王炸组合：微信接入满血DeepSeek R1，背后的Agentic RAG技术~

原创 PaperAgent PaperAgent 2025年02月16日 14:25 河南

终于，昨天微信以“AI 搜索”的形式接入了满血版DeepSeek R1，目前灰度测试ing，其中“深度思考”模式由DeepSeek-R1模型经过长思考而提供的更全面的回答：

目前DeeSeek-R1是**不支持function call**的，微信接入DeepSeek-R1可以采用**Agentic RAG**的方式，那么一个通用的AI Agentic（RAG）框架如何设计尼？本文进行专门剖析：

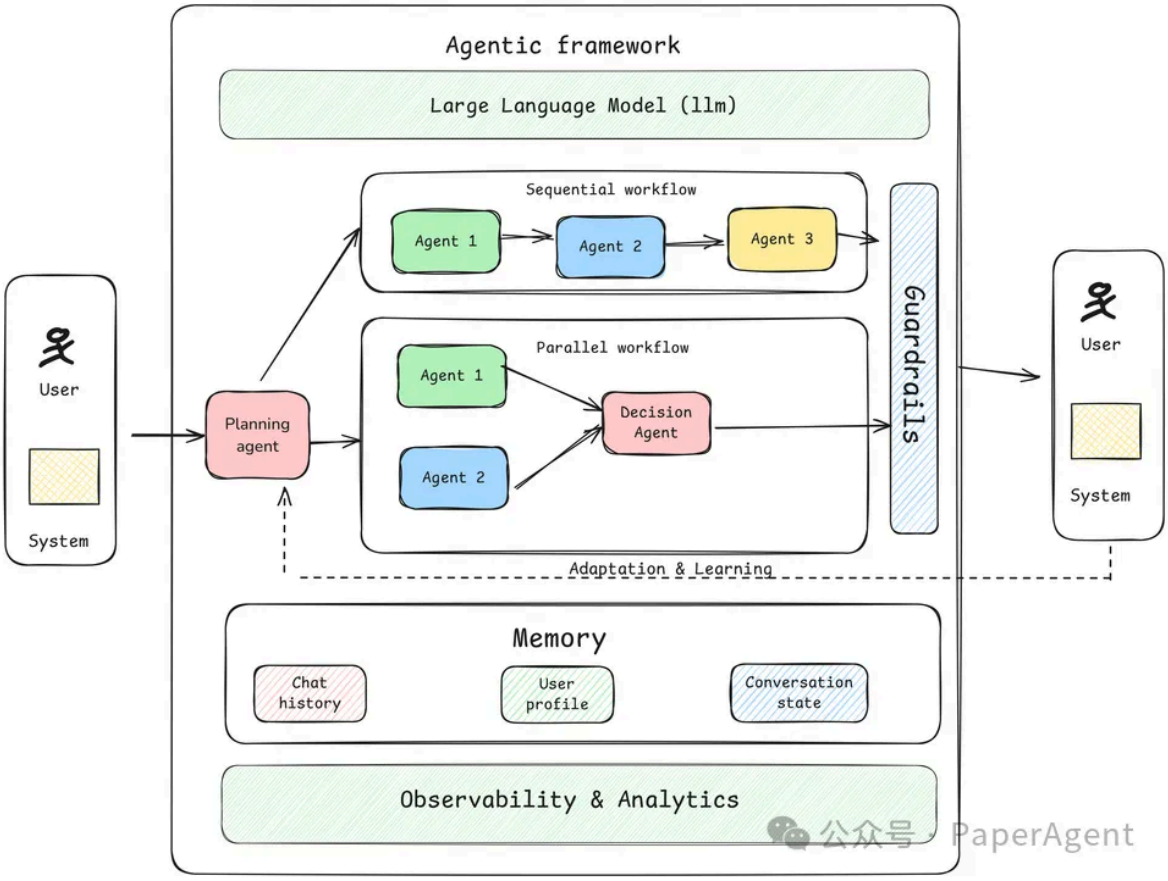
另外，如果有小伙伴想做自己的满血R1+搜索，可以结合DeepSeek官方发布的搜索接入中文版Prompt + Agentic RAG来设计，可参考之前的专栏文章：

