# 用RAG提高LLM事实准确性以对抗幻觉：私人知识库中特定领域的案例研究

RAG最近被认为是一个很有前途的解决方案，通过引入外部数据集的上下文信息，可以扩展模型在需要高真实性的任务上的生成能力。将模型连接到外部数据集，RAG可以帮助LLM处理知识密集型任务，并允许持续的知识更新和特定领域信息的集成。

用户查询

→ 数据爬取 → 清洗分块 → QA标注 → 知识库

→ 检索模块 → 上下文检索 → 生成模块 → 答案生成

→ 评估指标 → 优化反馈（微调）

→ 最终答案输出

我们的项目旨在构建一个性能优于基准开源解决方案的系统。具体来说，我们的贡献如下。

1.创建专业数据集：我们使用新颖的网络爬虫和自动注释过程，策划了一个全面准确的数据集，重点关注卡内基梅隆大学（CMU）和语言技术研究所（LTI）。

2. RAG管道的开发：我们利用SOTA的研究成果创新了RAG管道，确保了先进的语言模式处理能力。

3.广泛的基于实验的评估：通过严格的实验，包括消融研究和案例研究，我们证明了我们的系统的性能。

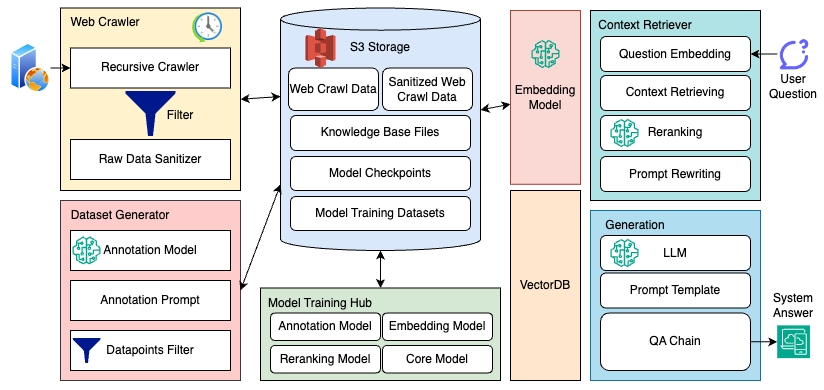


图1：我们的系统设计概述。该过程从Web Crawler开始，它由递归Crawler和过滤器组成，用于收集原始数据。这些原始数据经过清理并存储在S3存储桶中。数据集生成器（图示DG）依赖于注释模型来提示注释和过滤数据点，为系统中的所有模型生成微调数据集（MTH绿色部分），即注释，嵌入，重新排序和核心模型。对于查询处理，上下文检索器（CR）嵌入用户问题并检索相关上下文，必要时重新排序和重写。然后生成模块（蓝色）利用核心模型和提示模板通过QA链生成系统答案，从而产生利用检索信息和生成能力的答案，以获得准确和上下文感知的响应。

1.输入阶段（Input）​

​输入形式：用户提出的自然语言查询​（如“CMU 2024秋季学期何时开学？”）。

​输入预处理：

对查询进行基础清洗（去除特殊符号、拼写纠错），但不做复杂语义解析（依赖后续模块处理）。

​2. 数据处理与知识库构建（Data Pipeline）​

​步骤1：数据爬取（Web Crawling）​

​工具：基于Selenium + BeautifulSoup的定制爬虫。

​数据源：

CMU/LTI官网页面（BFS深度2遍历链接）

学术论文（通过Semantic Scholar API获取LTI教授2023年发表的论文PDF）。

​输出：原始HTML/PDF文件，按类型存储为html/、pdf/目录。

​步骤2：数据清洗与分块（Preprocessing）​

​清洗规则：

删除JavaScript代码、HTML标签、页眉页脚。

过滤不含关键词（如cmu, carnegie mellon）的文档。

移除短文本（<200字符）和无效页面（如“Page\_not\_found”）。

​分块处理：

使用unstructured库将文本转换为纯文本。

按1000词长度分割长文本（适应LLM上下文窗口）。

​输出：结构化文本块，存储为sample/目录下的清洗后文件。

​步骤3：自动化QA对生成（Annotation）​

​标注工具：WizardLM（开源模型，用于生成问答对）。

​流程：

​种子数据生成：从sample/选取高质量文本块，人工验证生成结果。

​批量生成：对全部文本块生成10个QA对，格式为{"question": "...", "answer": "...", "context": "..."}。

​质量控制：计算Cohen's Kappa评分（0.67），过滤低质量样本。

​输出：34,781个QA对，按8:2划分为训练集（27,824）和测试集（6,957）。

​3. 检索增强生成（RAG Pipeline）​

​步骤4：检索模块（Retriever）​

​嵌入模型：mixedbread-ai/mxbai-embed-large-v1

​微调方法：使用训练集QA对，通过SentenceTransformer+MultipleNegativesRankingLoss优化语义匹配。

​输入：用户查询 + 知识库所有文本块。

​流程：

计算查询与文本块的余弦相似度，返回Top-10候选。

​重排序（Reranker）​：使用BAAI/bge-reranker-large对Top-10精排，输出Top-5最相关上下文。

​输出：拼接Top-5文本块作为生成模块的上下文。

​步骤5：生成模块（Generator）​

​核心模型：meta-llama/Llama-2-7b-chat-hf

​微调方法：

​参数：LoRA（rank=16），4-bit量化，batch\_size=8，学习率2e-4，训练5 epochs。

​输入格式：

f"""[INST] <<SYS>>

You are an assistant for question-answering tasks.

Use the retrieved context to answer. Keep answers concise.

<</SYS>>

Question: {question}

Context: {context}

Answer: [/INST]"""

​生成策略：贪心解码（Greedy Decoding），限制最大长度50词。

​输出：生成的自然语言答案（如“Fall 2024学期课程于8月26日开始”）。

​4. 评估与优化（Evaluation）​

​评估指标：

​检索阶段：Recall@5、MRR（Mean Reciprocal Rank）。

​生成阶段：F1 Score、BLEU、余弦相似度（对比参考答案）。

​实验设计：

​Ablation Study：对比不同配置（Baseline/RAG/微调组合）。

​Case Study：分析典型成功/失败案例（如时间敏感问题、复杂领域问题）。

​5. 输出阶段（Output）​

​最终输出：结构化JSON格式

json

复制

{

"question": "When will the classes begin in Fall 2024?",

"answer": "Classes for Fall 2024 will begin on August 26, 2024.",

"context": [".../academic\_calendar\_2024.txt: Fall semester starts on 8/26/2024..."],

"confidence": 0.89 // 基于生成概率的置信度

}

3.

3.1 **Web Crawler**

使用Selenium和BeautifulSoup库构建了一个自定义的Web Crawler来抓取CMU网站。

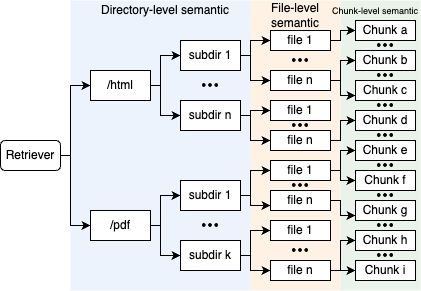
具体来说，我们首先存储说明页面上提供的所有链接。我们还从每个网页中运行深度2的广度优先搜索，所以我们的程序也会访问并存储嵌入其中的链接。

然后，我们对所有链接运行**爬虫程序**。爬虫程序会自动提取所有网页的HTML源代码。由于对潜在问题有用的信息可能很稀疏，因此还进行了**预处理**，包括但不限于删除JavaScript代码和HTML标签以及删除网站文本标题，以确保数据质量。

3.1.2数据组织和后处理抓取的数据**最初以两种格式存储，HTML和PDF，**基于原始内容格式。然后处理这些数据以**提取文本信息**，从而产生组织到html/和pdf/目录中的**文本文件**集合。

html页面和pdf都是从互联网上获得的，我们用这种方式区分数据，以便为检索器提供更多关于文件层次结构的语义信息，如图2所示。

此外，创建了一个样本/目录，以容纳**根据其清洁度和信息量选择的数据点子集，作为初始问答对生成的种子。**



图二：层次知识库文件系统保持了原始文件之间关系的结构信息，丰富了对检索器的语义提供。为了确保数据集的相关性和质量，我们实施了一个**后处理步骤**，**删除了内容或标题中缺少与CMU和LTI相关的关键字的文本文件**（例如，cmu，卡内基梅隆大学，格子呢，完整列表见附录C）。

我们还过**滤掉**了小于200个字符的文件和标题为“Page\_not\_found”的文件，确保我们的数据集既相关又高质量。Python包非结构化用于执行后处理。

这种细致的数据创建和管理过程为我们的检索增强生成系统提供了强大的知识库，使其能够提供与语言技术研究所和卡内基梅隆大学有关的精确和信息丰富的响应。

3.2研究论文另一方面，由于大多数问题可能与CMU和LTI进行的研究相关，我们手工制作了LTI教师名单作为参考，并将其用于Semantic Scholar API中，以搜索和过滤2023年发表的所有开源论文。所有获取的论文都被下载。纸质PDF文件，沿着我们从CMU网站下载的其他PDF文件，转换成纯文本并存储。

