arxiv.org/abs/2508.10874

SSRL: Self-Search Reinforcement Learning

SSRL：通过llm Self-Search来降低Web Agent训练成本

Yuchen Fan[1,3,*] **Kaiyan Zhang**[1,*,†] **Heng Zhou**[3,*] **Yuxin Zuo**[1,3] **Yanxu Chen**[1]
Yu Fu[4] **Xinwei Long**[1] **Xuekai Zhu**[2] **Che Jiang**[1] **Yuchen Zhang**[3] **Li Kang**[3]
Gang Chen[5] **Cheng Huang**[1] **Zhizhou He**[1] **Bingning Wang**[6]
Lei Bai[3,‡] **Ning Ding**[1,3,‡] **Bowen Zhou**[1,3,‡]

[1] Tsinghua University [2] Shanghai Jiao Tong University [3] Shanghai AI Laboratory
[4] University College London [5] CSCEC Third Bureau [6] WeChat AI
*Equal contributions †Project leader ‡Corresponding author

zhang-ky22@mails.tsinghua.edu.cn TsinghuaC3I/SSRL

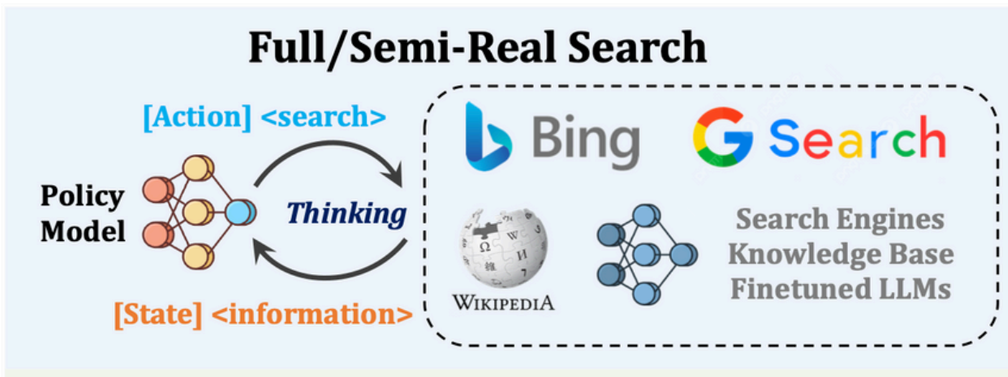
简介

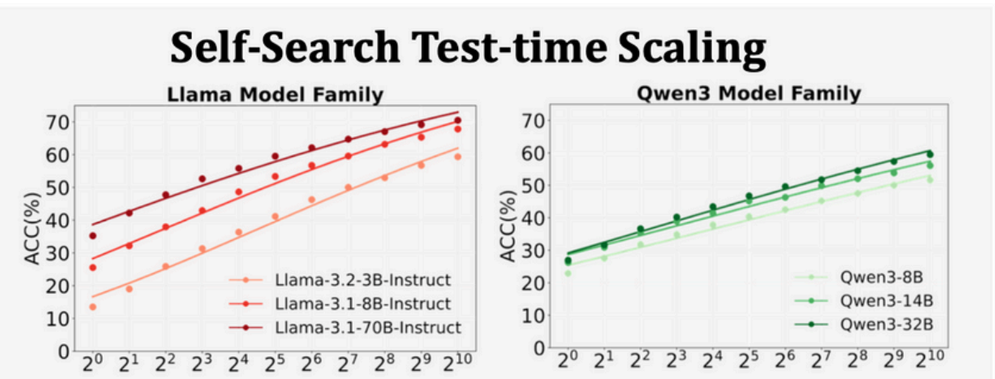
本文提出Self-Search Reinforcement Learning(SSRL)，目的是摆脱web agent训练阶段依赖外部搜索引擎api，将训练成本降下来，当然训练后模型效果也要好。简单来说，就是在TIR with RLVR训练时，llm既思考又生成搜索query又生成假的搜索结果，这就是on-policy self-search。
首先，需要说明llm是否包含充足的知识来模拟搜索引擎，作者做了一系列实验：增加candidate solution数量，计算pass@k指标(从K个候选里随机选出k个至少包含一个正确答案的概率)发现显著提升，说明llm内部是蕴藏着这些知识的，但同时发现major@k指标(对K个候选投票得到answer)却随着K的增加提升很小。这两个指标结果对比说明llm具备成为world model的潜质，但直接generate正确知识和答案也并不容易，因此需要RLVR训练来优化llm的自搜索与推理能力。

