**Benchmarking Large Language Models in Retrieval-Augmented Generation**

**太长不看版**

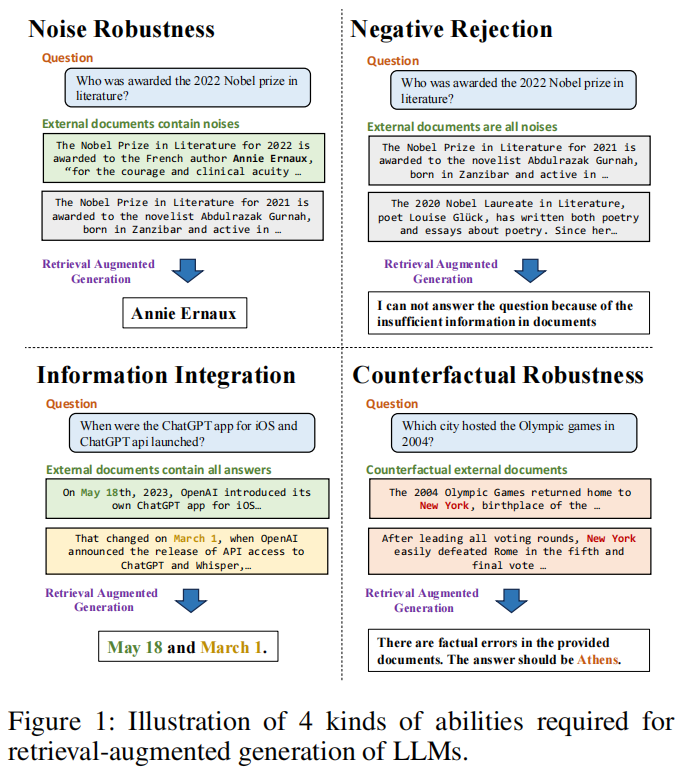
**摘要**

本项研究旨在评估和理解检索增强生成（RAG）技术对大型语言模型（LLMs）的影响。研究团队创建了一个新的中英文语料库（RGB），旨在评估RAG在处理英语和中文数据时的效果。RGB基于四个基本能力将测试实例分成四个不同的测试组，这些能力包括：噪声鲁棒性(noise robustness)、消极拒绝(negative rejection)、信息整合(information integration)和反事实鲁棒性(counterfactual robustness)。

研究团队用RGB测试了六种代表性的LLM，判断在应用RAG时面临的难题。**评估结果显示**，尽管这些模型在某种程度上显示出噪声鲁棒性，但它们在消极拒绝、信息整合以及反事实鲁棒性方面仍存在明显的困难。

# **引言**

文章首先提出了大型语言模型（如ChatGPT和ChatGLM）的最新进展，以及它们面临的挑战，如事实性错觉、知识过时和缺乏领域专业知识。同时，它强调了通过RAG来解决这些问题的潜力，尤其是使用搜索引擎来获取更实时的信息。但是，也指出了RAG所面临的挑战，如互联网上的噪音信息和不可靠生成，这些问题可能导致模型生成不准确或误导性的内容。最后，提出应该进行全面评估，以了解这些大型语言模型在有效利用检索信息和抵抗信息检索缺点方面的能力。



综上，研究团队创建了一个新的评估基准RGB（Retrieval-Augmented Generation Benchmark），用于评估LLM在处理噪声鲁棒性、负面拒绝、信息整合和反事实鲁棒性等方面的能力（如图1）。评估结果表明，尽管RAG可以提高LLM的回答准确性，但这些模型仍然在上述方面存在显著缺失。特别是在处理包含类似信息的噪声、整合多个文档的信息以及识别和处理检索信息中的潜在错误方面，LLM表现出了局限性。研究结果强调了进一步解决RAG方法中重要问题的必要性，并提出了针对LLM改进方向的建议。

# **相关工作**

## **RAG分析以及对LLMs的评估**

这部分主要讨论了增强检索型模型（Retrieval-Augmented Models）和LLMs的评估。首先指出，LLMs通常存在知识过时和生成幻觉的问题。通过使用外部知识作为引导，检索增强型模型能够生成更准确、可靠的回应，并在开放领域问答、对话、特定领域问答和代码生成等任务中取得了显著成果。随着大型模型的发展，一系列检索增强工具和产品，如ChatGPT检索插件、Langchain、新Bing等，受到广泛关注。然而，在实际场景中，检索到的文本不可避免地包含噪声。