从学术论文中生成的文本文件通常长度超过10，000个单词，并且对于LLM上下文窗口来说通常太长。

因此，**所有原始数据都被分割成1000个单词的块，这些块可以在标记化后由大多数LLM处理。**

3.3**注释自动化：**在构建外部数据集之后，我们继续进行注释。我们有304个html文档，12个pdf和245篇论文。虽然人工注释可以达到最高的准确性，但不幸的是，我们没有足够的资源以这种方式注释我们的数据。因此，我们希望探索使用**预训练的LLM**作为数据注释器。我们的三个选项是GPT 4All（Anand等人，2023），LLaMA-2（Touvron等人，2023）和WizardLM（Xu et al，2023），这些数据都是在HuggingFace上开源的。我们从每个数据类别中选择了一些数据点，包括，html，pdf和论文，这些样本的信息量更大，更清晰，可以作为代表。初始QA对是在所有样本的这个选定子集上生成的。这在以下两个方面有意义：（1）我们对三种模型生成的3个样本QA对进行了人工检查，并**选择了WizardLM，**（示例输出A）（2）我们使用人工验证的QA对作为少数例子，以指导未来的QA生成更大的数据集，提高整体注释质量。

总的来说，对于每个数据块，我们使用WizardLM（Xu et al，2023）帮助生成10个QA对。我们总共生成了34，781个QA对，其中27，824对用作训练数据，6，957对用作随机分割后的测试数据。

3.4数据集评估：为了评估数据集内注释的可靠性和一致性，我们采用了**注释者间一致性（IAA）指标，特别关注Cohen's Kappa评分的计算**。该指标对于评估注释者之间的一致性程度至关重要，超出了仅凭偶然性的预期，从而更可靠地了解数据集中的注释质量。

Cohen's Kappa分数（κ）是用公式计算的：κ = po − pe /1 − pe其中po表示注释者之间观察到的相对一致性，pe是偶然一致性的假设概率。使用这个度量，κ分数为1表示完全一致，而分数为0表示没有超出偶然的一致性。另一方面，负值表示注释者之间的不一致。

为了计算po，我们首先统计注释者在数据集的每个类别上达成一致的实例数量。然后，我们将其除以注释的总数。对于pe，我们根据每个类别的注释在注释者之间的分布来计算期望的偶然一致性。

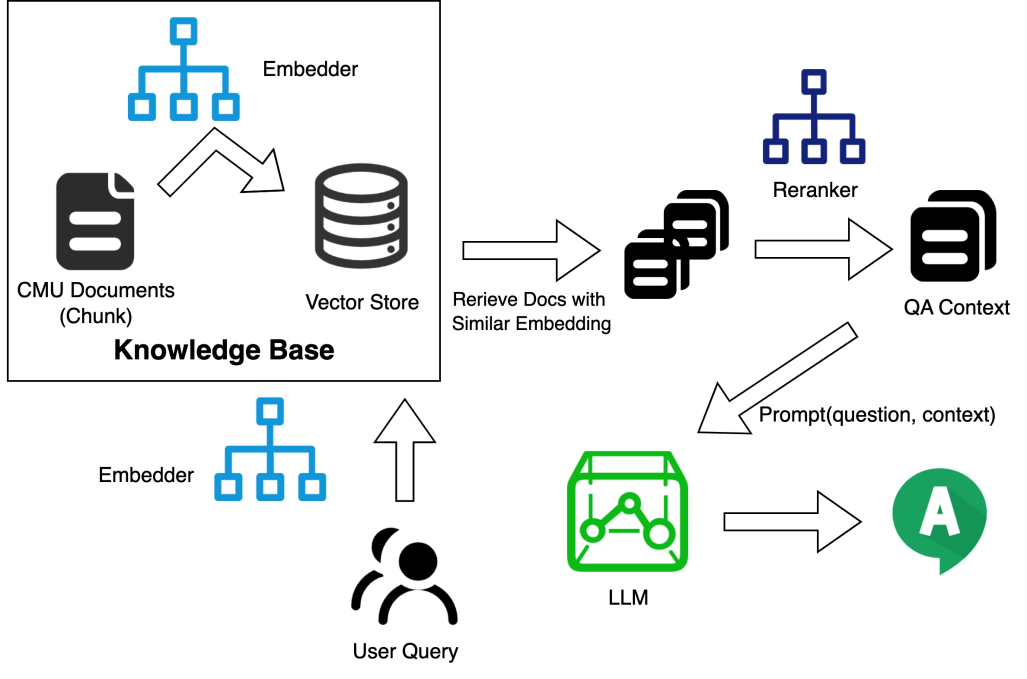
对于我们的数据集，两个注释者独立地对数据的一个子集进行分类，并计算Cohen's Kappa评分以评估其注释的一致性。结果发现，所得的κ评分为0.67，表明与数据集的基本一致性（83.33%）。

确保注释者的高度一致性，我们可以自信地利用这个数据集来训练和微调我们基于RAG的问答系统，旨在提高处理特定领域和时间敏感查询的准确性和可靠性。

1. **RAG QA Pipeline概述**，

分为检索阶段和生成阶段。在检索阶段，检索器获取mmr评分相似度最高的前5个参考块，然后将其发送到重排序模型，以优先考虑给定用户查询中最相关的信息。

在生成阶段，生成模型将重写的提示作为输入并完成答案。



如图3所示，我们展示了我们策划的RAG QA管道。我们首先组装了一个包含CMU和LTI相关信息的自定义数据集，我们在第3节中对此进行了详细说明。除了外部数据集外，我们的RAG管道还包括一个Retriever模型，该模型将从数据集中检索最相关的信息，并将其作为上下文返回，以供以后生成。生成器模型，通常是LLM，将获取返回的信息并将其作为上下文附加到QA提示中以生成答案。

所有模块都是端到端训练的，我们将在下面说明我们的模型规范。

4.1对于检索部分，我们使用了来自www.example.com的**预训练嵌入模型Mixedbread.ai**（HuggingFace模型ID：mxbai-embed-large-v1）（Lee et al，2024），因为它在MTEB基准测试（Muennighoff et al，2022）上的紧凑尺寸和高性能，该测试测量了跨分类，聚类，重新排名等任务的嵌入模型。

它受益于专有的数据集，其中包含超过7亿对使用对比训练，尽管需要额外的3000万个三元组进行微调。数据域的广泛覆盖确保了其在预训练和微调设置上的性能。因此，除了预训练模型外，我们还在QA对数据集的训练分割上对嵌入模型进行了微调，共17，390步，批量大小为8。

4.2重新排序模型：在我们的项目中，我们通过一个名为**BgeRerank的自定义重新排序器来增强查询结果的相关性**，它采用了基于CrossEncoder架构的BAAI/bge-reranker-large模型。该模型复杂地评估查询和文档对，分配反映每个文档与查询意图一致性的相关性得分。通过利用先进的自然语言处理技术，BgeRerank对初始的大量文档进行重新排序，优先考虑前N个（在我们的实验中检索到的10个文档中的前5个）与用户查询最相关的文档。我们将这种重新排序过程集成到检索管道中，确保提供与查询的语义细微差别紧密匹配的结果。

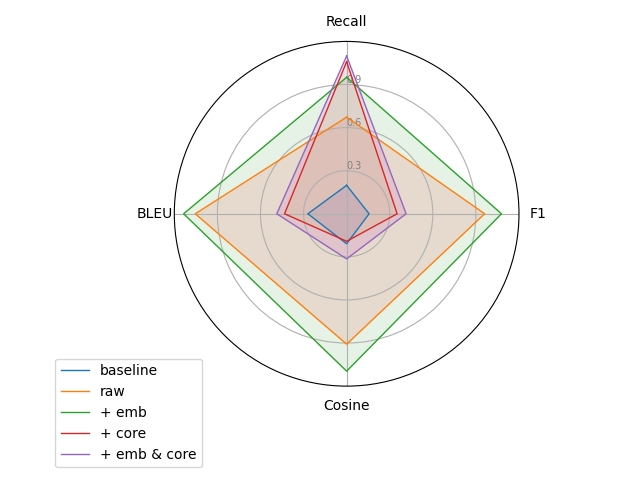
4.3核心型号我们使用**LLaMA-2（Touvron et al，2023）作为核心生成模型**。这个由Meta AI开发的最先进的大型语言模型在广泛的自然语言处理基准测试中表现出令人印象深刻的性能。LLaMA-2在2个Triumph令牌上进行预训练，并在10万人类注释数据上进行进一步微调，可以捕获语言模式和领域知识，使其能够生成高度流畅和连贯的文本。重要的是，LLaMA-2是一个在许可证下发布的开源模型，这使得我们能够对其进行持续的训练和实验。LLaMA-2已经针对可扩展性和效率进行了优化。它使用了组查询注意力（GQA）（Ainslie等人，2023年）。减少Transformer头的数量，GQA显著减少了Transformer块中键值对的数量，从而降低了内存需求。使用我们的定制数据集调整基线。此外，与LLaMA-1相比，其上下文窗口翻了一番，达到4，096个。

**5实验**

5.1设置在我们的实验中，我们广泛测试了对LLaMA-2模型进行微调并在精心构建的QA对数据集上嵌入模型所带来的性能提升。对于核心模型微调，利用HuggingFace的meta-llama/Llama-2- 7 b-chat-hf检查点，我们开始了对QA数据集进行微调的旅程，其中包括问题，答案，和参考文本。微调过程通过将模型量化为INT 4并将LoRA纳入16级以保持计算效率来优化。训练方案跨越5个时期，最大步骤为1，000，批量大小为8，学习率为2 e-4，确保表现和训练时间之间的平衡。

