






RAG 技术详解与实践应用

第1讲：RAG原理解读：让检索增强生成不再是黑盒



目录






-  1. 上节回顾
-  2. 大模型的简介及其基本概念
-  3. 大模型的局限及其应对策略
-  4. RAG的基本概念和工作流程
-  5. RAG的更多可能性



- **RAG** 的应用落地性最强，学习成本相对可控，因此是当前**最实用、最具商业价值**的方案。
- 对于初学者而言，**基于已有框架**是明智的选择：
 - 1.快速落地，先见成效**：已有框架提供了完备的检索、生成、调用接口，大大缩短开发周期。
 - 2.成熟度高，性能优化好**：框架解决了很多复杂的工程问题，避免重复“造轮子”。
 - 3.聚焦业务，注重成效**：专注应用场景的业务逻辑优化，而非底层细节，聚焦关键目标。



目录

-  1. 上节回顾
-  2. 大模型的简介及其基本概念
-  3. 大模型的局限及其应对策略
-  4. RAG的基本概念和工作流程
-  5. RAG的更多可能性





豆包



+ 新对话

Ctrl K

🔍 AI 搜索

✍️ 帮我写作

🖼️ 图像生成

📖 AI 阅读

💻 AI 编程

📞 语音通话 新

💬 最近对话 >

📁 AI 云盘

🌟 我的智能体 >

📁 收藏夹

⬇️ 下载电脑版 ×



早上好, Kitty

发消息、输入 @ 或 / 选择技能



🖼️ 图像生成

✍️ 帮我写作

🔍 AI 搜索

📖 AI 阅读

🔍 学术搜索

📖 解题答疑

⚙️ 更多



简介:

大模型是一种基于深度学习的人工智能技术，它通过在海量文本数据上进行训练，学习语言的语法规则、语义含义以及上下文关系，从而能够对自然语言进行高效建模和**生成**。这种能力使得大模型可以理解和生成类似人类的语言，完成诸如文本生成、问答、翻译等多种任务。

特点:

大模型通常由**数十亿甚至数万亿个参数**构成，这使得模型能够捕捉语言的复杂性和细微差别。此外，大模型展现出强大的文本生成、理解和推理能力。



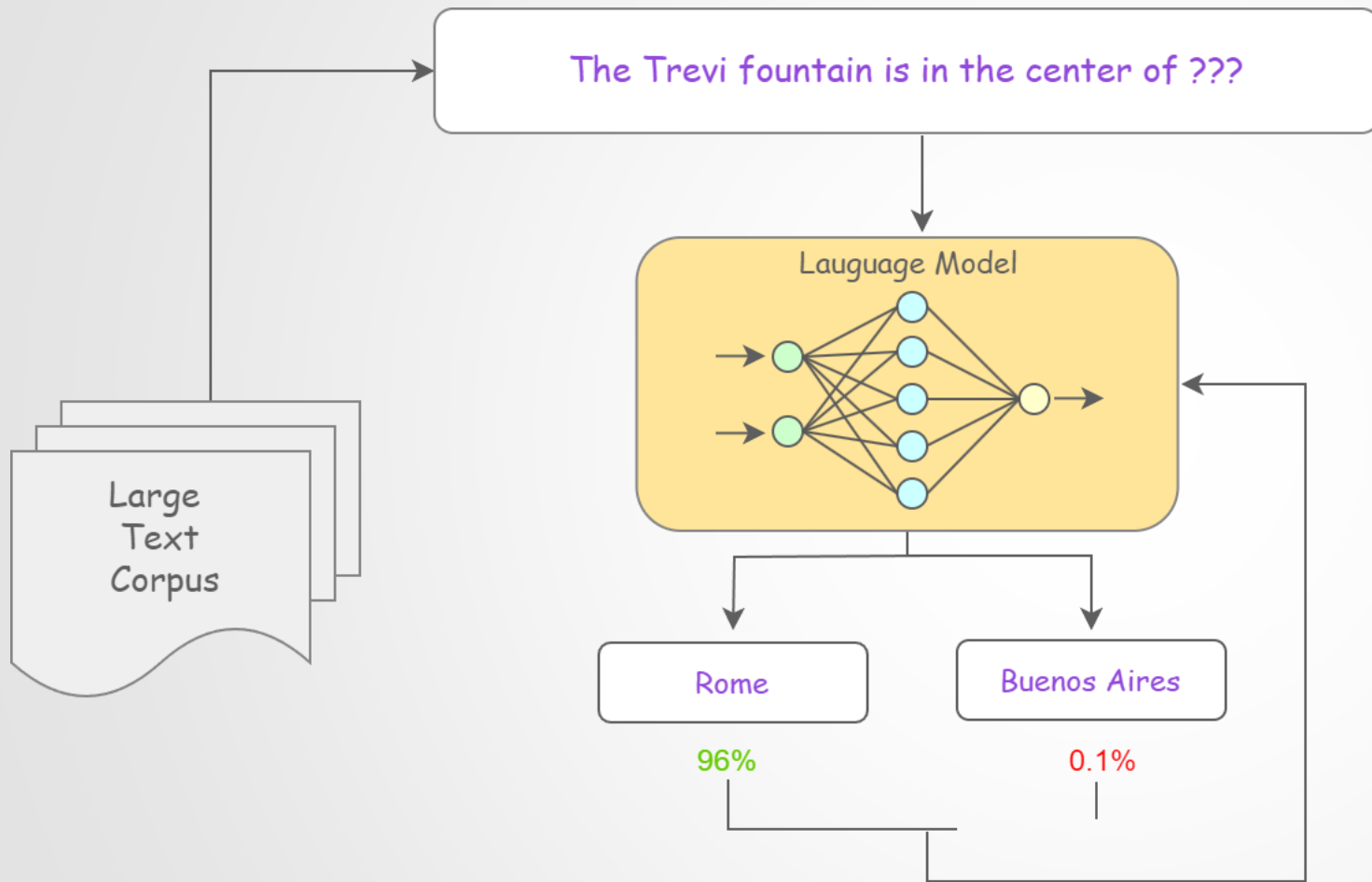
	生成式模型	判别式模型
目标	生成新的样本，即在给定条件下 生成数据	目标是进行分类或回归等判别任务，准确 预测标签
途径	建立联合概率分布 $P(X, Y)$ ，即模型学习数据特征 X 和标签 Y 之间的联合分布	建立条件概率分布 $P(Y X)$ ，即模型直接学习输入 X 到输出 Y 之间的映射关系
优势	<ul style="list-style-type: none">- 能够生成新样本，适合数据增强、缺失值填充- 能进行无监督学习和密度估计	<ul style="list-style-type: none">- 分类性能更好，泛化能力强- 训练速度更快，参数更少
劣势	<ul style="list-style-type: none">- 分类性能通常不如判别式模型- 模型复杂，计算量大	<ul style="list-style-type: none">- 无法生成新样本- 无法直接理解数据生成过程
场景	图像生成、文本生成、语言建模等	语音识别、图像分类、情感分析

