阅读随笔

1.

RAG：Retrieval-Augmented Generation检索增强生成

LLMs：Large Language Models大语言模型

### ****2.RAG的演变****

**基础RAG**：最初的RAG方法，主要包括索引、检索和生成三个步骤。

**高级RAG**：通过优化检索质量和引入预检索和后检索策略来克服基础RAG的局限。

**模块化RAG**：提供更大的灵活性，允许模块的替换和配置，支持多种检索策略。

检索（Retrieval）、生成（Generation）和增强（Augmentation）

3.RAG & ChatGPT

### RAG模型

**原理**: RAG模型结合了预训练的生成模型（如seq2seq结构）和检索机制。它可以从外部知识库（如维基百科）中检索相关信息，以增强生成的内容。

**优势**: 能够生成更准确和基于事实的回答，尤其在知识密集型任务中表现更好。它可以实时访问和利用更新的知识。

**应用**: 通常用于开放域问答、文本生成、事实验证等任务。

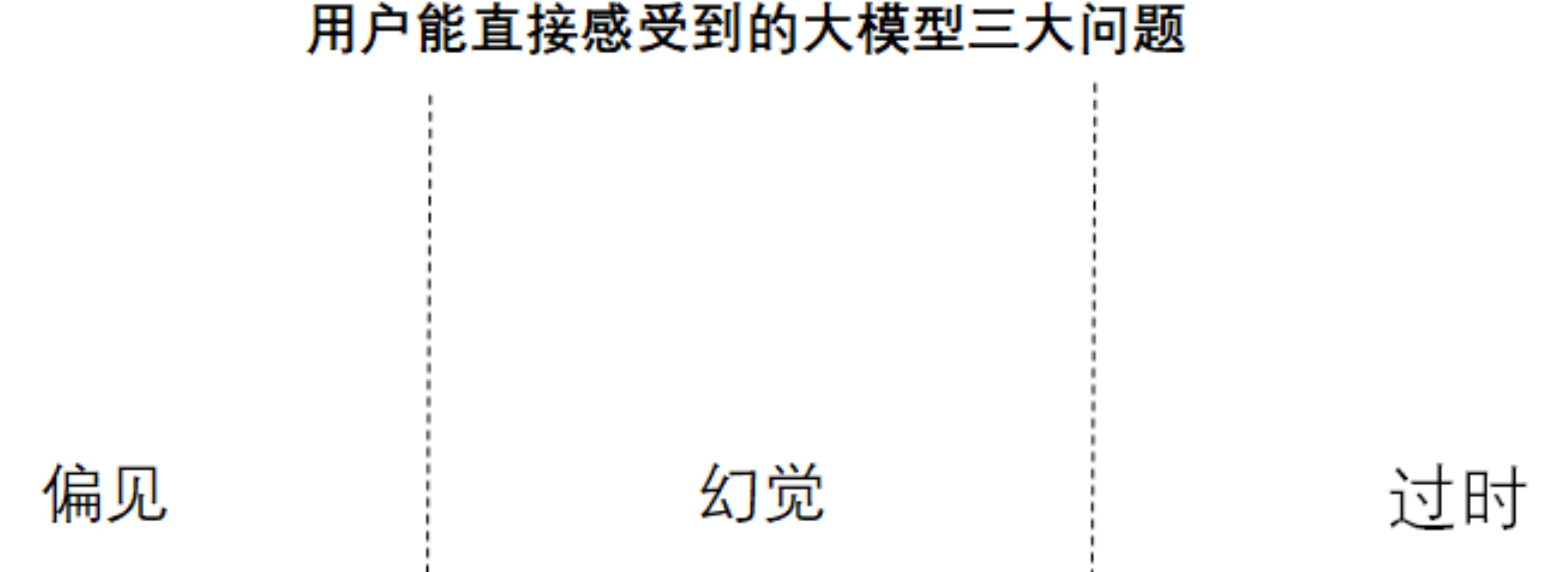
### ChatGPT

**原理**: ChatGPT是基于GPT（Generative Pre-trained Transformer）架构的语言模型，主要依赖于在大量文本数据上进行的无监督学习。它生成的内容主要基于已学习的知识，而不是实时检索外部信息。

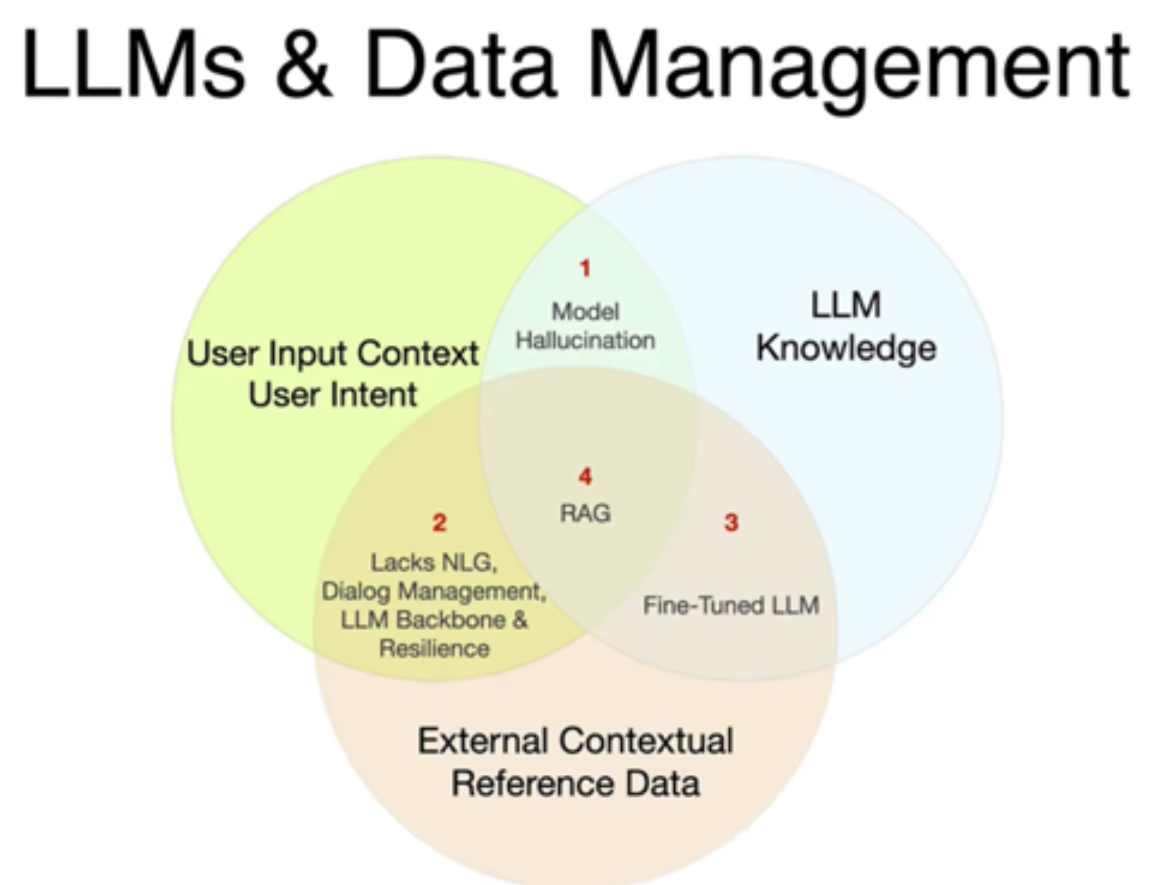
**优势**: 在对话生成和上下文理解方面表现出色，可以进行流畅的交流和多轮对话，但在某些情况下可能会生成不准确或过时的信息。

**应用**: 普遍用于聊天机器人、内容创作、教育辅导等场景。

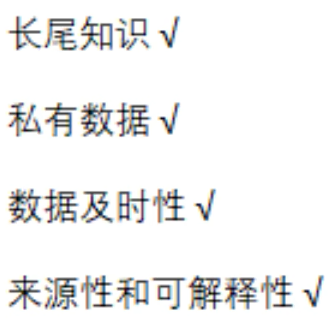
4.



5.