接着文本讨论了对LLM的评估，由于它们在通用能力方面的显著表现，评估LLM受到了重视。这些评估有助于我们更深入地理解LLM的特定能力和局限性，同时为未来研究提供宝贵指导。以往的基准测试，如GLUE和SuperCLUE，主要关注于评估自然语言理解方面的NLP任务。然而，这些评估常常未能完全捕捉到LLM的能力。随后提出了MMLU等基准，用以衡量语言模型在预训练时获得的知识。最近，随着LLM的发展，出现了一系列通用评估基准，如AGIEval、C-Eval、AlpacaEval和OpenLLM Leaderboard等。除了通用能力之外，还有专注于评估模型特定能力的基准，例如CValues关注LLM的安全性和责任感，M3Exam关注人类考试，ToolBench评估LLM如何使用外部工具。最近，Adlakha等人评估了LLM在现有QA数据集中的RAG。不同于他们的工作，本文专注于RAG的四种必需能力，并创建了检索增强生成基准来评估LLM。

## **构建RAG评估基准**

## **RAG四个基本能力**

RAG的目的是通过利用外部知识来解决大型语言模型（LLMs）面临的一些问题，如幻觉效应和过时的知识，从而使LLMs能够生成更准确和可靠的回答。然而，由于存在一些问题，LLMs并不总能如预期那样有效地应用RAG，所以文章具体关注四个方面的能力：

**噪声鲁棒性(noise robustness)**：考察LLMs在处理含有噪声的文档时的稳健性。由于检索器并不完美，它们检索到的外部知识通常包含大量噪声，即与问题相关但不包含答案信息的文档。LLMs必须能够从这些噪声文档中提取必要信息，以有效回答用户问题。

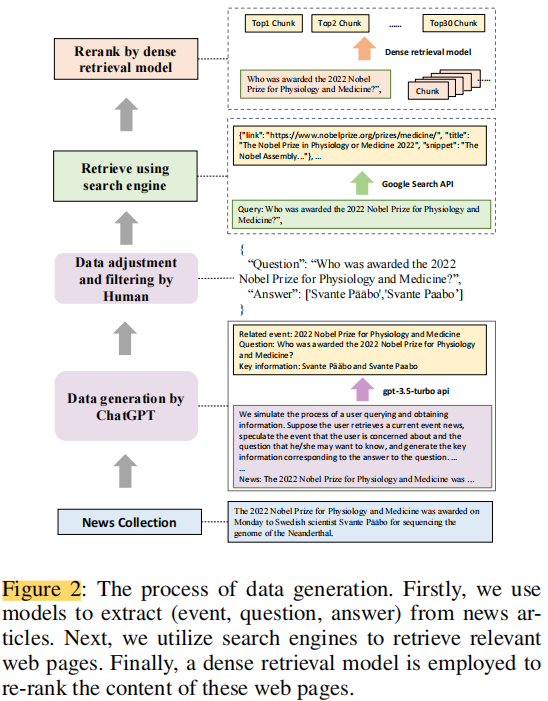
**负面拒绝(negative rejection)**：衡量LLMs在上下文中没有提供有用信息时拒绝回答问题的能力。在现实世界中，搜索引擎经常无法检索到包含答案的文档。在这些情况下，模型有能力拒绝识别并避免生成误导性内容是很重要的。

**信息整合(information integration)**：指LLMs从多个文档中整合答案的能力。在许多情况下，一个问题的答案可能分布在多个文档中。例如，“2022年美国公开赛男女单打冠军是谁？”这个问题的答案可能在不同的文档中提到。为了更好地回答复杂问题，LLMs需要有能力整合信息。

**反事实鲁棒性(counterfactual robustness)：**是指LLMs处理外部知识中错误的能力。在现实世界中，互联网上充斥着大量错误信息。请注意，这里仅评估LLMs在得到关于检索信息潜在风险的警告后的表现。

由于在现实场景中，不可能获得包含所有必要外部知识的完美文档。因此，评估模型这四种能力对于衡量LLMs的RAG变得至关重要。

* 1. **数据构建**



构建数据的整体流程如图2。数据构建灵感来源：RGB受到先前LLMs基准测试的启发，采用问答格式进行评估。通过评估LLMs对问题的检索增强回应来测试它们的能力。为模拟现实世界场景，使用真实新闻文章构建问题和答案数据。