同样，对于被标识为mixedbread-ai/mxbai-embed-large-v1的嵌入模型，我们采用了一种新的方法来使用从我们的数据集中提取的不同QA对集合进行微调。该方法专注于通过SentenceTransformer和MultipleNegativesRankingLoss来增强模型对上下文相关性的理解。嵌入模型在10个epoch上进行了微调，特别强调预热步骤，以使模型逐渐适应任务，从而改进其识别和编码数据中细微语义关系的能力。

这两个微调过程对于根据我们的特定数据集和目标定制模型至关重要，从而显著提升了模型生成上下文相关和准确答案的能力。所有评估方法及其详细信息均在附录D部分中列出。



5.2消融研究

**Recall（召回率）​**

**定义：检索模块从知识库中找到所有相关文档的比例，衡量检索系统的覆盖能力。召回率越高，系统漏检关键信息的风险越低，对时效性查询（如2024年课程时间）尤为重要。**

**F1 Score（F1分数）​**

**定义：精确率（Precision）和召回率（Recall）的调和平均数，综合评估答案的准确性与完整性。F1是平衡“答错”与“漏答”的核心指标，高F1表明系统在生成答案时既精准又全面。**

**Cosine Similarity（余弦相似度）​**

**定义：生成答案与参考答案在语义向量空间的相似性，衡量语义一致性。该指标揭示生成答案是否“正确传达意图”，而不仅是字面匹配。**

**BLEU Score（双语评估替补分数）​**

**定义：通过n-gram重叠度评估生成答案与参考文本的表面文本匹配度。BLEU适用于评估翻译或模板化答案，但对开放域生成任务的解释力有限。**

图4：在不同设置下，我们的本地测试问答数据集的召回率，F1得分，余弦相似度和BELU。注意图表中的数据在0和1之间进行了归一化，以获得更好的可见性。原始实验输出请参见表1。

我们的综合分析集中在对嵌入和核心模型进行微调对特定领域训练数据集的影响上。我们使用附录D中提到的指标来揭示这些微调过程对系统性能的单独和组合影响。

根据表1和图4，在RAG给出的性能提升之上，分析显示，单独和组合地微调嵌入模型和核心模型会导致不同的召回性能轨迹。

对嵌入模型进行微调，可以提高对数据中细微语义关系的理解，这一点可以从召回率和F1分数的提升中得到证明。同样，使用我们的数据集对核心模型进行微调，可以提高其生成精确和上下文相关答案的能力，这反映在召回率的提高上。微调两个模型的累积效果使F1分数下降，特别是从微调核心模型。我们对LLaMA-2- 7 B的微调在模型本身的参数大小和当前计算资源的总计算步骤方面都受到限制。虽然我们努力进一步减少微调的验证损失，数据集也可能很小，并且相对有偏差。微调可能会降低模型在语言生成方面的性能，并破坏我们管道的性能。

我们可以观察到语言生成的影响。我们检查了预测，它们往往很长，并且有“context：“，“answer：“，甚至““这样的标记<INSTR>，这是我们的数据集布局和LLaMA-2预训练中使用的不同提示。

我们的模型再次成功地回答了这个问题，并且比前一个例子更流畅。然而，模型没有使用解释的答案来回答问题，而是重述了从包含问题答案的摘要中提取的信息。这可能意味着微调可能会阻碍模型在进行域转移时在语言生成任务上的性能。另一方面，LLM似乎对文本中的数字很敏感。

1个问题：CMU创建了第一个高速计算机网络的项目的名称是什么？2 3回答：安德鲁项目，由卡内基梅隆大学和IBM的研究人员在20世纪80年代初创建，是开发出第一个高速计算机网络的CMU项目，综上所述，安德鲁计划是一个重要的研究前沿，大学和计算机科学系与IBM在20世纪80年代初建立。2它旨在提供一个高速计算机网络和GUI，的计算环境。安德鲁项目使卡内基梅隆大学成为世界上连线最多的校园，它展示了联网PC或工作站在解决这个问题上的潜力。这是一个更为内在的问题，它需要一个强大的检索模型，才能在所有的文本中找到相应的信息。

我们的模型具有很好的流畅性和上下文布局，可以对中等水平的问题提供满意的答案。

由于我们的模型参数只有70亿个，因此容量有限。有些答案是重复的，与问题无关。然而，我们的例子应该证明，检索准确率足以使LLM在某些情况下回答中等水平的问题

1. 上游数据处理模块

功能：构建结构化知识库

步骤顺序：

​数据爬取​（Web Crawler）

输入：CMU/LTI官网、学术论文（PDF）

输出：原始HTML/PDF文本（含噪声）

工具：Selenium + BeautifulSoup（爬虫）、Semantic Scholar API（论文获取）

​预处理与分块

输入：原始文本

操作：

去噪（删除HTML标签、脚本代码、页眉页脚）

过滤（关键词筛选、短文本过滤）

分块（按1,000词切分长文本）

输出：结构化文本块（存储为知识库）

​自动化标注

输入：文本块

操作：WizardLM生成问答对（QA pairs）

输出：34,781个标注QA对（训练/测试集）

​2. RAG核心管道

功能：检索增强的问答生成

步骤顺序：

​检索模块（Retriever）​

输入：用户查询（Query）

流程：

​嵌入编码：使用mxbai-embed-large-v1模型将查询和知识库文本编码为向量

​相似度计算：余弦相似度匹配Top-K候选文档

​重排序：BgeRerank模型对候选文档精排（优化Top-5）

输出：最相关的文本片段（Context）

​生成模块（Generator）​

输入：查询（Query） + 上下文（Context）

流程：

​提示工程：拼接指令模板（见附录B）

​生成控制：LLaMA-2读取上下文生成答案（限制50词）

输出：初步答案

​3. 下游评估与优化模块

功能：验证系统性能与迭代

步骤顺序：

​多指标评估

输入：生成答案 vs 标注答案

指标：

检索质量：Recall@K、F1

生成质量：BLEU、余弦语义相似度

输出：量化评分（如表1）

​反馈循环

输入：评估结果 + 用户交互数据

操作：

​嵌入模型微调：用训练集优化检索精度

​生成模型微调：LoRA适配器调整LLaMA-2（受限于数据规模）

输出：更新后的模型参数

​4. 系统输出与知识库更新

功能：闭环数据流

​最终输出：用户端返回答案（含上下文引用）

​知识库更新：新爬取数据（如更新论文）增量并入知识库

​关键承接关系

​数据驱动闭环：

知识库构建 → 检索增强生成 → 评估反馈 → 模型迭代 → 知识库更新

​模块依赖：

检索质量直接依赖知识库的覆盖度与标注准确性

生成质量受检索结果和LLM微调策略双重影响

评估结果反向驱动嵌入模型（优先）和生成模型（谨慎）优化

​错误隔离设计：

若检索失败（低相似度），系统直接返回“未知”答案，避免LLM自由发挥导致幻觉。

​LLM的局限性：尽管LLM（如ChatGPT、LLaMA-2）在问答任务中表现优异，但其训练数据的局限性和参数记忆能力不足可能导致“幻觉”，尤其在处理私有知识库​（如企业内部数据）或时效性强的查询时（如最新学术成果）。

​目标：构建一个端到端系统，通过RAG整合外部数据，提升LLM在特定领域（以卡内基梅隆大学（CMU）及其语言技术研究所（LTI）为例）的答案准确性。

方法创新

​定制化数据集构建：

​数据爬取：使用自动化爬虫抓取CMU官网、LTI相关网页及2023年发表的学术论文，并通过关键词过滤、文本分块等处理，构建结构化知识库。

​自动化标注：利用WizardLM模型生成34,781个问答对（QA），并通过Cohen's Kappa评分验证标注一致性（评分0.67，显示较高可靠性）。

​RAG流程设计：

​检索模块：采用Mixedbread.ai的嵌入模型（mxbai-embed-large-v1）进行语义检索，结合BgeRerank模型重排序，精准提取相关上下文。

​生成模块：基于LLaMA-2模型，将检索到的上下文融入生成过程，确保答案的准确性和相关性。

​模型微调：对嵌入模型和生成模型分别进行微调，适配特定领域数据。

​实验与评估：

​性能指标：通过召回率（Recall）、F1分数、余弦相似度和BLEU评分等评估系统表现。

​关键发现：

RAG显著提升答案准确性（F1分数从0.186增至0.304）。

​嵌入模型微调有效提升检索质量，但生成模型微调因数据规模小和偏差导致生成质量下降。

​案例验证：系统能准确回答时效性问题（如课程时间）和复杂领域问题（如CMU历史项目“Andrew Project”），但存在重复生成和上下文冗余现象。

结论与启示

​RAG的有效性：结合外部知识库的RAG系统能显著缓解LLM的幻觉问题，尤其适合知识密集型任务。

​微调的局限性：小规模、有偏差的数据集可能损害生成模型的通用能力，需平衡领域适应与模型泛化。

​开源贡献：公开了代码、模型和数据集，为后续研究提供基准。

# CRAG -综合RAG基准（Benchmark

检索增强生成（RAG）是近年来出现的一种很有前途的解决方案，以缓解大型语言模型（LLM）缺乏知识的缺陷。

然而，现有的RAG数据集并不能充分反映现实世界中问答（QA）任务的多样性和动态性。为了弥补这一差距，我们引入了综合RAG基准（CRAG），一个4，409个问答对和模拟API，用于模拟Web和知识图谱（KG）search.CRAG被设计为封装跨越五个领域和八个问题类别的各种各样的问题，反映从流行到长尾的各种实体流行度，和时间动态范围从几年到几秒。我们对这个基准的评估突出了与完全值得信赖的QA之间的差距。而大多数先进的LLM在CRAG上实现了34%的准确率，以直接的方式添加RAG仅将准确率提高到44%。最先进的行业RAG解决方案仅回答了63%的问题而没有任何幻觉。CRAG还揭示了在回答关于具有较高动态性，较低流行性，或更高的复杂性，建议未来的研究方向。CRAG基准为KDD Cup 2024挑战奠定了基础，吸引了数千名参与者和提交。我们致力于维护CRAG，以服务于研究社区，推进RAG解决方案和一般QA解决方案。CRAG可在www.example.com上获得https://github.com/facebookresearch/CRAG/