厨师：创造美食

评委：点评美食



工作原理 – 预测下一个字符

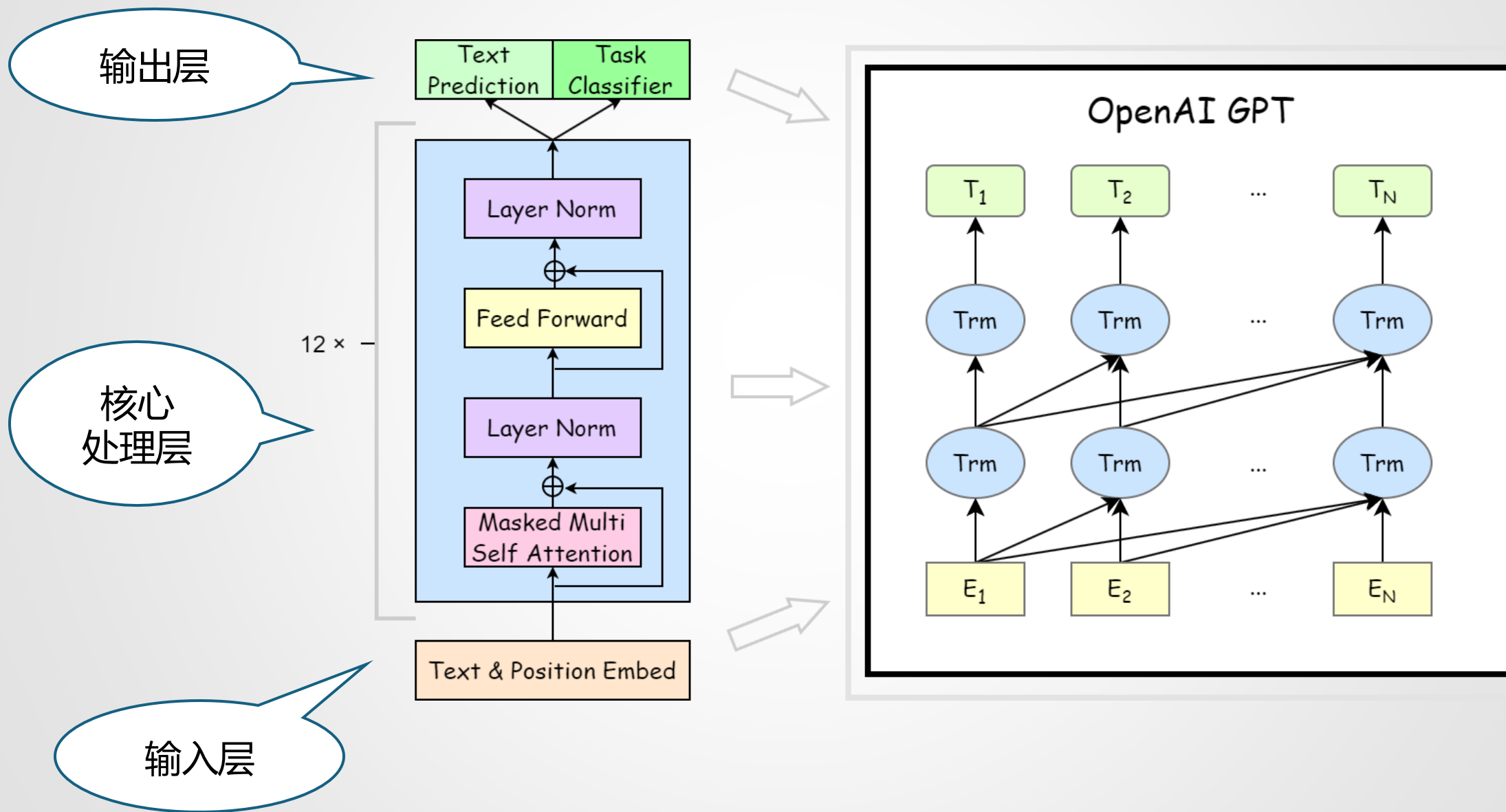


训练过程中，大模型借助海量文本数据，基于上下文尝试**预测下一个词或字符的概率分布**，并通过不断调整参数来最小化预测误差，以学习语言中的统计规律和模式。

可以简单理解为，大模型通过训练语料进行了**语感培养**。



模型结构



指数筛选 请输入筛选指数的问题句，如：“涨停家数大于2，涨幅

股票筛选 请输入筛选股票的问题句，如：“涨幅大于3%，市值大

搜索

重置

选出指数 110 (包含成分股 5320)

选出A股 0

指数列表

股票列表

涨停家数大于2(110)

涨幅大于3(921)

我的股票(0)

添加

问小达 AI选股

导出数据 设置表头

点赞 0 评论 0 提建议






锁定	序号	指数代码	指数简称	现价	涨跌幅(%)	最新涨停家数 2025.03.03
	1	880534	锂电池概念	2286.32	3.57	20
	2	880861	连续亏损	277.17	1.04	19
	3	880948	人工智能	1609.12	1.77	13
	4	880531	低安全分	1310.28	1.31	11
	5	880742	固态电池	904.84	5.92	11
	6	880865	近期新高	1346.08	0.38	11
	7	880957	工业互联	1850.16	1.73	11
	8	880880	近期强势	216.63	0.04	10



应用场景	具体功能	描述
内容创作	文本生成	根据主题或情境生成创意内容，如新闻、小说、诗歌、广告文案等，提高内容创作效率
	内容改写与润色	对文本进行改写、优化或润色，提升表达质量
智能客服	自动化客户服务	通过语言理解和生成能力，构建24小时在线客服系统，提高客户服务效率
机器翻译	跨语言交流	将文本从一种语言翻译为另一种，捕捉语言细节，提高翻译质量
问答系统	智能问答	解析用户提问，检索相关信息，生成准确答案，应用于客服、知识查询等
对话系统	自然语言交互	通过模拟人类对话，实现流畅、自然的互动，如聊天机器人
文本分类与情感分析	信息管理与分析	用于情感分析、主题分类等，帮助用户理解和管理大量文本数据
个性化推荐	定制化体验	分析用户数据，提供个性化推荐，广泛应用于电商、流媒体、社交媒体
教育助手	个性化学习	根据学习历史和表现，制定个性化学习计划，提供练习和反馈
医疗助手	辅助诊断	分析病历信息，结合医疗指南提供诊断建议
	治疗方案推荐	根据病情和历史数据推荐个性化治疗方案
	医疗文档生成	自动生成病历摘要、研究报告、出院小结等，提高医生工作效率
	医疗数据分析与预测	预测疾病风险，分析治疗方案效果，提供预防建议
金融助手	风险控制与决策支持	进行金融分析、风险评估、交易决策和欺诈检测
	个性化金融咨询	根据用户需求提供投资建议和理财规划
	运营优化	处理金融文档、合同审核、报表生成，提升运营效率



目录

-  1. 上节回顾
-  2. 大模型的简介及其基本概念
-  3. 大模型的局限及其应对策略
-  4. RAG的基本概念和工作流程
-  5. RAG的更多可能性

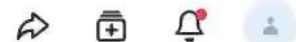




豆包



颤振的分类



+ 新对话

Ctrl K

AI 搜索

帮我写作

图像生成

AI 阅读

AI 编程

语音通话 新

最近对话

AI 云盘

我的智能体

收藏夹

下载电脑版

正常方式进行提问

颤振分为哪些类型?

