梁昊

hao.liang@stu.pku.edu.cn · (+86) 132-4193-5113 · https://haolpku.github.io/

教育背景

北京中关村学院, 北京 2025.03 – 至今

博士研究生 联合培养，导师: 董彬　研究方向: 强推理数据准备算法与系统

北京大学, 北京 2023.09 – 至今

博士研究生 数据科学，导师: 张文涛　研究方向: 大模型数据准备算法与系统

牛津大学, 牛津 2022.09 – 2023.07

访问学生 计算机科学，所有课程成绩均为 A。

北京理工大学, 北京 2019.09 – 2023.06

理学学士 信息与计算科学，专业排名 1/30。

实习经历

上海算法创新研究院　上海 2025.02 – 2026.02

大模型算法实习生

主要研究 Data-Centric Systems，包括数据准备系统 DataFlow 和数据训练系统 DataFlex。

百川智能　北京 2024.09 – 2025.02

大模型算法实习生

主要研究 Data Centric AI，产出 5 篇 CCF-A 论文。

Apple　北京 2024.03 – 2024.09

Machine Learning Research Internship

主要研究 Data Centric AI，产出 1 篇 CCF-A 论文。

开源项目经历

大模型数据准备系统 DataFlow(3k star)(DataFlow Technical Report) 2024.09 – 2025.11

Project Leader

1. 算法层面: Dataflow 在 Reasoning，Text2SQL 等领域取得世界领先效果。使用 Dataflow 准备的数据仅 10k 条超越最好的指令微调数据集 Infinity-Instruct 的 1M 数据的微调效果。Dataflow 团队使用 Dataflow 在LIC · 2025 语言与智能技术竞赛以及ICML Seephy 比赛中获得第一名。
2. 系统设计层面: Dataflow 系统主要分为三层设计，从上到下分别是 workflow 层，算子层，以及 prompt 层。我们对于算子和 prompt 进行注册，并且对算子进行描述，方便 Data Agent 进行调用。
3. Data Agent 层面: Data Agent 可以实现自动的数据处理工作流搭建，以及自动设计的数据处理算子，以及自动基于大模型的算子的 Prompt 设计。

大模型数据训练系统DataFlex 2025.09 – 2026.02

Project Leader

1. 功能与算法层面: DataFlex 目前作为 LLaMAFactory 项目的数据模块，目前适配 LLaMAFactory-v0，主要支持 (1) 数据选择，(2) 数据配比, (3) 数据权重调整。目前正在适配 LLaMAFactory-v1 以达到更好的解耦程度。
2. 系统层面: 目前数据选择与配比领域的仓库没有统一标准，而且年久失修往往难以复现。DataFlex 是 Unified Framework 可以支持更好的数据选择，配比以及权重调整论文复现与迭代。

Camel × DataFlow × LLaMAFactory 2025.09 – 2026.02

Project Leader

与 Camel AI(16k star) 团队, MinerU 团队 (55k star) 以及 LLaMAFactory(67k star) 团队合作，实现全流程的指令-> 自动化数据搜索-> 自动化数据准备-> 模型训练全流程工具。

大模型数据评估模块 2026.01 – Ongoing

Project Leader

目前的数据评估仅限于打分，但无法反应下游任务表现。我们通过计算 Transfer Learning 以及 Data Attribution 的指标实现数学数据集评估分数 R^2 达到 0.93。

论文发表

围绕 Data-Centric AI 发表 9 篇 CCF-A 一作/共同一作论文，一篇第二作者 ICLR 论文。下面展示部分论文，其余详见Google Scholar。

- **Text2SQL-Flow: A Robust SQL-Aware Data Augmentation Framework for Text-to-SQL**
  Qifeng Cai*, **Hao Liang***, Chang Xu*, Tao Xie, Wentao Zhang, Bin Cui
  Accepted by KDD 2026 (CCF-A)
- **Let's Verify Math Questions Step by Step**
  Chengyu Shen*, Zhen Hao Wong*, Runming He*, **Hao Liang*(project lead)**, Meiyi Qiang, Zimo Meng, Zhengyang Zhao, Bohan Zeng, Zhengzhou Zhu, Bin Cui, Wentao Zhang
  Accepted by KDD 2026 (CCF-A)
- **MathScape: Benchmarking Multimodal Large Language Models in Real-World Mathematical Contexts**
  Hao Liang*, Linzhuang Sun*, Minxuan Zhou*, Zirong Chen, Meiyi Qiang, Mingan Lin, Tianpeng Li, Fan Yang, Zenan Zhou, Wentao Zhang
  Accepted by ACM MM 2025 (CCF-A)
- **MM-Verify: Enhancing Multimodal Reasoning with Chain-of-Thought Verification**
  Linzhuang Sun*, **Hao Liang***, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, Wentao Zhang
  Accepted by ACL 2025 (CCF-A)

获奖情况

兴业奖学金, 北京大学 2025.09
萨师煊优秀学生论文奖, NDBC 2025.08
SeePhy 挑战赛第一名, 团队中排名第一, ICML 2025.07
博士研究生校长奖学金, 北京大学 2025.06
10/20000, 徐特立奖学金 (校最高奖), 北京理工大学 2023.06
8/500, 国家奖学金，北京理工大学 2020.10

其他信息

- GitHub: https://github.com/haolpku
- 语言: 汉语 - 母语, 英语 - 流利（托福 115 分）