大型语言模型（LLM）已经改变了自然语言处理（NLP）任务的面貌，特别是在问答（QA）中[20，22，38，39]。尽管取得了进步，但幻觉问题仍然是一个重大挑战; LLM可能会产生缺乏事实准确性或基础的答案[14，27，30，32].研究表明，GPT-4在回答涉及缓慢变化或快速变化事实的问题时，准确率低于15% [36];即使对于稳定的（永远不变的）事实，GPT-4在回答涉及躯干到尾部（不太流行）实体的问题时的准确率也低于35% [29]。因此，克服幻觉成为构建可靠QA系统的优先事项[13，14]。

检索增强生成（Retrieval-Augmented Generation，RAG）[6，8，12，19]是近年来出现的一种有前途的解决方案，可以缓解LLM在知识缺乏方面的不足，并吸引了学术界和工业界的大量关注。给定一个问题，RAG系统搜索外部资源以检索相关信息，然后提供有根据的答案[7，12，13]。19]（见图1）。

尽管RAG具有巨大的潜力，但它仍然面临着许多挑战，例如选择最相关的信息，减少问题回答延迟，以及综合信息来回答复杂的问题。

目前缺乏一个全面的**基准**来推动这一领域的持续研究工作。

我们的目标是建立一个基准，可以提供一个重要的功能和快速，但可靠的评估RAG推进该领域的发展的整体视图。什么是一个好的基准QA超过LLM？我们考虑五个关键特征。

**1.现实主义：**首先，一个好的基准测试应该最好地反映真实的用例。换句话说，在基准测试中达到高指标的解决方案在真实的场景中也应该表现得很好。例如，RAG基准测试中的问题应该与人们在真实的QA场景中提出的问题相似。

**2.富裕**：基准测试应包含一组不同的实例类型，涵盖常见用例和一些复杂和高级用例，以代表现实世界的挑战并揭示现有解决方案的可能局限性。

**3.洞察力：**基准测试应便于理解不同数据切片的性能，反映解决方案应对不同类型挑战的能力。

**4.可靠性：**基准应允许对指标进行可靠的评估：基本事实应准确;指标应很好地捕捉模型的性能;评估应简单可靠，计算的指标应具有统计意义。

**5.寿命**：最后，为了能够长期进行研究和实验比较，基准中的场景和数据不应很快过期，理想情况下应随着时间的推移进行更新和改进。

我们努力创建一个具有上述所有功能的基准，我们称之为CRAG -RAG的综合基准。

我们的第一个贡献是数据集本身（第3节）. CRAG包含来自五个领域的4，409个问答对：金融，体育，音乐，电影和开放领域。除了简单的事实问题外，（询问实体的属性），CRAG包含七种类型的复杂问题以覆盖真实的用户查询：**带条件的问题、比较问题、聚合问题、多跳问题、集合查询、后处理繁重问题和错误前提问题。**

CRAG反映了从流行到长尾的**各种实体流行度**，**时间跨度**从数秒到数年不等，可以轻松深入洞察。在生成问题时，我们参考了智能助手用例以确保问题的现实性，解释了问题以增加表达的多样性，并手动验证了基本事实以确保可靠性。

除了QA对之外，CRAG还提供了模拟API来模拟从各种可用信息中检索。这包括从真实世界的搜索引擎Brave Search API [5]返回的每个问题的多达50个完整的HTML页面，以及具有260万个实体的模拟KG。对于模拟KG，我们故意确保检索候选项反映现实环境中的噪声。

我们的第二个贡献是**评估机制**，以允许可靠的比较。我们设计了**3个任务**来测试RAG解决方案中的不同组件：检索摘要、知识图和Web检索以及端到端检索增强生成（第2节）。我们的评分系统不是计算正确回答问题的百分比，而是区分幻觉答案和缺失答案，并给予前者更高的惩罚，因为它对破坏用户的危害更大

我们还设计了一个有效的**自动评估机制**，以允许快速评估和迭代（第4节）。

我们的第三个贡献是对简单的RAG解决方案和行业最先进的RAG解决方案进行了**全面评估**（第5节）。尽管大多数先进的LLM在CRAG上的准确度达到34%，但以简单的方式添加RAG仅将准确度提高到44%。

最先进的行业RAG解决方案只能回答63%的问题，没有任何幻觉，在回答具有较高动态性，较低流行性或较高复杂性的事实问题时，准确率仍然很低。这些评估有两个作用：第一，它们表明CRAG具有，并允许从基准包含的不同维度中获得见解;其次，他们强调了一个完全值得信赖的QA系统的差距和研究方向。

CRAG基准为2024年KDD Cup挑战赛奠定了基础2，在比赛的前50天内吸引了数千名参与者和提交。我们致力于维护CRAG，以服务于研究社区，推进RAG解决方案和通用QA解决方案。

与现有基准的比较。表1将CRAG与现有的事实问题回答基准进行了比较。传统的QA基准，如自然问题（NQ）[18]，TriviaQA [16]，MS MARCO [4]和QALD-10 [35]在过去十年中已经推进了QA，但只考虑Web检索或KG检索的内容，并不能充分代表RAG面临的多样化和动态的挑战。LLM或RAG的新基准通常针对QA系统的某些功能。研究人员创建了基准来评估系统回答简单知识问题的能力[23，29，30]和处理更高级的场景。这些包括回答问题与改变答案[36]，整合来自多个文档的信息[6]，解决多跳问题[33]，以及回答问题与长文本[26]。

此外，传统的QA基准测试通常采用基于匹配的度量，如ROUGE [21]或F1来评估响应的质量[16，18]。这些度量虽然适用于提取方法，但已知对于生成自由形式响应的LLM来说**不是**非常有效[11]。

尽管CRAG比MS MARCO和NQ更小，但它提供了几个独特的优势：**全面的覆盖范围，模拟API的真实测试，动态问题处理，多样化的事实流行度，维基百科之外的广泛内容，以及快速而可靠的评估。**这些功能使CRAG成为测试RAG系统和广泛QA系统的强大而通用的基准，提供了一个共享的测试平台来评估这些系统如何处理真实世界，动态的，多样化的信息检索和可靠的基于LLM的问题回答合成的挑战。

2 Problem Description

RAG QA系统以问题Q为输入，输出答案A;答案由LLM根据从外部来源检索的信息或直接从模型中内化的知识生成。答案应提供回答问题的有用信息，而不添加任何幻觉

我们设计了**三个任务。它们共享相同的（问题，答案）对，**但在可用于检索以增强QA的外部数据方面有所不同。在这里，我们提供了可以在QA中使用的内容，以确保公平比较。我们在第3节中描述了我们如何生成数据。

**任务1：检索摘要**。在任务1中，我们为每个问题提供最多五个网页。

这些网页很可能（但不保证）与问题相关。本任务旨在测试RAG系统的答案生成能力。

第二项任务：**KG和Web检索增强**。在任务2中，我们还提供了模拟API来访问底层模拟KG中的信息。模拟KG存储与问题相关的结构化数据;问题的答案可能存在于模拟KG中，也可能不存在。模拟API采用输入参数，通常从问题中解析，该任务测试RAG系统1）查询结构化数据源和2）综合来自不同来源的信息的能力。

任务三：**端到端RAG**。与任务2类似，任务3也提供了Web搜索结果和模拟API作为检索的候选项，但提供了50个网页作为候选项，而不是5个。较大的网页集合更有可能提供回答问题所需的信息，但同时也更有可能包含噪音。因此，任务3还测试了RAG系统如何对大量检索结果进行排序。

这三个任务**都是在前一个任务的基础上增加**的，可以测试RAG系统的不同功能。RAG系统中唯一没有被这些任务覆盖的组件是**搜索检索**。人们可以很容易地扩展这些任务，使用我们基准测试中的所有22万网页作为搜索语料库，进行完全的端到端测试。

