爆改 RAG 提示词，让大模型秒变“解题高手”

原创 乘风破浪jxj 鸿煊的学习笔记 2025年03月16日 00:00 北京

从去年到现在，检索增强生成（RAG）在人工智能领域火得一塌糊涂！简单来讲，它能让大型语言模型（LLM）结合外部数据，这样一来，模型给出的回答不仅更加准确，还能紧跟最新信息，杜绝“胡说八道”，有效减少“幻觉”情况的出现。

不过，大家要知道，光有 RAG 还远远不够，提问时使用的“提示词”才是重中之重！就像在 Stack Overflow 上看到的文章指出的那样，如果提示词太笼统，系统可能会搜出一堆没用的信息，白白浪费大量 token（也就是处理能力）。根据笔者的经验，优化提示词能带来巨大提升，像之前参加的一些评测，仅仅调整 Prompt（提示词），就能让分数大幅变动。今天就毫无保留地给大家分享 5 种超实用的 RAG 提示词模板，让 RAG 给出的答案又稳又准!

1. 为啥提示词对 RAG 这么重要?

提示词就像是给大模型（LLM）下达的“指令”，你与 RAG 交流的方式，直接决定了它回答的质量。一个好的提示词，需要精准传达这 3 件事：检索到的信息该怎么用（RAG 不能光拿到数据，还得知道怎么整合进回答里）、你的具体需求是什么（RAG 可猜不透你的心思，得明确告诉它方向）、推理逻辑该如何展开（RAG 得清楚怎么组织信息，避免胡编乱造）。

此外，LLM 处理文本是有“容量限制”的（它只能处理一定数量的 token，也就是文本片段）。这意味着，你没办法把一整个资料库都丢给它，而是要靠 RAG 系统先筛选出最相关的数据，再通过提示词引导 RAG 使用这些信息。要是提示词不清晰，RAG 很可能还是会给出错误或不完整的答案。所以，设计好提示词真的太关键啦!

2. 五大实用 RAG 提示词模板来袭

2.1 提示词 #1：精准提问 + 关联信息

这招看似简单，实则非常好用！核心思路就是先把用户的问题提炼得更加精准，然后结合相关的知识库信息，这样就能让 RAG 生成最优答案。

- **示例提示词**：“用户想了解 [X]。请先提炼问题的核心意思，然后结合以下知识库中的内容，给出最清晰、最准确的答案。” 这里的 [X] 就是用户提出的具体问题，比如 “苹果的营养价值”。
- **为啥有效**：一方面，优化问题表述能减少歧义，让 RAG 更聚焦，避免误解问题；另一方面，提示 RAG 必须基于检索到的信息回答，这样就能有效避免它 “脑补” 或乱编内容，回答自然更精准。
- **咋操作**：
  - 优化用户问题：可以用简短的 NLP（自然语言处理）处理或摘要方法，提炼出最核心的意思。比如用户问 “现在市场上哪种手机性价比最高，拍照功能还要好”，就可以提炼为 “求推荐高性价比且拍照好的手机”。
  - 检索相关内容：从知识库中找到匹配的信息片段，像手机评测数据、产品参数信息等。
  - 组合提示词：把优化后的问题和检索到的内容放进提示词里。
  - 让 RAG 生成答案：确保 RAG 回答时紧扣检索内容，不要添油加醋。
- **适用场景**：当用户问题比较笼统或模棱两可时，这个方法能帮 RAG 更好地理解需求。要是你希望 RAG 直接抓住问题重点，不绕弯子，这招也很合适。尤其是各种开放式提问，当用户的问题过于宽泛时，先 “精炼问题 + 提供上下文”，回答质量会大大提升！

2.2 提示词 #2：用 “思维链” 提升 RAG 逻辑性

当遇到复杂问题，信息量大的时候，RAG 很容易给出混乱甚至错误的答案。这时候，“思维链”（Chain-of-Thought，CoT）技巧就派上用场啦！它的核心思路是让 RAG 按照清晰的步骤推理，而不是直接跳到答案，这样不仅逻辑更清楚，还能减少胡编乱造的情况。

- **咋用**：可以让 RAG 按照下面的提示词来一步步思考：“这是用户的问题和相关文本。请按照以下步骤来回答：用更简单的语言总结用户问题；挑选出最相关的文本片段；把这些片段整理成逻辑清晰的大纲；基于大纲撰写一个完整、连贯的答案。请展示你的推理过程，并提供最终精炼后的答案。”
- **为啥有效**：首先，强制 RAG 先分析问题，再回答，避免它 “想当然” 地生成错误内容；其次，先列大纲能保证答案结构清晰，更易读，逻辑也更顺畅；最后，推理过程透明可见，如果答案有错，你能轻松找到问题出在哪。
- **适用场景**：在需要准确推理的任务中，比如金融、法律、医学等领域，RAG 可不能胡编乱造，这时候 “思维链” 就非常有用。另外，当检索信息量很大时，它可以帮 RAG 过滤掉无关信息，专注于重要部分。看过 RAG 论文的朋友都知道，让 RAG 逐步推理，比直接生成答案更精准，特别是在需要组合多个信息块的情况下！所以，下次遇到复杂问题，不妨试试让 RAG 按步骤来思考，说不定答案质量会大幅提升。

