马 培贤

86-137-1074-7969 pma929@connect.hkust-gz.edu.cn 广东省 • 广州市 • 南沙区 • 东涌镇
香港科技大学（广州）HKUST(GZ) 数据科学与分析 (Data Science and Analytics) • 哲学硕士 (M.Phil.)
https://mpx0222.github.io/ github.com/mpx0222

香港科技大学（广州）2025 届数据科学与分析专业（NLP 和 AI4DB 方向）硕士研究生，有扎实的数学与统计学基础，擅长数据科学、大语言模型开发与应用、数据库应用，热衷于探索和开发前沿 AI 技术，熟练掌握多种编程技能、开发工具和前沿算法。本科和硕士研究生在读期间参与并发表多篇相关方向文章。在 GitHub 上参与和贡献多个项目。

技能和语言

操作系统	Linux （4 年）
后端	Python, C
前端	JavaScript, Vue3, HTML, CSS, Markdown
中间件	Flask, MongoDB, Redis
数据科学	SQL, Pandas, Scikit-Learn, PyTorch, Transformers, Spark, Hadoop
开发工具	SSH, Git, Tmux, Vim
大语言模型	SFT, RLHF, Prompt, RAG
语言	**英语** — CET-6, CET-4

教育背景

2023.09 – 2025.08 | **香港科技大学（广州）** • 信息枢纽 (Information Hub)
数据科学与分析 (Data Science and Analytics) • 哲学硕士学位 (M.Phil.) - GPA: 3.8/4.0

2019.09 – 2023.07 | **暨南大学** • 信息科学技术学院/网络空间安全学院
智能科学与技术 • 工学学士学位 - GPA: 3.3/5.0

荣誉奖项

- 2024.09 | 2024-2025 红鸟硕士生全额奖学金，香港科技大学（广州）
- 2023.09 | 2023-2024 红鸟硕士生全额奖学金，香港科技大学（广州）
- 2023.06 | 优秀毕业生，信息科学技术学院/网络空间安全学院，暨南大学
- 2022.12 | 优秀学生二等奖学金，暨南大学
- 2021.12 | 优秀学生三等奖学金，暨南大学
- 2021.12 | 第 11 名，2021 年 CCF-BDCI 计算智能竞赛 • 情感分类 & 命名实体识别赛道
- 2021.05 | Honorable Mention，2021 年 MCM 美国大学生数学建模竞赛

实习经历

2025.01 – 2025.06 | 大模型数据合成工程实习生 @ 粤港澳大湾区数字经济研究院（IDEA）FinAI Lab
- 参与面向文本、图像等不同模态的自动化数据合成技术研发，用以合成高质量预训练数据，研究并开发面向复杂推理任务的自动化数据合成技术以及强化学习技术，用于大模型后训练。
- 参与面向大模型的数据匿名化、隐私保护、加密训练等模块的研发，将算法模块包装成 API，支持下游应用的调用。

2023.01 – 2023.06 | 数据挖掘实习生 @ 暨南大学计算传播研究中心
- 从国外新闻网页爬取相关舆情新闻信息，并进行数据清洗和数据预处理工作。（基于 Python, Requests, BeautifulSoup，Selenium 等框架工具实现）。
- 协助研究中心开发工程师开发和完善自动化舆情新闻信息采集分析系统。

科研成果

2025.04 SQL-R1: Training Natural Language to SQL Reasoning Model By Reinforcement Learning
- NLP & NL2SQL 方向研究长文，第一作者
- 本研究使用强化学习技术进行训练。与以前主要依赖于监督微调的方法不同，SQL-R1 利用强化学习来优化其在 SQL 生成过程中的决策，从而产生更准确的查询，并在复杂场景中更好地泛化。主要创新包括：显式推理过程，数据库执行反馈，GRPO 和合成数据工程
- [Paper] 目前正在投稿 CCF-A 会议中，文章可在 Arxiv 预览。

2024.08 A Plug-and-Play Natural Language Rewriter for Natural Language to SQL
- NLP & NL2SQL 方向研究长文，第一作者
- 本项目基于大语言模型（LLM）开发了一个即插即用的自然语言重写器，通过自动检测和修复用户问题（NL）中的潜在语义性缺陷（如语义歧义、实体错误、表述不完整等），提高现有的 NL2SQL 模型的准确性。本项目的技术栈包括：Prompt Engineering, Self-Reflection, LLM-based Multi-Agent System, Supervised Finetuning。
- [Paper] [Github] 目前正在投稿 CCF-A 会议中，文章可在 Arxiv 预览。

2024.08 A Survey of NL2SQL with Large Language Models: Where are we, and where are we going?
- NLP & NL2SQL 方向综述长文，第四作者 & 第一硕士作者
- 本项目为 NL2SQL 系统发展综述，提供了近 20 年来该领域最全面的技术总结和方向洞察，能够让读者轻松跟踪文献中最新的 NL2SQL 技术，并为研究人员和从业者提供实用指导。
- [Paper] [Github] 文章可在 Arxiv 预览。目前投稿至 IEEE Transactions on Knowledge and Data Engineering (TKDE，CCF-A 期刊) 接受评审。

2023.07 Development and Validation of a Deep-broad Ensemble Model for Early Detection of Alzheimer's Disease
- ML & 医学影像分析应用方向研究长文，第一作者
- 本项目基于三维卷积结构和宽度学习算法（BLS）开发了一个改进的宽度-深度集成模型，并将其应用于阿尔兹海默症医学影像早期检测任务中。该模型能够在无需预训练的情况下有效提取三维医学影像中的复杂特征，并在低计算成本的情况下比其他方法的检测效果更好。
- [Paper] [Github] 文章已发表于 Frontiers in Neuroscience（JCR Q2 期刊）。

项目经历

2023.12 – 2024.12 StorytellingAgent
- 方向 & 技术栈：HCI、Prompt Engineering、Multi-Agent System
- [Github] 本项目致力于构建一个基于时间循环机制的交互式冒险游戏。该游戏还包括一个基于角色扮演技术、大语言模型驱动的多智能体 NPC 系统。玩家必须在有限的时间内解决谜题，与游戏中的智能体 NPC 角色进行深度的交互，从而找寻出逻辑链，完成拯救同学的最终任务。

服务经历

2024.08 – 2024.08 学生志愿者 @ 50th International Conference on Very Large Databases (VLDB'25)
- 广东省 • 广州市

2019.09 – 2020.07 学生干事 @ 暨南大学学生会
- 广东省 • 广州市