3数据集描述

CRAG包含两部分数据：

QA对和检索内容。我们现在描述每部分数据。数据生成详细信息可参见附录A.1.1-A.1.6。

3.1问答对

CRAG涵盖五个领域：金融，体育，音乐，电影和开放领域，以及八种类型的问题，所有都是英语。问题类型列于表2。我们从底层KG和Web内容构建了问答对。

**从KG构建**的QA对。我们通过收集一组基于公开可用数据的实体，然后基于选定的实体类型和关系创建600多个问题模板，从KG构建QA对。接下来，我们从KG中采样具有不同流行度的实体（头部，躯干和尾部），以填充模板并生成完整的问题和答案。

**从网页内容构建**的问答对。我们要求注释者写下用户可能会问的问题（例如，“2023年最受欢迎的动作片”），并从相应的网页搜索结果中创建问答对。

知识图谱（KG）​ 是存储结构化数据的数据库，包含实体、属性及其关系。在 CRAG 中，KG 用于支持复杂问题的回答，例如查询股票价格、电影信息等。

使用上述方法，我们收集了2，425个Web问题和1，984个KG问题，其中661、658和665 KG问题分别包含头部、躯干和尾部实体。表3和表4总结了问题在不同维度上的分布。（例如，快速变化的事实）使我们能够获得误差幅度为1.5%的指标（95%置信水平）。动态分布大致反映了域的性质（例如，金融领域的实时问题要比其他领域的实时问题多得多）。有关动态类别的定义，请参见附录A.1.2。

3.2检索内容我们包括**两种类型的检索内容**，以模拟RAG的实际场景：Web搜索和**KG搜索**。

**Web搜索结果**。对于每个问题，我们使用问题文本作为搜索查询，并从Brave搜索API [5]中存储多达50个HTML页面。参见附录A.1.5中的表8。

我们用一种基于启发式的方法估计了网络搜索召回率：首先检查在页面中是否找到了地面实况答案URL;如果没有，则确定地面实况中的事实是否包含在具有LLM的页面片段或内容中。特别地，我们将问题、地面实况和页面传递给Llama 3 70 B指令，并要求它判断上下文是否足以回答问题。

图2显示了所有CRAG问题的估计网络搜索召回率曲线，当使用所有50个页面时，总体召回率为85%。它反映了基准测试的多个设计优势。首先，召回率曲线在开始时很尖锐，之后变平，前5页的召回率约为69%。这与我们在开发RAG系统时的实际观察结果相当。非完美覆盖率，尤其是对于任务1，允许我们测试当检索结果不包含必要的信息时，RAG解决方案是否承认“我不知道”。第二，与从网页片段的召回相比，完整网页的召回提高了约20%，强调了提取和理解HTML内容的重要性。此外，估计的网络搜索召回率（50个网页）对于网络问题是93%，对于KG问题是74%，表明**KG问题的召回率明显低于网络问题**。这与我们的观察结果一致，即躯干和尾巴实体的网络搜索召回率通常较低，强调在任务2和任务3中利用幼儿园的重要作用。

模拟KG。我们创建了模拟KG，其中包含用于生成问题的公开可用KG数据，随机选择的相同类型的实体，以及具有相似名称的“硬否定”实体（例如，“phantom”表示“phantom of the opera”）。

模拟API。我们创建了带有预定义参数的模拟API，以支持模拟KG中的结构化搜索。例如，对于询问股票价格的查询，示例模拟API的形式为get\_price\_history（ticker）。

我们同时收集KG和Web搜索数据的快照，同时提出实时和快速变化的问题。这种方法确保我们在回答问题时捕获信息世界的“快照”。在基准测试中表现良好的RAG解决方案还应该能够随着时间的推移进行推理，并推广到不断变化的问题。

总共，生成的数据包含22万个网页，260万个实体的KG和38个模拟API。有关模拟API的完整列表，请参见附录中的表9。

4测试和评估在本节中，我们介绍了评估RAG系统的指标，并在附录A. 2. 3中简要描述了2024年Meta KDD Cup挑战。

4.1**度量**我们使用评分方法来评估RAG系统的性能。对于评估集中的每个问题，我们首先根据以下标准将答案标记为完美，可接受，缺失或不正确。

完美。响应正确回答了用户的问题，并且不包含任何幻觉内容。

可接受。响应为用户的问题提供了有用的答案，但可能包含不损害答案有用性的小错误。

缺少。响应是“我不知道”、“对不起，我找不到......"、系统错误（如空响应）或系统请求澄清原始问题。

不正确。响应提供了错误或不相关的信息来回答用户的问题。

我们使用一种评分方法，对每个完美、可接受、缺失和不正确的答案分别打1分、0.5分、0分和-1分，其中我们惩罚幻觉答案，并优先选择缺失答案而不是不正确答案。然后，我们将真实性定义为给定RAG系统评估集中所有示例的平均得分。

4.2**评估**与以前的工作类似[37]，我们采用了**两种**评估方法，（人工评估）和基于模型的自动评估（自动评分）。在前者中，我们使用手动评分来判断每个答案的完美，可接受，缺失和不正确。在后者中，我们将完美和可接受合并，称之为准确，并使用三种评分系统，1，-1，0表示准确，不正确，和失踪的答案

我们设计了一个自动评估的两步方法：如果答案与地面事实完全匹配，则认为它是准确的;否则，我们使用LLM来确定响应是准确的，不正确的还是缺失的。为了避免自我偏好问题[25]，我们使用**两个LLM评估器**：聊天GPT（gpt-3.5-turbo-0125）[24]和Llama 3（llama-3- 70 B-instruct）[2]并报告平均准确率，幻觉率，缺失率，我们的离线实验表明，这两个模型对每个RAG系统的真实性得分。与人类评估相比，分步法得出ChatGPT的平均F1评分为94.7%，Llama 3的平均F1评分为98.9%。更多详情见附录A.2.2。

测试数据分割。我们将数据随机分割为验证集（30%）、公共测试集（30%）和私有测试集（40%），并发布验证集和公共测试集（附录A.2.3）。KDD杯挑战赛的参与者可以使用验证和公共测试集来开发和测试他们的模型，提交的解决方案在私有测试集上进行评估。CRAG的未来离线用户可以使用验证集进行开发，微调和验证，并使用公共测试集进行测试和结果报告。

5基准测试

在本节中，我们介绍了LLM和RAG系统在CRAG上的性能，证明CRAG具有合理的难度，可以帮助获得见解并显示RAG技术开发的方向。

5.1简单的RAG解决方案

实验设置：我们首先在CRAG公共测试集上运行LLM专用解决方案，该测试集有1，335个问题，使用简单的提示，鼓励简短的回答，并在信心较低时回答“我不知道”（附录A.3.1）。我们雇用了Llama 2 Chat（llama-2- 7 b-chat和llama-2- 70 b-chat）[34]，Llama 3指令（llama-3-8B-instruct和llama-3- 70 B-instruct）[2]，混合（Mixtral-8x7B-Instruct-v0.1）[15]，Falcon（40B）[3]，FLAN-T5（FLAN-T5-XXL）[9]和GPT-4 Turbo（gpt-4-turbo-2024-04-09）[1]。我们评估的仅Web RAG解决方案（任务1）使用固定长度的Web上下文窗口（Falcon和FLAN-T5的代币为1 K，Llama 2 Chat为2K，Llama 3 Instruct和GPT-4 Turbo为4K）;我们使用数据的原始顺序作为参考文本连接网页片段，直到填满窗口（类似于[17，23，36]）。我们基于KG的解决方案（任务2、3）另外使用了固定长度的KG上下文窗口（Falcon和FLAN-T5的令牌为0.5K，Llama 2 Chat的令牌为1 K，Llama 3 Instruct和GPT-4 Turbo的令牌为2K），以包含来自模拟API的结果;我们使用llama-3- 8B-instrument和上下文学习来提取相关查询实体（类似于[28]），并将所有适用的模拟API返回的结果连接在一起我们在附录A.3.2中提供了所有LLM的广泛比较，并重点关注表现最好的LLM（即Llama 3 70 B指令和GPT-4 Turbo）

表5显示了两个自动评估器（ChatGPT和Llama 3）的平均评估结果，并说明CRAG基准测试是不平凡的。首先，最好的仅限LLM的解决方案（GPT-4 Turbo）的准确率仅为34%，真实性为20%，显示出很大的改进空间。

第二，直接的RAG解决方案获得了高达44%的准确率，表明额外的信息确实有助于可靠地回答更多的问题，有趣的是，没有一个RAG解决方案获得高于20%的真实性，这是因为所有的RAG解决方案都引入了更多的不相关的检索结果产生的幻觉，显示了一个很大的挑战，**在RAG-如何明智地使用检索结果**，而不被检索噪音分心？第三，我们发现任务2的真实性得分高于任务1，这表明**KG知识有助于提高准确性**，但幻觉率相似甚至更低，因为KG知识通常是简短但精确的。不幸的是，改进是平庸的，这表明RAG的第二个挑战-如何最好地利用KG数据的力量？最后，任务3的真实性也高于任务2，因为搜索排名更好（回想一下，任务1和2提供了从前10个搜索结果中随机选择的5个页面）和更好的搜索回忆。特别是，我们发现，超过30%的问题的基本事实在Web检索结果中是可用的，但由于上下文窗口的限制而没有包括在提示中。2这表明搜索排名在RAG中的重要性。

图3显示了跨领域、动态性、流行度和问题类型维度的自动评估结果。结果揭示了许多有趣的观察结果，并表明CRAG基准测试可以得出更有见地的结论。首先，它显示了基准测试的哪些部分更难。例如，我们发现**金融和体育领域的RAG真实性要低得多**，对于real 8表6：使用行业SOTA RAG系统对CRAG问题进行基准测试。完美，可接受（Acc.），幻觉（Hall.），缺失率（Miss.）和人类评估报告的真实性（Truthh）以百分比表示。最好的系统实现了51%的真实性，并为高达63%的问题提供了完美的答案。

