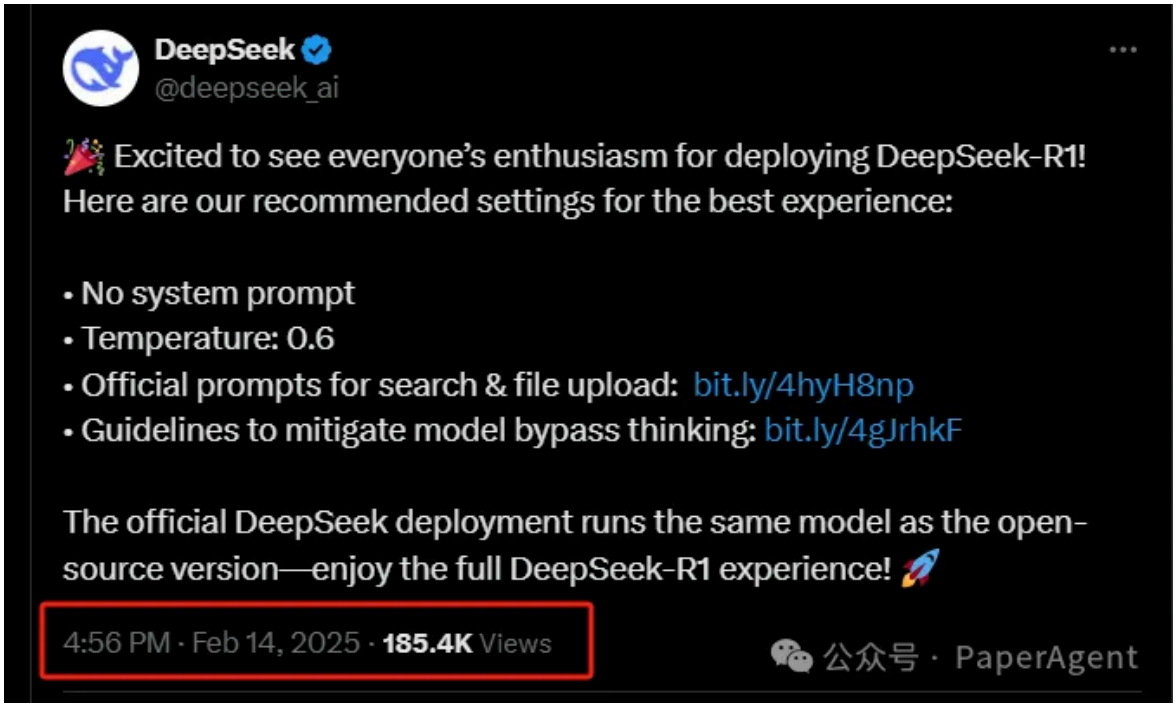
DeepSeek官方发布部署 R1 的正确打开方式~

PaperAgent 2025年02月14日 23:47 湖北

今天，DeepSeek官推发布了DeepSeek-R1部署最佳配置实践，并且强调**官方部署的DeepSeek**运行的模型与**开源版本相同：**

- 不使用系统提示
- 温度设置为0.6
- 搜索和文件上传的官方提示
- 避免模型跳过思考的指南



搜索的官方提示，可用于以Agentic RAG的方式设计自己的联网版DeepSeek-R1

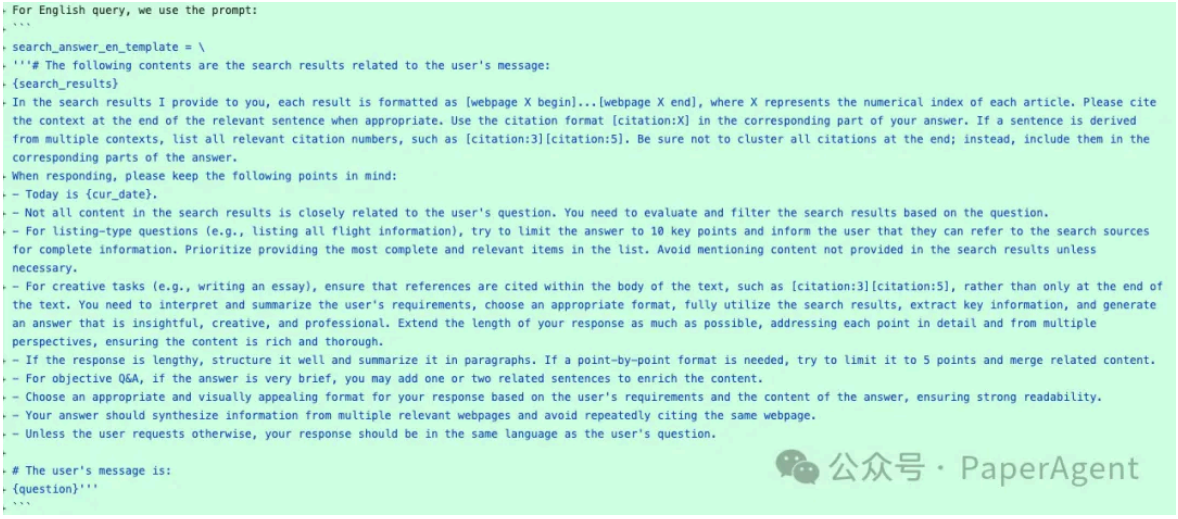
动手设计自己的满血版DeepSeek-R1+联网智能体

中文版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}'''
```
公众号 · PaperAgent

英文版Prompt：
```



```
For English query, we use the prompt:
```
search_answer_en_template = \
'''# The following contents are the search results related to the user's message:
{search_results}
In the search results I provide to you, each result is formatted as [webpage X begin]...[webpage X end], where X represents the numerical index of each article. Please cite the context at the end of the relevant sentence when appropriate. Use the citation format [citation:X] in the corresponding part of your answer. If a sentence is derived from multiple contexts, list all relevant citation numbers, such as [citation:3][citation:5]. Be sure not to cluster all citations at the end; instead, include them in the corresponding parts of the answer.
When responding, please keep the following points in mind:
- Today is {cur_date}.
- Not all content in the search results is closely related to the user's question. You need to evaluate and filter the search results based on the question.
- For listing-type questions (e.g., listing all flight information), try to limit the answer to 10 key points and inform the user that they can refer to the search sources for complete information. Prioritize providing the most complete and relevant items in the list. Avoid mentioning content not provided in the search results unless necessary.
