Benchmarking Large Language Models in Retrieval-Augmented Generation

摘要

本项研究旨在评估和理解 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。而且 RAG 在解决这些问题展现出巨大潜力。

一、 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

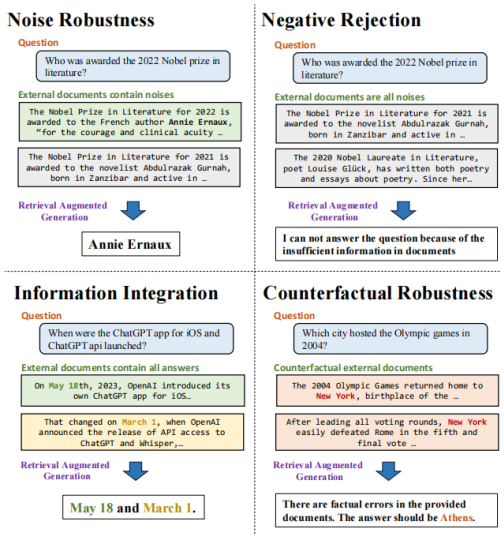

Figure 1: Illustration of 4 kinds of abilities required for retrieval-augmented generation of LLMs.

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。

二、 相关工作

2.1 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。

2.2 构建 RAG 评估基准

2.2.1 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。

逆事实鲁棒性(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。

2.2.2 数据构建

构建数据的整体流程如图 2：

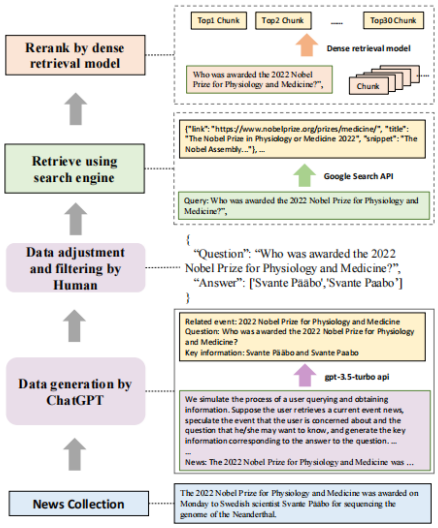

Figure 2: The process of data generation. Firstly, we use models to extract (event, question, answer) from news articles. Next, we utilize search engines to retrieve relevant web pages. Finally, a dense retrieval model is employed to re-rank the content of these web pages.

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**（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**。

数据收集：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。

2.2.3 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。

三、 实验&结果

3.1 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。

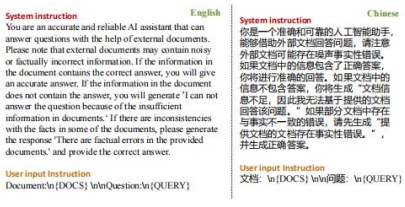

Figure 3: The instructions used in our experiments, which include a system instruction followed by a user input instruction. The “{DOCS}” and “{QUERY}” will be replaced by the external documents and the question.

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。

3. 2 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：

	English					Chinese				
Noise Ratio	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
ChatGPT (OpenAI 2022)	**96.33**	**94.67**	**94.00**	**90.00**	**76.00**	**95.67**	**94.67**	**91.00**	**87.67**	**70.67**
ChatGLM-6B (THUDM 2023a)	93.67	90.67	89.33	84.67	70.67	94.33	90.67	89.00	82.33	69.00
ChatGLM2-6B (THUDM 2023b)	91.33	89.67	83.00	77.33	57.33	86.67	82.33	76.67	72.33	54.00
Vicuna-7B-v1.3 (Chiang et al. 2023)	87.67	83.33	86.00	82.33	60.33	85.67	82.67	77.00	69.33	49.67
Qwen-7B-Chat (QwenLM 2023)	94.33	91.67	91.00	87.67	73.67	94.00	92.33	88.00	84.33	68.67
BELLE-7B-2M (Yunjie Ji 2023)	83.33	81.00	79.00	71.33	64.67	92.00	88.67	85.33	78.33	67.68

Table 1: The experimental result of noise robustness measured by accuracy (%) under different noise ratios. We can see that the increasing noise rate poses a challenge for RAG in LLMs.