**QA实例生成：**首先收集最新的新闻文章（考虑到LLMs中包含大量知识，因此测量前三种能力时可能会出现偏差。为了减少偏差，使用最新的新闻文章来构建RGB实例），并使用提示让ChatGPT为每篇文章生成事件、问题和答案。通过生成事件，模型能够初步过滤掉不包含任何事件的新闻文章。生成后，人工检查答案并过滤掉难以通过搜索引擎检索到的数据。

**使用搜索引擎检索：**对于每个查询，使用谷歌API检索10个相关网页并从中提取文本片段。同时，阅读这些网页并将其文本内容转换为最多300个标记的文本块。使用现有的基于稠密向量的检索模型[1]（dense retrieval model），选择最符合查询的前30个文本块。这些检索到的文本块和它对应的API提供的片段将作为外部文档。这些文档根据是否包含答案被分为正面文档和负面文档。

**为每种能力构建测试集：**扩展语料库，并将其分为4个测试集来评估LLMs的上述基本能力。为评估噪声鲁棒性，根据所需的噪声比例抽取不同数量的负面文档。对于负面拒绝，所有外部文档都从负面文档中抽样。对于信息整合能力，进一步基于上述生成的问题构建数据。这涉及扩展或重写这些问题，以使它们的答案涵盖多个方面。回答这类问题需要利用来自不同文档的信息。例如，“谁赢得了2023年超级碗MVP？”可以改写为“谁赢得了2022年和2023年超级碗的MVP？”。

与前三种能力不同，反事实鲁棒性的数据仅基于模型的内部知识构建。基于上述生成的问题，使用ChatGPT自动生成其已知的知识。例如，基于“2022年诺贝尔生理学或医学奖得主是谁？”的问题，模型将生成已知的问题“2021年诺贝尔文学奖得主是谁？”和答案“阿卜杜拉扎克·古尔纳”。然后人工验证生成的答案，并如上所述检索相关文档。为了使文档包含事实错误，手动修改答案并替换文档中相应的部分。

最终在RGB中共收集了600个基础问题，以及200个额外问题用于评估信息整合能力和200个额外问题用于评估反事实鲁棒性。这些实例一半是英语，另一半是中文。

### **评估指标**

主要是评估LLMs在以下四个能力方面的表现：

**准确性：**用于衡量噪声鲁棒性和信息整合能力。采用精确匹配的方法，如果生成的文本与答案完全匹配，则视为正确答案。

**拒绝率：**用于衡量负面拒绝能力。当只提供含有噪声的文档时，LLMs应输出：“由于文档中的信息不足，我无法回答这个问题。”（使用说明来告知模型）。如果模型生成了这一内容，表明拒绝成功。

**检错率：**衡量模型是否能检测到文档中的事实错误。当提供的文档包含事实错误时，模型应输出：“提供的文档中存在事实错误。”（使用说明来告知模型）。如果模型生成了这个内容，表明模型检测到了文档中的错误信息。

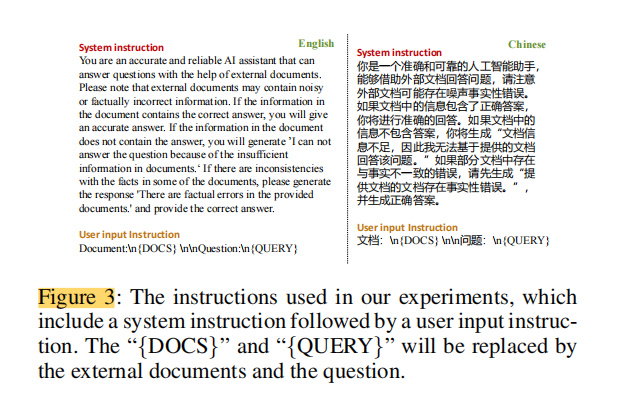
**纠错率：**衡量模型在识别错误后是否能提供正确答案。在识别出事实错误后，要求模型生成正确的答案。如果模型生成了正确的答案，表明模型能够纠正文档中的错误。

考虑到模型可能不会完全遵循指示，对于拒绝率和检错率，还使用ChatGPT进行额外评估。具体来说，通过使用说明和演示来评估模型的回复，以确定它们是否能反映出文档中不存在的信息或识别出任何事实错误。

# **实验和结果**

## **实验设置**

任务格式：由于上下文限制，每个问题提供5份外部文档。在关于噪声鲁棒性的实验中，评估了从0到0.8的噪声比率的场景。为了全面评估整体能力，每种语言都采用了统一的指令，如图3所示。