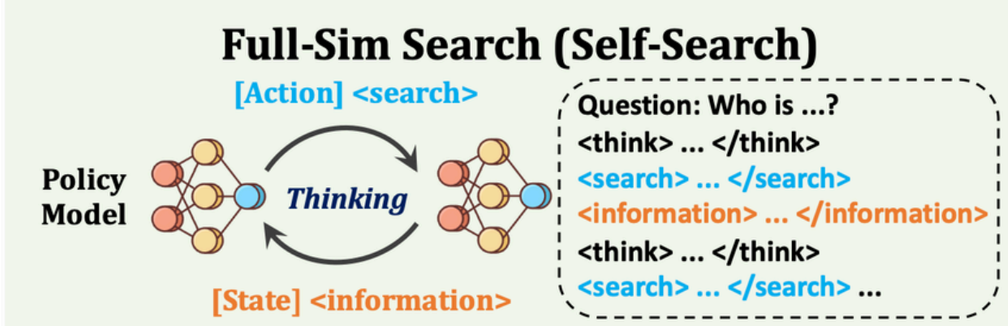
背景

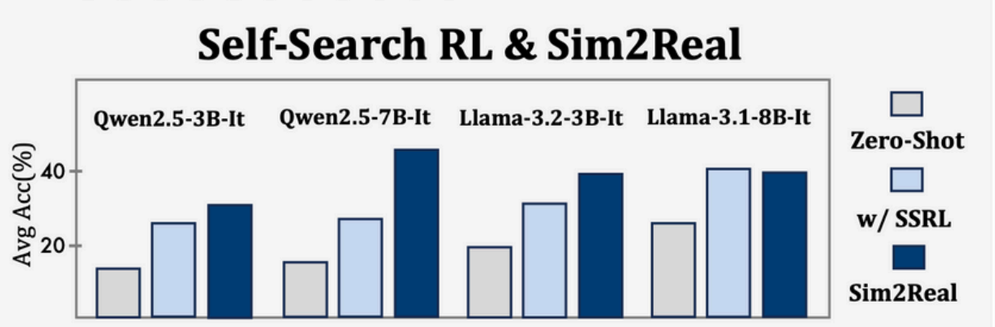
本文属于TIR(tool-integrated reasoning)方向的工作，再具体一点属于web agent方向，当搜索引擎作为tool时，llm能够获得时效性与覆盖面更强的外部知识，但也有代价：搜索引擎api的调用成本非常非常高。如果inference阶段调用搜索引擎不可避免，那么训练阶段能否少调用一些呢?
我们之前读过的ZeroSearch就研究过这个问题，在RLVR训练TIR阶段用一个sft tuning的llm模拟搜索引擎，但是sft数据集仍然需要调用真正的搜索引擎，并且作者认为sft llm和policy model不一致，这种off-policy训练可能不稳定，为此本文更进一步，提出SSRL，就是RLVR，只不过policy这个llm不仅生成搜索query，还直接生成假的“搜索结果”(self-search)，作者称为on-policy self-search，不但在训练阶段彻底摆脱搜索引擎api，实验效果还不错。

Full-Sim Search (Self-Search)









思考

1. 最好结合我们之前读过的ZeroSearch一起看，ZeroSearch在训练阶段还是依赖一点搜索引擎的，用来创建sft数据集，SSRL则在训练阶段彻底摆脱了搜索引擎，其次就是本文说SSRL属于on-policy
2. 本文的实验做的很充分，工作量真不小，感兴趣的朋友可以仔细阅读下