基于 5 个搜索来源

颤振根据不同的领域和研究方向，有多种分类方式，常见的分类如下：

按结构动力学特性分类

- 硬颤振**：其特点是净阻尼非常突然地下降，非常接近颤振点。在硬颤振发生时，结构的振动状态会出现较为突然的变化，从稳定状态迅速转变为颤振状态，振动幅值可能会急剧增大。
- 软颤振**：软颤振中净阻尼是逐渐减小的。与硬颤振相比，软颤振的发生过程相对较为平缓，结构的振动幅值是逐渐增大的，系统从稳定状态到颤振状态的过渡较为平滑。

按机械振动分类

图像生成 帮我写作 AI 搜索 AI 阅读 学术搜索 解题答疑 音乐生成 更多

发消息、输入 @ 或 / 选择技能



受限于训练数据的局限性、上下文理解的缺陷或模型推理能力的不足，大模型经常出现“答非所问”的现象，通常情况下我们称之为“**幻觉**”。

下面简单介绍一下大模型中的幻觉问题。关于幻觉问题的分类，大体上可分为事实性幻觉（Factuality Hallucination）和忠实度幻觉（Faithfulness Hallucination）两种类别。

类别	问题	描述	示例
事实性幻觉	事实不一致	模型输出的信息可在现实中验证，但与真实情况相矛盾。	问“谁是第一位登月的人”，大模型回答加加林而非阿姆斯特朗。
	事实捏造	模型生成的信息完全虚构，无法通过已知知识验证。	询问独角兽的起源，大模型编造了一段毫无依据的历史。
忠实度幻觉	指令不一致	模型的输出未能遵循用户指令。	让大模型翻译句子，结果它直接回答问题而未翻译。
	上下文不一致	模型未能正确利用用户提供的上下文信息，导致输出内容与上下文矛盾。	依据上下文要求总结文章，但大模型遗漏了关键内容。
	逻辑不一致	模型的输出在逻辑上自相矛盾。	解方程 $2x+3=11$ ，大模型正确求得 $2x=8$ ，但随后错误地输出 $x=3$ 。



针对大模型存在的**数据依赖性强、幻觉严重**等问题，检索增强生成（Retrieval-Augmented Generation, RAG）技术在生成阶段引入了外部知识库，使得大模型的回答基于**真实数据**，而非**仅依赖训练数据推理**，从而减少幻觉，提高了信息的准确性。借助RAG技术，大模型在企业应用、专业知识辅助等场景中的表现更加**稳定可靠**。

当大模型面对特定难题或需要最新信息来作答时，RAG 就会发挥作用，从外部数据库或知识库中**检索相关内容**，提供给大模型。如此一来，大模型便能生成**更精准、更具权威性**的回答，如同学生借助书本在考试中取得更好的成绩一样，大模型结合 RAG 后，其表现也得到了显著提升。








RAG技术与大模型的结合，不仅提高了模型在处理知识密集型任务时的准确性和可靠性，还增强了模型处理最新信息和保护数据安全的能力。

优势	描述
知识增强的响应	RAG 利用大量语料库中的最新信息，提供更准确和信息丰富的回答，弥补大模型在特定领域知识或最新信息上的不足。
可扩展性	RAG 可随检索语料库的规模扩展，而无需为每个新主题微调 大模型，使其更容易适应新领域。
减少训练需求	只需训练或更新检索器，而无需微调整个 大模型，降低计算资源消耗，提高训练效率。
减少幻觉问题	通过外部知识库支持，减少大模型生成虚假或不准确信息的风险，提高回答的可靠性。
实时更新	RAG 可访问外部最新知识，而大模型训练数据是静态的，从而确保信息的时效性。
可解释性	RAG 的回答可追溯至具体的数据来源，增强可信度、可解释性和可追溯性。
安全性	无需重训练 大模型，即可调整检索系统以满足企业的安全要求。
灵活性	RAG 可灵活集成更多工具、更新数据或适配不同应用场景，增强大模型的适应能力。

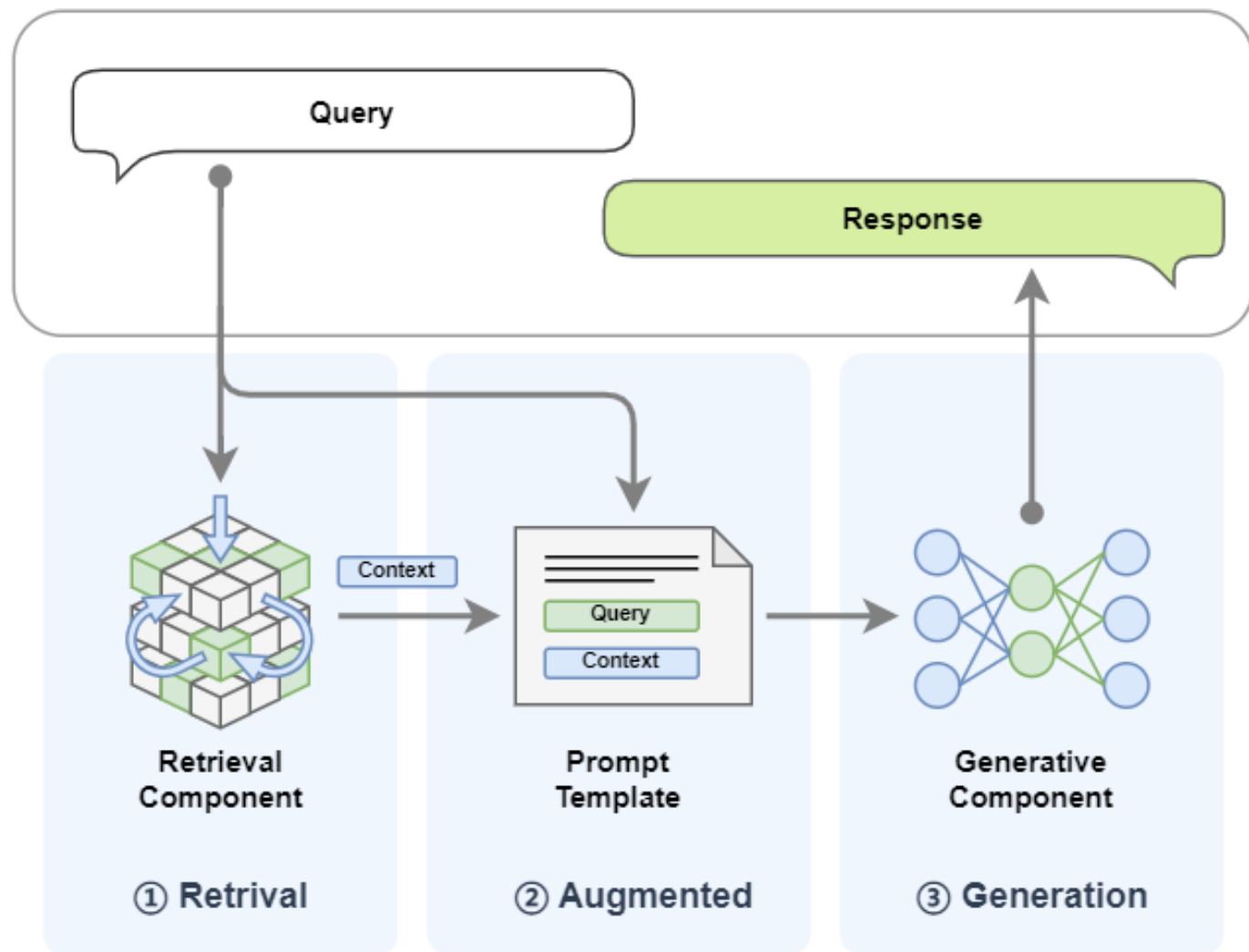


目录

-  1. 上节回顾
-  2. 大模型的简介及其基本概念
-  3. 大模型的局限及其应对策略
-  4. RAG的基本概念和 workflows
-  5. RAG的更多可能性