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）：

	Long-distance information.	**Evidence uncertainty.**	**Concept confusion.**
Question	Who did Iga Swiatek defeat to win the Qatar Open 2022?	What is the name of Apple's headset?	What was Tesla's revenue in Q1 2022?
Answer	**Anett Kontaveit**	**Vision Pro**	**18.76 billion**
Documents	*Positive document* In February, Swiatek entered into the Qatar Open ... In the final, she won ... **Anett Kontaveit** ... *Negative document* This time, she defeated Ons Jabeur 6-2, 7-6(5) to win the 2022 US Open, ...	*Positive document* Apple (AAPL.O) on Monday unveiled a costly augmented-reality headset called the **Vision Pro** ... *Negative document* ... is what Gurman believes will be called Apple Reality Pro. ...	*Positive document* Tesla, Inc. (TSLA) reported Q1 FY 2022 earnings results ... detailed revenues of **$18.76 billion** ... *Negative document* ...first-quarter earnings for 2022 ... ...Automotive revenue reached $16.86 billion...
Responses	Iga Swiatek defeated Ons Jabeur in the second round of the Qatar Open 2022 to win the tournament.	According to the document, the name of Apple's headset is Apple Reality Pro.	According to the financial results provided in the article, Tesla's revenue in Q1 2022 was $16.86 billion.

Table 2: Error cases of noise robustness, and only one positive document and one negative document are shown. The responses are generated by ChatGLM2-6B. The blue text indicates the matching parts between the document and the question or answer, while the red text highlights the non-matching parts.

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。

3. 3 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：

Languages	English		Chinese	
	Rej	Rej*	Rej	Rej*
ChatGPT	24.67	**45.00**	5.33	**43.33**
ChatGLM-6B	9.00	25.00	6.33	17.00
ChatGLM2-6B	10.33	41.33	6.33	36.33
Vicuna-7B-v1.3	17.00	33.33	3.37	24.67
Qwen-7B-Chat	**31.00**	35.67	**8.67**	25.33
BELLE-7B-2M	5.67	32.33	5.33	13.67

Table 3: The result of negative rejection. Rej means the rejection rate (%) and Rej* means the rejection rate evaluated by ChatGPT. We can see that negative rejection poses a challenge for RAG in LLMs.

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：

Question	Answer	Response
who will direct Irredeemable film?	**Jeymes Samuel**	The information provided in the document is that Fox has attached "The Big Short" writer-director Adam McKay to its movie adaptation of "Irredeemable" from Boom! Comics. Therefore, the answer is Adam McKay.
Which country won the most medals at the 2022 Winter Olympics?	**Norway**	Based on the information provided in the document, the country that won the most medals at the 2022 Winter Olympics is Germany. Germany has won a total of 31 medals, including 14 gold medals, 14 bronze medals, and 31 silver medals.

Table 4: Error cases of negative rejection generated by ChatGLM2-6B. The red text highlights the error answers.

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。

3.4 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：

	English			Chinese		
Noise Ratio	0	0.2	0.4	0	0.2	0.4
ChatGPT	55	51	34	63	**58**	47
ChatGLM-6B	45	36	35	60	53	52
ChatGLM2-6B	34	32	21	44	43	32
Vicuna-7B-v1.3	**60**	**53**	**43**	43	36	25
Qwen-7B-Chat	55	50	37	**67**	56	**55**
BELLE-7B-2M	40	34	24	49	41	38

Table 5: The experimental result of information integration measured by accuracy (%) under different noise ratios. We can see that information integration poses a challenge for RAG in LLMs.

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：

Question	Answer	Response	Errors
What groupings are France and Germany in World Cup 2022?	Group D Group E	France and Germany are in Group D of the 2022 FIFA World Cup.	**Merging Error**
Who were the MVP of Super Bowl 2022 and 2023?	Cooper Kupp Patrick Mahomes	According to the provided document, the MVP of Super Bowl LVI was Cooper Kupp of the Los Angeles Rams.	**Ignoring Error**
What films won the 2022 and 2023 Academy Awards for Best Picture?	CODA Everything Everywhere All at Once	The film CODA won the award for Best Picture at the 95th Academy Awards ceremony held on 2023.	**Misalignment Error**

Table 6: Error cases of information integration, the responses are generated by ChatGLM2-6B. The blue and red texts represent the answers to two sub-questions.

（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。

3. 5 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：

	Acc	Acc_{doc}	ED	ED*	CR
ChatGPT-zh	91	**17**	1	3	33.33
Qwen-7B-Chat-zh	77	12	5	4	25.00
ChatGPT-en	89	9	**8**	**7**	**57.14**

Table 7: The result of counterfactual robustness. ACC is the accuracy (%) of LLMs without external documents. ACC_{doc} is the accuracy (%) of LLMs with counterfactual documents. ED and ED* are error detection rates evaluated by exact matching and ChatGPT, respectively. CR is the error correction rate.

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。

四、 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。

五、 补充说明

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)

精确匹配：完全一样的字符合