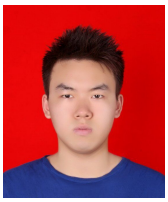
李承奥

电话： 15713816972 | 邮箱： lichengao22z@ict.ac.cn | 现居城市： 北京



教育经历

中国科学院大学-计算机软件与理论 博士 中国科学院 计算技术研究所 2022.09 - 2027.07

- GPA：3.73 / 4.0 加权平均分：85
- 专业课程：计算机算法设计与分析，模式识别与机器学习，高级人工智能，自然语言处理，深度学习

中国科学技术大学-数据科学与大数据技术 本科 少年班学院 2018.09 - 2022.06

- GPA：3.63 / 4.3 加权平均分：86.9
- 专业课程：数据结构，算法基础，人工智能基础，机器学习概论，深度学习导论

实习经历

华为-实习生 华为云计算 2024.07 - 至今

- 基于OpenRLHF框架实现基于梯度正则和多梯度下降的大语言模型的多目标RLHF算法
- 基于PKU-SafeRLHF开源数据集以安全性为主要目标对Pangu-7B-V3模型进行RLHF，保持金融通用能力的同时提高安全性，在金融合规指标上实现提升

浙江盈阳资产管理股份有限公司-实习生 2021.10 - 2022.03

- 利用Python进行数据处理和因子计算，开发基于深度学习的低频股票选股模型，在历史市场交易数据上测试复利达到1.7%
- 分析和优化股票选股策略，包括高低频数据结合和风险控制等。基于强化学习方法建立交易模型，进行模拟交易和回测

研究工作

Gradient-Adaptive Policy Optimization: Towards Multi-Objective Alignment of Large Language Model. ACL 2025

- 提出 GAPO（Gradient-Adaptive Policy Optimization）多目标 RLHF 范式，通过梯度正则和多梯度下降算法解决大语言模型的多样化偏好分布对齐。
- 引入 P-GAPO（Preference-based GAPO），整合用户在不同目标上的偏好，能够得到更符合用户特定需求的帕累托解，进一步提升了大语言模型与多样化人类偏好的对齐效果。

Controlling Large Language Models Through Concept Activation Vectors. AAAI 2025

- 提出轻量级框架 GCAV（Generation with Concept Activation Vector），用于大模型的可控生成。该框架无需对大模型进行微调，即可控制大模型响应内容安全性、情感和主题等概念。
- GCAV 通过收集少量数据训练概念激活向量，并在推理过程中将其注入模型，从而引导生成内容对特定概念进行加强或减 弱。该方法不仅实现可控生成，还支持对单个样本进行调整，实现高效、灵活的定制化文本生成。

《生成式大模型安全评估白皮书》 2024.12

- 参与编写2024年度《生成式大模型安全评估白皮书》。
- 白皮书全面梳理生成式大模型的发展现状与安全风险，从安全评估方法到实践案例，深入剖析技术面临的关键挑战及应对策略。

社团和组织经历

中国科学院计算技术研究所智能信息处理实验室团支部-团组织委员 2023.09 - 至今

- 团支部工作的组织和统筹安排，团员代表大会代表选举，团支部优秀团干与优秀共青团员选举

中国科学院计算技术研究所2022704班-副班长 2023.09 - 2025.07

- 班级内部工作的组织和统筹安排

中国科学院大学计算机学院2022704团支部-团组织委员 2022.09 - 2023.07

- 团支部工作的组织和统筹安排，团支部优秀共青团干与优秀共青团员选举，团支部成员团课学习，发展优秀共青团员入党

荣誉奖项

中国科学院大学 优秀共青团员， 三好学生 2023 - 2025

中国科学技术大学 全国大学生数学竞赛（非数学组）三等奖，暑期社会实践征文比赛 一等奖 2018 - 2022

中国科学技术大学 “拔尖计划”奖学金，优秀学生奖学金 2018 - 2022

技能/证书及其他

- **技能：** Python, C, Mathematica, Office
- **语言：** 英语 TOFEL 95 | GRE 313 (141+169+3) | CET-4 610 | CET-6 522