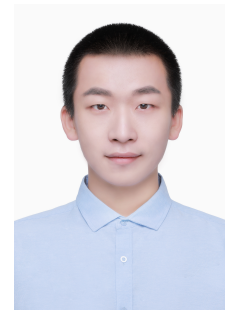


刘议骏

Homepage | Email: yijunliu@ir.hit.edu.cn | Phone: 177-4513-0877



教育背景

哈尔滨工业大学

2024 - 至今

- 专业：计算机科学与技术，直博生
- 导师：车万翔教授（社会计算与交互机器人研究中心 SCIR）
- 荣誉：校级优秀团员（2026）
- 研究方向：大语言模型的高效推理、长文本优化、Agent 记忆

哈尔滨工业大学

2020 - 2024

- 专业：计算机科学与技术，工学学士
- 成绩：90.135 (24/123)
- 荣誉：黑龙江省三好学生（2023）、黑龙江省优秀毕业生（2024）、校级优秀学生干部（2021）、校级优秀团员（2022）、校级优秀团干部（2023）、一等人民奖学金（2021、2022、2023）
- 职务：未来技术学院学生会生活部部长、班级生活委员

学术成果

[1] **Yijun Liu**, Yixuan Wang, Yuzhuang Xu, Shiyu Ji, Yang Xu, Qingfu Zhu, Wanxiang Che. "Judge Q: Trainable Queries for Optimized Information Retention in KV Cache Eviction". *AAAI 2026*. (CCF A, 一作) 针对 KV Cache 驱逐中局部信息偏重的问题，我们提出 Judge Q 轻量训练框架，通过可学习查询对齐解码阶段注意力分布，从而在 prefilling 阶段捕获全局上下文信息并优化缓存保留决策。该方法仅需微调少量 soft token 参数，在多项长文本基准上，相同缓存预算下，提升了模型性能与关键值命中率。

[2] Yuxuan Wang, **Yijun Liu**, Fei Yu, Chen Huang, Kexin Li, Zhiguo Wan, Wanxiang Che, Hongyang Chen. "CVLUE: A New Benchmark Dataset for Chinese Vision-Language Understanding Evaluation". *AAAI 2025*. (CCF A, 二作) 现有中文视觉-语言评测数据存在文化适配不足的问题，我们构建了 CVLUE 中文视觉-语言理解基准。该数据集包含 92 类中国文化代表性视觉对象，并覆盖图文检索、视觉问答、视觉定位与视觉对话等任务，为中文多模态模型评测提供了系统化资源。

[3] **Yijun Liu**, Jinzheng Yu, Yang Xu, Zhongyang Li, Qingfu Zhu. "A Survey on Transformer Context Extension: Approaches and Evaluation". *preprint*. (一作, 在投) 我们系统整理了 Transformer 为基础的模型长上下文扩展方法与评测体系，归纳长文本建模中的 OOD 问题、Lost in the Middle 及注意力平方复杂度等核心挑战。进一步从位置编码、上下文压缩、检索增强和注意力优化四类方法，

最后从数据、任务、指标三个评测维度进行系统梳理，并列举出当前尚待解决的问题。

[4] Yixuan Wang*, Shiyu Ji*, **Yijun Liu**, Yuzhuang Xu, Yang Xu, Qingfu Zhu, Wanxiang Che. “Lookahead Q-Cache: Achieving More Consistent KV Cache Eviction via Pseudo Query”. *EMNLP 2025*. (CCF B, 二作) KV Cache 驱逐方法存在预填充查询与解码查询分布不一致的问题，对此，我们提出 Lookahead Q-Cache 框架，利用低预算前瞻解码生成伪查询以指导缓存压缩。该方法提升了低缓存预算下的驱逐一致性与推理性能，并可与现有驱逐策略正交互补提升性能。

[5] Yixuan Wang, **Yijun Liu**, Yuzhuang Xu, Yang Xu, Qingfu Zhu, Wanxiang Che. “Think Before You Accept: Semantic Reflective Verification for Faster Speculative Decoding”. *preprint*. (二作, 在投) 现有投机解码验证过程缺乏语义约束，对于这一问题，我们提出无需训练的 Reflective Verification 方法，通过模型自反思信号融合原始 logits 与反思 logits，构建语义感知的验证分布。该方法在保持生成质量的同时提升端到端推理吞吐量，并可与现有统计验证策略兼容。

项目经历

南方网-文本校对项目 2023.4 - 2024.4

主要负责项目三期通用领域与政务场景实体纠错数据集自动化构建、模型训练以及项目相关材料书写等工作。该系统的主要功能包括文本纠错、实体纠错（例如地名和时间）以及各类实体搭配纠错等。主要涉及基于 BERT 模型的分词、词性标注和命名实体识别等技术，同时结合现有的语义分析工具（LTP）辅助完成。

华为小艺-大模型长期记忆和个性化能力技术合作项目 2024.3 - 2025.3

作为主要开发人员，负责项目的方案设计、数据构建、系统搭建与最终文档和 demo 交付等工作。该项目旨在完成对用户关键信息、事件的记忆，满足用户个性化需求。整套系统涉及 prompt 设计、分类模型训练以及 soft token 压缩等技术，最终得到自动化数据批量生成与标注流程和具备自适应记忆提取、记忆压缩以及基于记忆增强回复的完整系统。

竞赛获奖

- 2022-11 全国大学生数学建模竞赛**国家一等奖**
- 2022-05 美国大学生数学建模竞赛 **M 奖**
- 2021-12 全国大学生数学竞赛（非数学类）**省二等奖**
- 2023-08 CCL 2023 评测比赛中小学作文流畅性评价赛道**第一名**
- 2024-08 CCL 2024 评测比赛中文图文多模态理解评测比赛**负责人**