RAG的基本流程

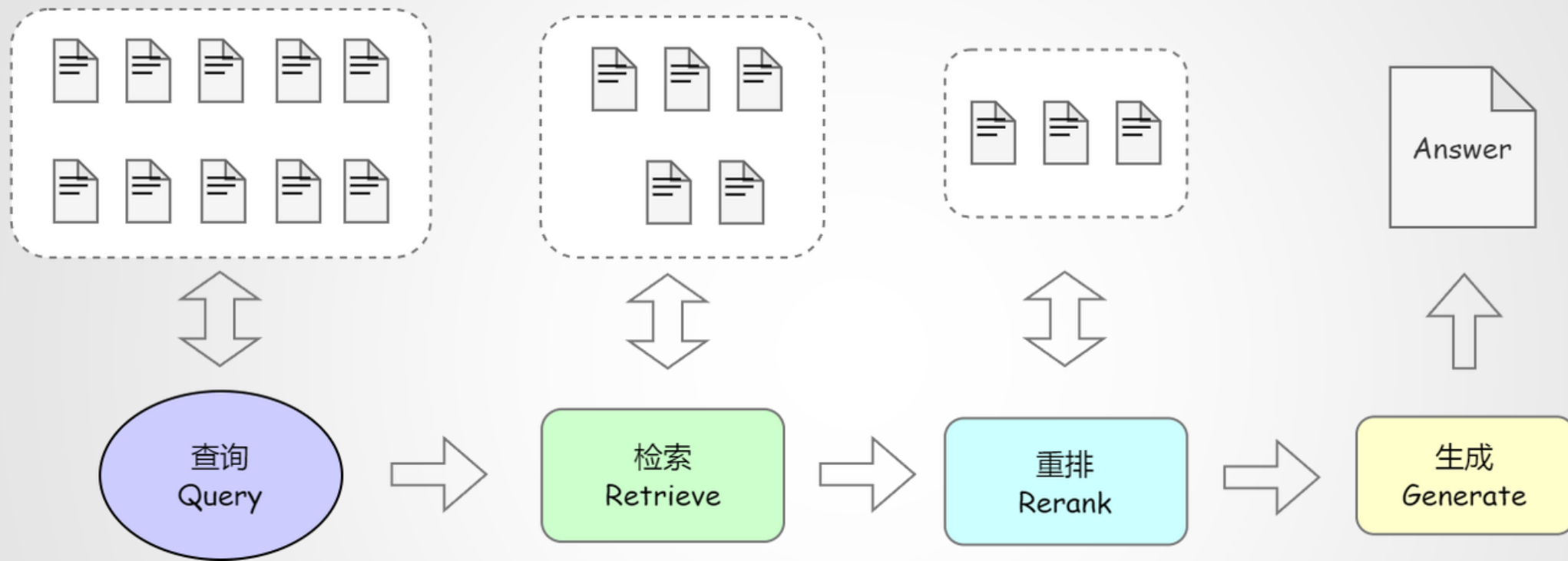


RAG（检索增强生成，Retrieval-Augmented Generation）是一种结合信息检索（Retrieval）和文本生成（Generation）的技术，旨在提高大型语言模型（大模型）的准确性和实用性。

顾名思义，RAG的流程应该是**先对海量的文本进行检索**，再结合检索到的文段，进行相关文案的**生成**。

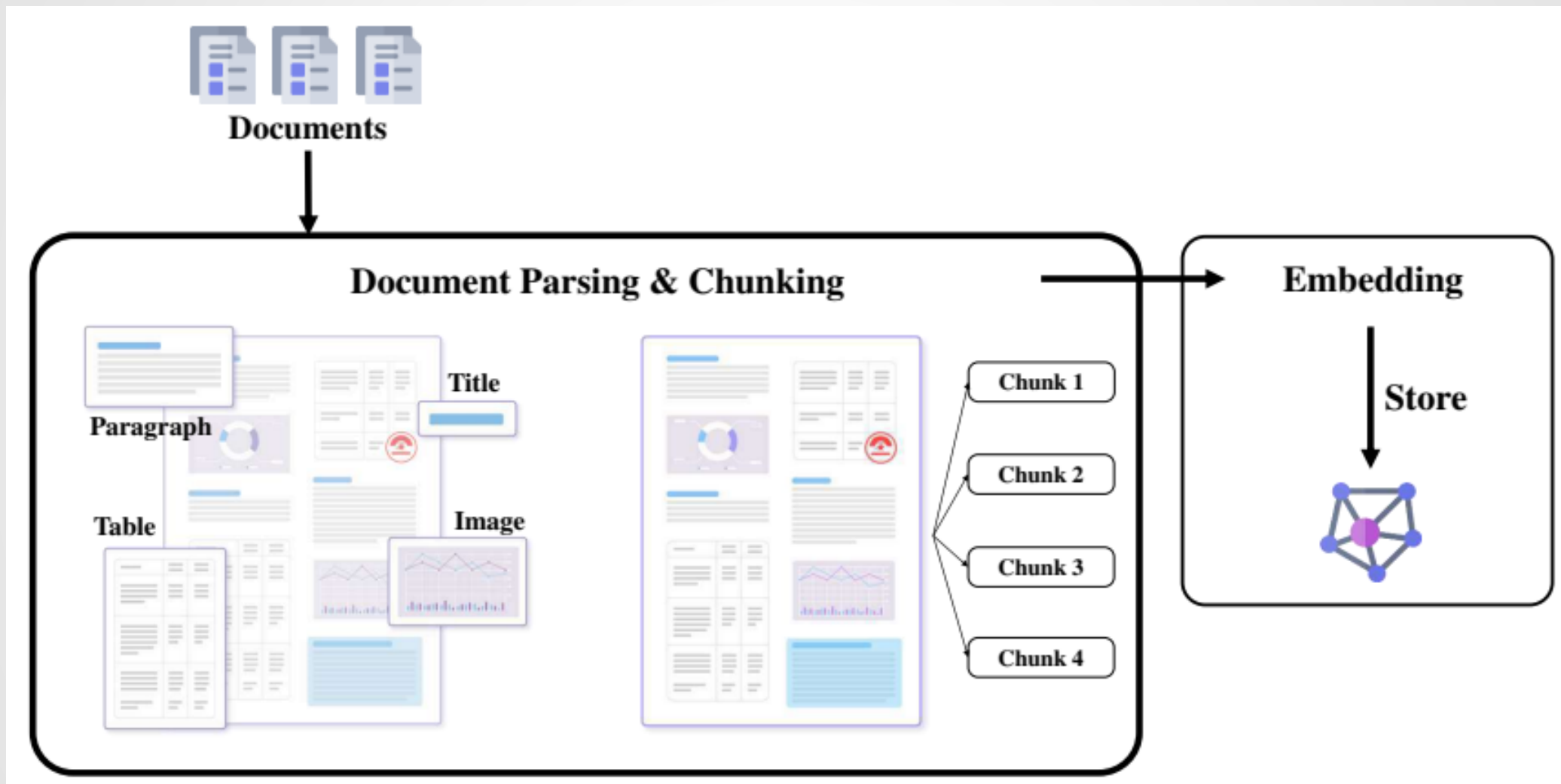


RAG的基本流程



该流程图展示了知识库问答的关键流程：知识库中共有10 个片段，Retrieval 阶段召回 5 个相关片段，重排序后筛选出 3 个，最终由大模型结合筛选内容生成回答。





