arxiv.org/abs/2508.02694

Efficient Agents: Building Effective Agents While Reducing Cost

Efficient Agents：专注于性价比的agent框架

开源代码：https://github.com/OPPO-PersonalAI/OAgents

OPPO AI Agent Team

简介

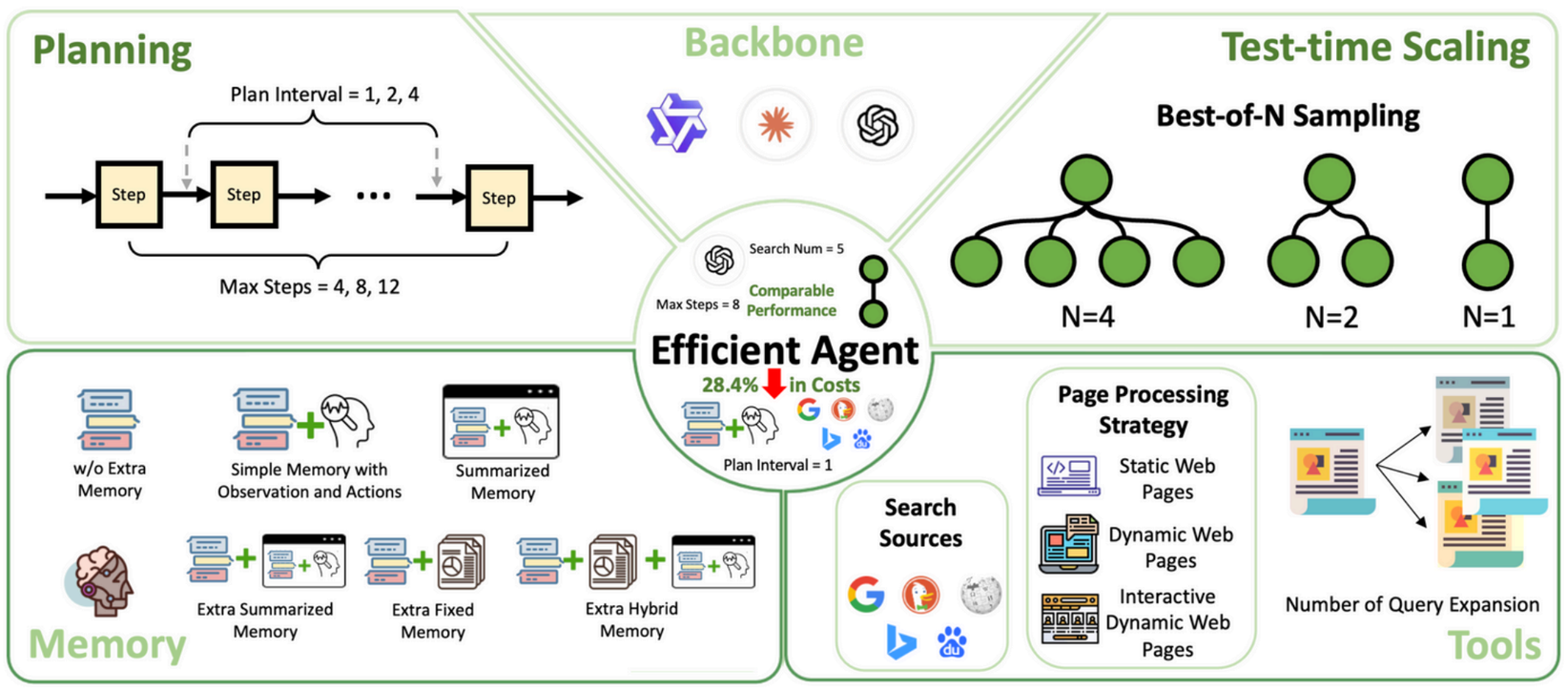
本文关注在保留agent效果的前提下，如何系统性地降低agent系统的运行成本。作者以GAIA数据集为基础，分析了能够影响agent效果和效率的关键模块，包括llm模型选择、planning策略、tool use使用方式、短期memory机制等，根据各个实验结论，提出了Efficient Agents高性价比框架。

背景

本文属于Agent框架方向的工作。回顾科技史，基本上一项新技术的诞生，前期是以性能导向的技术突破为主，也就是尽力把效果做到最好，比如ChatGPT系列，接下来的阶段是技术落地应用，这个时候就要考虑成本问题了，转而研究如何以最低的成本部署技术来服务社会。

本文认为llm agent也到了要考虑成本的阶段了，作者系统性地分析了agent系统中各种组件的“性价比”，并据此构建性价比很高的框架Efficient Agents。

Effective Agents框架



实验结论：

- LRM能力虽然强，但在面对复杂任务时，推理(reasoning)成本飙升导致效率大幅下滑
- Best-of-N用更多计算量获取高性能，但它的计算成本上涨幅度远远超过效果提升的幅度，性价比很低
- 动态plan比静态plan好，并且根据reasoning step实时更新plan是最好的
- reasoning step不能太大
- .....

Table 6 The Configuration of EFFICIENT AGENTS . The choice of each component is conducted by the observation from the previous empirical studies.

Component	Backbone	Max Step	Plan Interval	Search Source	Search Num	BoN	Memory
Settings	GPT-4.1	8	1	Multi	5	1	Simple

本文memory指的是multi-step reasoning内部的step-level memory

思考

本文对Agent系统中各模块进行了很好的性价比分析，结论很实用，但需要提醒的是：这些结论是基于GAIA数据集得出的，是否可以应用到你的任务场景，一定要考虑清楚。如果你的任务类型或应用场景与GAIA相差较大，完全可以借鉴本文的实验思路，针对自己的任务重新测试各模块的性价比。