时间和快速变化的事实，对于尾部实体，以及需要设置答案，后处理和错误前提的复杂问题。其次，它显示了利用检索结果的难度。以流行度切片为例，我们观察到GPT-4 Turbo的真实性从头部（21%）到躯干（11%）再到尾部（8%）下降，与过去的观察结果一致[29];然而，基于GPT-4 Turbo的直接RAG解决方案提高了躯干（+7%）和尾部实体（+6%）的QA质量，但降低了头部（-4%）的质量。最后，尽管我们的目标不是比较不同的LLM，但不同的维度使我们能够了解每种方法的优点和缺点。例如，尽管基于Llama 3 70 B指令的RAG系统具有比基于GPT-4 Turbo的RAG系统更低的总体真实性分数，但是其在回答简单和比较问题时具有相似或稍高的真实性，而在回答集合和后处理问题时真实性低得多，建议对推理能力进行调查。

5.2最先进的行业解决方案

接下来，我们在CRAG公共测试集上评估了**行业最先进的（SOTA）RAG**解决方案。我们选择了五个基于SOTA LLM和搜索引擎构建的RAG系统，使用CRAG问题对其进行查询，收集响应，并应用手动评分（详见附录A.4）。

此外，我们对问题应用了流量权重，以了解真实用例中的解决方案。流量权重反映了真实的QA流量中每个问题类型的频率，如表2所定义。我们对所有域赋予相同的权重，并报告了跨域的宏观平均值。

这是因为我们在不同的用例中观察到了非常不同的域级分布，但在查询类型级别上观察到了类似的分布。

表6和图4显示了SOTA系统的总体性能及其在不同维度上的性能。评估结果证实了我们的信念，即CRAG基准测试揭示了有趣的见解，并显示了现有RAG解决方案的改进空间。首先，SOTA解决方案的结果实现了**更好的真实性（最高51%）**。**然而，幻觉率从16%到25%不等，所以答案仍然不可信**。请注意，SOTA解决方案和直接解决方案之间的真实性得分并不完全可比，因为它们对检索内容具有不同的访问（附录A.3和A.4.1），前者使用auto-eval，后者使用human-eval;然而，这种趋势是有效的。第二，我们观察到非常不同的延迟，范围从3.4秒到11.6秒，反映在权衡延迟和质量方面的不同设计选项;例如，Copilot Pro具有最高的真实性，但同时具有最高的延迟，而Meta SG [10]具有中等真实性，但具有最低的延迟。（更多结果以及我们如何测量延迟，请参见附录A.4.2。）第三，我们在简单解决方案中看到的最困难的切片对于SOTA解决方案来说仍然很困难：实时和快速变化的查询，以及关于躯干和尾巴实体的问题，显示当系统依赖于检索结果来回答问题时处理检索噪声所需的改进;作为另一个例子，我们看到需要聚合、多跳推理或后处理的查询的真实性较低，显示了推理的改进空间

最后，与简单的解决方案相比，SOTA解决方案中关于集合和错误前提问题的真实性显著提高，显示了RAG系统在提供准确和完整的集合答案和检测错误前提方面的进步。

6结论本文提出了CRAG，这是一个丰富而全面的基准，旨在推进检索增强生成（RAG）的研究。通过详细的实证研究，CRAG回顾了现有RAG解决方案的差距，并为未来的改进提供了有价值的见解。我们计划继续改进和扩展多语言问题，多模态问题，多轮对话等基准，确保CRAG保持在最前沿，推动RAG研究，适应新兴的挑战，并为新的研究需求而发展。

# 检索增强生成（RAG）的综合研究：发展、现状和未来方向

本文对检索增强生成（Retrieval-Augmented Generation，RAG）进行了全面的研究，追溯了它从基本概念到当前技术发展的历程。RAG将检索机制与生成语言模型相结合，以提高输出的准确性，解决了LLM的关键局限性。

本文探讨了RAG的基本架构，重点研究了如何将检索和生成相结合来处理知识密集型任务，并详细回顾了RAG的重大技术进步，包括检索增强语言模型的关键创新以及在问答、摘要和基于知识的任务等各个领域的应用。

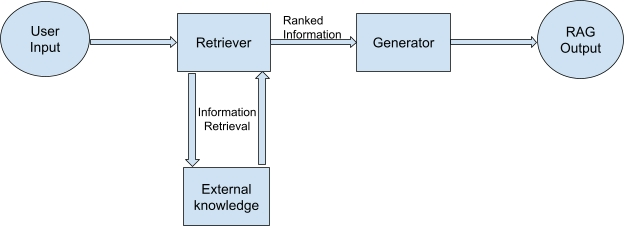
本文讨论了RAG模型的最新研究进展，重点介绍了提高RAG模型检索效率的新方法，分析了RAG模型在部署过程中面临的可扩展性、偏见和伦理问题等挑战，并提出了未来的研究方向，重点是提高RAG模型的鲁棒性，扩大RAG模型的应用范围，本次调查旨在为研究人员和实践者提供基础资源，以了解RAG的潜力及其在自然语言处理领域的发展轨迹。

**自然语言处理（Natural Language Processing，NLP）**已经成为人工智能（Artificial Intelligence，AI）中的一个重要领域，其应用范围从简单的文本分类到更复杂的任务，如摘要、机器翻译和问题回答。NLP的一个重要分支是**自然语言生成（Natural Language Generation，NLG）**，其目标是使机器能够生成连贯、相关和上下文感知的文本，从而改善人机交互（Gatt et. al. 2018）.随着人工智能的发展，对更具有情境感知和基于事实的生成内容的需求不断增加，这给NLG带来了新的挑战和创新。

传统的NLG模型，尤其是序列到序列架构（Sutskever et al 2014），在生成流畅连贯的文本方面取得了显著进步。然而，这些模型往往严重依赖于训练数据，在为需要超出训练集的知识的查询生成事实准确或上下文丰富的内容时，往往会遇到困难。因此，GPT之类的模型（拉德福et al 2019）或基于BERT（Devlin et al 2019）文本生成器易于产生幻觉，它们产生了看似合理但不正确或不存在的信息（Ji等人，2022年）这一局限性促使人们探索将检索机制与生成能力相结合的混合模型，以确保流利性和RAG是一种新的算法和方法，它在许多领域都得到了广泛的应用，但是，由于RAG的复杂性，在对这一领域进行充分的调查以跟踪这一领域的演变和最近的变化方面存在着差距。2目前的调查旨在填补这一空白间隙

1.2检索-扩充生成（Retrieval-Augmented Generation，RAG）是一种新兴的混合体系结构，旨在解决纯生成模型的局限性。（i）检索机制，其从外部知识源检索相关文档或信息，以及（ii）生成模块，这种结合使得RAG模型不仅能够生成流畅的文本，还能够将其输出建立在真实世界的最新数据基础上。

RAG中的检索模块通常**利用密集的矢量**表示来识别来自大型数据集（如Wikipedia或专有数据库）的相关文档。一旦检索到这些文档，就会将其传递给生成模块，该模块通常使用基于转换器的体系结构来构建。这种方法有助于减轻幻觉问题，并确保生成的文本更真实，在此期间，RAG模型已在多个领域得到应用，包括开放域问题回答（Karpukhin et al，2020）、会话代理（Liu et al，2021）和个性化推荐。



1.3在引入RAG之前，NLP模型主要依赖于检索或生成方法，每种方法都有其自身的优点和局限性。基于检索的系统，如传统的信息检索引擎（Salton等人，1975年），有效地提供了相关的文件或片段，以回应查询，但不能综合新的信息或提出的结果，在一个连贯的另一方面，随着Transformer架构的兴起而流行的纯生成模型（Vaswani et al 2017）提供了流畅性和创造性，但通常缺乏事实准确性。

随着研究人员认识到两种方法的互补优势，结合检索和生成的**混合系统**的发展开始获得动力。（Chen et al 2017），它采用检索技术来获取相关文档以进行问答任务。然而，此类系统中的生成组件很少，通常仅限于直接从检索到的文档中选择文本。同样，在信息检索（Dai et al 2019）等模型中，检索被视为不同的独立组件。

真实的创新来自于对检索和生成可以紧密集成的认识。

模型如REALM（Guu et al，2020）代表了一个关键的里程碑，因为他们联合训练了检索和生成组件，使检索的信息和生成的输出之间能够更好地对齐。RAG（刘易斯et al 2020）通过使用密集通道检索进一步扩展了这一范例（Karpukhin等人，2020年）获取相关文件和BART等变压器（刘易斯et al，2020）。这种架构提供了检索和生成的更无缝集成，允许模型以流利和事实为基础回答开放式问题。

1.4基于事实的语言生成的重要性

开发RAG的主要**动机**之一是对事实准确、上下文相关和最新生成内容的需求不断增长。在许多应用中，如客户服务、医疗诊断或法律的咨询系统，对可靠和有根据的反应的需求是至关重要的。产生幻觉或不准确信息的生成模型可能导致严重的后果，如传播错误信息或提供错误建议（Ji et al 2022）。RAG模型通过将其生成过程建立在外部最新的知识源中，直接解决了这些问题。这种基础提高了输出的事实准确性，并通过合并与查询直接相关的真实世界数据来增强响应的相关性。此外，RAG模型不太可能传播静态训练数据中存在的偏见，因为它们可以从外部来源检索更多样化和平衡的信息1.5 RAG模型的应用RAG模型已应用于事实准确性和上下文理解至关重要的广泛领域。最突出的应用之一是开放域问题回答，模型必须根据广泛的主题生成答案。RAG已被证明可以有效地通过检索相关信息并根据该数据生成响应来提高答案的准确性（Izacard等人，2021年）。密集通道检索（DPR）等模型（Karpukhin等人，2020）和解码器融合（Izacard等人，2021）在这种情况下已经取得了很大的效果，与传统的生成或仅检索模型相比有了显着的改进。

