

# 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、长上下文



## 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 - 2025.12
• 业务模型优化：主导 Qwen-Doc 数据挖掘模型 RL 优化，面向 B 端复杂 JSON 字典输出，设计递归字典匹配细粒度奖励函数；针对多任务训练“跷跷板效应”，引入动态均衡采样、checkpoint 合并与分阶段课程学习。基于 Qwen-Doc-SFT，RL 后模型在 16 个 B 端评测集平均提升 7.8 分，并上线阿里云百炼平台，API 日调用量达千万级		
• 基础模型研发：深度参与 QwenLong-L 系列长上下文 RL 项目，贡献多智能体自进化数据合成链路以及长文本任务统一评测框架；负责 RL 训练，搭建多阶段渐进式上下文扩展训练框架，引入领域均衡采样、领域批次标准差优势估计与自适应负梯度裁剪策略优化（AEPO）等算法优化。QwenLong-L1.5-30B-A3B 在 OpenAI-MRCR 达 SOTA，综合性能达到开源 SOTA，媲美 Gemini-2.5-Pro		
• 前沿技术探索：面向长上下文任务高质量数据稀缺痛点，提出单模型多角色自进化强化学习框架 SPELL，实现无需人工标注的持续自我提升；相关一作论文已被机器学习顶会 ICLR 2026 接收		

## 科研经历

聚焦大模型后训练，首次提出基于偏好优化的隐式模型融合研究问题，并在自我提升、自适应思考与长上下文强化学习方向持续产出；相关成果发表于 ICLR / ACL / EMNLP 等顶级会议。

### 隐式模型融合

- WRPO: 基于加权奖励偏好优化的隐式模型融合 [\[ICLR 2025\]](#) / [\[Github\]](#) / [\[HF\]](#)

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）：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，并在 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 将 WRPO 落地为可复用的融合范式：扩大融合数据的领域覆盖与规模，并针对数学/代码任务引入规则/程序验证的数据合成，同时在损失中加入长度约束，增强融合模型能力。该方法使 Llama-3.1-8B-Instruct 在 14 个基准测试平均提升 16.8%，登顶 AlpacaEval-2 与 Arena-Hard，成为当时最强 8B 模型；成果以第一作者发表于 SCI-FM @ ICLR 2025。

- FuseRL: 面向异构模型融合的密集奖励偏好优化 [\[Preprint\]](#)

FuseRL 围绕最大化利用多源回复重构融合训练：在 SFT 与 DPO 阶段分别引入来自多个源模型的回复/偏好对，并结合奖励分数进行加权优化，以更高的数据利用率显著提升模型能力。

### 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 [\[ICLR 2026\]](#) / [\[Github\]](#)

SPELL 通过单模型三角色闭环实现无监督自我提升：提问者生成高质量问题，回答者完成解题，验证者产出奖励信号，从而驱动三角色协同进化。进一步引入自动课程学习和历史记忆，让问题难度随模型能力同步提升，实现持续而稳定的自我进化训练。实验在 12 个模型上取得一致增益，尤其使推理模型 Qwen3-30B-A3B-Thinking 的 pass@8 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”带来的高成本问题，ThinkSwitcher 提出动态思维链切换框架，使模型可按任务难度自适应选择快思考/慢思考，在保持复杂任务高准确率的同时降低 20%-30% 计算成本。

## 项目经历

### QwenLong-L1.5 : 长上下文推理后训练方案 [Paper] / [Github] / [Daily Papers] / [r/LocalLLaMA]

#### • 项目内容：

- 针对长上下文场景标注昂贵且不可靠导致的高质量数据稀缺，构建结构化多跳问答数据合成流水线，并引入**多智能体自进化**机制提升问题覆盖与难度分布，实现自动化高质量数据合成
- 针对长上下文 RL 训练中策略熵爆炸导致的不可持续训练，提出**渐进式输入/输出扩展**的多阶段 RL 后训练框架，并引入自适应负梯度裁剪策略优化 (AEPO) 以稳定训练过程
- 面向多任务 RL 中奖励增长不均衡与相互干扰，设计**任务均衡采样与任务特定优势估计**策略，实现多任务性能协同提升

#### • 项目成果：

- 基于 Qwen3-30B-A3B-Thinking 在 6 个长文档推理基准平均提升 9.9 分；在 **OpenAI-MRCR** 达到 SOTA，综合性能超过 **DeepSeek-R1-0528**、**Qwen3-Max-Thinking**，媲美 Gemini-2.5-Pro
- 项目 GitHub 仓库收获超 **500 stars**，在大模型社区内受到一致好评

### FuseChat: 基于成对蒸馏与参数合并的大模型融合

[EMNLP] / [Blog] / [Github] / [mergekit]

#### • 项目内容：

- 针对传统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
- 针对现有跨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 **600 stars**，SCE 算法被合入知名模型合并库 **mergekit (6.7k+ stars)**
- 基于 SCE 合并得到 FuseO1-32B 模型一度成为最强 **32B 推理模型**，AIME24 超越 **OpenAI o1-mini**；发布 3 天内登上 **HuggingFace 首页**，累计下载超 **10 万次**
- SCE 算法被美团引入，用于 LongCat-Flash-Thinking 多任务 RL 训练阶段的模型融合

## 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**
- [4] Weizhou Shen, Ziyi Yang (co-first author), Chenliang Li, ..., Fei Huang, Jingren Zhou, Ming Yan. QwenLong-L1.5: Post-Training Recipe for Long-Context Reasoning and Memory Management. Tech report.
- [5] 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
- [6] Longguang Zhong, Fanqi Wan, Ziyi Yang, Guosheng Liang, and Xiaojun Quan. FuseRL: Dense Preference Optimization for Heterogeneous Model Fusion. **ICLR 2026**, under review
- [7] Fanqi Wan, Longguang Zhong, Ziyi Yang, Ruijun Chen, Xiaojun Quan. FuseChat: Knowledge Fusion of Chat Models. **EMNLP 2025** Main
- [8] Guosheng Liang, Longguang Zhong, Ziyi Yang, Xiaojun Quan. ThinkSwitcher: When to Think Hard, When to Think Fast. **EMNLP 2025** Findings
- [9] 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