

徐浩铭

📞 17341432003 ✉️ haomingxu2003@gmail.com 🧑‍🎓 导师: 张宁豫; 邓淑敏

🏠 haomingx.github.io 📖 Google Scholar



教育背景

浙江大学	2025.09 – 2030.06
人工智能 · 硕博连读	中国杭州
哈尔滨工业大学	2021.09 – 2025.06
英才学院-计算机科学与技术 · 本科	中国哈尔滨

个人情况

奖励情况: 蒋震奖学金; 浙江省新苗人才计划

研究兴趣: 大模型机理研究和可持续学习

技术栈: Python/C++, PyTorch, HuggingFace Transformers, VLLM, DeepSpeed, Linux, Git

科研成果 (Selected Publications)

- When Should Models Change Their Minds? Contextual Belief Management in LLMs** EMNLP 2026 Main
第一作者 – 提出上下文信念管理 (CBM) 范式, 构建 BeliefTrack 闭世界基准诊断 LLM 的信念保持、更新与隔离失效; 引入信念状态奖励 RL 平均降低失败率 70.9%。
- Illusions of Confidence? Diagnosing LLM Truthfulness via Neighborhood Consistency** ACL 2026 Main
第一作者 – 提出邻域一致性信念 (NCB) 度量, 揭示 LLM 的信念脆弱性; 引入 SAT 方法提升稳健性约 30%。
- Relearn: Unlearning via Learning for Large Language Models** ACL 2025 Main
第一作者 – 提出基于“学习”的大模型遗忘新范式, 实现对特定知识的彻底擦除同时不损害模型通用推理能力。
- ZJUKLAB at SemEval-2025 Task 4: Unlearning via Model Merging** ACL 2025 Workshop
第一作者 – 国际评测比赛的技术报告, 创新性地将模型融合引入知识遗忘任务, 结合多种遗忘策略和正则化策略。
- Towards Principled Knowledge Editing Methods for Large Language Model Reasoning** 子刊 NMI 2026
核心贡献者 – 系统分析推理型大语言模型知识编辑面临的挑战, 探讨现有知识编辑方法的局限性, 并提出面向模型推理能力的原则性知识编辑研究方向。
- EasyEdit2: An Easy-to-Use Steering Framework for Editing LLMs** EMNLP 2025 Demo
核心贡献者 – 整合多种主流知识编辑算法, 为社区提供高效、统一的 LLM 知识调控工具。

重点项目与竞赛

- LightMem: 轻量级记忆增强生成系统** ICLR2026, GitHub 1.2k+ Stars
实现了高效的长期记忆管理机制, 解决了长文本场景下的显存冗余问题。参与 Memory 仓库基线的构建以及后期维护。
- EasyEdit: 大语言模型知识编辑框架** GitHub 2.8K+ Stars
知识编辑领域影响力最大的开源工具之一。作为核心维护者, 主导了 DPO、QLoRA 等主流对齐与微调算法模块的集成与架构设计; 负责 MEMIT、ROME 等核心算法的性能优化与迭代, 保障了框架在新架构模型上的扩展性。
- 第 19 届国际语义评测研讨会 (SemEval-2025 Task 4)** 第二名
带队获得 LLM 敏感内容遗忘任务第二名: 结合多种 Unlearning 方法和模型融合的策略, 实现了遗忘和保留的权衡。
- ASC24 世界大学生超级计算机竞赛** 全球二等奖
在 vllm 基础上, 优化大语言模型推理管道, 利用量化、早停等技术提升超算环境下的吞吐量。

实习经历

阿里巴巴安全 AGI Lab · Yuvion Team	2026.06 – 至今
算法实习生	中国
乐聚机器人 (LEJU Robotics)	2022.07 – 2022.08
AI 技术部实习生	中国苏州