- For creative tasks (e.g., writing an essay), ensure that references are cited within the body of the text, such as [citation:3][citation:5], rather than only at the end of the text. You need to interpret and summarize the user's requirements, choose an appropriate format, fully utilize the search results, extract key information, and generate an answer that is insightful, creative, and professional. Extend the length of your response as much as possible, addressing each point in detail and from multiple perspectives, ensuring the content is rich and thorough.
- If the response is lengthy, structure it well and summarize it in paragraphs. If a point-by-point format is needed, try to limit it to 5 points and merge related content.
- For objective Q&A, if the answer is very brief, you may add one or two related sentences to enrich the content.
- Choose an appropriate and visually appealing format for your response based on the user's requirements and the content of the answer, ensuring strong readability.
- Your answer should synthesize information from multiple relevant webpages and avoid repeatedly citing the same webpage.
- Unless the user requests otherwise, your response should be in the same language as the user's question.

The user's message is:
{question}'''
```



### 避免DeepSeek-R1跳过思考指南

- DeepSeek-R1 系列模型在响应某些查询时倾向于跳过思考模式（即输出“think/think”），这可能会对模型的性能产生不利影响。
- 为了确保模型进行彻底的推理，建议在每次输出的开始强制模型以“think/think”启动其响应。

```
https://github.com/deepseek-ai/DeepSeek-R1/pull/399/files
```

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
223篇原创内容

公众号

LLM热点Paper 379

LLM热点Paper · 目录

上一篇
DeepSeek异构&分布式部署：全平台+国产GPU支持，你值得拥有！

下一篇
王炸组合：微信接入满血DeepSeek R1，背后的Agentic RAG技术~
```