李庆尧

邮箱：ly890306@sjtu.edu.cn

主页：simonlqy.github.io ◇ 地点：上海

研究概述

研究主线：让大语言模型在复杂任务与真实环境中学会**推理与决策**——以可执行、可验证的环境为杠杆：同一份**验证信号**，训练时是强化学习的 reward，推理时是搜索的 verifier。问题域上聚焦 **code intelligence** 与 **general agent**，方法上贯通**强化学习**（训练阶段）与 **test time scaling**（推理阶段）。

上海交通大学计算机科学与技术专业直博生（**致远荣誉计划**），师从张伟楠教授与俞勇教授，并于新加坡南洋理工大学联合培养（导师安波教授）。已发表**一作/共同一作论文 8 篇**，覆盖 **ICML**、**ICLR**、**EMNLP（2 篇）**、**ACL**、**KDD**、**CIKM**、**AAMAS**。研究沿两个问题域展开（已发表工作集中于前者，当前重心正推进后者）：

- **Code Intelligence（代码生成、验证与调试）**：代码任务天然可执行、可验证，测试与执行反馈是核心杠杆。一方面构建更强、更难被“骗”的验证器，直面 reward hacking 与“伪正确性”——对抗强化学习训练测试生成器 ATGen（ICLR 2026）、对抗测试搜索 AdverMCTS（ICML 2026）、执行反馈过程奖励模型 CodePRM（ACL 2025 Findings）；另一方面以执行反馈驱动推理时搜索与修正——思维级搜索 RethinkMCTS、自然语言中介调试 NL-Debugging（均 EMNLP 2025）。
- **General Agent 与 Agentic RL（真实环境的长程决策）**：当前在小红书用 **GRPO** 在 **SWE-bench**、**Terminal-Bench 2.0** 等真实执行环境上训练 Code / Terminal Agent（进行中），研究长程任务中 skill 的**选择性调用**与 token 级**信用分配**；博士早期的序贯决策工作——分层强化学习 GEHRL（CIKM 2023）、特权信息蒸馏 PKSD（KDD 2024）——是这条线在 LLM 之前的方法积累。

方法上，强化学习（GRPO 及 PPO / DAPO 等算法）有从算法设计到大规模训练系统（Megatron + SGLang + Ray）落地的一手实践，test time scaling（MCTS / PRM / Best-of-N）有从搜索算法、PRM 训练到完整评测链路的一手实践。当前研究重心：真实执行环境下 Code / Terminal Agent 的 Agentic RL。

