中国科学 : 信息科学 2025 年 第 55 卷 第 6 期: 1548–1549
SCIENTIA SINICA Informationis

刊讯

大模型使能技术与前沿应用专题简介

景丽萍[1], 刘淇[2], 张敏灵[3*]

1. 北京交通大学, 北京 100044
2. 中国科学技术大学, 合肥 230026
3. 东南大学, 南京 210096
* 通信作者. E-mail: zhangml@seu.edu.cn

大模型作为人工智能领域的重要突破, 展现出跨任务、跨模态的卓越泛化能力. 尽管其在通用场景中表现出色, 但在多样化领域中, 其技术与应用拓展仍有待深入研究. 为推动相关研究的发展, *SCIENCE CHINA Information Sciences* 于 2025 年 68 卷第 6 期组织出版了 “大模型使能技术与前沿应用” 专题 (Special Topic: Enabling Techniques and Cutting-Edge Applications of Foundation Models), 共收录 10 篇具有创新性与前瞻性的研究成果.

Cui 等在 “Dynamic prompt allocation and tuning for continual test-time adaptation” 中针对大模型在跨域测试自适应 (test-time adaptation) 中面临的灾难性遗忘问题, 提出了一种动态提示分配与调优方法. 该方法通过引入可学习的领域特定提示, 有效解耦不同目标域的参数空间, 从而降低跨域干扰, 提升模型在持续学习场景中的适应能力.

Jie 等在 “Agent4Vul: multimodal LLM agents for smart contract vulnerability detection” 中聚焦智能合约安全中的漏洞检测问题. 针对现有大模型方法在处理多样化漏洞特征方面效果有限的难题, 作者提出了包含 Commentator 和 Vectorizer 两个智能体的联合框架, 同时融合语义信息与图结构, 实现了对多类型漏洞的稳健检测.

Zhao 等在 “Visual and text prompt learning for multi-modal brain disease diagnosis” 中, 针对医学诊断中脑疾病检测常受限于数据稀缺或不完整的问题, 提出了一种融合专家知识的多模态提示学习框架. 该方法通过提示机制从视觉数据中提取语义感知特征, 并调适预训练的单模态模型以应对多模态任务, 从而保持优越的性能.

Wang 等在 “DiagLLM: multimodal reasoning with large language model for explainable bearing fault diagnosis” 中, 为实现轴承故障的准确及时诊断、保障机械系统可靠运行, 提出结合大型语言模型推理能力的多模态诊断方法, 融合包络谱图像与专家知识, 实现了高准确率且具有可解释性的故障识别.

Huang 等在 “The superalignment of superhuman intelligence with large language models” 中, 针对大模型如何实现与人类价值观对齐的核心问题, 提出 “超级对齐 (superalignment)” 概念. 文章深入剖析了弱到强泛化、可扩展监督等关键挑战, 并构建了由攻击者、学习者和评论者组成的对齐框架, 推动了超人类智能系统的发展.

Tong 等在 “MindScore: quantifying human preference for text-to-image generation through multi-view lens” 中, 针对文生图 (text-to-image, T2I) 任务中生成结果难以精准对齐人类意图的问

引用格式: 景丽萍, 刘淇, 张敏灵. 大模型使能技术与前沿应用专题简介. 中国科学: 信息科学, 2025, 55: 1548–1549, doi: 10.1360/SSI-2025-0209
Jing L P, Liu Q, Zhang M L. Special topic: enabling techniques and cutting-edge applications of foundation models. Sci Sin Inform, 2025, 55: 1548–1549, doi: 10.1360/SSI-2025-0209

题, 提出了一种基于多视角评估的人类偏好量化方法, 从匹配度、忠实性、质量与真实感 4 个维度评估生成图像, 为文生图任务提供了更具人类感知对齐性的评价体系.

Wang 等在 “RAG-leaks: difficulty-calibrated membership inference attacks on retrieval-augmented generation” 中, 针对检索增强生成 (retrieval-augmented generation, RAG) 在处理敏感信息时易遭受成员推理攻击 (membership inference attack, MIA) 的问题, 重新定义了 RAG 场景下的 MIA 问题, 提出将高相似度样本视为成员样本, 并通过似然比检验对原始相似度接近的样本进行成员性得分校准, 从而提升了攻击的准确性与有效性.

Yuan 等在 “On the encryption for graph foundation model inference of sparse graph” 中, 针对云端服务中图大模型推理存在的隐私泄露风险, 提出了一种融合图加密与边采样技术的提示方法, 用于实现私密的图大模型推理. 该方法不仅具备实用性, 还辅以理论分析验证其有效性, 为图模型推理的隐私保护提供了切实可行的解决方案.

Tao 等在 “LEDNet: a multimodal foundation model for robust deepfake detection” 中提出了一种新型多模态检测框架, 通过引入对 “真实” 与 “伪造” 的源不变语言描述, 辅助视觉特征学习, 有效提升了模型对多种生成方式的泛化能力, 增强了深度伪造检测的稳健性.

Duan 等在 “Trustworthy forgery detection with causal inference” 中进一步引入因果推理机制, 通过构建结构化因果模型与轻量级插件, 有效解耦了伪造特征中的虚假相关性, 实现了高精度的可信伪造内容检测.

在专题筹备与出版过程中, 我们诚挚感谢所有为专题撰写高质量论文的作者, 感谢审稿专家们始终如一的严谨评审与高效反馈, 也感谢广大读者长期以来的关注与支持.