

廖桓萱

✉ huanxuanliao@gmail.com · ☎ 15575986157 · 🌐 Xnhyacinth · 🌐 https://xnhyacinth.github.io/

🎓 教育背景

中国科学院大学自动化研究所, 北京	2023 – 至今
在读博士 模式识别, 自动化研究所模式识别国家重点实验室, 导师何世柱 and 刘康	
华北电力大学, 北京	2019 – 2023
学士 智能科学与技术 (创新), 1/13	

💡 研究兴趣

- 长上下文建模:
 - 综述: Awesome-LLM-Long-Context-Modeling (GitHub Star > 1.7K)
 - 文本压缩: HyCo2 (ArXiv 2025)
 - KV 压缩: QuarK (ArXiv 2025)
- 高效推理和知识激活:
 - 指令学习: TAGI (NeurIPS 2024)
 - 知识唤醒: AAG (COLING 2025)
 - 知识内化: SKIntern (COLING 2025)
 - 知识协同: NesyCD (AAAI 2025)
 - 知识迁移: DATA (Arxiv 2025)
- 高效多模态模型及推理

📖 主要文章 (部分) [FULL LIST]

(* stands for equal contribution; Listed in reverse chronological order.)

- [Huanxuan Liao](#), Yixing Xu, Shizhu He, Guanchen Li, Xuanwu Yin, Dong Li, Emad Barsoum, Jun Zhao, Kang Liu. SparK: Query-Aware Unstructured Sparsity with Recoverable KV Cache Channel Pruning. (Preprint, 2025)
- [Huanxuan Liao](#), Shizhu He, Yupu Hao, Jun Zhao, Kang Liu. Beyond Hard and Soft: Hybrid Context Compression for Balancing Local and Global Information Retention. (Preprint, 2025)
- [Huanxuan Liao](#), Shizhu He, Yupu Hao, Jun Zhao, Kang Liu. DATA: Decomposed Attention-based Task Adaptation for Rehearsal-Free Continual Learning.
- [Huanxuan Liao](#), Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, Jun Zhao. Neural-Symbolic Collaborative Distillation: Advancing Small Language Models for Complex Reasoning Tasks. (AAAI 2025, CCF-A)
- [Huanxuan Liao](#), Shizhu He, Yupu Hao, Xiang Li, Yuanzhe Zhang, Jun Zhao, Kang Liu. SKIntern: Internalizing Symbolic Knowledge for Distilling Better CoT Capabilities into Small Language Models. (COLING 2025, CCF-B)
- [Huanxuan Liao](#), Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, Shengping Liu, Jun Zhao. Awakening Augmented Generation: Learning to Awaken Internal Knowledge of Large Language Models for Question Answering. (COLING 2025, CCF-B)
- [Huanxuan Liao](#), Shizhu He, Yao Xu, Yuanzhe Zhang, Yanchao Hao, Shengping Liu, Kang Liu, Jun Zhao. From Instance Training to Instruction Learning: Task Adapters Generation from Instructions. (NeurIPS 2024, CCF-A)

👨‍💻 实习/项目经历

字节跳动北京	2025 年 6 月 – 至今
实习 研究型实习生 (高效多模态模型和长视频理解)	

实习 研究型实习生 (长上下文)

基于全局和局部的长上下文压缩

- 实现了基于 Qformer 和 MLP 的动态路由 soft token 全局压缩和基于分类的 hard token 局部压缩

长上下文建模综述

- 基于自己的长上下文建模仓库完成综述撰写

KV Cache Channel 压缩

- 对 token 进行动态 channel 裁剪 (topk 和 topp) 而不是只针对长度压缩, 同时考虑恢复裁剪的 channel

♡ 获奖情况

奖学金

攀登二等奖学金	2024 年 12 月
北京优秀毕业生	2023 年 6 月
国家奖学金	2022 年 12 月
国家奖学金	2021 年 12 月
四方社会奖学金	2020 年 12 月

竞赛

国家三等奖, 信息安全竞赛	2022 年 8 月
国家级优秀, 全国大学生创新创业项目	2021 年 12 月
北京市三等奖, 互联网 +	2021 年 8 月

💡 专业服务

会议审稿: ACL ARR 2024, 2025. NeurIPS 2025.

开源社区: Awesome-LLM-Long-Context-Modeling 1.7k+ Stars.