曹邹颖

电话: 18362155818
邮件: zouyingcao@sjtu.edu.cn
Github: github.com/zouyingcao
主页: zouyingcao.github.io
谷歌学术: zouyingcao

"I am an AI researcher. My wish is: efforts can be in proportion to gains!"

教育经历

2023.09-至今 **学术型硕士，计算机科学与技术专业，上海交通大学**
研究兴趣：大语言模型的效率与性能优化研究，绩点：3.9/4.0，导师：赵海教授

2019.09-2023.06 **本科，计算机科学与技术专业，东南大学**
东南大学优秀毕业生，国家奖学金，校长奖学金，绩点：3.96/4.0, 专业排名：5/95

实习经历

2025.06-至今 **阿里巴巴通义实验室**，暑期实习生，研究内容: 参与ReMe框架的开源工作，围绕 LLM Agents 的经验 (程序化记忆) 系统，探索优化“经验收集、经验复用和经验管理”三大模块。

2024.07-2025.06 **阿里巴巴淘天集团**，科研实习生，研究内容: 聚焦于 LLM Agents 的 Planning 能力，探究了伪代码格式相较于自然语言格式在构建 Agent Plan 上的优势，一作论文PGPO已被 ACL2025 接收。

科研经历

1. **PGPO: Enhancing Agent Reasoning via Pseudocode-style Planning Guided Preference Optimization**.
曹邹颖; 王润泽; 杨逸飞; 马欣贝; 朱晓勇; 郑波; 赵海.
ACL-2025 (CCF-A) | 代码：zouyingcao/PGPO | 论文：arxiv.2506.01475
贡献: 我们聚焦于 LLM Agents 的 Planning 能力，实验发现将 Agent Plan 设计为伪代码格式相较于传统自然语言格式更具优势，因其结构化与简洁性更能提高 LLM Agents 的任务泛化能力，减少平均环境交互步骤。进一步，我们设计了两个面向规划的奖励信号，提出一种基于伪代码格式规划的智能体偏好优化方法 PGPO，增强了任务推理能力。

2. **SCANS: Mitigating the Exaggerated Safety for LLMs via Safety-Conscious Activation Steering**.
曹邹颖; 杨逸飞; 赵海.
AAAI-2025-Oral (CCF-A) | 代码：zouyingcao/SCANS | 论文：arxiv.2408.11491
贡献: 我们通过分析模型隐层状态在面对有害查询时如何变化，挖掘出模型表征空间的安全防御机制。设计了 SCANS 方法，首先提取代表拒绝行为的表征转向向量，并发现中间层为影响模型拒绝行为的特定安全层，从而利用转向向量在这些层干预模型的激活，达到缓解过度安全问题与保持模型基本安全性之间的权衡。

3. **LESA: Learnable LLM Layer Scaling-Up**.
杨逸飞; **曹邹颖**; 马欣贝; 姚杳; 覃立波; 陈志; 赵海.
ACL-2025 (CCF-A) | 代码：yangyifei729/LESA | 论文：arxiv.2502.13794
贡献: 通过将每一层的参数拼接起来应用 SVD，在降维空间观察到层与层之间的权重存在连续性等潜在模式，这表明层间参数的可学习性。因此，LESA 方法通过一个轻量级神经网络来预测插入到相邻层之间的参数，得到了更好的扩展模型层数方法，有效地提高了扩展后模型继续预训练的收敛速度。

4. **Head-wise Shareable Attention for Large Language Models**.
曹邹颖; 杨逸飞; 赵海.
EMNLP-2024 (清华推荐 A / CCF-B) | 代码：zouyingcao/DirectShare | 论文：arxiv.2402.11819
贡献: 探索了大语言模型在注意力头部之间进行权重共享的可行性，提出两套 head-wise 权重共享方案，DirectShare 无需后训练直接进行注意力头间的权重共享，PostShare 则是通过后训练的方式完成权重共享，两者在时间和性能方面相互补。