≈外挂一个数据库？

6.RAG解决的问题



9个问题回答

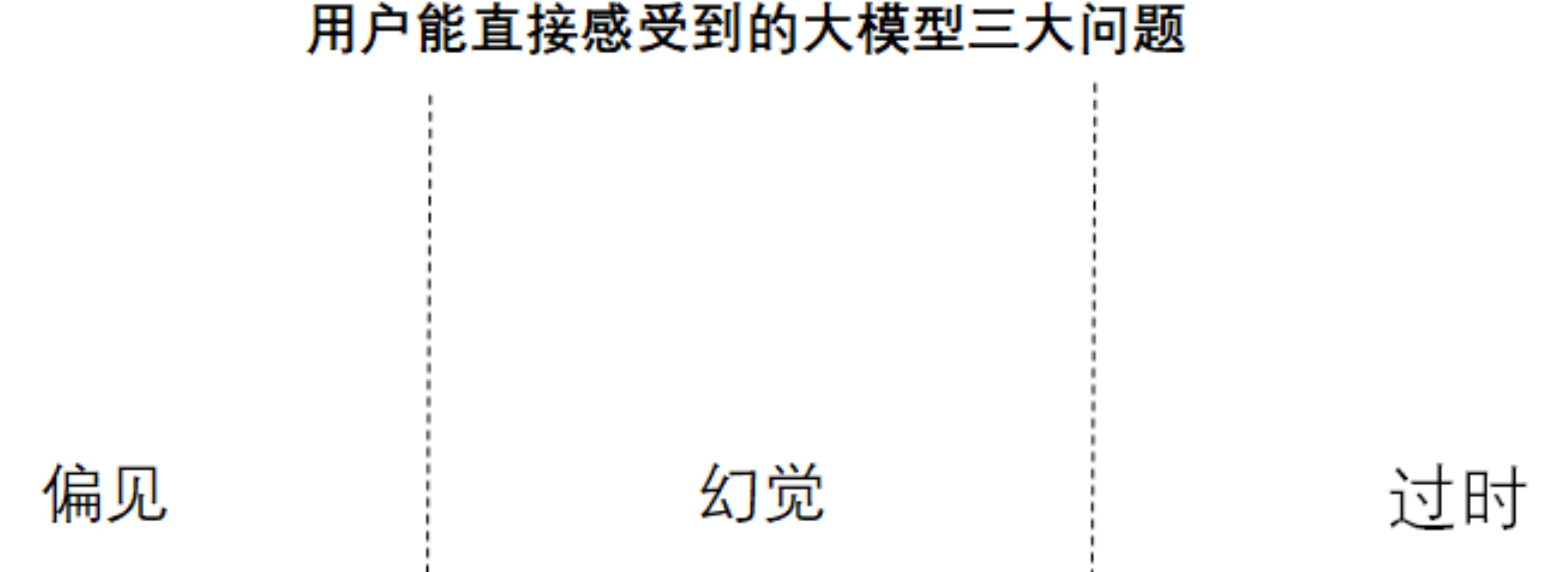
### 1. 什么是RAG

RAG（Retrieval-Augmented Generation）是一种结合了信息检索和文本生成的模型，旨在提高语言模型在知识密集型任务中的表现。它通过从外部知识库中检索相关信息，然后利用这些信息生成更准确和上下文相关的回答。

### 2. RAG是哪一年提出的概念，在什么背景下诞生了RAG

时间：2020年提出

背景：



大型预训练语言模型在处理知识密集型任务时表现出以上局限性，尤其是在访问和操作知识方面的不足。研究者们希望通过引入检索机制来增强模型的知识获取能力，从而提升其在问答和文本生成等任务中的表现。

### 3. 什么是KG和LLM

**KG（知识图谱）**: 知识图谱是一种以图形方式表示知识的结构，节点代表实体，边代表实体之间的关系。KG用于存储和组织信息，以便于检索和推理。

eg：假设一个城市的政府推出了一个在线政策问答系统，旨在为市民提供关于城市政策、法规和服务的信息：

1）用户在系统中输入问题，例如：“我可以申请失业救济金吗？需要哪些条件？”

2）知识图谱应用:

系统利用知识图谱来理解“失业救济金”相关的政策、条件和流程。知识图谱中包含以下信息：

实体: 失业救济金、申请条件、资格要求、相关法规。

关系: 失业救济金与申请条件之间的关系、申请流程的步骤等。

3）信息检索:

系统从知识图谱中检索与“失业救济金”相关的信息，包括：

资格要求（如工作时间、收入水平等）。

申请流程（如需要提交的文件、申请截止日期等）。

4）回答生成:

系统结合检索到的信息，生成清晰的回答。例如：

“您可以申请失业救济金，只要您在过去的12个月内工作超过6个月，且失业不是由于您的过错。申请时需要提供身份证明、工作记录和收入证明。”

5）用户反馈:

用户可以对回答进行反馈，系统根据反馈不断优化知识图谱，提升回答的准确性和相关性。

**LLM（大型语言模型）**: 大型语言模型是通过在大量文本数据上进行训练的深度学习模型，能够理解和生成自然语言文本。LLM在多种自然语言处理任务中表现出色，但在知识更新和准确性方面存在挑战。

### 4. RAG的工作原理是什么

【RAG系统主要包括两个部分：检索器和生成器。其中生成器可以理解为就是大语言模型（货架商品或深度定制），而检索器有的是基于语言模型、有的是基于词频算法、有的是两者结合使用】

RAG的工作原理主要包括两个阶段：

**1）检索阶段**: 使用编码模型根据用户的查询从知识库中检索相关文档。

**2）生成阶段**: 将检索到的文档与用户的查询结合，输入到生成模型中，生成最终的回答。

### 5.RAG工作流程

用户查询→检索器（从知识库中检索相关文档）→生成器（结合检索到的文档生成回答）→用户回答

？

### 6. 现在的RAG已经用了什么技术，解决了什么场景的问题

1）整合了多种先进的检索和生成技术（深度学习模型、向量检索、知识图谱等）。

#### 2）解决了问答系统、内容生成、教育辅助、研究与开发等场景问题。

### 7. 现在的RAG面临什么挑战，在哪方面做的有欠缺

RAG目前面临的挑战包括：

* **检索质量**: 检索结果的相关性和准确性仍然是一个问题，可能导致生成的回答不准确。
* **生成质量**: 模型可能在缺乏足够上下文时生成错误或不相关的答案。
* **上下文融合**: 如何有效地将检索到的信息融入生成过程中，避免信息冗余和不一致性

8.当前最新技术

**Dense Retrieval**: 使用深度学习模型（如BERT、RoBERTa）进行语义搜索，提高检索的相关性。

**Multimodal RAG**: 结合文本、图像和其他数据类型，增强生成模型的能力。

**Fine-tuning与Transfer Learning**: 在特定领域进行微调，提高模型在专业领域（如医学、法律）的表现。

**End-to-End RAG**: 一些新模型实现了端到端的训练，使检索和生成过程更加紧密结合，提高整体效率。

9.？