

刘议骏

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

个人简介

具备从问题建模、算法设计、数据构建、模型训练到系统落地的完整研发经验，曾围绕 KV Cache 驱逐、Agent 记忆等领域开展研究与项目实践，熟悉大模型推理优化、Agent 记忆等任务。

教育背景

哈尔滨工业大学	计算机科学与技术 直博生 导师：车万翔教授（SCIR 实验室） 研究方向：LLM 高效推理、长文本优化、多模态理解与评测、Agent 记忆	2024 - 至今
哈尔滨工业大学	计算机科学与技术 工学学士 成绩：90.135 (24/123)	2020 - 2024

学术成果

LLM 高效推理

- 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) 现有 KV Cache 驱逐方法易偏向局部上下文而遗漏关键全局信息，我们提出 Judge Q 轻量化训练框架，通过可学习查询对齐解码阶段注意力分布，仅微调 embedding 层极少量参数提升长文本推理性能与关键值命中率。
- Shiyu Ji*, Yixuan Wang*, **Yijun Liu**, Qingfu Zhu, Wanxiang Che. “Seer Self-Consistency: Advance Budget Estimation for Adaptive Test-Time Scaling”. *ACL 2026 findings*. (CCF A, 二作) 推理时扩展难以同时降低 token 消耗与推理延迟，SeerSC 通过 System 1 答案熵前置估计查询扩展潜力，指导 System 2 并行分配预算，在保持准确率的同时降低计算开销。
- 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 驱逐中的 prefilling 查询与 decoding 查询存在分布不一致现象，提出 Lookahead Q-Cache 框架，通过低预算前瞻解码生成伪查询指导缓存压缩，提升低缓存预算下的驱逐一致性与推理性能。
- 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主要依赖分布一致性而缺少语义约束，提出无需训练的 Reflective Verification 方法，通过融合模型自反思信号构建语义感知验证分布，在保持生成质量的同时提升端到端推理吞吐量。

长文本优化

- Yijun Liu**, Jinzheng Yu, Yang Xu, Zhongyang Li, Qingfu Zhu. “A Survey on Transformer Context Extension: Approaches and Evaluation”. *preprint*. (一作, 在投) Transformer 长上下文综述对方法分类与评测体系仍较为分散，我们系统梳理位置编码、上下文压缩、检索增强和注意力优化等技术路线，并从数据、任务和指标维度总结评测体系。

多模态理解与评测

- 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 CVLUE 中文视觉-语言理解基准，覆盖 92 类中国文化代表性视觉对象与多类多模态任务，为中文多模态模型评测提供系统化资源。

Agent 记忆

目前研究兴趣从模型层面以 Efficient 为核心的探索逐步延续到围绕 Agent 记忆机制，重点包括：

- 高效记忆优化** 面向现有开源记忆系统，设计通用优化框架，在保持记忆效果同时降低 LLM 调用与资源消耗。
- 多模态主动记忆** 针对流式长视频场景，探索主动记忆机制，使系统能够根据当前记忆库状态选择性提取关键记忆，而非被动等待外部输入。
- 自进化记忆系统** 为解决不同任务下的记忆系统适配需求，研究无需人工规则设计的自进化记忆框架，提升记忆机制的自动化构建能力。

项目经历

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
--------------------------	-----------------

- 自动化记忆数据构建与标注** 设计基于 prompt 的自动化数据构造流程，模拟用户口吻批量生成个性化对话数据，并围绕记忆类别、非记忆扩写及相关/无关提问构建多样化训练样本。
- 记忆分类决策** 将记忆识别建模为 token 级标签预测与分句级类别聚合任务，将记忆触发判断建模为句子级二分类任务，并结合触发结果设计差异化 prompt 回复模板。
- 记忆文本压缩与高效调用** 设计基于向量表示的记忆压缩与存储机制，以更短的 token 序列替代原始记忆文本，降低多条相关记忆引入时的上下文开销。
- 模型训练与微调优化** 负责项目中分类模型与记忆压缩模型的训练和优化，基于构造数据完成超参数调优、指标评估和错误分析；其中，针对记忆识别与触发判断任务，训练 RoBERTa 模型提升记忆分类与记忆时机判别效果；针对记忆压缩表示学习任务，采用 LoRA 等轻量化训练方式优化压缩模型表现，支撑长期记忆系统的稳定落地。

竞赛获奖

2022-11	全国大学生数学建模竞赛 国家一等奖 (0.6%, 全校唯一, 队长身份)
2022-05	美国大学生数学建模竞赛 M 奖 (7%)
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