教育经历

上海交通大学 *2022.9 – 2027.6（预计）*

博士在读，计算机科学与技术专业（致远荣誉计划）

导师：张伟楠教授、俞勇教授

GPA: 3.91/4.00

新加坡南洋理工大学 *2025.9 – 2026.3*

博士联培，计算与数据科学学院

导师：安波教授

西安交通大学 *2016.9 – 2022.8*

本科，自动化专业（钱学森实验班）；经少年班（2016 级）提前两年入学

GPA: 4.05/4.3（专业排名：1/25）

科研实习

小红书 · ace 团队 2026.3 至今

算法研究实习生 · Code / Terminal Agent 强化学习 导师：焦文祥、陆远

- 研究真实执行环境下 Code / Terminal Agent 的长程 Agentic RL：在 **SWE-bench**、**Terminal-Bench 2.0**、Skills-Bench 等 **5 个 benchmark** 上，用 **GRPO** 训练小型开源 LLM（9B）在真实 Docker Agent Runtime 中学会调用大规模 skill library（约 600 个 skill）、完成长链路终端任务。
- 从零搭建统一的 Agentic RL 训练—评测系统（**Megatron + SGLang + Ray**），核心是保证 reward 信号纯净——verifier 工程决定 RL 能不能训：设计 abort/refill 机制区分基础设施失败与任务失败（避免污染 GRPO advantage）、补齐镜像预构建、重写 verifier 高并发执行，将单条验证由 **300s+ 超时压到秒级**、每步掉样由 **≈50 降为 0**，此后产出项目首条健康的 reward 训练曲线。
- 针对核心问题“长程 agent 如何学会**选择性调用 skill**（skill gate，即通用 agent 的 tool / skill selection 问题）”——发现 selector 与 executor 的**信用纠缠**：任务结果信用会污染选择行为的学习（读对 oracle 但任务失败被错误惩罚、误选干扰 skill（misleading）但侥幸成功被错误奖励），且选择动作 token 在长轨迹中被严重稀释——设计 **token 级信用分离**的 GRPO：selector advantage 仅作用于决定 skill 身份的 action token（首次读取 oracle 为正、误读/重复读取在组内获得相对负 advantage），task advantage 作用于其余执行 token 并排除全部 skill 读取调用，使「选择」只对选了谁负责、「执行」只对任务做得怎样负责（oracle 身份仅用于计算训练期 advantage，对模型不可见；模型需在混合候选中自行判别）。
- 在 **280 任务 × 5 benchmark × 3 评测条件**（oracle / no-skill / mixed-slate）的网格上系统评测：在每任务 **1 个 oracle + 15 个干扰 skill** 的 mixed-slate 设定下，相对同初始化、同评测的仅任务奖励 GRPO 对照，oracle 读取率从 54.3% 提升至 **78.9%**、misleading 误读率从 69.6% 降至 **34.3%**；训练出的 9B 模型在 oracle 与 no-skill 条件下均超过 27B base（oracle 条件 **50.7% vs 41.4%**、no-skill 条件最高 **41.1% vs 37.1%**）。
- 以消融与**梯度 / NLL 探针**系统诊断训练信号，根因定位多类训练崩溃（如违反 on-policy 假设导致约 40 步后崩溃），保障大规模 RL 训练的稳定性。

华为诺亚方舟实验室 2024.7 – 2025.5

研究实习生 · 代码智能：验证器与执行反馈 导师：夏伟、戴心仪、唐睿明

- 围绕执行反馈与验证独立推进代码智能方向研究，实习期间产出 **4 篇一作/共同一作论文**（ICLR 2026、ACL 2025 Findings、EMNLP 2025 两篇）。
- 四篇工作各司其职：测试生成（ATGen）、过程奖励模型（CodePRM）、执行反馈搜索（RethinkMCTS）、执行错误归因调试（NL-Debugging），搭建了从数据构造、PRM 训练到搜索评测的完整实验链路。
- 荣获 **“华为优秀实习生”** 称号。

研究工作与论文

一作/共同一作论文 8 篇（另含领域综述 1 篇），以下按问题域组织：Code Intelligence → 序贯决策与强化学习 → 其他工作。

模块一 · Code Intelligence（代码生成各个部分）

构建验证器 + 执行反馈驱动的搜索与修正

AdverMCTS：双智能体对抗 MCTS，治理代码生成中的“伪正确性” **ICML 2026 · 一作 [Paper]**

搜索的瓶颈不在搜索本身，而在验证信号稀疏且可被过拟合——“过拟合 public test、却在 hidden test 上失败”的伪正确性，正是 reward hacking 在推理时的形态。设计求解器（Solver）与攻击者（Attacker）双树对抗架构：攻击者针对代码池的逻辑分歧持续生成靶向测试，并以硬约束形式注入求解器搜索，使验证环境逐步严苛；同时以 LLM-Arbiter 替代多数投票，将有效测试率从 38% 提升至 **80%**。方法 training-free、可直接作用于闭源与超大模型：将 DeepSeek-V3.2（671B）在 APPS 上的 pass@1 从 49.7% 提升至 **73.3%**；Qwen3-4B 从 38.7%（思维级 MCTS 基线）提升至 **49.7%**。

ATGen：对抗强化学习训练测试用例生成器 **ICLR 2026 · 一作 [Paper]**

训练一个“验证器生成器”：以 IO 准确率与攻击成功率双目标奖励（GRPO 优化）训练测试生成器，并由对抗代码生成器持续构造“能逃过当前测试”的 bug，二者协同进化形成动态难度课程、突破静态数据集的固定难度天花板。产出的测试身兼三职——作为**验证器**：7B 模型 attack rate 达 **36.99%**（UTGen-7B 为 16.24%），发现 bug 能力较 GPT-4-turbo 高 **60%**；作为 **Best-of-N 过滤器**：pass@1 从 30.67 提升至 **35.00**，逼近人类专家测试；作为**下游模型的 RL 奖励**：显著优于其他测试来源，直接打通「验证器 → 训练信号」的闭环。

