杨子逸

手机：19924680829
邮箱：yangzy39@mail2.sysu.edu.cn
个人主页：https://yangzy39.github.io
研究方向：模型融合、自进化强化学习

教育背景

中山大学　计算机学院 2023.09 - 2026.06
硕士研究生　计算机技术　导师：权小军教授

中山大学　计算机学院 2019.09 - 2023.06
本科　计算机科学与技术

实习经历

通义实验室　NLP 文档智能团队 2025.05 - 至今

- **业务模型优化**：主导 Qwen-Doc 模型的 RL 流程优化，为适配 B 端任务复杂 JSON 字典输出格式，引入自顶向下递归匹配的奖励函数；针对多任务 RL 训练时的“跷跷板效应”，设计动态均衡采样器，根据模型在各任务上的平均奖励动态调整相应任务的采样权重。基于 DAPO 优化算法，在 16 个 B 端业务评测集上实现**平均性能提升 7.8 分**，成功推动 Qwen-Doc 上线，当前模型客户**日调用量达千万级**
- **基础模型研发**：作为核心成员参与 Qwen-Long 基座模型研发，实习初期参与 L1 模型技术报告撰写，该模型性能媲美 o3-mini；现阶段作为核心成员参与 L1.5 系列模型全链路研发，贡献自进化数据合成方案，领域批次优势估计，以及自适应负梯度裁剪策略优化等算法，并搭建了长文本任务模型性能统一评测框架。基于 Qwen3-30B-A3B-Thinking 优化后的模型在 **OpenAI-MRCR 达到 SOTA**
- **前沿技术探索**：针对长文档场景人工标注困难的问题，提出了一个“提问-解答-校验”一体化的单模型多角色**自进化强化学习**框架 SPELL，实现模型无监督自我提升，相关论文投稿 ICLR2026

科研经历

首次提出了**基于偏好优化的隐式模型融合**研究问题，其余研究方向包括大模型自我提升，自适应思考，长上下文强化学习。相关研究成果发表于机器学习顶会 ICLR, 自然语言处理顶会 ACL, EMNLP。

隐式模型融合

- **WRPO: 基于加权奖励偏好优化的隐式模型融合** [ICLR 2025] / [Github] / [HF] / [AI Time]
  针对模型融合中词表与分布对齐困难、效率低下的问题，提出加权奖励偏好优化方法（WRPO），通过让目标模型隐式学习源模型与自身输出的差异，并结合内部奖励加权与渐进式调整策略，有效缓解分布偏差、提升融合稳定性。实验表明，WRPO **显著优于**同规模融合方法，性能**媲美 106 倍参数量的集成模型**，在 AlpacaEval-2 上超越所有源模型；相关成果以第一作者身份发表于 **ICLR 2025**。
- **FuseChat-3.0: 偏好优化邂逅异构模型融合** [ICLR SCI-FM] / [Github] / [HF] / [魔搭社区]
  FuseChat-3.0 是对 WRPO 方法的实践拓展，通过扩大融合数据的领域与规模，并针对数学与代码任务引入**规则验证的数据合成**机制及在损失中引入长度约束项，进一步提升融合模型能力。该方法帮助 Llama-3.1-8B-Instruct 在 14 个基准测试中**平均性能提升 16.8%**，登顶 AlpacaEval-2 与 Arena-Hard 榜单，成为当时**最强 8B 模型**；相关成果以第一作者身份发表于 SCI-FM @ **ICLR2025**。
- **FuseRL: 面向异构模型融合的密集奖励偏好优化** [Preprint]
  FuseRL 核心思想是最大化隐式模型融合中不同源模型回复的利用率，通过在 SFT 和 DPO 过程中分别引入来自源模型的多个回复或偏好对，并结合奖励分数进行加权优化，显著提升模型能力。

大模型自我提升

- **SPELL: 长上下文语言模型的自进化强化学习方法** [Preprint] / [Github]
  在 SPELL 中，单模型通过循环**扮演三种角色实现无监督自我提升**：作为提问者提出问题，作为回答者解答问题，以及作为验证者提供奖励信号，以指导三个角色的协同训练。通过引入**自动课程学习机制**，SPELL 逐步增加提问者出题难度，以适应回答者不断提升的能力。SPELL 在 12 个模型上取得一致提升，尤其帮助**强推理模型 Qwen3-30B-A3B-Thinking 实现 pass@n 平均提升 7.6 个点**。

- **Mutual-Taught: 策略与奖励模型协同适应的互教学习** [ACL Main]