动手设计自己的满血版DeepSeek-R1+联网智能体

DeepSeek官方推荐的搜索接入Prompt

```
+ For Chinese query, we use the prompt:
+ ```
+ search_answer_zh_template = \
+ '''# 以下内容是基于用户发送的消息的搜索结果:
+ {search_results}
+ 在我给你的搜索结果中，每个结果都是[webpage X begin]...[webpage X end]格式的，X代表每篇文章的数字索引。请在适当的情况下在句子末尾引用上下文。请按照引用编号[citation:X]的格式在答案中对应部分引用上下文。如果一句话源自多个上下文，请列出所有相关的引用编号，例如[citation:3][citation:5]，切记不要将引用集中在最后返回引用编号，而是在答案对应部分列出。
+ 在回答时，请注意以下几点：
+ - 今天是{cur_date}。
+ - 并非搜索结果的所有内容都与用户的问题密切相关，你需要结合问题，对搜索结果进行甄别、筛选。
+ - 对于列举类的问题（如列举所有航班信息），尽量将答案控制在10个要点以内，并告诉用户可以查看搜索来源、获得完整信息。优先提供信息完整、最相关的列举项；如非必要，不要主动告诉用户搜索结果未提供的内容。
+ - 对于创作类的问题（如写论文），请务必在正文的段落中引用对应的参考编号，例如[citation:3][citation:5]，不能只在文章末尾引用。你需要解读并概括用户的题目要求，选择合适的格式，充分利用搜索结果并抽取重要信息，生成符合用户要求、极具思想深度、富有创造力与专业性的答案。你的创作篇幅需要尽可能延长，对于每一个要点的论述要推测用户的意图，给出尽可能多角度的回答要点，且务必信息量大、论述详尽。
+ - 如果回答很长，请尽量结构化、分段落总结。如果需要分点作答，尽量控制在5个点以内，并合并相关的内容。
+ - 对于客观类的问答，如果问题的答案非常简短，可以适当补充一到两句相关信息，以丰富内容。
+ - 你需要根据用户要求和回答内容选择合适、美观的回答格式，确保可读性强。
+ - 你的回答应该综合多个相关网页来回答，不能重复引用一个网页。
+ - 除非用户要求，否则你回答的语言需要和用户提问的语言保持一致。
+
+ # 用户消息为:
+ {question}'''
+ ```
```