该示意图展示了一个简单的PDF文档读取与解析的示例，向我们直观地展示了文档读取与解析的流程



① 文档读取与解析

这是整个流程的**基础**，为后续检索任务提供结构化的数据。文档读取就是把各种格式的文档加载到系统中。常见的文档格式包括**PDF、Word、Html、PPT**等，面对文档格式多样性、内容复杂性和非结构化数据等挑战，可以借助**开源工具**(如MinerU)来提高解析的准确率。解析后的文档内容会被转换为统一的格式，然后进一步处理和标准化。

② 分块和向量化

对收集到的原始数据进行**清洗、去重、分块**等预处理工作，去除无关内容和噪声，确保数据的质量。预处理过程中还需要对文本进行**向量化编码**，使其能够被高效检索。这一阶段通常使用深度学习模型或其他文本编码方法（如Word2Vec、BERT等）来生成文本的向量表示，便于后续的检索工作。

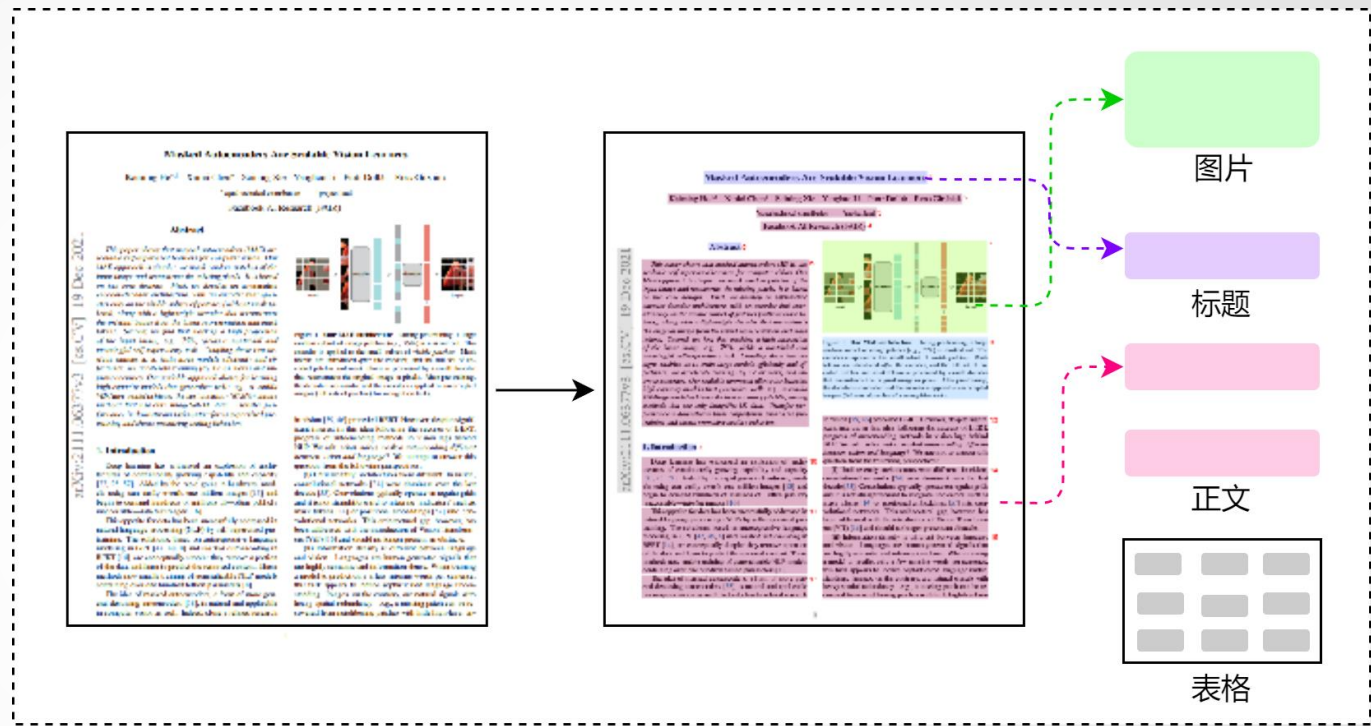
③ 索引构建和存储优化

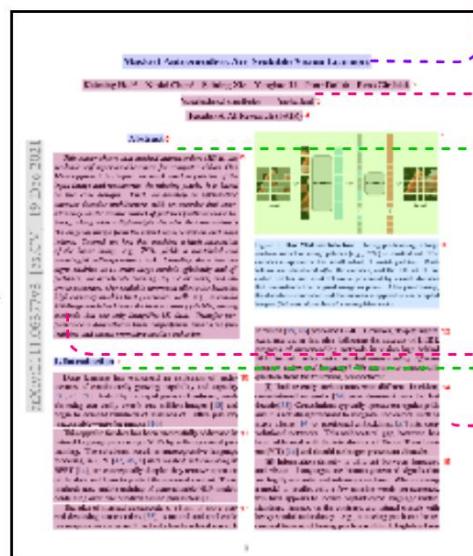
利用**向量数据库**（如FAISS、ElasticSearch、Milvus等）或其他高效的向量检索工具，将处理后的文本数据进行高效的存储和索引。文档解析后的信息会被索引系统存储，并且定期更新，使得后续能动态的高效**存储新的知识**，并且保证检索操作能够高效快速地找到与用户查询相关的文档片段，以保持信息的时效性。



① 文档读取与解析

这是整个流程的**基础**，为后续检索任务提供结构化的数据。文档读取就是把各种格式的文档加载到系统中。常见的文档格式包括**PDF**、**Word**、**Html**、**PPT**等，面对文档格式多样性、内容复杂性和非结构化数据等挑战，可以借助**开源工具** (如MinerU)来提高解析的准确率。解析后的文档内容会被转换为统一的格式，然后进一步处理和标准化。





Masked Autoencoders Are Scalable Vision Learners

标题

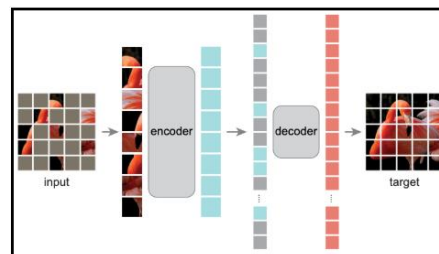
Kaiming He, Xinlei Chen, Saining Xie.....

作者

Abstract

1. Introduction

标题



图片

This paper shows that masked autoencoders.....