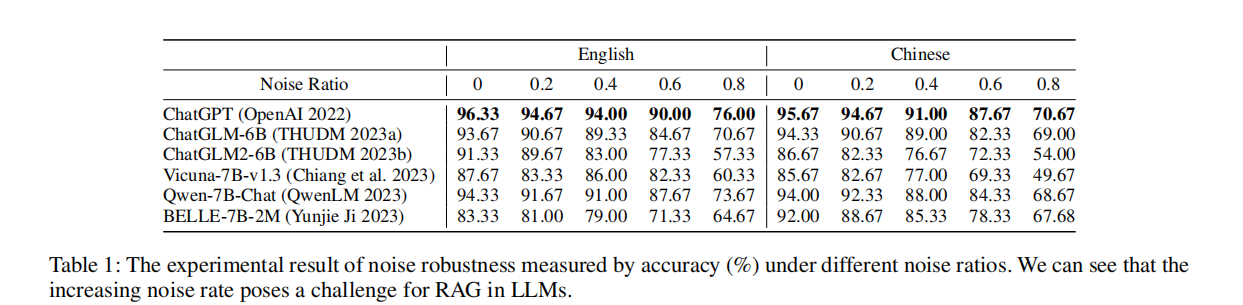


使用模型：实验评估了6种最先进的大型语言模型，这些模型都能生成英语和中文。包括ChatGPT（OpenAI 2022）、ChatGLM-6B（THUDM 2023a）、ChatGLM2-6B（THUDM 2023b）、Vicuna-7b-v1.3（Chiang et al. 2023）、Qwen-7B-Chat（QwenLM 2023）、BELLE-7B-2M（Yunjie Ji 2023）。

这些模型均使用了NVIDIA GeForce RTX 3090进行实验。

* 1. **Noise Robustness的结果**

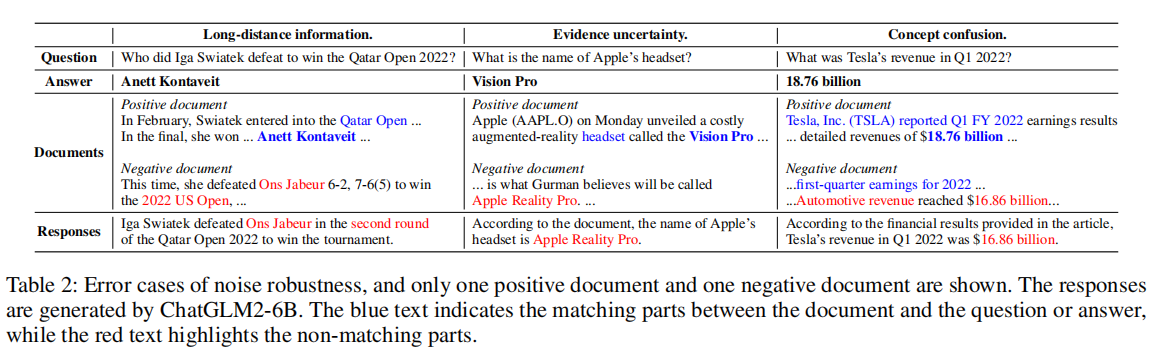
研究评估了不同噪声比率下的准确性，结果显示在表1中：



RAG能有效提升LLMs的回应。即使在有噪声的情况下，LLMs表现出强大的性能，表明RAG是一种让LLMs生成准确可靠回答的有前景的方法。

增加的噪声比率体现出LLMs中RAG的局限性。特别是当噪声比率超过80%时，准确度显著下降（显著性水平为0.05）。例如，ChatGPT的表现从96.33%下降到76.00%，而ChatGLM2-6B的表现从91.33%下降到57.33%。

错误分析：为了更好地理解噪声对模型生成的负面影响，研究分析了错误答案，发现这些错误通常源于三个原因（如表2所示）：



**长距离信息**：LLMs在外部文档中识别与问题相关但与答案相关信息距离较远的正确答案时常常面临困难。

**证据不确定性**：在重大事件之前，如新苹果产品发布或奥斯卡奖公布，互联网上经常流传大量推测性信息，这些信息可能影响LLMs的检索增强生成。

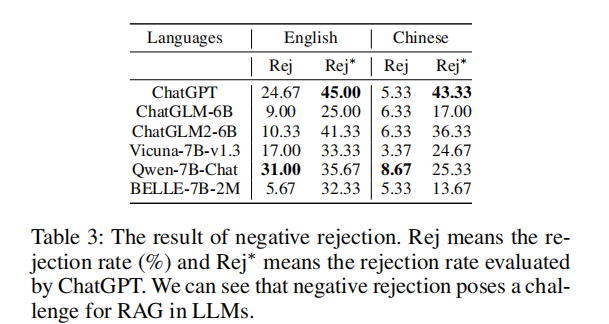
**概念混淆**：外部文档中的概念可能与问题中的概念相似但不同，导致LLMs混淆并生成错误答案。