一个通用的**AI Agentic（RAG）框架组件设计**如下：

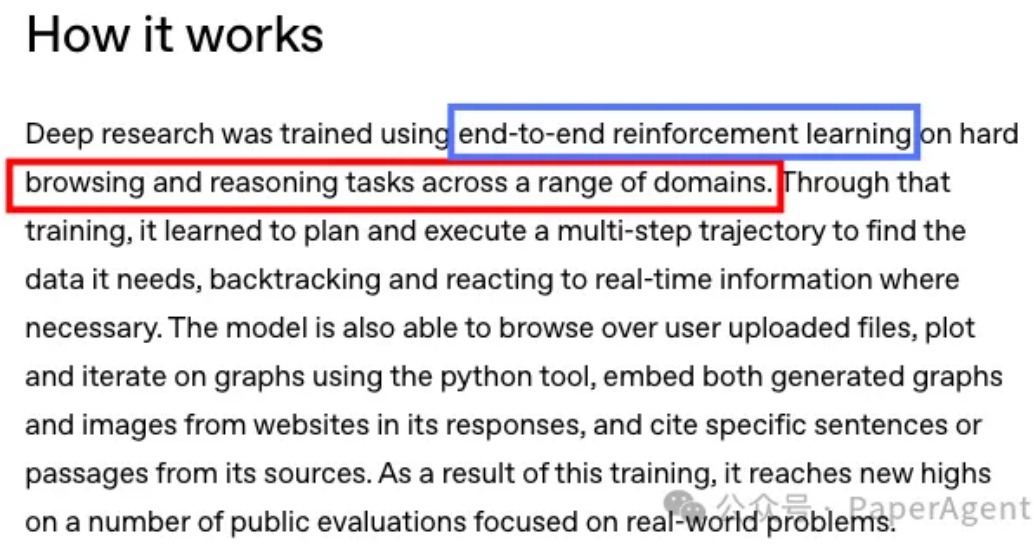


语言模型（LLM）

每个智能体框架都依赖于语言模型（LLM）。每个组件可以访问相同或不同的语言模型来完成其目标，从而在处理各种任务时提供灵活性和可扩展性。

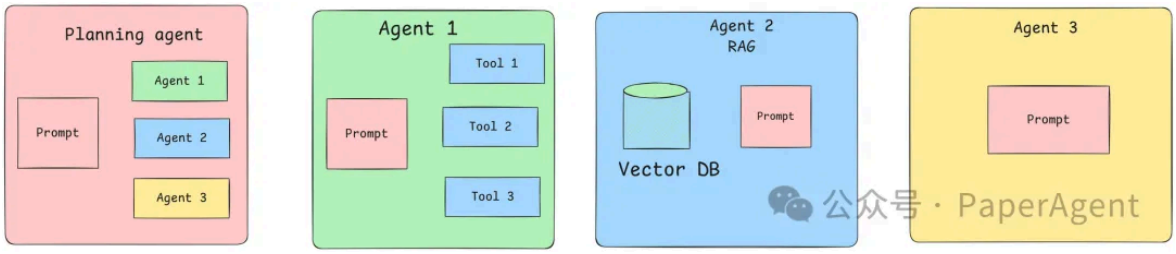
Planning Agent

- **规划智能体是协调组件，包括推理、规划和任务分解。**该智能体了解所有其他智能体，并通过合理的规划、推理和任务分解，决定执行哪些智能体以及执行顺序。
- 推理大模型之前，该Agent通常是具有function call的大模型来实现，而当前具有推理能力的语言模型（例如 **OpenAI o1/o3、DeepSeek R1**）最适合此类智能体，综合平衡智能体系统的其他考虑因素来决定是否采用。
- **OpenAI发布的智能体Deep Research**便是依托OpenAI o3在多个领域的**复杂浏览**和**推理任务**上进行**端到端强化学习而训练**的



Agents

智能体封装了一组指令和工具，用于完成特定任务（**图3 Agent2 RAG**）：



- **提示**：向语言模型发出的命令以及其可访问的工具。
- **工具**：执行动作的代码块，例如简单的代码块、API 调用或与其他系统的集成。
- **环境**：工具还可以关联特定的执行环境，例如集成开发环境（IDE）或通用计算机使用。
- **复杂智能体**：智能体也可以是整个架构，**例如检索增强生成（RAG）**，其中包括嵌入和向量数据库。
- **记忆**：智能体 AI 中的记忆功能允许智能体存储信息，并在未来交互中回忆这些信息。记忆始终对所有组件可用，并包括以下不同类型：
- **用户画像**：用户特定的信息，帮助智能体创建个性化体验。
- **聊天历史**：对话的历史记录，允许智能体从过去的交互中提取上下文。
- **聊天状态**：跟踪已执行的工作流程，避免重复任务。

安全护栏

安全护栏是防止有害行为的安全机制，同时确保在处理不可预见的输入或场景时的鲁棒性。例如，“确保回复中不提及竞争对手”等规则应存在于框架级别。这些约束对于在动态环境中部署智能体至关重要，提供了可编辑的默认安全检查。

智能体可观察性

可观察性允许开发者和用户了解智能体正在做什么以及为什么这样做。提供智能体行为的透明度有助于诊断问题、优化性能，并确保智能体的决策与期望结果一致。

适应与学习适应

涉及智能体根据环境反馈调整其行为的能力。这包括强化学习或其他自适应技术，使智能体能够随着时间优化其决策。例如，营销智能体可以根据客户偏好的变化调整其策略。

AI Agentic系统生态

在过去的一年中，智能体 AI 基础设施取得了巨大发展，并预计将继续快速演变，这种增长带来了许多新的工具和组件，有助于构建更好、更强智能体 AI 系统。

推荐阅读

- 动手设计AI Agents：Coze版（编排、记忆、插件、workflow、协作）
- DeepSeek R1 + Agent 的下半场
- RAG全景图：从RAG启蒙到高级RAG之36技，再到终章Agentic RAG！
- Agent到多模态Agent再到多模态Multi-Agents系统的发展与案例讲解（1.2万字，20+文献，27张图）

欢迎关注我的公众号“**PaperAgent**”，每天一篇大模型（LLM）文章来锻炼我们的思维，简单的例子，不简单的方法，提升自己。

PaperAgent

日更，解读AI前沿技术热点Paper

224篇原创内容

公众号

LLM热点Paper 379