正文

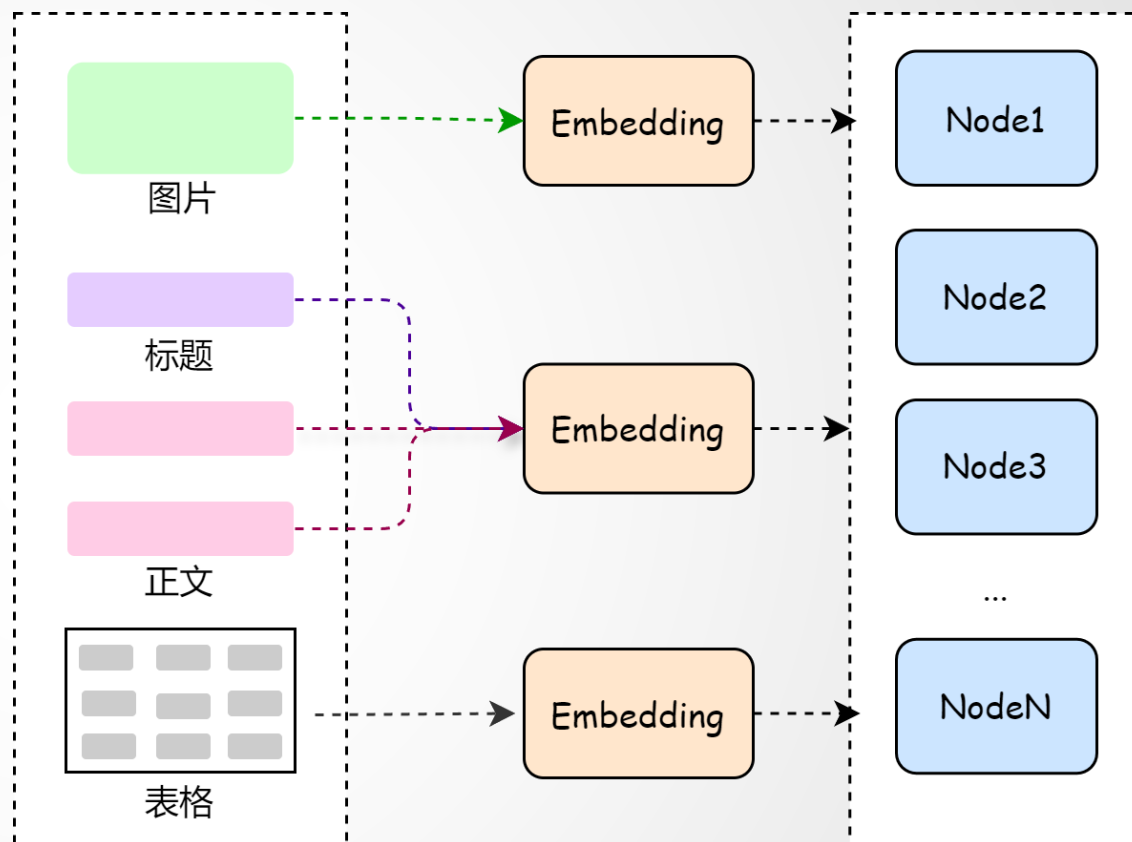
Deep learning has witnessed an explosion of

文档解析



② 预处理

对收集到的原始数据进行**清洗、去重、分块**等预处理工作，去除无关内容和噪声，确保数据的质量。预处理过程中还需要对文本进行**向量化编码**，使其能够被高效检索。这一阶段通常使用深度学习模型或其他文本编码方法（如Word2Vec、BERT等）来生成文本的向量表示，便于后续的检索工作。



标题

Masked Autoencoders Are Scalable Vision Learners

作者

Kaiming He^{✉,†} Xinlei Chen[✉] Saining Xie[✉].....

标题

Abstract

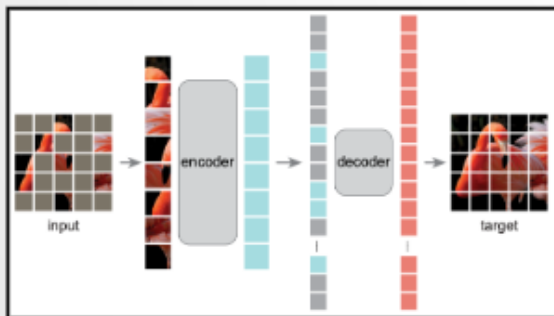
1. Introduction

正文

This paper shows that masked autoencoders.....

Deep learning has witnessed an explosion of

图片



Embedding

Embedding

[-0.019483, -0.00124,
-0.046196, -0.01168,...]

[-0.03, -0.004, 0.020993,
0.00462152, 0.0158111,...]

[0.0020972, 0.0454002,
-0.004875471,...]

[-0.00524014, -0.03730351,
0.0365762747,...]






[0.010654919, -0.00192339,
-0.0061225499,...]

[-0.0033790657, -0.022992,
-0.009422365,...]

.....



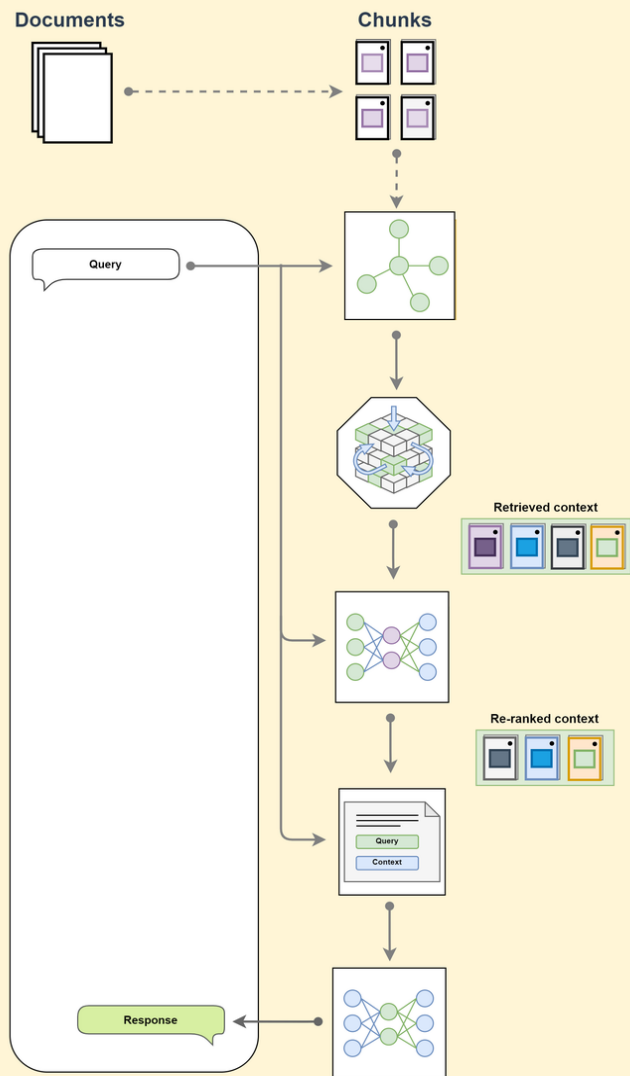
目录

-  1. 上节回顾
-  2. 大模型的简介及其基本概念
-  3. 大模型的局限及其应对策略
-  4. RAG的基本概念和工作流程
-  5. RAG的更多可能性



Retrieve-and-rerank RAG

Retrieve-and-rerank



通过引入**重排序** (Reranking) 步骤, 更好地筛选检索结果, 提高传递给生成模块 (Generator) 上下文的质量。

流程:

用户输入一个查询 → 检索模块从知识库中找到一批初步相关的文档 (向量检索) → 对检索到的文档进行重排序, 以筛选最相关的信息 → 生成模块使用重排序后的文档作为上下文, 生成最终回答。

优点:

- a.提升检索精度: 初步检索模块通常快速但粗略, 重排序能更精准地选择最相关的文档。
- b.减少生成错误: 提供高相关性上下文, 避免生成模块在不相关或错误信息基础上生成答案。
- c.适配长尾查询: 对于少见或复杂的查询, 重排序能进一步优化初步检索效果。

