李远鹏

+(86) 136-6256-6313 liyp2001@outlook.com yuanpeng-li.github.io github.com/Yuanpeng-Li lyp13662566313

求职方向： LLM 后训练（RLHF / RLAIF / GRPO）/ 多目标对齐与奖励优化 / 强化学习算法

教育背景

加州大学尔湾分校（UC Irvine） — **美国加州，尔湾**
统计学博士（*PhD*），在读 — *2024.09 – 至今*

加州大学尔湾分校（UC Irvine）暑期研究 — **美国加州，尔湾**
研究助理（*RA*），统计系；导师：*Annie Qu* 教授 — *2023.06 – 2023.09*

吉林大学 · 数学学院 — **中国，长春**
理学学士，数学与应用数学专业 — *2020.08 – 2024.07*

代表性论文

TOPPO: Rethinking PPO for Multi-Task Reinforcement Learning with Critic Balancing. — 审稿中，*2026*
Yuanpeng Li, Gefei Lin, Annie Qu, Rui Miao. arXiv:2605.11473，第一作者。

科研经历

面向多目标对齐的大语言模型强化学习：梯度冲突感知的公平优化（Probe-FairGRPO） — **美国加州，尔湾**
研究生研究员，加州大学尔湾分校 — *2026.05 – 至今*

- 将正确性、格式、长度等多奖励 GRPO 微调建模为多任务强化学习，把公平梯度合并（FairGrad）引入 LLM post-training；为使其在 LLM 上算得起，提出轻量梯度探针——仅在 lm_head、末层或 LoRA 适配器等小参数子集上构造逐奖励 $K \times K$ Gram 矩阵，低成本估计奖励目标间的梯度冲突与一致性。
- 用 FairGrad 求解器将 probe Gram 映射为逐奖励权重 α（下调冲突目标、上调一致目标）；设计「逐目标独立裁剪后加权」损失，单次标量反传即在 PPO/GRPO 裁剪下保持精确加权梯度，并以单元测试验证其避免「先混合优势再裁剪」的目标偏置。
- 在 verl 0.8 / vLLM / FSDP 上搭建并打通可复现的多目标 GRPO 多卡训练与评测栈（算法核心独立、有单元测试覆盖），配 W&B 诊断；已在 Qwen2.5-0.5B/7B、Llama-3.1-8B-Instruct 上配置 GSM8K / DAPO-Math 训练与 AIME / MATH-500 / OlympiadBench / AMC23 / GPQA 数学评测。

高效多任务强化学习研究（TOPPO） — **美国加州，尔湾**
第一作者；导师：*Annie Qu* 教授、*Rui Miao* 教授 — *2023.07 – 2026.04*

- 提出以价值网络为中心的诊断框架，逐任务记录策略网络 / 价值网络梯度，发现多任务 PPO 的主要瓶颈来自价值网络侧的数值病态（梯度尺度差异、任务梯度共线塌缩、聚合权重不公平），而非单纯策略网络冲突。
- 实现 *Critic Balancing* 训练方法：结合 PopArt 价值归一化、价值网络 LayerNorm、价值网络侧公平加权（FairGrad）与策略网络侧梯度投影（PCGrad）；设计投影牛顿求解器，使 $K = 50$ 任务的每个 mini-batch 梯度加权可实际运行。
- 在 Meta-World MT50 上，以 71.7 万参数达到 $90.88 \pm 1.59\%$ 平均成功率和 $56.50 \pm 7.76\%$ 最差 10 任务成功率；参数量仅为已发表强基线 ARS-LN-1024 的 $1/22.7$，尾部任务表现较 ARS-LoRA 相对提升 27.2%。

项目经历

gradescope-mcp | *Python, Model Context Protocol, uv*

- 构建面向 AI 智能体的 Gradescope MCP 服务器，封装课程管理、提交批改、评分量规（rubric）增删改查与统计分析等 34 个工具；所有写操作均设置人工确认，降低自动化评分风险。
- 沉淀可复用的辅助批改工作流，支持人工审核后的批量评分；实现自定义花名册解析器，并以 30 个自动化测试覆盖核心工具调用与边界输入。

StudyGround | *JavaScript, Node, Claude Code Plugin, Pyodide*

- 开发 AI 辅助自学 Claude Code 插件：在网页阅读器中嵌入 *Ask* 入口，通过 Pyodide 支持浏览器内 Python 运行，并可一键深链至 VSCode 完成练习。
- 实现 NotebookLM 式课程资料 RAG：解析 PDF/OCR 文本，融合 BM25 与向量检索，返回带页码定位的引用，提升长文档学习与溯源效率。

实习经历

Synkrotron — **中国，西安**
实习生，*AI* 研究部 — *2022.12 – 2023.01*

- 参与自动化道路巡检项目，负责远程 Linux 设备接入、网络配置与仿真数据集调研，支持自动驾驶算法测试与真实场景泛化评估。

获奖荣誉

- 2022–2023 学年奖学金 — *2023.10*
- 中国大学生数学竞赛：**全国三等奖** — *2021.12*
- 吉林省大学生数学竞赛：**省级二等奖** — *2021.12*

技能

- **LLM 后训练与强化学习：** PyTorch, Hugging Face Transformers, vLLM, verl, FSDP, PPO/GRPO, RLHF/RLAIF, LoRA/PEFT
- **智能体与检索工具链：** Model Context Protocol (MCP), RAG, BM25, 向量检索, Claude Code Plugin, Pyodide
- **训练基础设施与工程：** Python, Git, Docker, Linux, SLURM, Ray, W&B, uv, pytest, 远程服务器训练与实验管理