闫禹杭 (Henry)

YanY_Henry
+86 18945222225
https://yany-henry.github.io
yanyuhang2002@link.cuhk.edu.hk

教育经历

香港中文大学 2021.09 - 2025.07 (预期)
计算机科学理学士 中国，香港

- GPA：**3.757**/4.000（前 10%）；院长嘉许名单（Dean's List）获奖
- 工程学院创新科技工程领袖培训专修（ELITE Stream）学员；善衡书院（S.H. Ho College）成员
- 2023-24 学年春季学期**洛桑联邦理工学院**计算机与通信科学学院（IC School, EPFL）交换
- 2022-23 学年春季学期**北京大学**元培学院（Yuanpei College, Peking University）交换

科研经历

当下生成模型的漏洞评估 2024.02 至今
指导教授：*Prof. Sabine Süsstrunk*（*EPFL*） 瑞士，洛桑

- **"越狱"分析与安全增强**：探索了十余种大型语言模型（LLM）的"越狱"技术，以发现潜在的安全漏洞。这项研究为技术公司提供了支持，使其能够开发更为健壮的防御机制，提高 LLM 的安全性，以及对有害内容生成的抵御能力。
- **防御绕过策略**：运用强化学习和贪婪搜索等方法，在潜在空间内寻找敏感词汇的替代词，以规避 LLM 的对齐过程。这些技术绕过了传统的对齐机制，揭示了 LLM 中关键的安全漏洞。
- **注意力分析与转移**：设计了一系列复杂的输入提示词，并在有害语句的数据集上进行了广泛实验。这有效地分散了 LLM 的注意力，有助于准确地识别其弱点，并设计有针对性的攻击策略。

高效视频分析 2023.06 - 2023.09
指导教授：*Prof. Eric Chi Lik LO*（*CUHK*） 中国，香港

- 在 58 个本科生项目中荣获 **2023 最佳项目奖**（Best Project Award 2023）
- **香港国际机场机场失物招领系统开发**：利用 CLIP 和 OWL-ViT 等模型，应用视觉和自然语言处理技术，构建了多模态失物招领系统，通过图像和文字信息检索大量监控视频数据，提升了机场服务的智能化水平。
- **数据集构建与评估标准建立**：创建了机场视频数据集，并建立了多模态算法评估标准。筛选出高性能的多模态 Zero-Shot 目标检测和自然语言处理模型，提升了系统处理效率，将机场失物招领系统的处理时间节省 5-8 小时。
- **概率片段检测算法**：开发了用于识别失物关键帧的概率模型，提高了系统的分析效率。在确保高准确率和高召回率的条件下，使系统在处理大规模视频素材时更为迅速。

校园网网络流量特征分析 2023.04 - 2023.06
指导教授：*Prof. Tong YANG*（*PKU*） 中国，北京

- **流量特征分析平台构建**：开发了高性能流量分析平台，细致分析校园网络用户的行为模式，包括链路、流和包级别的网络流量特征分析，了解校园网络用户的行为模式。
- **TCP/IP 和 DNS 数据包分析**：利用 C++ 编程，从校园网络流量数据中提取并解析 IP 地址到域名的映射关系，重点关注 DNS 数据包结构，为发现网络异常行为和计算用户使用偏好提供了技术手段。
- **网络流量分析**：分析网络使用模式，包括高请求时间和高流量站点等。识别了具有最高流量的前 K 个域名，深入分析校园网络用户的偏好和行为习惯，为网络管理和优化提供了重要参考。

工作经历

AIJobTech 职能科技 2023.10 至今
创始人兼首席科技官（*CTO*） 中国，香港

- **初创公司的启动和融资**：参与公司的创立阶段，负责早期资金的筹集和团队组建。通过开发 AI 推荐系统算法，实现了针对不同背景求职者的精准职位匹配，为公司的快速发展奠定了基础。
- **利用大语言模型进行简历润色**：带领团队部署大型语言模型，利用其对不同岗位信息的理解能力，对求职者的简历和个人陈述进行润色优化。这一举措构建了一个全方位的求职平台，极大提升了求职者的成功率和用户满意度。
- **领导技术团队开发**：执掌技术团队，主导开发了基于人工智能驱动的职位匹配平台。着重关注技术细节，确保工作高效推进，并不断优化产品质量，提升市场竞争力。

获奖经历

院长嘉许名单（Dean's List） *CUHK 工程学院*
2023 本科生暑期最佳科研项目奖 *CUHK 工程学院*
美国大学生数学建模竞赛 H 奖（Honorable Mention） *COMAP*
2022-23、2023-24 书院杰出学生奖学金（10,000HKD） *CUHK 善衡书院*
本科生杰出交换奖学金（35,900HKD） 香港中文大学
全国高中数学联合竞赛、全国中学生物理竞赛省级二等奖

工作技能

语言 英语 (流利)、普通话 (母语)、粤语（中级）
编程 C/C++、Java、JavaScript、MATLAB、Python{PyTorch, TensorFlow}、Scala、SQL
其它 LaTeX、Anaconda、Git、Linux、Spark