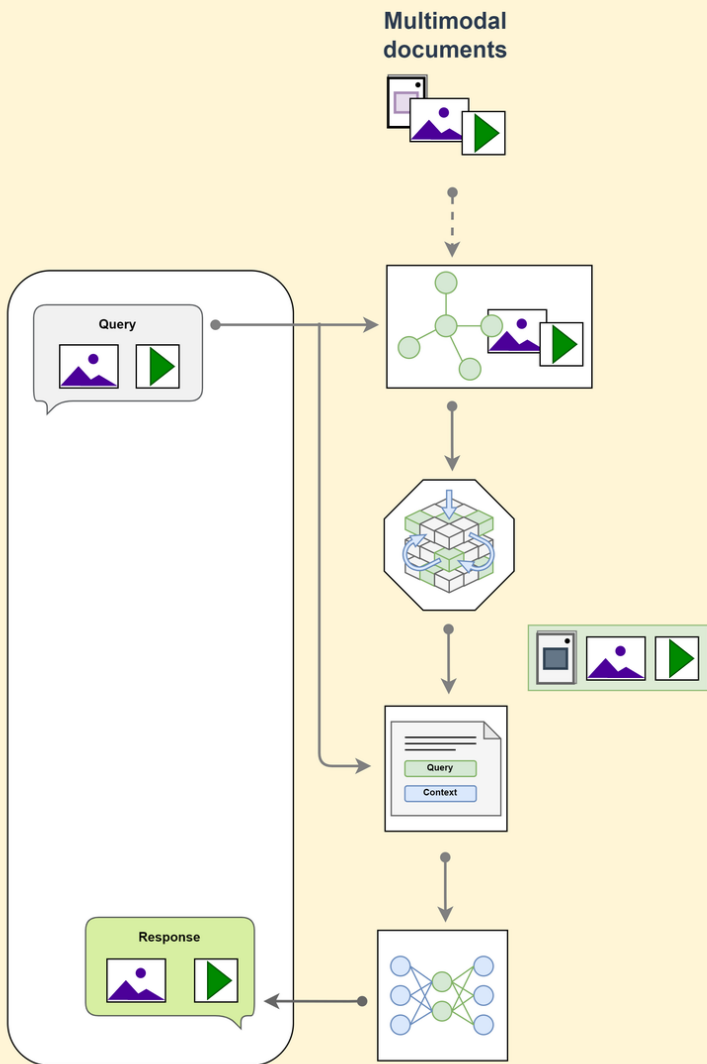
应用场景:

- a.推荐系统: 在搜索和推荐场景中, 重排序步骤可以显著提高最终推荐内容的相关性和用户满意度。
- b.技术支持: 从文档中筛选最相关的答案, 减少生成模块的错误回答率。



Multimodal RAG

Multimodal RAG



也是由检索模块和生成模块组成，但增加了对**多模态数据**的支持。

流程：

用户输入可以是文本或其他模态（图像等） → 多模态检索模块找到与输入相关的多模态上下文 → 将检索结果传递给生成模块 → 结合上下文信息生成多模态回答。

优点：

- a.支持多种输入类型：除文本问题外，还能处理图像、视频等相关的查询。
- b.增强的上下文理解：将文本、图像等模态上下文结合起来，生成更准确、更有深度的内容。

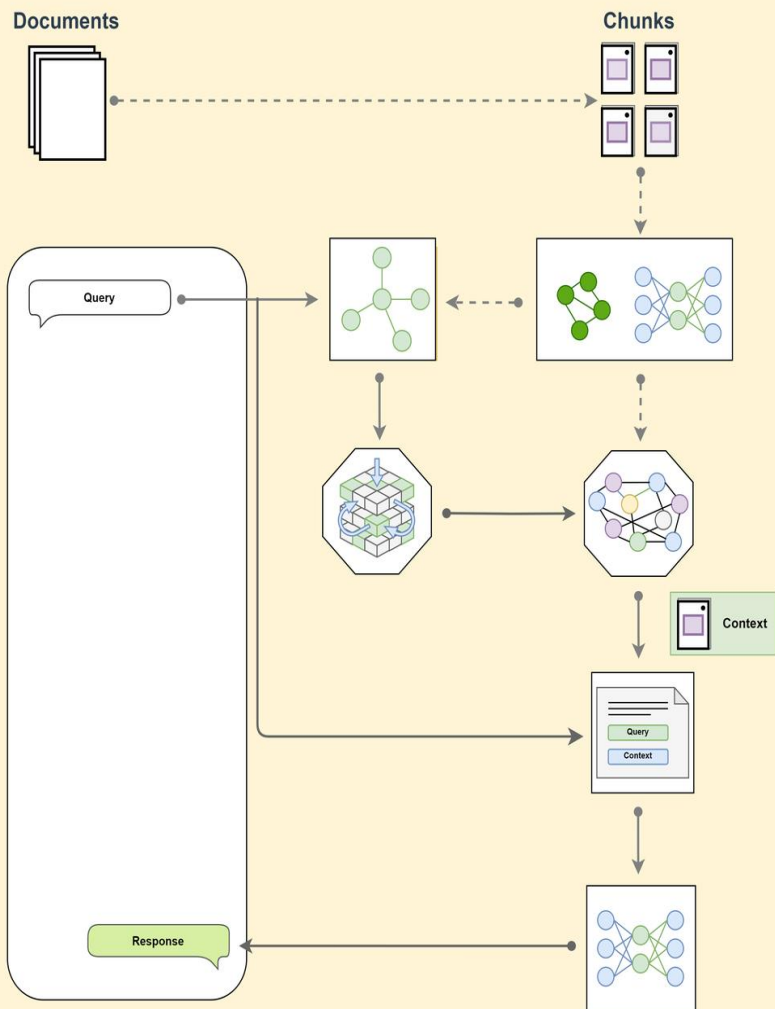
应用场景：

- a.医疗诊断：结合医学文本和影像数据，生成诊断报告或建议。
- b.内容生成：从视频或音频中提取关键信息并生成摘要或分析报告。
- c.图像描述生成：为图像生成自然语言描述，适用于教育或辅助工具。



Graph RAG

Graph RAG



该流程是对基础 RAG 架构的一种扩展，通过引入图数据库来增强知识点之间的关联和文档间关系的理解。

流程：

知识建模（从知识库或文档集合中提取实体、关系和文本内容，构建图数据库）→ 用户查询（用户输入问题，将查询转换为图查询）→ 检索与用户问题相关的子图 → 上下文扩展（将检索到的子图中的信息转化为文本上下文，并传递给生成模块）→ 内容生成。

