

孙一丁

+86 15522420639 emanual20.sun@gmail.com [Github: Emanuel20](#)

教育经历

中国人民大学 2022.09 – 2025.06
人工智能 硕士 高瓴人工智能学院 GPA: 3.85/4.0 北京
南开大学 2018.09 – 2022.06
计算机科学与技术 本科 计算机学院 GPA: 3.84/4.0 专业排名: 6/122 天津

科研经历

Yulan大模型 2023.02 – 至今
基石模型数据组负责人 高瓴人工智能学院, 中国人民大学

- 为了解决大模型社区缺乏统一预训练数据清洗框架的问题, 搭建数据通用清洗框架ZHEM。该框架实现数据过滤、清洗、去重、检索、评价等模块, 用户可以根据语料特点自由组合预定义模块进行质量提升。经ZHEM清洗后数据训练的 GPT-2 表现与 OpenAI 发布的版本可比, 在 GPT-4 质量标注和 ppl 指标下同样优于清洗前的数据。相关成果以第一作者拟投稿 CCF-A 类会议的 **Demonstration Paper Track**。
- 作为主要作者(数据组负责人)参与 Yulan-Chat-2 项目 [\[Code\]](#), 以 LLaMA-2 为基座模型扩充中文词表与上下文长度, 并使用高质量中英文双语指令进行指令微调, 在 LLaMA-2 的基础上提升了模型的中英文基础语义和理解能力。相较于同期基于LLaMA-2微调的模型, Yulan-Chat-2 具有显著性能优势。该项目在 Github 已获得 320 Stars。
- 作为主要作者参与 Yulan-LLM 项目, 从头预训练 12B 的中英双语基石模型。项目使用ZHEM清洗后的高质量中英文双语数据进行预训练, 并利用检索器链接外部知识库知识增强的方式进行继续预训练。通过高质量双语指令微调, 使模型在 MMLU/C-EVAL 等主流大模型评测基准上与同规模模型达到可比水平。

实习经历

美团外卖推荐算法组 - 推荐算法实习生 2022.02 – 2023.12

- 为了缓解美团线上解释和推荐模型评分与物品特征对齐不显著的问题, 提出了一种基于最大化解释和评分/物品特征互惠信息的强化学习框架改善了基石模型的性能。相关成果以第二作者在投于 CCF-A 类会议。
- 为了解决现有可解释推荐中评价解释过度依赖模拟场景下用户实验的问题, 提出了一种包括三个自动化指标的解析评价框架, 通过在线日志数据实验与用户实验, 验证了提出指标的区分和诊断能力。相关成果以第二作者拟投稿 CCF-A 类期刊。
- 分析线上日志, 开展用户实验, 验证用户对线上推荐解释标签的感知能力, 从多方面分析用户对解释标签的偏好, 就改进标签线上展出策略提出建议。相关成果以第一作者在 **QUARE 2022 (CCF-A workshop)** 上报告。
- 设计了一种综合考虑有用性和真实性的线上推荐解释标签的定量评价框架, 在线上日志数据上验证了评价框架的有效性。相关成果以第二学生作者身份在 **QUARE 2022 (CCF-A workshop)** 上报告。

项目经历

nkucs.icu 经验交流平台 [\[Code\]](#) 2020.01 – 至今

- 为了打通学校信息交流壁垒, 作为网站创始人使用 Github Pages + Docsify 轻量解决方案搭建并长期维护平台。以南开大学师生为用户群体, 累计访问量达 20k 人次。该项目在 Github 已获得 97 Stars, 23 Forks。

NKU-TEDA ARM编译器 [\[Code\]](#) 2021.04 – 2021.08

- 在2021年全国大学生计算机系统能力大赛编译系统设计赛(华为毕昇杯)中作为主要参与者, 在Raspberry Pi上使用 C++ 实现类 C 语言从词法分析、语法分析、中间代码生成及优化和目标代码生成的完整编译过程。
- 设计基于 llvm 的 SSA 中间代码形式思想进行代码优化, 实现数据流分析、循环展开、自动向量化等优化Pass, 在某些性能测试样例上达到可比gcc-O2的效率。该项目获得比赛全国二等奖。

获奖情况

中国人民大学研究生新生奖学金, 二等学业奖学金 2022, 2023
南开大学本科优秀毕业生 2022
中国大学生程序设计竞赛 (CCPC-2019) 秦皇岛站优胜奖 2020