在对话式人工智能中，RAG模型通过确保响应一致并基于事实信息来增强对话系统的能力（Roller等人，2020）。例如，用于客户服务的聊天机器人可以受益于RAG从产品数据库或文档中检索特定细节的能力，从而为最终用户提供更准确和有用的响应。

其他应用包括医疗诊断系统，其中RAG可以检索和整合最新的研究结果或患者特定的数据，以生成准确的诊断建议，以及法律的咨询系统，其中该模型可以检索相关的判例法或法规，以提供法律上合理的建议。

此外，RAG在个性化推荐系统中也有应用，它可以检索用户偏好或过去的交互，并生成个性化建议

1.6 RAG的挑战和局限性

尽管RAG模型前景光明，但仍存在一些挑战需要注意。尽管检索机制功能强大，但在检索最相关的文档时仍然存在困难，特别是在处理模糊查询或利基知识领域时。对密集向量表示的依赖，例如DPR中使用的那些，有时会导致检索到不相关或偏离主题的文档。**努力改进检索技术**，包括引入更复杂的查询扩展和上下文消歧，在这些领域中，需要提高性能。检索和生成之间的集成虽然在理论上是无缝的，但在实践中有时会失败。例如，生成模块可能并不总是有效地将检索到的信息合并到其响应中，**研究更好的对齐机制**，如改进的注意力模型或层次融合技术，可能有助于缓解这些问题（Izacard et. al. 2021）。此外，RAG模型的计算开销是一个问题，因为它们需要每个查询的检索和生成步骤。这种双重过程可能是资源密集型的，特别是对于大规模应用（Borgeaud et al 2021）。

**模型修剪（Han et al 2015）或知识蒸馏**（Sanh et al，2019）等技术可以提供在不牺牲性能的情况下减少计算负担的方法。最后，RAG模型的部署存在**伦理问题**，特别是在偏见和透明度方面。

AI和LLM中的偏见一直是一个经过充分研究和不断发展的领域，研究人员发现了不同类型的偏见，不仅限于性别，社会经济阶层，甚至教育背景（Gupta等人，2024;虽然RAG有可能通过检索更平衡的信息来减少偏差，但仍然存在放大检索源中存在的偏差的风险（Binns，2018）.此外，确保检索结果如何选择和用于生成的**透明度**对于保持对这些系统的信任至关重要。

1.7调查范围

本文旨在提供对RAG模型的全面调查，涵盖其演变，关键架构组件，该领域的最新研究，当前RAG的挑战和局限性，以及未来的研究方向。

### 第二章：RAG系统的核心组件和架构概述

2.1 RAG模型概述检索增强生成（RAG）是一种高级混合模型架构，它通过外部检索机制增强自然语言生成（NLG），以增强模型的知识库。

传统的大型语言模型（LLM），如GPT-3和BERT，是在庞大的语料库上预先训练的，完全依赖于它们的内部知识表示，这使得它们容易受到幻觉等问题的影响-模型生成看似合理但不正确的信息。这些模型在没有重新训练的情况下无法有效地更新它们的知识库，这使得它们在动态，知识密集型任务，如开放域问题回答和事实验证（Brown，T.等人

为了克服这些局限性，论文（刘易斯等人，2020年）提出了RAG架构，该架构检索实时的相关外部文档，以事实信息为基础生成文本。

RAG模型包含两个关键组件：**1.检索器：**使用密集段落检索（DPR）（Karpukhin et. al. 2020）或传统BM 25算法等技术从语料库中检索最相关的文档。

**2.生成器**：它将检索到的文档合成为连贯的、与上下文相关的响应。

RAG的优势在于它能够动态地利用外部知识，使其能够超越GPT-3等生成模型和BERT等依赖静态数据集的知识基础系统。

在开放域问题回答中，RAG已被证明是非常有效的，一致地检索相关信息，并提高了生成的响应的事实准确性（Guu，K.，et al 2020）。除了知识检索，RAG模型擅长**更新知识库**。由于模型为每个查询获取外部文档，它不需要重新训练就可以包含最新的信息。这种灵活性使得RAG模型特别适合于信息不断发展的领域，例如医学研究、金融新闻和法律的诉讼。此外，研究表明，RAG模型在各种知识密集型任务中实现了上级结果，包括文档摘要和基于知识的对话2.2 RAG系统中的检索器机制RAG系统中的检索器对于从外部语料库中提取相关文档是必不可少的。

有效的检索确保模型的输出基于准确的信息。常用的**检索机制**有几种，从传统的方法如BM 25到更复杂的技术如密集通道检索（DPR）。

2.2.1 BM 25

**BM 25**是一种成熟的信息检索算法，它使用词频-逆文档频率（TF-IDF）根据相关性对文档进行排名。尽管是一种经典方法，BM 25仍然是许多现代检索系统（包括RAG模型中使用的系统）的强大基线。

BM 25基于查询词在文档中出现的频率来计算文档的相关性得分，同时调整文档的长度和词在语料库中的频率（Robertson等人，2009）。虽然BM 25对于**关键字匹配**是有效的，但是它在理解语义含义方面具有局限性。例如，BM 25无法捕捉单词之间的关系，并且在需要理解上下文的更复杂的自然语言查询上表现不佳。

尽管有这样的限制，BM 25仍然被广泛使用，因为它的简单性和效率。BM 25是有效的任务，涉及更简单的，基于关键字的查询，虽然更现代的检索模型，如DPR往往优于它在语义复杂的任务。

2.2.2密集通道检索（DPR）

**密集通道检索（DPR）**由Karpukhin等人（2020）提出，代表了一种更现代的信息检索方法。它使用密集向量空间，其中查询和文档都被编码为高维向量。DPR采用**双编码器**架构，其中查询和文档分别编码，允许有效的最近邻搜索（Xiong et. al. 2020）。与BM 25不同，DPR**擅长捕捉查询和文档之间的语义相似性**，使其对于开放域问题非常有效-回答任务。DPR的优势在于它能够**基于语义**而不是关键字匹配来检索相关信息。通过在大型问题语料库上训练检索器，答案对，DPR可以找到与查询上下文相关的文档，最近的研究通过将DPR与预先训练的语言模型集成来进一步改进DPR，一个例子是适用于密集RetrievAl方法的LLM（Li et. al. 2023）

2.2.3 **REALM（**

检索增强语言模型）RAG模型检索机制的另一个重要进步是REALM（Guu et al（2020））。

REALM将检索集成到语言模型的预训练过程中，确保检索器与生成器一起优化下游任务。REALM的关键创新在于，**它学习检索文档，从而提高模型在特定任务（例如问答或文档摘要）上的性能。**在训练期间，REALM更新检索器和生成器，确保检索过程针对生成任务进行了优化。REALM的检索器**经过训练**，能够识别不仅与查询相关，而且有助于生成准确和连贯响应的文档。因此，REALM显著提高了生成响应的质量，特别是在需要外部知识的任务中。最近的研究表明，REALM在某些知识密集型任务中的**表现优于BM 25和DPR，特别是当检索与生成紧密耦合时。**

RAG的核心在于**检索文章的质量**，但目前的许多方法都依赖于基于相似性的检索（Mallen等人，2022年）。（Asai等人，2023 b）和REPLUG（Shi等人，2023）通过**利用LLM来增强检索能力，实现更自适应的检索**而取得了进展。交叉编码器模型用于通过联合编码查询和每个检索到的文档来计算相关性分数，从而对检索到的结果进行重新排序。这些模型以更高的计算开销为代价来提供更多的上下文感知检索。逐点和成对排序，通常基于学习排序（LTR）算法，用于为检索到的文档分配相关性分数，独立地（逐点）或通过比较文档对RAG系统利用LLM内的自我注意力来管理输入和检索文本的不同部分的上下文和相关性。当将检索到的信息集成到生成模型中时，使用注意机制，确保在生成期间强调最相关的信息片段。

2.3 RAG系统中的**生成器机制**

在检索-增强生成（RAG）系统中，生成器机制在通过将检索到的信息与输入查询相结合来产生最终输出方面起着至关重要的作用。在检索组件从外部源提取相关知识之后，生成器将这些信息合成为连贯的、上下文适当的响应。大语言模型（LLM）充当生成器的主干，这确保所生成的文本流畅、准确并且与原始查询一致。

2.3.1 **T5（文本到文本转换器Transformer）**

**T5**（文本到文本转换Transformer）（Raffel et al 2020）是RAG系统中最常用的生成任务模型之一。T5的方法是通用的，将每个NLP任务框定为文本到文本任务。这种统一的框架允许T5针对广泛的任务进行微调，包括问答，摘要，通过将检索与生成相结合，基于T5的RAG模型已被证明在几个基准测试中**优于传统的生成模型**，如GPT-3和BART，包括Natural Questions数据集和TriviaQA数据集。