CodePRM：执行反馈增强的代码过程奖励模型 **ACL 2025 Findings · 一作 [Paper]**

过程级验证器：发现代码思考步骤的质量必须结合其衍生代码与执行结果判断、仅凭文本判据不可靠——执行反馈是可靠过程监督的必要成分。收集（思考步骤、衍生代码、执行反馈）三元监督信号，以通过率作为连续标签（而非二元 0/1）训练 PRM；推理时作为生成—验证—修正（GVR）流程中的过程验证器，实现搜索中的动态纠错，在不同生成预算下稳定优于普通 PRM 等基线。

RethinkMCTS：思维级 MCTS 搜索 + 执行反馈修正错误思考 **EMNLP 2025 · 一作 [Paper]**

执行反馈不只用来打分，而是直接改写推理过程：将 MCTS 从代码空间提升至思维（thought-level）空间展开，更契合 LLM 的语义推理优势；用块级执行反馈直接改写错误思考，使搜索沿修正后的路径继续探索，阻止错误沿路径持续传播。在 HumanEval 上将 pass@1 从 70.12 提升至 **89.02**（GPT-3.5-turbo）、从 87.20 提升至 **94.51**（GPT-4o-mini），token 成本较 Reflection 类方法最高降低 **35.7%**。

NL-Debugging：以自然语言为中间表示的代码调试 **EMNLP 2025 · 共同一作 [Paper]**

回答“拿到验证信号之后，修哪一层”：传统调试直接在代码层修补，容易陷入局部 patch、难以处理深层算法逻辑缺陷。本工作从执行错误反推算法层面的逻辑误解，在自然语言层重写解法草图后再回写代码——在「问题表征」而非「代码符号」上修正，其反馈改写思想与 RethinkMCTS 一脉相承。实验表明自然语言中介调试在复杂算法逻辑缺陷上显著优于代码层修补方法。

模块二 · 序贯决策与强化学习（2023–2024）

分层决策、特权信息与信用分配——*general agent* 的问题雏形

PKSD：特权信息蒸馏——训练期特权特征，部署期无特权策略 **KDD 2024 · 一作 [Paper]**

环境隐状态不可观测且各不相同，导致 RL 训练不稳定：将隐状态作为特权特征（privileged feature）训练强 RL agent，再经知识蒸馏迁移至仅依赖可观测信息的 Adapter，部署时不依赖额外特征即具备自适应能力，显著提升 5 个代表性模型的推荐性能。

GEHRL：长程稀疏奖励下的图增强分层强化学习 **CIKM 2023 · 一作 [Paper]**

面对长程、稀疏奖励的目标导向规划问题：高层模型负责分解整体目标、低层模型基于测试驱动的内在奖励推荐具体知识点，并以知识图谱编码状态、约束合理顺序；在基于真实数据与规则构建的 3 个模拟基准上全面达到 SOTA。其「高层负责选择、低层负责执行」的分解，是我今日研究 selector / executor 信用分配问题的早期形态。

其他工作

推荐系统 × LLM、领域综述

SSRec：结构与知识感知的 LLM 表征用于概念推荐 **AAMAS 2026 EA · 一作 [Paper]**

以 LLM 世界知识构建文本表征引入概念推荐，设计基于知识图谱的文本适配器消除表征各向异性，全面超越 ID-Text / Text-only / ID-only 基线，推荐结果更符合人类认知结构。

Adapting LLMs for Education：大模型教育应用综述 **arXiv 2024 · 一作 [Paper]**

将 LLM 解决教育问题概念化为数学推理、写作、编程、逻辑推理与知识问答五项核心能力的整合，系统总结增强方法论、发展潜力与挑战。

获得奖项

西安交通大学优秀毕业生 *2022.7*

百度大数据竞赛国际优秀奖 *2020.9*

数学建模国家二等奖 *2019.10*

邱昌荣二等奖学金 *2019.9*

专业服务

会议审稿人： ICML 2026、KDD 2024、CIKM 2023、AAAI 2023

专利

融合人类知识结构与学生知识状态的知识点推荐系统（申请中） *2024*

一种心脏起搏器的无线充电器（实用新型专利） *2022.12*