任韬

生日：2001 年 5 月 6 日
手机：13823983901
邮箱：rtkenny@stu.pku.edu.cn

教育背景

哈尔滨工业大学 - 机械工程/人工智能 - 主修/辅修 - 本科 2019.09 - 2023.06

北京大学 - 人工智能研究院/光华管理学院 - 运筹学 - 直博（导师：彭一杰） 2023.09 - Now

研究方向与专业技能

强化学习，后训练，随机优化，随机梯度估计，风险敏感优化

发表论文

Half-order Fine-Tuning for Diffusion Model: A Recursive Likelihood Ratio Optimizer **ICLR 2026 Oral** (一作)

- 提出了“半阶梯度估计量”一种针对 diffusion 的无偏梯度估计方法用于后训练
- 融合视觉多尺度信息，在视频和图像生成模型的后训练阶段实现有效偏好对齐

OVLR: Efficient, and Robust Training via Output-Level Variance-Reduced Likelihood Ratio **ICML 2026** (二作)

- 提出输出级方差缩减似然比（OVLR）框架，通过注入结构化对称噪声降低梯度方差，可直接优化 0-1 损失、截断损失等不可微分目标
- 研究在图像分类、生成建模、语言建模和机器人模仿学习等任务中开展实验

RiskPO: Risk-based Policy Optimization via Verifiable Reward for LLM Post-Training **ICLR 2026** (一作)

- 提出基于风险度量优化的深度强化学习方法用于大模型后训练
- 提升模型复杂问题推理能力，性能超越 GRPO 及其变体

FLOPS: Forward Learning with Optimal Sampling **ICLR 2025** (一作)

- 提出最优采样器，解决了基于蒙特卡洛梯度估计的前向学习算法（BP-free）训练效率低下的问题
- 在大模型黑箱微调和跨模态对齐的任务上，使用最优采样器加速的算法展现出优秀的表现

Exploring and Exploiting Model Uncertainty in Bayesian Optimization **NeurIPS 2025** (共一)

- 提出了基于非参数方法的混合高斯代理模型用于贝叶斯优化
- 在重尾、非平稳的黑箱优化中超越传统的高斯代理模型，在 NAS，大模型提示词场景下取得显著效果

SCOUT: Teaching Pre-trained Language Models to Enhance Reasoning via Flow Chain-of-Thought **NeurIPS 2025**

- 通过逐步蒸馏与跨注意力回溯模块实现递归式、阶段化的潜在推理
- 在八个大模型推理基准上持续提升准确率与解释质量

RiskMiner: Discovering Formulaic Alphas via Risk Seeking Monte Carlo Tree Search
发表在金融工程会议 ACM International Conference on AI in Finance 2024 （一作）

- 提出了一种基于强化学习的因子挖掘算法，在 A 股市场上进行了实验
- 设计了一种结合风险度量的蒙特卡洛树搜索算法求解奖励密集的 MDP 进行因子挖掘

Deep Reinforcement Learning for Solving Management Problems: Towards A Large Management Model
发表至自动化领域会议 IEEE Conference on Automation Science and Engineering 2025 （共一）

- 提出了一种通用的强化学习框架用于解决供应链管理问题（库存管理、定价、推荐）
- 通过预训练决策大模型对库存、定价、推荐的协同优化决策，取得了对基于传统强化学习算法的优势

实习经历

Agentic RL 2026.06 - Now 九坤 IQuest 后训练团队

- Scaling Agentic RL，Autoresearch，Search and Code Agent

Self-Evolving Agent 2025.12 - 2026.05 通义实验室 Qwen-Character 团队

- 通过测试时训练（test time training），将 agent 的记忆内化至模型参数空间中，实现自进化
- 在对话、推理、搜索的数据集上验证算法

具身智能 VLA 训练 2025.03 - 2025.11 清华 AIR/光象科技联合实验室

- 引入高斯泼溅作为 diffusion based vla 的输入模态，提升模型的空间感知能力和泛化能力
- 在 mujoco 仿真环境和松灵 piper 机械臂上部署并验证算法

工作论文

Scaling Self-Evolving Agents via Parametric Memory

Arxiv

- 使用 LoRA adapter 作为参数化记忆，通过特殊的初始化和测试时训练内化模型记忆
- 在 Locomo、Longmemeval、Hotpotqa、CL-Bench 等数据集上进行验证

Optimal Low-Rank Stochastic Gradient Estimation for LLM Training

二作投稿至 Operations Research

- 从随机梯度估计的角度设计了一种最优梯度估计量
- 在语言模型预训练，微调等任务上验证最优梯度估计量用于模型训练的有效性

Omni-Masked Gradient Descent: Memory-Efficient Optimization via Mask Traversal with Improved Convergence

共一投稿至 NeurIPS 2026

- 设计了一种基于不放回抽样的 Mask Gradient Training 的算法，具有更快的收敛率
- 在语言模型预训练、微调和图像分类的任务上超越了传统的高效参数微调方法

Adaptive Robust Estimator for Multi-Agent Reinforcement Learning

投稿至 NeurIPS 2026

- 提出融合双智能体问答 - 评判 - 重写（DACR）交互协议与自适应鲁棒估计器（ARE）的多智能体强化学习框架
- 在数学推理基准和无人机视觉语言导航（VLN）任务中开展实验

Perception Without Engagement: Dissecting the Causal Discovery Deficit in LMMs

投稿至 NeurIPS 2026

- 提出了 PROCAUEVAL，一种基于扰动的评估方案，用于诊断大多模态模型中的因果发现失效问题。
- 提出了 ADPO，一种强化学习框架，通过降低对文本先验的依赖，提升视频因果推理中的视觉定位能力。