语言模型智能体所面临的多重风险，并分享了相应的缓解策略。她重点介绍了团队在隐私保护、鲁棒性、防止幻觉生成和公平性等方面的研究进展，强调必须采取全生命周期的安全防护机制。具体的，在前沿科技探索阶段，需研究能够从根本上提升模型安全性的技术范式；在模型对齐阶段，应加强 AI 系统的防护能力，以有效抵御多种攻击方式；在模型评估阶段，需深化对模型能力和行为的理解来确保它们以可信的方式来运作。在模型部署阶段，必须保障系统按预期安全、稳定地运作。演讲者还指出，前沿技术的迅猛发展，正在促使网络攻击数量显著上升。在防御实践方面，演讲者分享了她们在使用深度学习检测物联网脆弱性、以及利用大型语言模型进行零日漏洞识别方面的最新研究成果。最后，演讲者指出要确保一个安全的 AI 未来，我们必须采取一种社会技术协同（sociotechnical）的方法。演讲者还分享了其最近提出的一项基于科学与证据的 AI 政策建议，重点阐述三个关键优先事项：加深对 AI 风险的理解、制定有效的缓解措施，以及推动制定更加健全的 AI 政策。

3.5 Danqi Chen: Training Language Models in Academia: Challenge or Calling?

演讲者陈丹琦副教授是普林斯顿大学计算机学院副教授，领导普林斯顿自然语言处理组，近期研究方向为训练、适配并理解语言模型，尤其注重如何让学术界更容易地训练和使用这些模型。此演讲的主题为如何在学术界训练语言模型。随着语言模型的发展，训练超大规模模型已成为机器学习的重要方向。然而，由于这一领域依赖大规模计算资源、专有数据和工程基础设施，因此几乎完全由工业界领导。相比之下，学术界面临着显著的劣势，包括计算资源有限、缺乏基础设施支持，以及对高质量训练数据的访问受到限制。这些劣势使得学术界难以直接训练大语言模型。尽管如此，演讲者强调，学术界在语言模型训练中仍具备不可替代的价值。通过从训练过程本身出发，学术研究可以帮助我们更深入地理解模型行为，并在资源受限的条件下开发出更高效、可解释的算法与系统。演讲者分享了在过去两年中，

其实验室在有限预算下进行的预训练与后训练工作。这些研究关注于训练动态的分析、模型性能与效率之间的权衡，以及资源受限条件下的技术创新，同时将所得模型与工具以开放形式贡献给社区。此外，演讲者还指出了三个值得学术界持续投入的研究方向：1. 开发小而强的模型，在模型尺寸受限的条件下实现良好性能；2. 理解与改进训练数据，探索数据质量对模型训练效果的影响；3. 推进基于开源模型的后训练方法，包括对齐、指令微调和偏好优化等技术。总体而言，该演讲不仅呈现了在学术资源限制下进行语言模型训练的现实路径，也强调了学术界在方法创新、理论探索与社区共享方面的独特贡献。演讲呼吁学术界继续参与这一领域的研究，并探讨与工业界协作的可能形式，以促进语言模型技术的持续发展与应用。

3.6 Tim Rocktaeschel: Open-Endedness, World Models, and the Automation of Innovation

演讲者 Tim Rocktaeschel 教授是 Google DeepMind 的首席科学家，也是伦敦大学学院计算机系人工智能中心的教授，工作主要重点在于通用人工智能、开放性和自我改进智能。此演讲探讨了通往人工超级智能（ASI）的一条关键路径——从传统的目标导向优化转向“开放性进化”（Open-Endedness）这一研究范式。“开放性进化”由 Stanley、Lehman 与 Clune 等学者提出，其核心理念是构建能够持续产出新颖且可学习成果的系统，而非仅仅追求单一性能指标的最优化。在这一框架下，系统不断自我生成和探索新的任务与解决方案，模拟出更接近自然智能的创新过程。演讲者重点介绍了其团队在大规模基础世界模型方面的研究。这类模型能够生成丰富多样的环境，为训练更具泛化能力和鲁棒性的智能体提供支持。这些世界模型不仅扩展了可供学习的任务空间，也为推动更广义人工智能奠定基础。进一步地，演讲者提出一个重要论点：开放性进化与基础模型的结合，可能是实现自动化创新的关键机制。这一融合已在实践中展现出初步成果，例如自动提示工程、自动红队测试、大语言模型中的 AI 辩论等。这些自我迭代、自我挑战的能力，标志着人工智能系统正在逐步具备推动自身进步与创新的潜力。总体而言，这场演讲不仅展示了基础模型和开放性研究的新进展，也提出了一个深远的愿景：未来的人工智能或将具备自我生成目标、自我改进能力，成为推动创新本身的主体力量。这一方向为构建更具创造力与自主性的智能系统提供了理论基础与实践路径，值得广泛关注。

四、 总结展望

此次 ICLR 会议上，“AI 安全”是一个备受关注的热点话题。事实上，杰出论文与大会演讲中 AI 安全话题占据 1/3，这足以说明此次 ICLR 对 AI 安全的重视。的确，随着大模型技术的迅速发展，如何有效管控 AI 系统、防范潜在风险，已成为技术进步所带来的关键社会议题之一。从数据收集的可靠性、到训练过程中的偏见控制、再到部署阶段的安全防护，AI 安全问题贯穿了整个技术链条。在当前监管框架尚未成熟的背景下，学术界的前瞻性探索显得尤为重要。ICLR 作为顶级学术会议，其重视安全议题不仅体现了研究界的高度敏感性，也预示着 AI 发展趋势正在从“能力竞赛”向“可信落地”转变。未来，技术与治理的深度融合，将成为 AI 安全发展的必经之路。

责任编委 张 杰

参考文献

[1] Qi X, Panda A, Lyu K, et al. Safety alignment should be made more than just a few tokens deep [C]//ICLR. 2025.
[2] Ren Y, Sutherland D. Learning Dynamics of LLM Finetuning [C]//ICLR. 2025.
[3] Fang J, Jiang H, Wang K, et al. AlphaEdit: Null-Space Constrained Knowledge Editing for Language Models [C]//ICLR. 2025.
[4] Wang J, Mittal P, Song D, et al. Data Shapley in One Training Run [C]//ICLR. 2025.
[5] Ravi N, Gabeur V, Hu Y, et al. SAM 2: Segment Anything in Images and Videos [C]//ICLR. 2025.
[6] Narasimhan H, Jitkrittum W, Rawat A, et al. Faster Cascades via Speculative Decoding [C]//ICLR. 2025.

[7] Kingma D, Ba J. Adam: A Method for Stochastic Optimization [C]//ICLR. 2015.
[8] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate [C]//ICLR. 2015.

徐逸峰

中国科学院计算技术研究所硕士研究生，导师为山世光研究员，主要研究方向为视觉内容生成。
Email:yifeng.xu@vipl.ict.ac.cn

王中琦

中国科学院计算技术研究所硕士研究生，导师为张杰副研究员，主要研究方向为人工智能安全。
Email: wangzhongqi23s@ict.ac.cn

何振梁

中国科学院计算技术研究所特别研究助理。主要研究方向为计算机视觉、生成模型、视觉的生成与理解协同。谷歌学术引用 1400 余次，AttGAN 工作曾入选 ESI 高被引论文。曾获得 2024 年中国图象图形学学会优秀博士论文提名、2020 年中科院计算所所长特别奖。
Email:hezhenliang@ict.ac.cn