arxiv.org/abs/2506.11763

DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents

博士级任务 + LLM-as-a-Judge，评测DeepResearch Agent的新基准

Mingxuan Du[1]*, **Benfeng Xu**[1,2], **Chiwei Zhu**[1], **Xiaorui Wang**[2], **Zhendong Mao**[1]†
[1]University of Science and Technology of China, [2]MetastoneTechnology, Beijing, China
{dumingxuan, benfeng}@mail.ustc.edu.cn

开源代码：github.com/Ayanami0730/deep_research_bench

简介

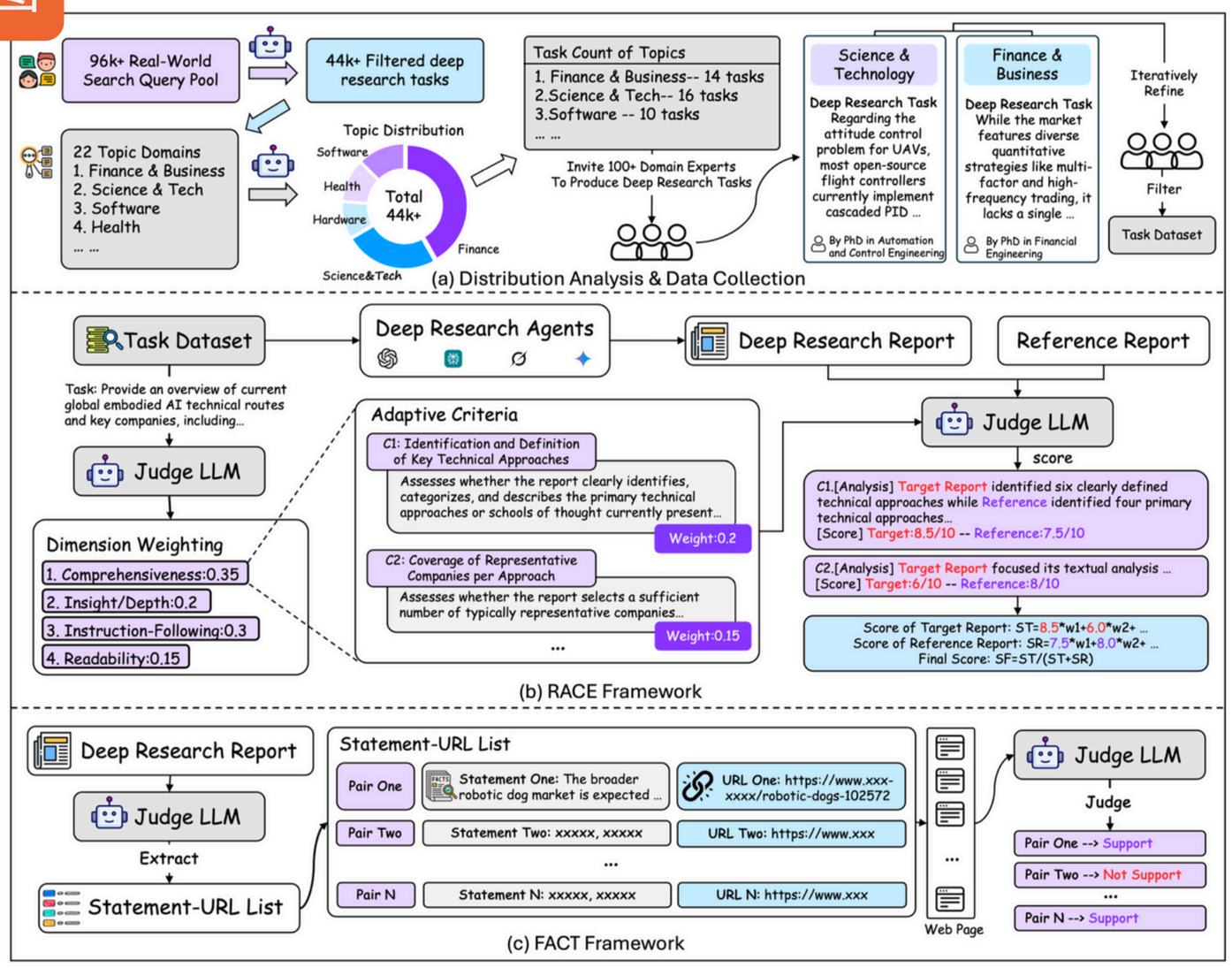
本文设计了DeepResearch Bench，用于评估DeepResearch Agent在22个领域中面对博士级别任务时的表现，具体来说，基准数据集覆盖22个领域、共计100个博士级的任务，然后以LLM-as-a-Judge为基础，以Agent生成的报告为输入，分别从生成报告的质量和Agent信息检索能力两大方面进行评估。
针对如何评估报告质量，作者设计了RACE(Reference-based and Adaptive Criteria-driven Evaluation)框架，在四个核心维度(全面性、深度、指令遵循度、可读性)下，让LLM生成分数。
针对信息检索能力的评估，作者提出了FACT(Factual Abundance and Citation Trustworthiness)框架，以生成报告中的引用为核心，主要判断引用是否能够从事实上支撑句子内容。

Note：RACE框架中用的reference report不是专家撰写的，是Gemini-2.5-pro-based Deep Research生成的

背景

DeepResearch Agent可以说是目前LLM Agent研究与应用中最活跃的方向之一，各家LLM大厂都推出了相应的产品，开源社区也在快速演进。DeepResearch Agent根据用户输入的query/任务，它能够自主完成多次网页检索和信息汇总，最终生成高质量的报告。然而，如何去评估到底哪个DeepResearch Agent效果最好，却并不容易。首先，它的输出是长报告，难以确定ground truth，不像有标准答案的数学题容易验证，其次DeepResearch会进行大量的思考和检索，中间过程不透明。

示意图



- 22个领域来自WebOrganizer
- 领域的任务数量分布来自业务真实数据
- 任务由博士或领域专家撰写
- RACE最终分数是相对打分，移除reference report对相对排序的影响
- Gemini-2.5 Pro/Flash作为Judge
- Gemini-2.5-Pro DeepResearch不错

对比