# Kimi-Researcher

## End-to-End RL Training for Emerging Agentic Capabilities

June 20, 2025 • 10 min read

## 简介

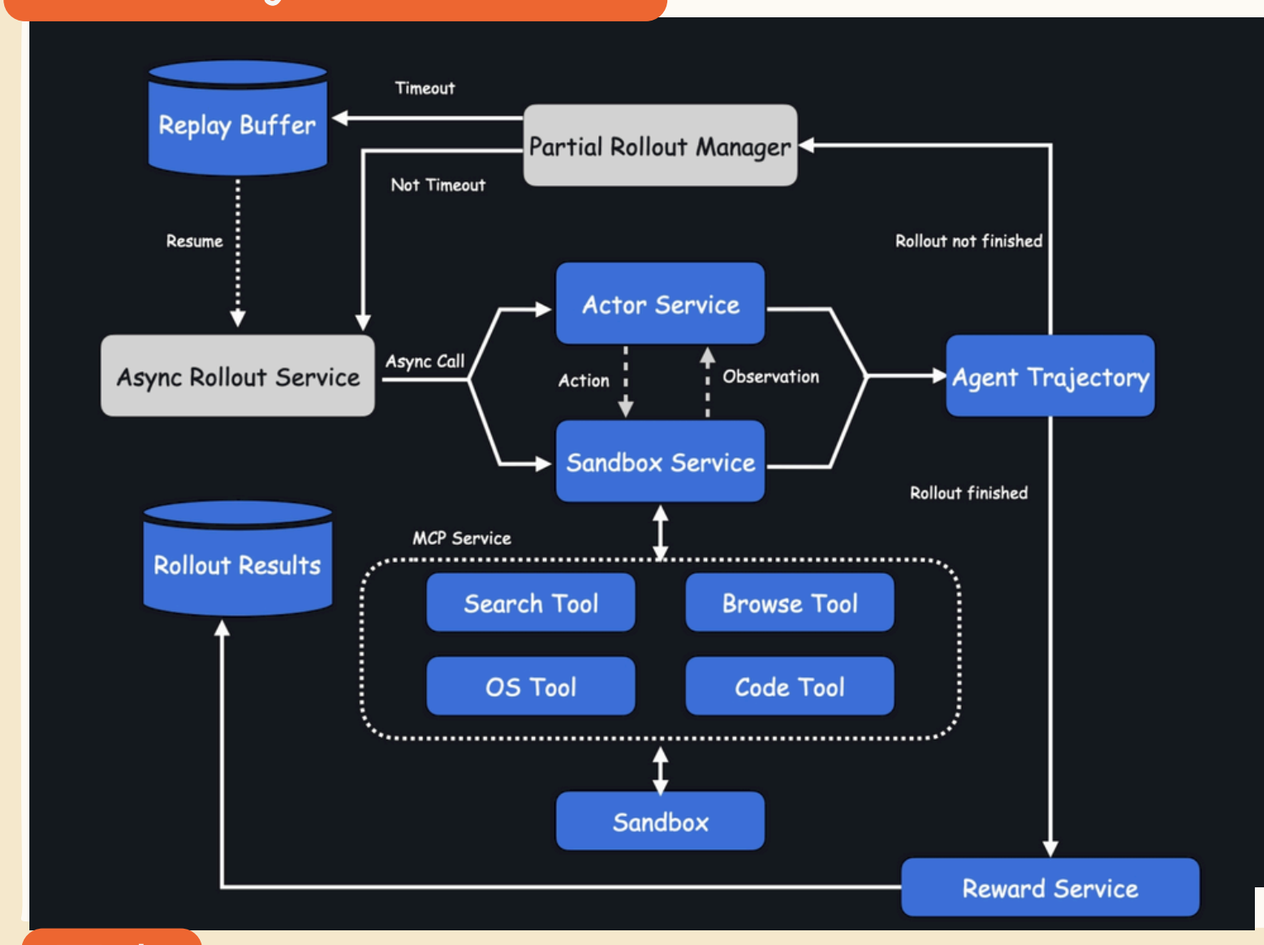
Kimi-Researcher 是Moonshot AI推出的DeepResearch产品，基于其内部的多模态推理模型 Kimi k-series 构建，并采用纯强化学习(RL)范式进行训练。出于对DeepResearch的兴趣，我们尝试结合Kimi 官方技术博客中披露的细节，对其训练数据、训练机制和奖励设计进行学习与分析，看看能有什么启发。

