方晗骏

电话：(+86)177-3766-3888 | 邮箱：17737663888@163.com

教育背景

北京大学

社会学系-法学硕士 2023.09-至今

- 研究方向：LLM in Social Science、计算社会科学
- 毕业时间：2026.06

社会学系-法学学士 2019.09-2023.06

- 总 GPA：3.72/4.00

学术经历

北京大学智能学院 2023.06~2024.12

大模型价值观项目组成员，宋国杰课题组 北京

- 共同构建用于评估大模型价值倾向和理解能力的**基准数据集 ValueBench（英文）**，从 44 个价值测量问卷中提取并标注了 453 个多维价值维度的数据。完成价值问卷的理论准备和内容搜集工作，负责 10^{+}价值测量问卷的数据整理工作，并对整体的基准数据集质量进行人类评估。
- 提出了创新的价值评估流程方法，利用“seek-suggestion”替代“self-report”构建**提示词模版**，完成了对 LLM 价值取向和理解能力的系统化评估；负责完成 6 个主流 LLM 的价值测试结果的标注，发现 LLMs 在有充分上下文时能达到 80%以上的专家级判断准确率。
- 负责 **GPV**（Generative Psychometrics for Values）**理论框架**的构建工作，基于“选择性感知”理论和社会学评估方法，完成大模型价值评估理论验证。并负责完成对 GPV 框架 Perception 提取工作的评估标注工作。
- 参与团队**基于 LLM 的自动化生成框架**的迭代工作，**主导**构建了可兼容 GPT、Claude 等 2 种市场主流模型，支持多种数据类型，支持提示词模版配置的自动化生成框架，完成“模型-数据-任务”的流程解耦。有效提升长链路（2^{+}steps）数据生成任务效率，承担团队内部分价值数据（1000^{+}）的测量工作。

实习经历

趣加公司（FUNPLUS） 2023.09-2023.11

GenAI 投资实习生，战略投资 北京

- 基于 Python 搭建自动化数据管道，通过 bs4 框架和 Github API 抓取每日的 Arxiv 论文和 Github 热门项目，利用 GPT-3.5 对项目&论文内容进行简单分类和关键信息提取，缩短项目筛选时长，将投研周报制作时间从 3^{+}天压缩到 **1** 天。
- 投研周报提炼多个技术趋势（2023.09 推荐 Suno.ai；2023.10 推荐 Heygen.ai），支撑公司高层快速定位高潜方向，成为公司 AI 投研重要信息参考。

北京大学武汉人工智能研究院 2023.06-2023.08

实习研究员，智能治理研究中心 武汉

- 基于 Schwartz 价值体系、Rokeach 价值体系搭建适用于大模型价值观体系建模的**理论框架**。进行大模型论文整理，收集整理 AI 仿真层、智能治理及价值伦理等最新的研究报告，进行汇总分类、总结报告，在大模型价值观研究在理论层面为项目组提供材料支撑。
- **对比中西方传统价值体系**的不同，发现西方价值体系偏重于分析性，和系统化的组织建构（建立价值和行为的逻辑联系）；以儒家为代表的中国价值体系，更注重实证性，重视伦理道德和个人修养。

会议发表

CCF-A 类会议

- Ren, Y., Ye, H., ***Fang, H.***, Zhang, X., & Song, G. (2024). ValueBench: Towards Comprehensively Evaluating Value Orientations and Understanding of Large Language Models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2015-2040). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.111

（* ACL 2024）

- Ye, H., Xie, Y., Ren, Y., ***Fang, H.***, Zhang, X., & Song, G. (2025). Measuring Human and AI Values Based on Generative Psychometrics with Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence.

（* AAAI 2025）