秦 若愚

132-2390-6909 qry6909@163.com github.com/chestnut-Q
清华大学 计算机科学与技术系 • 博士

清华大学计算机系一年级博士研究生，研究方向为高效的机器学习系统，有扎实的数理基础和专业知识。本科期间曾在 THUMT 实验室进行科研且有相关论文产出。在推荐直博至清华大学高性能所 MADSys 实验室后，加入月之暗面 Infra 团队，负责大语言模型高性能推理框架 Mooncake 的研发。

教育经历

- 清华大学，计算机科学与技术系，工学博士 2024-09 至 2029-06（预计）
  MADSys 实验室，导师：章明星
- 清华大学，计算机科学与技术系，工学学士 2019-09 至 2024-06
  GPA 3.95/4.0，Rank 10/182，**计算机系优秀毕业生**

荣誉奖励

- 清华大学优良毕业生（**前 5%**） 2024-06
- 清华大学优秀共青团员 2023-06
- 清华之友——灵均领航奖学金 2022-10
- 唐立新优秀奖学金 2020-10
- 清华大学新生二等奖学金（第 32 届中国化学奥林匹克决赛金牌） 2019-09

科研实习

- **Ruoyu Qin**, Zheming Li, Weiran He, Mingxing Zhang, Yongwei Wu, Weimin Zheng, and Xinran Xu. ”A KVCache-centric Disaggregated Architecture for LLM Serving.” arXiv preprint arXiv:2407.00079 (2024).
  作为月之暗面 Infra 团队大模型推理优化实习生，负责大语言模型高性能推理框架 Mooncake 的研发、实验以及论文撰写。目前该开源项目在 Github 上已获得 2.2k stars。 2023-09 至 2024-06
- Chen, Chi, **Ruoyu Qin**, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. ”Position-enhanced visual instruction tuning for multimodal large language models.” arXiv preprint arXiv:2308.13437 (2023).
  在 THUMT 研究组刘洋老师指导下进行多模态大模型 Instruction Tuning 的研究，设计出高效的高质量多模态数据生成 Pipeline，极大提升多模态大模型 Instruction Tuning 效果。 2023-02 至 2023-08

项目经历

- `[Attention]` **FlashAttention from Scratch** MADSys，2024-12
  不依赖任何算子库的 FlashAttention 前向传播的实现，实现分为两个版本，分别支持在单卡 GPU 和多卡 CPU 集群上运行。该实现代码为 FlashAttention 初学者提供了一个简洁易懂的实现参考。
  技术栈：`FlashAttention / CUDA / MPI`
- **GQA Triton Kernel 优化** 月之暗面，2023-09
  在大模型推理的 Decoding 阶段，注意力机制中读取 KVCache 的速度为性能瓶颈。对于使用了 GQA 的模型，我优化了 triton kernel 实现，在保证并行性的同时共享 KV heads 的读取，从而大大节省了显存带宽，使在长上下文的场景下模型推理提速约 100%。
  技术栈：`Pytorch / Triton`
- `[JNeRF]` **第二届计图人工智能挑战赛可微渲染新视角生成赛道三等奖** 清华大学-腾讯，2022-06
  利用 Jittor 框架以及开源 JNeRF 代码，设计高效率剪枝方法，完成三维物体新视角图片渲染任务，并基于光线跟踪算法重建三维物体模型。
  技术栈：`Jittor / NeRF`
- `[SnakeGo]` **第 26 届智能体大赛播放器项目** 清华大学计算机系学生科协，2021-09 至 2022-02
  为清华大学第 26 届智能体大赛开发比赛播放器，开发使用 Cocos 2D 引擎，首次实现了在网页平台嵌入播放器。自己负责播放器部分逻辑以及与平台的通信对接。
  技术栈：`TypeScript / WebSocket`