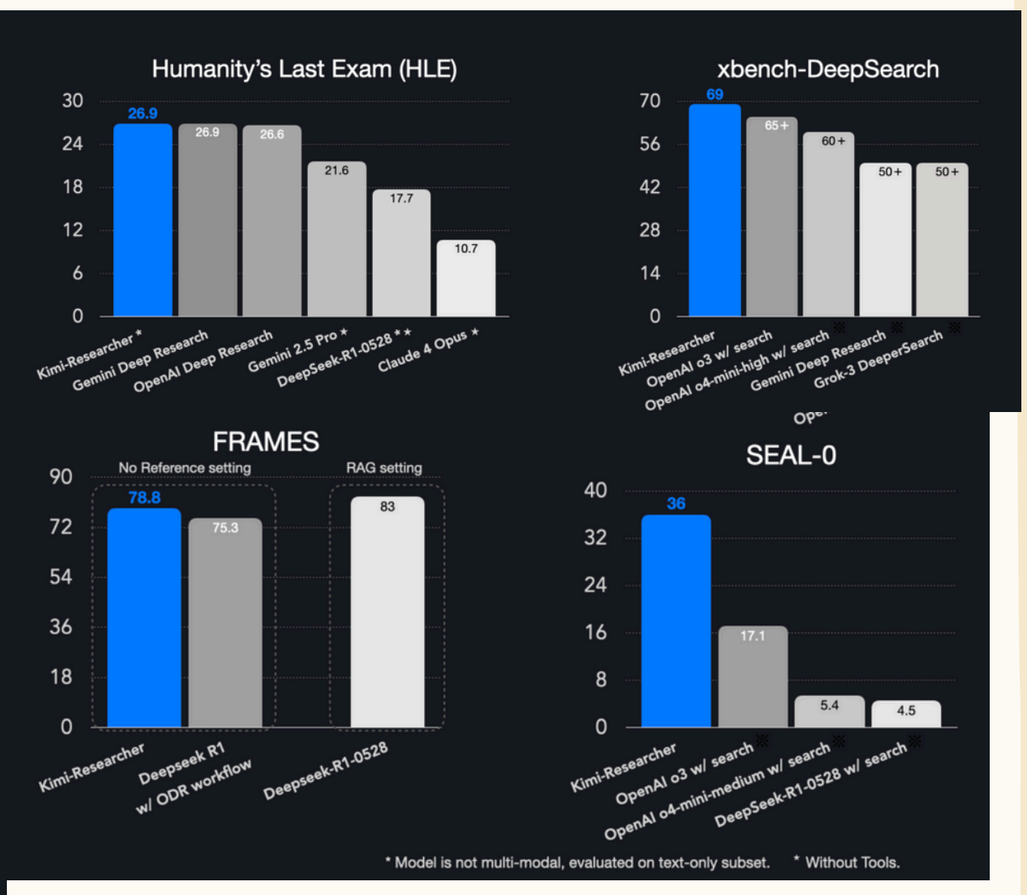
## 实验设置

- 训练数据：格式(query task, 可验证的answer)，通过自动化流程大规模合成，query task主要聚焦工具调用依赖型任务和推理密集型任务
- 底座：Kimi k系列模型 (多模态reasoning模型)
- RL算法：REINFORCE。训练过程遵循 On-Policy 策略，并引入负样本控制机制：对生成的rollout进行质量判断，识别为
- 负样本的某些数据将不参与模型更新，以提升训练稳定性。
- 三个tool：search、基于文本的browser、code执行
- reward function：基于PRM的RLVR，reward包括format reward和根据answer reward计算的每一个step的reward，并且只有format正确才计算answer reward

$$\text{step\_reward}_i = r \times \gamma^{T-i}$$

## 大规模Agent RL Infra




## 思考

尽管Kimi没有发布详细的技术报告，但这篇技术博客并非仅用于PR宣传，还是提供了很多有价值的训练细节。相比一些相对粗浅的开源方案，Kimi所采用的技术路径毕竟经过实际产品验证，因此更具参考价值与可信度。

目前，我们基本可以将DeepResearch理解为两部分：DeepSearch + 长文生成。其中，DeepSearch本质上是以搜索工具为核心的工具整合推理(Tool-Integrated Reasoning, TIR)，通过RLVR方式进行训练，关键在于如何构造高质量、具备工具依赖的训练数据集。Kimi的这篇技术博客对如何DeepSearch很有参考价值，当然博客中并未涉及长文本报告生成的具体做法。