  互教学习（Mutual-Taught）通过策略模型与奖励模型的协同进化实现无监督自我提升：策略模型生成数据以优化奖励模型，而奖励模型则提供更精准的反馈来改进策略模型。

自适应思考

- **ThinkSwitcher: 何时深入思考，何时快速决策** [EMNLP Findings]

  针对推理模型过度思考的问题，ThinkSwitcher 设计了一种动态思维链切换框架，使推理模型能根据任务难度自适应调整推理模式，在保持复杂任务高准确率的同时**降低了 20% 至 30% 的计算成本**。

项目经历

FuseChat: 基于成对蒸馏与参数合并的对话大模型融合 [EMNLP Main] / [Github] / [mergekit]

- 项目内容：
  - 针对以往多教师蒸馏融合可扩展性差的问题，提出成对教师蒸馏加模型合并的两阶段融合框架
  - 针对现有模型词表对齐方法准确率和效率低的问题，提出基于统计的全局映射矩阵对齐方法
  - 针对模型合并中的参数知识干扰问题，提出基于**权重矩阵参数单元粒度**的模型合并算法 SCE
- 项目成果：
  - 项目 Github 仓库收获超过 **600 stars**，SCE 算法贡献在知名模型合并仓库 **mergekit (6.3k+ stars)**
  - SCE 算法近期被美团引入作为 LongCat-Flash-Thinking 模型 RL 阶段模型融合算法

FuseO1: 推理大模型融合 [Blog] / [Github] / [HF]

- 项目内容：
  - 在 DeepSeek-R1 发布 24 小时内，通过 SCE 算法将 R1-Distill-Qwen-32B, QwQ-32B 和 Sky-T1-32B 进行合并得到 FuseO1-32B 模型，验证 SCE 模型合并算法在推理领域的适用性
- 项目成果：
  - 融合模型 FuseO1-32B 在 AIME、LiveCodeBench、GPQA 等多个主流推理评测集上超所有合并前的模型，成为当时**最强的 32B 推理模型**，AIME24 性能**超越 OpenAI o1-mini**，接近 OpenAI o1
  - 模型发布 3 天内登上 **HuggingFace 首页**，总下载量**超 10 万次**

QwenLong-L1：推理模型长上下文强化学习 [Paper] / [Github] / [HF] / [Daily Papers]

- 项目内容：
  - 通过强化学习算法增强模型从长上下文定位相关知识并进行多步复杂推理的能力
  - 针对长上下文强化学习中训练不稳定的问题，提出渐进式上下文扩充和难题回顾采样策略
- 项目成果：
  - QwenLong-L1-32B 在 7 个长文档问答任务中**超过 o3-mini**、**Qwen3-plus**, 媲美 Claude-3.7-Thinking
  - 项目 Github 仓库收获近 **300stars**，被机器之心、量子位等公众号报导

发表论文

[1] **Ziyi Yang**, Fanqi Wan, Longguang Zhong, Tianyuan Shi, and Xiaojun Quan. Weighted-reward preference optimization for implicit model fusion. **ICLR 2025**

[2] **Ziyi Yang**, Fanqi Wan, Longguang Zhong, Canbin Huang, Guosheng Liang and Xiaojun Quan. FuseChat-3.0: Preference Optimization Meets Heterogeneous Model Fusion. SCI-FM @ **ICLR 2025**

[3] **Ziyi Yang**, Weizhou Shen, Chenliang Li, Fanqi Wan, Ming Yan, Xiaojun Quan, and Fei Huang. SPELL: Self-Play Reinforcement Learning for evolving Long-Context Language Models. **ICLR 2026**, under review

[4] Tianyuan Shi, Canbin Huang, Fanqi Wan, Longguang Zhong, **Ziyi Yang**, Weizhou Shen, Xiaojun Quan, Ming Yan. Mutual-Taught for Co-adapting Policy and Reward Models. **ACL 2025**, main

[5] Longguang Zhong, Fanqi Wan, **Ziyi Yang**, Guosheng Liang, and Xiaojun Quan. FuseRL: Dense Preference Optimization for Heterogeneous Model Fusion. **ICLR 2026**, under review

[6] Fanqi Wan, Longguang Zhong, **Ziyi Yang**, Ruijun Chen, Xiaojun Quan. FuseChat: Knowledge Fusion of Chat Models. **EMNLP 2025** Main

[7] Guosheng Liang, Longguang Zhong, **Ziyi Yang**, Xiaojun Quan. ThinkSwitcher: When to Think Hard, When to Think Fast. **EMNLP 2025** Findings

[8] Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, **Ziyi Yang**, Ji Zhang, Fei Huang, Jingren Zhou, Ming Yan. QwenLong-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning. Tech report