2.3 提示词 #3：让 RAG 学会说 “不知道”

RAG 能帮助大模型（LLM）检索外部数据，提高回答的准确性。但现实情况是，如果知识库里没有相关内容，RAG 可能还是会给出 “不完整” 甚至 “瞎编” 的答案。那怎么让它在 “缺数据” 的情况下做出更靠谱的回应呢？答案就是引导 RAG 诚实地承认 “我不确定”！

- **咋用**：使用这个提示词可以让 RAG 在没有足够信息时，谨慎作答：“以下是与用户问题最相关的内容。如果你发现其中有足够的信息来回答，就请据此作答；如果没有，请直接说：‘我没有关于此问题的完整信息。’摘要：（插入检索到的相关文本或要点）现在你的最终答案是什么？”
- **为啥有效**：这样做能有效减少 “幻觉” 问题，RAG 只会用现有数据回答，避免胡乱生成内容；同时，提示 RAG 在回答前先确认信息是否足够；而且，如果 RAG 频繁回答 “我不确定”，还能发现知识库可能存在数据缺口，需要补充新内容。
- **咋落地**：
  - 优化数据分块方式：确保 RAG 返回的是简明、有用的知识点，方便后续判断信息是否足够。
  - 定期更新知识库：如果某些问题 RAG 经常回答 “我不知道”，那就说明可能数据不足，需要补充新资料啦。
- **适用场景**：在智能客服场景中，能避免 RAG 胡乱回答，而是礼貌地承认 “没有完整信息”；在研究分析场景中，能确保 RAG 只在有足够依据的情况下给出答案，不随意推测。RAG 的作用是增强信息获取，它可不能凭空创造内容，与其误导用户，不如让它学会 “坦诚不知”！

2.4 提示词 #4：让 RAG 先写再改，答案更精准！

有些任务，比如总结技术文档、改写政策文件、生成详细报告，让 RAG 一次性给出完美答案并不现实。这时候，多步骤修订的方法就特别有用 —— 让 RAG 先写初稿，然后自己检查、修正，最后再输出完整答案，还能提供来源列表，增强可信度。

- **咋用**：用下面这四步提示词引导 RAG 进行 “先写后改”：
  - 第一步：“根据用户请求，生成一份完整的草稿，并结合下方 RAG 检索的所有相关段落。”
  - 第二步：“现在重新检查初稿，看是否遗漏了任何有价值的上下文，并进行修订。”
  - 第三步：“提供最终版本，确保内容连贯、精准。”
  - 第四步：“标明引用的所有来源。”
- **为啥有效**：RAG 先写初稿再自查修订，能进行自我审查，减少遗漏，确保充分利用所有关键信息；如果 RAG 检索到的内容较多，这种方式能帮助它全面整合，不遗漏重要细节，提高准确性；最后，提供数据来源就像研究论文引用参考文献一样，能增强读者对答案可靠性的信任。
- **适用场景**：像政策文件、人力资源指南、法律声明等正式文档，对内容准确性要求极高，就很适合用这种方法；在多来源汇总场景，比如营销文案，需要从多个产品页面提取信息并整合，用 “先写后改” 也很合适；还有复杂知识库，如果数据库信息较多，单次生成可能会遗漏关键内容，多步骤审查能保证完整性。所以说，让 RAG 先写后改，比一次性生成更靠谱！想要高质量内容，就别怕 “多走一步”。

2.5 提示词 #5：用 “对比查询” 让 RAG 更聪明！

想让 RAG 更精准地回答问题？试试 “对比查询” 法！这个方法不是简单地抛出一个问题，而是给 RAG 两个相关但有差异的问题，让它在回答时学会分辨，并明确引用不同的信息来源。