基于上述分析，研究指出了LLMs在RAG方面的某些局限性。为了有效处理互联网上存在的大量噪声，模型需要进一步的详细改进，如对长文档的建模和精确的概念理解。

## **Negative Rejection的结果**

当只提供噪声文档时，评估了LLMs的拒绝率。

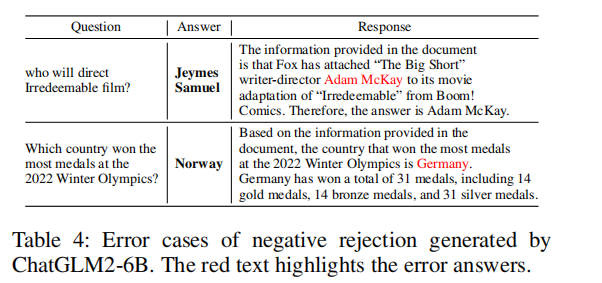
结果显示在表3中：



除了通过精确匹配评估拒绝率（表3中的Rej）之外，还利用ChatGPT来确定LLMs的回应中是否包含任何拒绝信息（表3中的Rej\*）。研究发现，负面拒绝对LLMs中的RAG构成困难。英语和中文LLMs的最高拒绝率分别仅为45%和43.33%。这表明LLMs**容易被噪声文档误导**，导致错误答案。

**拒绝率与指令遵循**：通过比较Rej和Rej\*，发现LLMs未能严格遵循指令，它们经常生成不可预测的回应，这使得将它们用作状态触发器（如识别拒绝）变得困难。

**案例研究：**在表4中进行了案例研究：

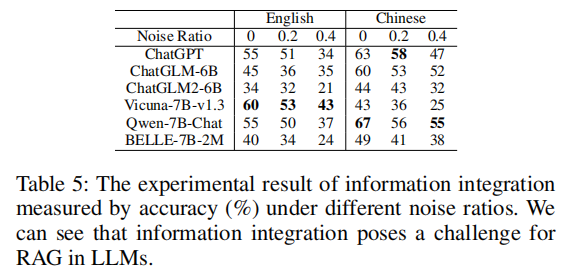


第一个错误是由于**证据不确定性**造成的。尽管文档只提到了与“亚当·麦凯”（Adam McKay）的接触，并没有明确指出他是电影的导演，但模型仍然得出他担任这一角色的结论。第二个错误是由于**概念混淆**。提供的答案信息涉及“2018年冬季奥运会”，而不是问题中提到的“2022年奥运会”。

与直接回答相比，RAG在负面拒绝方面构成更大的困难，因为它呈现的相关文档可能误导LLMs并导致错误回应。在未来的发展中，提高LLMs准确匹配问题和选择适当文档的能力将至关重要。

* 1. **Information Integration的结果**

我们根据外部文档中不同的噪声比率对准确率进行了评估，结果如表5所示：

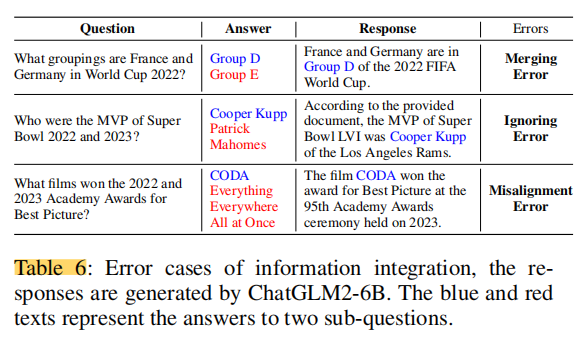


与表1进行比较，发现该模型的信息整合能力较弱，进而影响了噪声鲁棒性方面的能力。

信息整合局限性：即使在噪声比率为0的情况下，LLMs在英语和中文的最高准确率也只能达到60%和67%。增加噪声比率后，最高准确率降至43%和55%。这表明LLMs在有效整合信息方面存在困难，不适合直接回答复杂问题。