5. **LaCo: Large Language Model Pruning via Layer Collapse**.
杨逸飞; **曹邹颖**; 赵海.
EMNLP-2024 (清华推荐 A / CCF-B) | 代码：yangyifei729/LaCo | 论文：arxiv.2402.11187
贡献：聚焦于 Transformer layer 级的权重压缩，基于相邻层的权重相似性发现，搜索了一种将后几层权重融合进前一层后剪枝的非训练压缩策略，达到近 30% 的压缩比例下，仍可保持大语言模型的 80% 平均性能。

6. **AutoHall: Automated Hallucination Dataset Generation for Large Language Models**.
   曹邹颖; 杨逸飞; 赵海.
   TASLP (清华推荐 A / CCF-B) | 代码：zouyingcao/AutoHall | 论文：arxiv.2310.00259
   贡献: AutoHall 方法基于已有的大规模事实核查数据集，自动化地快速收集大语言模型事实性幻觉数据，缓解手工标注的压力。同时，基于该数据集验证了提出的基于一致性估计的幻觉检测方案的有效性，并针对当前主流的大语言模型，分析了各自的事实性幻觉属性。
7. **KVSharer: Efficient Inference via Layer-Wise Dissimilar KV Cache Sharing**.
   杨逸飞; 曹邹颖; 陈麒光; 覃立波; 杨东杰; 赵海; 陈志.
   ICLR-2025-审稿中 | 代码：yangyifei729/KVSharer | 论文：arxiv.2410.18517
   贡献: KVSharer 方法基于一个反直觉的现象“在层间共享相似性较低的 KV Cache 更能保持模型性能”，能够减少 30% 的 KV Cache 计算以及内存开销，而对模型性能影响不大，同时还能实现至少 1.3x 的生成加速。

荣誉奖励

2025-09 **2025 年度杨元庆教育基金-优秀硕士生奖学金 (前 0.3%)**
2024-11 **2024 年度华泰证券科技奖学金 (前 0.5%)**，上海交通大学硕士研究生学业奖学金 (一等)
2023-06 东南大学优秀毕业生，东南大学 2023 年优秀本科毕业设计 (论文) (前 3%)
2022-10 2022 年度华为“智能基座”奖学金
2021-12 **2020-2021 学年度本科生国家奖学金 (前 2%)**
2021-05 2021 年全国大学生英语竞赛 (NECCS) C 类二等奖
2020-2022 连续三学年的东南大学“三好学生”
2020-12 **2019-2020 学年东南大学“校长奖学金”(前 0.9%)**，2019-2020 学年东南大学“社会工作优秀奖”
2020-05 东南大学 2019 年度“优秀团干”，第十七届东南大学大学生程序设计竞赛-三等奖

学校项目

2022.09-2022.12 **计算机系统综合课设《嵌入式计算机系统 Minisys-1A SoC》** | 代码：zouyingcao/minisys-1A
团队组长，负责 *MiniSys-1A* 流水型 *CPU*，*CP0*（中断与异常），*Verilog* 语言

2022.04-2022.05 **软件工程课设《河马先生酒店平台》** | 代码：zouyingcao/hotel_bankend
一个支持酒店预订和酒店管理的在线网站，负责后端部分，技术栈：*SpringBoot & MySQL*

2020.11-2021.11 **校级创新训练项目《基于 BiLSTM 的动态行驶时长估计模型》** | 代码：zouyingcao/ETA_Project
该项目基于 *BiLSTM* 模型改良，设计了一种数据驱动的深度学习模型完成行驶时长的估计.

国家专利

2023-09-26 **基于多方公平的即时配送派单系统** | 公开号：CN116805227A
2023-05-09 **一种面向移动端的手势识别和跟踪方法及系统** | 公开号：CN116092178A

其他

学术兼职 ARR (ACL Rolling Review)、AAAI 学术会议审稿人
专业技能 CET6-568 / CET4-643, 熟悉 Python, C++, 数据结构和算法
学生工作 2023.09-至今 上海交通大学电院 23M 硕士生计算机第一党支部书记
2021.08-2023.07 东南大学计算机科学与工程学院、软件学院、人工智能学院 2021 级本科生班指导
2020.09-2023.07 中国红十字会南京分会会员
2019.09-2020.09 团支部书记 & 学生会干事，东南大学
志愿服务 本科新生迎新志愿者, 学校急救志愿者, 抗击疫情志愿服务活动志愿者
兴趣爱好 绘画, 音乐, 美食