# 佘帅杰

shesj@smail.nju.edu.cn · (+86) 188-5116-0388 ·

## 教育背景

**南京大学** 2022 – 至今

在读直博二年级 计算机科学与技术系，导师黄书剑
发表 AAAI2023 CCF-A 论文一篇。一年级获得国家奖学金，南京大学 2023 年度优秀研究生。
预计 2027 年 6 月毕业

**南京大学** 2018 – 2022

学士 计算机科学与技术
人民奖学金一等奖一次，人民奖学金三等奖三次

## 科研经历

**生成式摘要的事实一致性评估** 一作，发表于 AAAI-23

本研究提出了一种利用 prompt 来控制模型偏好来检测生成式摘要中的事实不一致框架 (CoP)。该方法通过分离无关偏好，不需要训练就可以精确的检测出事实不一致。此外还可以衡量特定类型的偏好以检测具体不一致类型。我们还探索了结合 prompt tuning 来强化偏好，高效的从少量真实数据中学习。我们在三个不一致检测任务上取得了 SOTA 结果，证明了方法的有效性。该论文发表于 AAAI2023 (CCF-A)，并在大会上做 Oral presentation。论文地址

**探究大模型的对话理解一致性** 一作，2023 年

本研究基于现有的对话摘要数据集提出了一个新的数据集，用于评估主流大模型的对话理解一致性。分析结果表明现有的大模型的对话理解一致性依旧存在比较严重的缺陷，特别是在主客体的理解以及幻觉问题。为了解决这个问题，我们提出了一种多任务伪数据构建策略，并对大模型进行微调。实验结果表明我们构造的数据有效的提升了模型的对话理解能力，在我们提出的 DIAC-QA 上取得了 11% 的错误率改进。论文 NAACL 在投，已公开 arxiv。

**思维链能力和偏好对齐研究** 第四范式算法实习生，2023 年 4 月

在第四范式实习期间，继续大模型思维链推理能力以及大模型对齐相关研究，有成熟的使用 DeepSpeed，PEFT 等相关工具监督训练多个 7-33B 中文、英文模型的经验，以及有比较丰富使用 PPO 和 DPO 成功对齐 7B 模型的经验，目前有一篇相关的计划在 ACL2024 的投稿。

## 项目经历

**开源和学术社区贡献**

积极在开源社区贡献自然语言处理相关的代码，github 主页：Ricardokevins。目前共计在开源社区 github 获得了 700+ stars。公开的训练模型权重获得超过 6000 次下载。担任 EACL23, ACL23, EMNLP24 的审稿人，参加了 NIPS22, ICLR23, IJCAI23 的审稿工作。

**2022 阿里天池全球人工智能技术创新大赛**

阿里天池全球人工智能技术创新大赛赛道三: 短文本语义匹配比赛。在团队中主要承担算法研发和实现，使用了多种预训练语言模型。采用 MLM 预训练，知识蒸馏，自蒸馏，对抗训练，多模多折融合等技术。最终排名 44/5345 (Top 1%)

## 其他

英语四级 615 分，六级 594 分
2022 美国大学生数学建模 M 奖，2022 全国大学生数学建模江苏省一等奖