吴烜圣

手机： +01 (213)-376-9286 **| 邮箱：** wuxsmail@163.com **| 主页:** jacksonwuxs.github.io

教育经历

2021.08 - 2025.12 博士 University of Georgia 计算机科学 校级奖学金1次

研究方向：LLM 解释性[1, 2, 3, 5], LLM 评估[1, 3, 4, 5], In-context Learning[2, 3, 6, 8, 10, 12], 推荐系统[2, 7]。

2016.09 - 2020.06 本科 上海对外经贸大学 应用统计学 校级奖学金2次

主要荣誉和奖项

- Google Scholar 论文累计获得 **170+** 引用量 2024.07
- Github 第一作者开源项目累计获得 **800+** Stars 2024.07
- 美国科学基金会（NSF）- 学生旅行奖学金 2024.05
- 腾讯广告算法大赛 - 第5名（**Top 2%**） 2021.08
- 上海市爱建奖学金 - 三等奖 （**Top 2%)** 2020.01
- 中国大学生计算机应用能力大赛 - 二等奖（**Top 4.5%**） 2019.08
- 第一届亚洲大学生射箭锦标赛 – 男子团体**第3名** 2017.04

出版物

[1] Understanding the Behavior Shift in LLMs after Instruction Tuning (一作), **NAACL (*Oral*)**, 2024.
[2] Could Small Language Models Serve as Recommenders (一作), **WWW (*Oral*)**, 2024.
[3] Unveiling Scoring Processes: Dissecting Differences between LLMs and Human Graders (一作), ArXiv, 2024.
[4] InFoBench: Evaluating Instruction Following Ability in Large Language Models, **ACL Findings**, 2024.
[5] Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era, ArXiv, 2024.
[6] Retrieval-Augmented In-context Model Editing for Multi-hop Question Answering, ArXiv, 2024.
[7] DIRECT: Dual Interpretable Recommendation with Multi-aspects Word Attribution (一作), **TIST**, 2024.
[8] Applying Large Language Models and Chain-of-Thought for Automatic Scoring, **CEAI**, 2024.
[9] Black-box Backdoor Defense via Zero-shot Image Purification, **NIPS**, 2023.
[10] MeNSP: Matching Exemplar as Next Sentence Prediction (一作), **AIED (*Oral*)**, 2023.
[11] NoPPA: Non-Parametric Pairwise Attention Random Walk Model (一作), ArXiv, 2023.
[12] A survey of graph prompting methods: techniques, applications, and challenges (一作), ArXiv, 2023.
[13] Rethinking Impacts of Overfitting and Feature Quality on Video Classification (一作), **ACM MM (*Oral*)**, 2021.
[14] Lifestyle-based Approach for Cervical Cancer Screening (一作), **ICDATA**, 2018.
[15] Optimization of Value Average Strategy in China Stock Market (一作), **China Collective Economy**, 2018.

实习经历

三星 - Samsung Ads 研究实习生 2024.06 - 至今

- 提出利用 Reward Model 内部表征在 helpful/harmful 等目标概念的激活度以构建更忠实的LLM自动评估系统。

腾讯 - Tencent AI Lab 研究实习生 2023.05 - 2023.08

- 提出一套通用的LLM解释性方法，应用于一对预训练/微调模型以研究指令微调的影响，成果发表于 NAACL 2024。

百度 - 自然语言处理部 算法工程师实习生 2020.09 - 2021.04

- 提出长文本语义分段任务以减少错误分段信息对下游任务的负面影响，提升FAQ挖掘服务准确率 **7.71%**。
- 设计基于贪心策略的多任务GPU调度算法并应用于FAQ挖掘系统，GPU利用率提高**2.9**倍、QPS提高**4.1**倍。

中国平安 - 金融壹账通 - 大数据研究院 算法工程师实习生 2019.01 - 2019.06

- 提出使用 Feature Context-Free Grammar 解决 Seq2SQL 任务，其支持**34**种额外句式、速度提升**30%**。

融昊投资有限公司 数据分析师实习生 2018.01 - 2018.02

- 统计分析4000万条毫秒级期货数据的分布规律，基于分析结果设计短时交易策略并进行回测。

其他项目经历

复现 Forward-Forward 优化算法 2023.01

- 复现 Geoffrey Hinton 于2022年NIPS会议上发表的Forward-Forward算法，Github收获 **140+** stars。

DaPy: 一个易用的数据分析框架 2017.09 - 至今

- 基于 Python 内置数据结构设计了比 Pandas 更人性化的APIs同时保持近似效率，Github收获 **580+** stars。

分布式射箭比赛管理系统 2017.12 - 2018.12

- 该系统服务于2018中国大学生射箭锦标赛，在线成绩查询峰值访问量**1430**人次，自动生成报表**33**页。

技能和兴趣

- Python、Linux、SQL、SPSS、C、MATLAB。
- 射箭 (1年助教及专业比赛经历)、跑步 (10KM配速5’40’’)、篮球、和家人/女朋友旅游。