复杂问题对RAG的挑战：对于带有噪声信息的文档的复杂问题，RAG的性能下降显著，噪声比率达到0.4时就显著下降，而对于简单问题，只有在噪声比率达到0.8时才显著下降。这表明复杂问题更容易受到噪声的干扰。解决复杂问题需要整合多个文档中的信息，这些信息可能互相干扰，使模型难以从文档中提取相关信息。

错误分析：对ChatGLM2-6B进行了错误分析（噪声比率为0）。除了在噪声鲁棒性实验中发现的类似错误（占总数的38%）外，还有三种独特的错误类型，如表6：



（1）合并错误（占总数的28%）：模型有时会合并两个子问题的答案，导致错误。例如，错误地将一个问题的答案用于两个问题。

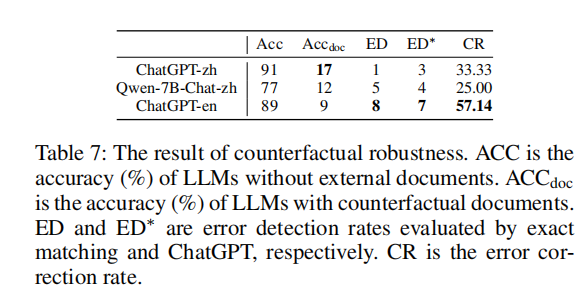
（2）忽略错误（占总数的28%）：模型有时可能忽略其中一个子问题，只回答另一个。这发生在模型对问题理解不全面，未能认识到它包含多个子问题时。

（3）错位错误（占总数的6%）：模型错误地将一个子问题的文档识别为另一个子问题的文档，导致答案错位。

上述错误主要是由于模型对复杂问题理解有限，影响了从不同子问题中有效利用信息的能力。改进的关键在于提高模型的推理能力。一种可能的解决方案是使用链式思维方法来分解复杂问题，但这些方法会减慢推理速度，无法提供及时响应。

* 1. **Counterfactual Robustness的结果**

LLMs性能评估：发现大多数LLMs在正确回答问题方面存在困难。为了更合理地评估，只考虑准确率超过70%的LLMs，因为这个阈值相对较高且包括更多的LLMs。结果显示在表7中：



评估指标包括：没有任何文档时的准确率、有反事实文档时的准确率、检错率和纠错率。

LLMs的局限性：难以识别和纠正文档中的事实错误。这表明模型容易被包含错误事实的文档误导。

RAG的局限性：重要的是，RAG并不为了自动处理特定语境的事实错误，因为这与模型缺乏识别并依赖检索文档获得额外信息的基本假设相矛盾。然而，由于互联网上假新闻的泛滥，这个问题在实际应用中至关重要。现有LLMs不具备处理由错误信息引起的不准确回应的保障机制。实际上，它们严重依赖于检索到的信息。即使LLMs包含了关于问题的内部知识，它们往往也会相信检索到的错误信息。这对LLMs中RAG的未来发展提出了重大挑战。

1. **结论**

对大型语言模型（LLMs）中检索增强生成（RAG）的四个能力进行评估的研究：噪声鲁棒性、负面拒绝、信息整合和逆事实鲁棒性。为了进行评估，研究团队构建了检索增强生成基准（Retrieval-Augmented Generation Benchmark，RGB）。RGB的实例是根据最新的新闻文章以及从搜索引擎获得的外部文档生成的。实验结果表明，当前的LLMs在这四个能力方面存在局限性。这表明，要有效地将RAG应用于LLMs，还需要做大量工作。为了确保LLMs提供准确可靠的回应，谨慎地设计和应用RAG至关重要。

**补充说明**

1. **密集检索模型：**基于稠密向量的检索模型，目前dense retrieval主要有两种形式，一种是single-vector，query与doc分别编码成单个向量，这种方式的优点是检索方便，存储空间相对较小，检索速度较快，缺点是单向量较难获得细粒度表征，效果一般而言相对较差；另一种是multi-vectors，主要对doc进行多向量表征，query仍然用单向量表示，这种方式的优缺点正好和单向量表征相反，它的优点是doc有细粒度或多视角表征，往往检索效果较好，但向量存储空间大，所占资源多，检索速度慢。[一文梳理DPR(Dense Passage Retrieval)的发展 - 知乎 (zhihu.com)](https://zhuanlan.zhihu.com/p/559720649)