

刘琦

北京, 中国 ◊ +86 137-9812-2637 ◊ qiliu6777@gmail.com
个人主页 ◊ Google Scholar ◊ GitHub



教育背景

中国人民大学, 高瓴人工智能学院 人工智能专业, 博士研究生; 导师: 毛佳昕教授	2024.09 - 至今 北京
中国人民大学, 高瓴人工智能学院 人工智能专业, 工学学士	2020.09 - 2024.06 北京

研究方向

方向一：大语言模型与搜索

多向量检索。聚焦多向量检索的效果与效率分析; 同时主导开发开源长文本多向量检索模型 jina-colbert。

稠密向量检索模型。熟悉主流向量模型训练范式与流程并参与 jina-embeddings-v2 开发, 并进行向量模型范式扩展研究 (E2Rank 统一检索和排序模型, RELER 探索通过强化学习使用下游任务奖励训练检索模型)。

大语言模型重排序。开发 LLM4Ranking 框架, 进行使用大语言模型进行重排序的相关性机制探究、效率优化 (PE-Rank、E2Rank), 以及新范式研究 (DiffuRank)。

方向二：搜索智能体

搜索行为与失败诊断。通过长程搜索智能体行为分析, 区分未找到关键证据的检索缺口与找到却未正确利用证据的利用缺口; 分析搜索次数、累计证据召回与答案质量的关系, 定位冗余查询和低收益搜索, 为查询改写与搜索停止策略提供依据。

信息获取与证据管理。围绕搜索智能体的搜索与文档获取与访问的工具和框架, 对其机制与性能进行公平的对比与分析, 探究将文档选择与证据提取解耦后, 持久化工作区如何支持按需提取证据、跨轮次回访页面, 进一步考察这一机制对重复获取、上下文负担及首次阅读遗漏证据的影响 (FtE)。

Search as Code (近期主要工作)。开发 OpenSaC 框架, 提供可扩展的 Python 接口, 将搜索、页面抓取、重排序与信息提取组合为可通过代码编排的研究流程, 并支持独立配置服务提供方与不同智能体的调用方式, 在 Codex 环境下获得了比其原生搜索更好的端到端性能。正在探索使用强化学习训练代码搜索智能体。

技术能力

Coding: 熟练使用 Python 和 Linux 开发环境; 熟练使用 Codex、Claude Code 等工具。

AI 框架: 熟悉 DeepSpeed、vLLM、slime 等模型训练和推理框架; 熟悉 Harbor 等智能体评测框架, 具备丰富的搜索智能体开发与实践经验, 熟悉智能体基本机制与前沿进展。

研究与实习经历

新加坡国立大学 NExT++ 研究中心 访问博士生; 与朱凤彬博士、Tat-Seng Chua 教授合作开展搜索智能体研究。	2026.01 - 2026.07 新加坡
阿里巴巴通义实验室 研究实习生; 研究文本向量模型与重排序模型。	2025.05 - 2026.01 北京
Jina AI 研究实习生; 研究文本向量模型与多向量迟交互检索模型。	2023.11 - 2024.06 北京

1. **OpenSaC: A Python Library for Search as Code in Deep Research Agents**
Qi Liu, Jiaxin Mao
Github Repo.
2. **Learning to Retrieve via Reinforcement Learning in Embedding Space**
Qi Liu, Fengming Liang, Yiqun Chen, Erhan Zhang, Jiaxin Mao
Preprint, 2026.
3. **Fetch-then-Explore: Decoupling Selection from Extraction over a Persistent Workspace for Search Agents**
Qi Liu, Yiqun Chen, Zidan Chen, Yan Gao, Yi Wu, Yao Hu, Jiaxin Mao, Fengbin Zhu, Tat-Seng Chua
Preprint, 2026.
4. **Diagnosing Search Behavior and Failure Modes in Long-Horizon Search Agents**
Qi Liu, Jiaxin Mao, Fengbin Zhu, Tat-Seng Chua
SIGIR-AP 2026.
5. **DiffuRank: Effective Document Reranking with Diffusion Language Models**
Qi Liu, Kun Ai, Jiaxin Mao, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, Fengbin Zhu, Ji-Rong Wen
Preprint, 2026.
6. **E2Rank: Your Text Embedding can Also be an Effective and Efficient Listwise Reranker**
Qi Liu, Yanzhao Zhang, Mingxin Li, Dingkun Long, Pengjun Xie, Jiaxin Mao
EMNLP 2026.
7. **How do Large Language Models Understand Relevance? A Mechanistic Interpretability Perspective**
Qi Liu, Haozhe Duan, Jiaxin Mao, Ji-Rong Wen
ACM Transactions on Information Systems, 2025.
8. **LLM4Ranking: An Easy-to-use Framework of Utilizing Large Language Models for Document Reranking**
Qi Liu, Haozhe Duan, Yiqun Chen, Quanfeng Lu, Weiwei Sun, Jiaxin Mao
Preprint, 2025.
9. **Leveraging Passage Embeddings for Efficient Listwise Reranking with Large Language Models**
Qi Liu, Bo Wang, Nan Wang, Jiaxin Mao
WWW 2025.
10. **An Analysis on Matching Mechanisms and Token Pruning for Late-interaction Models**
Qi Liu*, Gang Guo*, Jiaxin Mao, Zhicheng Dou, Ji-Rong Wen, Hao Jiang, Xinyu Zhang, Zhao Cao
ACM Transactions on Information Systems, 2023.
11. **Understanding the Multi-vector Dense Retrieval Models**
Qi Liu, Jiaxin Mao
CIKM 2023.