- **咋用**：用这个结构化提示词来引导 RAG：“查询 A：(用户的第一个问题)；查询 B：(一个相似但角度不同的问题)；检索到的文本：(插入相关内容片段)；要求：针对每个问题单独作答，确保每个答案都引用最匹配的文本。回答完成后，解释你是如何决定哪些内容适用于哪个问题的。” 比如，查询 A 是 “苹果手机和安卓手机哪个系统更流畅”，查询 B 是 “苹果手机和安卓手机哪个拍照效果更好”，然后插入从手机评测报告、技术资料等获取的相关文本片段。
- **为啥有效**：这样能让 RAG 学会对比和归类，有时候知识库的内容涵盖多个话题，这种方法能帮助 RAG 选取最合适的文本回答不同的问题；同时，指定每个答案必须基于不同的来源，能减少 “答案混淆” 的情况；而且，让 RAG 自我解释推理逻辑，不仅能帮助调试，还能提高回答的透明度，让你清楚它是如何选择答案的。
- **适用场景**：在客户支持 & 销售场景中，比如一个客户问 “这个产品多少钱”，另一个问 “这个产品支持哪些功能”，RAG 需要从定价和技术文档中找出最匹配的内容，用 “对比查询” 就不会把答案混在一起；在内部培训 & 评测场景，用对比查询来测试 RAG 在不同问题上的表现，看看它是否真的能精准引用不同的文本来源；还有多主题知识库，如果数据库里内容交叉较多，这种方法可以帮 RAG 识别哪些信息适合回答哪个问题。“对比查询” 就像是给 RAG 施加 “压力测试”，逼它更精准地匹配问题和答案！试试这个技巧，让你的 RAG 更聪明、更精准。

3. 优化 RAG 提示词的 4 个关键技巧

想要让 RAG 给出更精准、可靠的答案，除了设计合理的提示词，数据质量、格式选择、Token 限制等因素同样重要。下面这四个实用技巧可以帮助你优化 RAG 提示词，提高整体生成效果。

- **清理和整理 RAG 数据源**：RAG 的输出质量取决于它能检索到的内容。如果知识库中存在不相关或低质量的文档，模型可能会被误导，给出错误或冗余的回答。因此，定期清理数据源至关重要。可以设定规则，确保检索出的信息足够精准，并过滤掉无关内容，提高系统整体的准确性。比如，在一个电商产品知识库中，定期删除已下架产品的信息，避免 RAG 在回答关于产品的问题时引用到无效内容。
- **控制 Token 限制，找到平衡点**：LLM 处理的信息量是有限的，过长或过短的提示词都会影响效果。过长的上下文可能会超出 Token 限制，让模型难以聚焦重点；过短的上下文则可能遗漏关键信息，导致回答不完整。最佳做法是使用摘要或预处理方式精简信息，确保 RAG 只接收最核心的数据。比如，在处理一篇长篇新闻报道时，先对报道进行摘要，提取关键信息后再提供给 RAG，这样既能保证信息完整，又不会超出 Token 限制。
- **选择合适的提示词格式**：不同的内容类型，适合不同的提示词格式。在某些情况下，调整提示词结构能显著提升 RAG 的回答质量。例如：
  - 要点式总结：适用于技术性内容，能让信息更清晰易读。比如总结代码技术文档时，用要点式列出关键代码功能、使用方法等。
  - 问答结构（Q&A）：适用于 FAQ 或知识库查询，便于模型精准匹配答案。像常见问题解答系统，直接采用一问一答的格式让 RAG 更高效地给出回应。

- 表格格式：适用于信息比对，比如产品参数、数据分析等场景。在对比不同品牌产品参数时，用表格呈现数据，RAG 能更直观地进行分析和回答。

- **加入审核机制，确保答案可靠**：在高风险场景（如法律、医疗、金融等领域），不能仅依赖 RAG 自动生成答案，而是需要增加审核机制。可以采用两种方式：
  - 辅助模型审核：用另一个 AI 先检查回答质量，发现问题后优化。比如先用一个专门的语言质量评估模型检查 RAG 生成的法律文件解读是否准确、通顺，再进行后续处理。
  - 人工复核：对于关键内容，增加人工审核流程，确保最终输出准确无误。像医疗诊断建议，在 RAG 生成初步结果后，由专业医生进行人工复核，保障患者安全。

4. RAG 模板

模板可见：https://github.com/gomate-community/TrustRAG/blob/main/trustrag/modules/prompt/templates.py

5. 结论：优化提示词，让 RAG 更智能

优化提示词的方式，直接决定了 RAG 的表现。从精简查询到思维链推理，每种策略都在解决同一个核心问题 —— 如何精准检索上下文，让模型正确整合信息，并合理应对不确定性。

大家要知道，不断试验是关键！甚至微小的提示词调整，都可能对最终结果产生显著影响。因此，在实际应用中，持续优化提示词设计，观察模型的反馈，再根据效果调整，能让 RAG 更加精准和高效。

如果你正考虑搭建新的 RAG 应用，或想优化现有的 RAG 方案，那么一个能整合检索、提示词优化和工作流管理的平台会极大提高效率。统一管理这些环节，可以帮助你更方便地调整提示词，并从用户互动中提取有价值的反馈。

```
1  笔者能力有限，欢迎批评指正或在留言区讨论。
```

RAG 10

RAG · 目录

上一篇 · 一文读懂 RAG 中的 embedding model