优点：

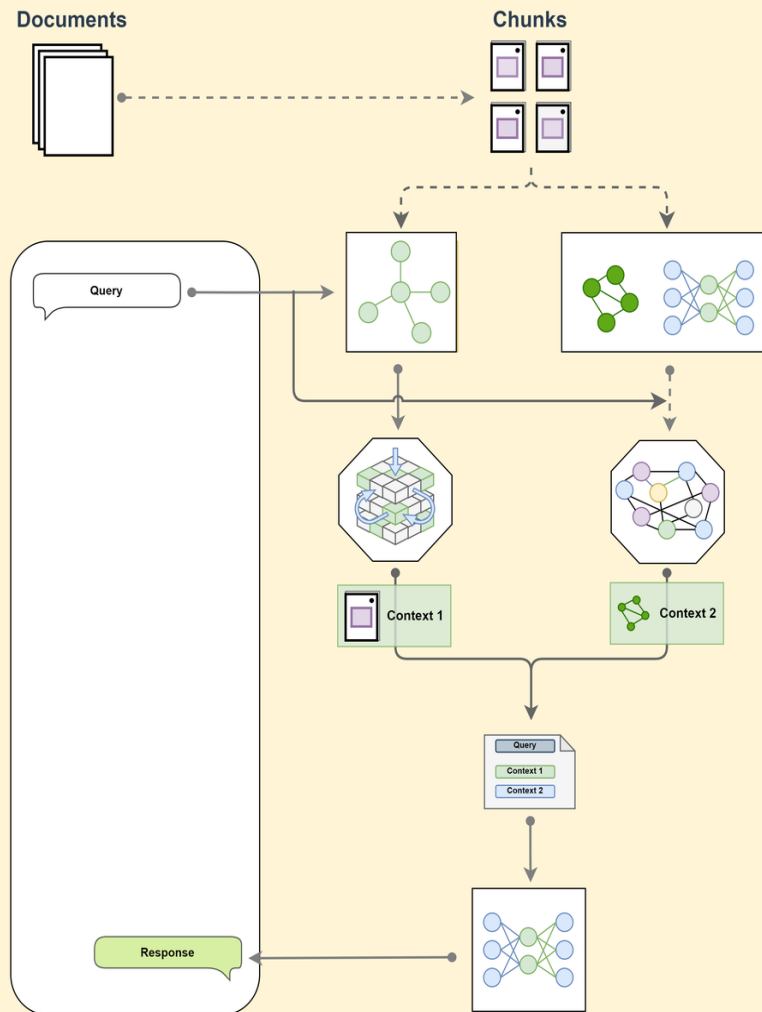
- a. 知识点间关系的深度挖掘：通过图结构，捕捉文档或知识点之间的复杂关系（如层次关系、因果关系等），提高检索结果的质量。
- b. 上下文的精准扩展：在检索阶段，图数据库可以帮助找到更相关的上下文，而不仅仅依赖向量相似性。
- c. 增强推理能力：利用图的结构化数据，可以进行关系推理，例如多跳检索（从一个节点找到间接相关的节点）。
- d. 动态更新与维护：图数据库支持动态更新，易于在知识库扩展时维护新数据的关系。

应用场景：

- a. 复杂问答和推理问题：需要跨文档或跨实体推理的问答任务，如法律问答或科技文献分析。
- b. 知识管理：在企业或科研机构中，利用图数据库管理和查询大量关联文档或研究成果。

Hybrid RAG

Hybrid RAG



这是一种结合了多种检索方法和生成方式的架构，旨在优化检索的覆盖率和生成的准确性。

流程：

输入处理（用户提出查询后，系统首先对输入的查询进行预处理） → 向量检索 + 图检索 → 将向量检索和图检索的结果进行融合 → 增强提示构建（构建一个增强的提示，结合用户的查询和检索到的信息，同时这个提示将被用作生成模型的输入） → 内容生成。

优点：

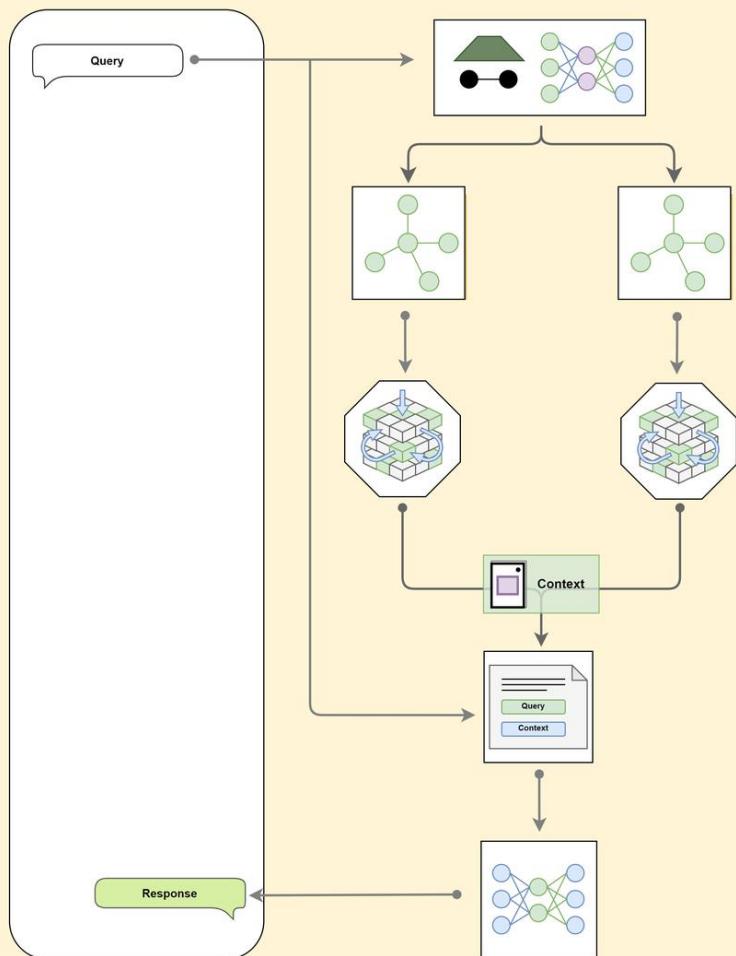
- a.高检索覆盖率：结合了向量检索和图检索，提高了检索的覆盖率和准确性。
- b.增强上下文关联性：通过整合不同检索系统的结果，更深入地理解实体间的关系及其出现的上下文，增强了内容关联性。
- c.动态推理能力：知识图谱可以动态更新，使系统能够适应新信息的可用性，增强了系统的推理能力。

应用场景：

- a.复杂问答系统：理解用户的查询，从文档中检索信息，并生成准确和详细的答案。这种系统特别适用于需要准确信息检索和生成的场合，如在线帮助中心、客户服务等。
- b.对话系统：在对话系统中，Hybrid RAG可以生成更自然、更相关的回复，提高用户体验。这对于聊天机器人和虚拟助手等应用尤为重要，它们需要提供信息丰富且连贯的对话。
- c.文档生成：利用检索到的信息和知识图谱的结构化数据来创建内容丰富、逻辑清晰的文档，适用于报告生成、内容创作等领域。
- d.内容推荐：Hybrid RAG可以分析用户的兴趣和偏好，检索和生成推荐内容，提升推荐的个性化和准确性。这对于新闻推荐、电商产品推荐等场景非常有用。

Agentic RAG

Agentic RAG(Router)



通过引入**AI Agent** 作为路由器，根据用户的查询动态选择最合适的处理路径或模块。它在复杂、多任务场景中具有明显优势，因为不同查询可能需要不同的数据源或处理逻辑。

流程：

用户输入问题或任务描述 → Router分析查询的意图和模态 → 模块选择（调用文本、图像检索模块，或同时调用多模态检索模块） → 内容生成。

优点：

a.查询分析与智能重构：Agentic RAG能够精细分析和重构原始用户查询，将模糊或复杂的查询转化为更精确、可检索的形式，并智能路由判断是否需要额外的数据源来全面回答问题。

b.多源数据检索：能够灵活地从多个数据源检索信息，包括实时用户数据、内部文档和外部数据源，打破信息孤岛，提供更全面的答案。

c.动态答案生成与优化：Agentic RAG不满足于仅仅给出一个答案，而是通过多轮迭代不断优化，生成多个候选答案并评估每个答案的准确性和相关性，必要时重新查询或调整生成策略。

应用场景：

a.医疗辅助：动态调用医学图像分析模块、文献检索模块或诊断生成模块。

b.教育内容生成：根据学生的问题选择合适的资料来源并生成解释。

c.自动化工作流：处理复杂查询时，调用外部工具（如计算器、翻译器、编程执行器）完成多步骤任务。



Q&A

1. RAG相比于将文档直接上传给大模型让它回答，有什么不同



感谢聆听
Thanks for Listening

