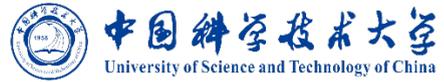




# 潘浩文

18575606474

phw1129@mail.ustc.edu.cn



## 教育经历

中国科学技术大学	数据科学与大数据技术 本科	2019.9 - 2023.6
成绩	3.54/4.3	
中国科学技术大学	大数据技术与工程 硕士	2023.9 - 至今
成绩	3.78/4.3	毕业时间: 2026年6月

## 技术能力

- 编程能力:** 熟练使用Python编程, 熟悉PyTorch, Transformers等深度学习框架; 熟练部署大模型与多模态大模型; 熟练使用Pandas等数据处理工具; 熟练使用Linux命令行。
- 科研能力:** 掌握深度学习基础知识, 关注前沿技术与最新发展。能够独立完成论文调研, 代码编写, 论文撰写等基础事务。对于科研成果与实际需求相结合有着浓厚的兴趣。

## 科研经历

多模态大模型的可解释性研究	第一作者	2023.6 - 2024.2
<b>Finding and Editing Multi-Modal Neurons in Pre-Trained Transformers (ACL 2024)</b>		
<ul style="list-style-type: none"><li>为多模态大模型提出一种定位关键神经元的方法, 在不使用梯度的情况下计算每个神经元对输出的贡献分数, 以此筛选出特定输出所对应的神经元, 并用于解释模型是如何建立视觉与文本信息之间的联系, 在定量指标下该方法显著优于先前工作。这些神经元被实验证实具有三个特性: 敏感性、特异性和因果性, 同时它们可以被用于类似知识编辑等下游任务中。</li></ul>		
大模型的知识编辑研究	第一作者	2024.4 - 2024.10
<b>Precise Localization of Memories: A Fine-grained Neuron-level Knowledge Editing Technique for LLMs (ICLR 2025)</b>		
<ul style="list-style-type: none"><li>为大模型提出一种细粒度的、在神经元层面的知识编辑方法, 用于解决现有“先定位后编辑”方法中过于关注待编辑主语的问题。该方法源于上篇定位神经元工作的扩展, 通过精准定位知识在模型中的对应的神经元并对其进行编辑操作, 性能在现有“先定位后编辑”方法中实现SOTA, 速度提高4~6倍。</li></ul>		

## 比赛经历

<b>Kaggle: U.S. Patent Phrase to Phrase Matching</b>	(91/1889 银牌)	2022.4 - 2022.6
<ul style="list-style-type: none"><li>训练多个Bert类模型, 模型中添加Bi-LSTM头, 使用BCE、MSE、Pearson loss等多个损失函数, 增强模型的多样性, 最终集成上述模型预测结果</li><li>对不同anchor下的target进行分组, 同一anchor的target添加进上下文中, 增强二者间的相关性</li></ul>		
<b>Kaggle: LLM - Detect AI Generated Text</b>	(41/4359 银牌)	2023.11 - 2024.1
<ul style="list-style-type: none"><li>利用GPT2-Large计算文本token困惑度以及token在模型概率分布中的rank来构造文本特征(GLTR)</li><li>采用TF-IDF辅助构造文本特征; 加权集成若干分类模型并对特征剪枝以加速推理</li></ul>		
<b>Kaggle: Chatbot Arena Human Preference Predictions</b>	(179/1688 铜牌)	2024.5 - 2024.8
<ul style="list-style-type: none"><li>采用LoRA微调LLaMA3作为奖励模型, 并改进ELO损失函数以接受平局偏好对</li><li>采用QLoRA微调Gemma2做三分类模型, 并做数据增强和测试时增强(TTA); 最终集成上述模型</li></ul>		

## 实习经历

科大讯飞股份有限公司	助理研究算法工程师 (核心研发平台)	2022.9 - 2023.3
<ul style="list-style-type: none"><li>搭建基于人类反馈增强的知识对话系统。通过引入人类偏好对话数据集, 训练Reward Model学习人类偏好信息, 引导模型生成更人性化、知识性更强的对话回复, 增强知识对话系统的可控性。</li></ul>		

## 获奖经历

校优秀学生奖学金铜奖、银奖, 地奥奖学金	2020 - 2022
----------------------	-------------