Google

arxiv.org/abs/2507.16075

Deep Researcher with Test-Time Diffusion

Rujun Han[*1], **Yanfei Chen**[*1], **Zoey CuiZhu**[2], **Lesly Miculicich**[1], **Guan Sun**[2], **Yuanjun Bi**[2], **Weiming Wen**[2], **Hui Wan**[2], **Chunfeng Wen**[2], **Solène Maître**[2], **George Lee**[1], **Vishy Tirumalashetty**[2], **Emily Xue**[2], **Zizhao Zhang**[2], **Salem Haykal**[2], **Burak Gokturk**[1], **Tomas Pfister**[1] **and Chen-Yu Lee**[1]

[1]Google Cloud AI Research, [2]Google Cloud

未开源

简介

本文提出TTD-DR(Test-Time Diffusion Deep Researcher)，一种未开源的基于multi-agent的Deep Research方案，作者说是受到人“先写草稿再查资料修正得到最终版本”的写作方式启发，并且将查阅资料修正的过程类比为扩散模型的去噪过程。TTD-DR主要包含了两个核心机制：1) 对各个单独的agent模块(比如生成写作plan、生成查询query、生成answer、修正写作)执行self-evolution，比如生成plan的agent不断修改plan；2)以当前版本草稿为核心，动态生成和query、plan相关的检索问题，然后调用搜索引擎获取answer，再根据(检索问题，answer)对草稿进行更新，这是一个迭代的过程。

背景

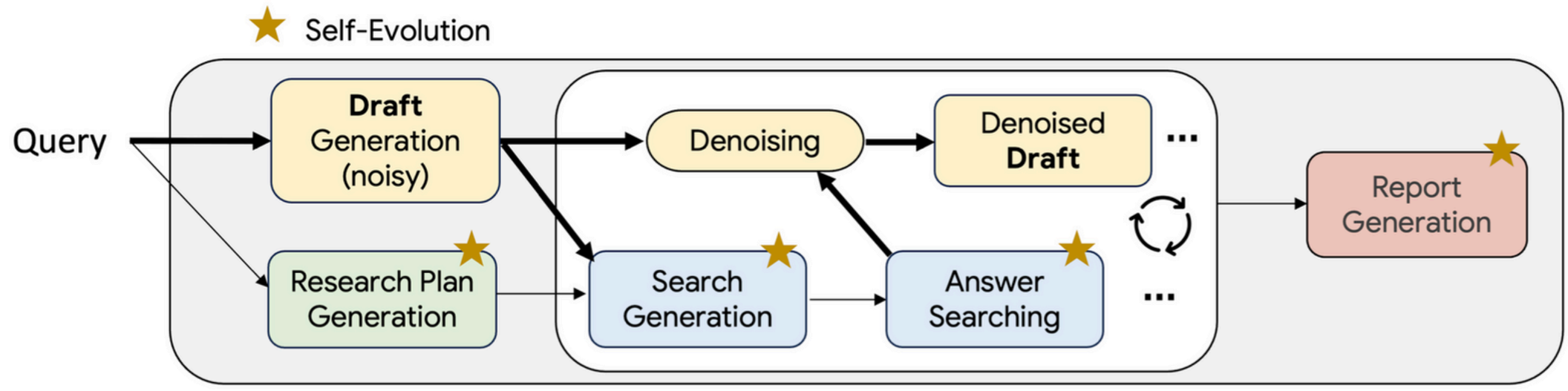
本文属于DeepResearch方向的工作，研究背景和意义就不多说了。
作者受到人写作的过程启发：先写出草稿，再通过从外部查阅信息反复修订完善，最终得到终稿。然后将将这一写作过程类比为扩散模型中的采样/去噪过程，即从噪声的draft出发，结合外部信息逐步“去噪”生成最终报告，提出了Test-Time Diffusion Deep Researcher(TTD-DR)框架。

实验设置

- 评估数据集：1) 生成报告类型的任务，LongForm Research和DeepConsult；2) 复杂的QA任务，HLE和GAIA
- multi-agent 框架：本文本质上是multi-agent system，用Google自家的Agent Development Kit (ADK)编程实现；用Gemini-1.5-pro as a judge

	LongForm Research	DeepConsult	HLE-Search	HLE-Full	GAIA
	Win Rate	Win Rate	Correctness	Correctness	Correctness
OpenAI Deep Research	-	-	29.1	26.6	67.4
LLM w/o agentic workflow					
Gemini-2.5-flash	21.0	16.7	2.8	11.6	31.5
Gemini-2.5-flash w/ search tool	27.8	17.6	14.6	14.6	57.6
Gemini-2.5-pro	31.0	17.6	8.6	20.9	57.0
Gemini-2.5-pro w/ search tool	35.0	19.6	20.0	21.6	61.8
Test-Time Diffusion Deep Researcher (ours)					
Backbone DR Agent	39.4	24.5	26.8	28.6	61.8
+ Self-evolution	60.9	59.8	30.6	29.4	63.0
+ Diffusion with Retrieval	**69.1**	**74.5**	**33.9**	**34.3**	**69.1**

TTD-DR



★标记的是支持self-evolve的agent子模块，由于未开源并且我认为写的也不算详细，所以感兴趣的朋友，只能自己思考如何实现了。

整体上来说，就是根据当前版本，集合搜索引擎生成(query, answer)获取外部知识来更新，得到新版本，这是一个迭代的过程。

思考

首先，作者假设人在写作时是先写一份完整的初稿，再不断去修正得到终稿，但是以我自己的写作习惯来说，我是按section/段落来写，每个section都反复经历润色，最终得到终稿。当然不是说假设不合理，顶多算是不能覆盖所有场景。
其次，所谓的扩散模型，我感觉更多的是一种写作包装，从模型结果和技术实现上来看，和扩散模型没关系。