此外，T5处理复杂多任务学习的能力使其成为需要处理各种知识密集型任务的RAG系统的热门选择。

2.3.2**BART**

BART由刘易斯等人（2020）引入的双向和自回归Transformer（BART）是RAG系统中使用的另一种**突出的**生成模型。BART**特别适用于涉及从噪声输入生成文本的任务，例如摘要和开放域问题回答**。作为去噪自动编码器，BART可以重建损坏的文本序列，使其对于需要**从不完整或噪声数据**中生成连贯的、真实的输出的任务是鲁棒的。当与RAG系统中的检索器配对时，BART已经被证明可以通过将其建立在外部知识的基础上来提高生成文本的事实准确性。研究表明，基于BART的RAG模型可以实现艺术产生了各种知识密集型任务，包括对话生成和新闻摘要

3. 跨不同模态的检索增广生成模型

3.1**基于文本的RAG模型**：

基于文本的RAG模型代表了最成熟和研究最广泛的类别。这些模型利用文本数据进行检索和生成任务，支持问答，摘要和会话代理等应用。

**Transformer架构**，如BERT（Devlin et al，2019）和T5（Raffel et al，2020），是基于文本的RAG模型的基础。这些模型利用自注意机制来捕获文本中的上下文关系，从而提高检索准确性和生成流畅性。

密集检索模型，例如使用BERT密集嵌入的模型，与TF-IDF等传统稀疏方法相比，具有上级性能。密集检索器（Karpukhin et al 2020），利用密集表示来**更有效地检索**相关文档。最近的进展集中在将检索和生成集成到单个训练管道中。REALM（Guu等人，2020）是这样一个端到端模型的一个例子，它联合优化了检索和生成过程，提高了整体任务性能。

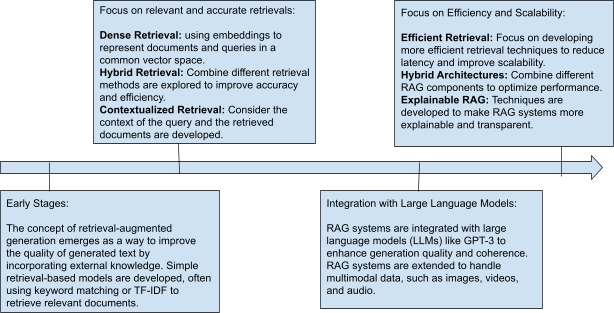
3.2**基于音频的RAG型号**：基于音频的RAG模型将检索增强生成的原理扩展到音频模态，支持语音识别、音频摘要和语音接口中的会话代理等应用。音频数据通常使用从预训练模型（如Wav 2 Vec 2.0）中派生的嵌入来表示（Baevski et al，2020）。这些嵌入作为检索和生成组件的输入，使模型能够有效地处理音频数据。

3.3**基于视频的RAG模型**：基于视频的RAG模型结合了视觉和文本信息，以增强视频理解、**字幕**和检索等任务的性能。

视频数据使用来自I3 D（Xie et. al. 2017）或TimeTransformer（Bertasius et al 2021）等模型的嵌入来表示。这些嵌入捕获了有效检索和生成所必需的时间和空间特征。

3.4**多模式RAG型号**：多模态RAG模型集成了来自多种模态（文本、音频、视频和图像）的数据，为检索和生成任务提供了更全面的方法。（Alayrac等人，2022年）将多种模态集成到一个统一的框架中，使文本，图像，跨模态检索的技术涉及跨不同模态检索相关信息（Li. et. al. 2023）。

多模式功能增强了RAG在各种应用中的**通用性和效率**。”检索即生成”（Wang et. al.2024）通过结合文本到图像和图像到文本检索将检索增强生成（RAG）框架扩展到多模态应用。当用户查询与存储的文本描述一致时，系统加速图像生成（“检索即生成”）。图像到文本功能允许用户参与基于输入图像的讨论。



4.该领域的**最新进展**：

在这一领域已经取得了重大进展，本节旨在捕捉一些重要的最近论文的关键发现。**一个新的agentic代理检索增强生成（RAG）框架**（Ravuru等人，2024）采用了一种**分层的多代理架构**，其中专门的子代理，使用**较小的**预训练语言模型（SLM），主代理将任务委托给这些子代理，这些子代理从共享知识库中检索相关提示。

在这种模块化的多代理方法中，作者实现了最先进的性能，证明了在时间序列分析中比特定任务方法更好的灵活性和有效性。

1. ，多模态检索增强生成（RAG）框架，旨在提高医学视觉语言模型的真实性（Med-LVLM），通过引入校准的选择策略来控制真实性风险，并且通过开发偏好优化策略来平衡模型的内在知识与检索的上下文，证明其在Med-LVLM系统中提高事实准确性的有效性。METRAG（Gan等人，2024），一个多层次的，思想增强的检索增强生成框架，整合了LLM监督，以产生实用导向的思想，并将文档相似性与实用性相结合，以提高性能。它还包括一个任务-自适应摘要器产生紧凑的想法。使用这些阶段的多层思想，LLM生成知识增强的内容，与传统方法相比，在知识密集型任务上表现出上级性能。

Distractor document is**检索增强微调（RAFT）**（Zhang et. al. 2024）的关键特征之一，其中模型被训练为**忽略不相关的干扰文档**，而是**直接引用相关来源**。这个过程与思维链推理风格相结合，增强了模型的推理能力。

RAFT在特定领域的RAG任务中表现出一致的性能改进，包括PubMed，HotpotQA和Gorilla数据集，作为LLM的训练后增强。**FILCO**（Wang et. al. 2023），一种旨在提高在开放域问题回答和事实验证等任务中提供给生成模型的上下文质量的方法，解决了对检索到的段落过度依赖或依赖不足的问题，该方法通过**词汇和信息论方法**识别有用的上下文，并在测试时训练上下文**过滤模型**来优化检索到的上下文，从而提高上下文质量

（**self——RAG**）（Asai et. al.2023），一种新颖的框架，旨在通过将**检索与自我反思相结合**来提高大型语言模型（LLM）的事实准确性。与检索和合并固定数量段落的传统方法不同，Self-RAG自适应地检索相关段落，并使用**反射令牌**来评估和改进其响应，允许模型根据特定任务的需求调整其行为，并在开放域问答，推理，事实验证，RAG的智能性和有效性在很大程度上取决于检索的质量，对存储库元数据的理解将提高RAG系统的有效性。

一种新的以数据为中心的检索-增强生成（RAG）工作流超越了传统的检索-读取模式，采用了一种**先读取-再重写-再检索-再读取**的框架，通过整合上下文相关，时间关键或特定领域的信息来增强LLM。关键创新包括生成元数据，合成问题和答案（QA），并介绍了Meta Knowledge Summary（MK摘要），用于文件组（Mombaidu等人，2024）。

最近的一篇论文介绍了**CommunityKG-RAG**（Chang等人，2024），一种**零触发框架**，将知识图（KGs）中的社区结构集成到检索增强生成（RAG）系统中。这种方法通过利用多个KGs内的跳连接，性能优于传统方法，而无需额外的特定领域的培训。

（Sarthi et. al.2024）介绍了一种用于检索增强语言模型的**分层方法**，解决了传统方法中仅检索短的连续文本块的局限性。**RAPTOR**通过递归嵌入，聚类，和摘要文本。实验证明了RAPTOR的上级性能，特别是在需要复杂推理的问答任务中。当与GPT-4配对时，RAPTOR将QuALITY基准的准确率提高了20%。

RAG的这一进步进一步证明了RAG系统的实用性，但最近推出的支持长期环境的LLM系统显着提高了性能。（Li等人，2024）**比较了检索增强生成（RAG）和长上下文（LC）大型语言模型（LLM）（如Gemini-1.5和GPT-4）的效率**。虽然LC模型在资源充足时优于RAG，RAG的成本效率仍然具有优势。为了平衡性能和成本，本文引入了Self-Route。该方法基于模型自反射动态地将查询定向到RAG或LC，优化计算成本和性能。这项研究为RAG和LC在处理长上下文任务中的最佳应用提供了有价值的见解。

Nguyen et. al.，2024引入**SFR-RAG**，一个小型但高效的检索增强生成（RAG）模型，旨在增强外部上下文信息到大型语言模型（LLM）的集成，同时最大限度地减少幻觉。LA-RAG（Li等人，2024），一种新的检索增强生成（RAG）范式，旨在增强大型语言模型（LLM）中的自动语音识别（**ASR）LA-RAG**的主要优点之一是它能够利用细粒度的令牌级语音数据存储以及语音到语音检索机制，通过结合LLM上下文学习（ICL）来提高ASR准确性。该研究侧重于**普通话和各种中国方言的数据集**，显示出显着的准确性提高，特别是在管理口音变化方面，这在历史上一直是现有语音编码器的挑战。研究结果突出了LA-RAG推进ASR技术的潜力，为各种声学条件提供了更强大的解决方案。由于过时的知识和幻觉，大型语言模型（LLM）在AI法律的和政策背景下面临挑战。**HyPA-RAG**（Kalra et. al.，2024），一个混合参数-自适应检索-增强生成系统，通过使用自适应参数调整和混合检索策略来提高准确性。在NYC Local Law 144（LL 144）上进行测试，HyPA-RAG**证明了增强的正确性和上下文精确性**，解决了法律的文本的复杂性。MemoRAG（Qian et. al.，2024）介绍了一种新颖的检索增强生成（Retrieval-Augmented Generation，RAG）范式，旨在克服传统RAG系统在处理模糊或非结构化知识方面的局限性。**MemoRAG**的双系统架构**利用轻量级远程LLM生成草稿答案和指导检索工具**，而更强大的LLM细化最终输出。该框架**，NLLB-E5**针对**更好的提示和记忆容量**进行了优化，在复杂和简单的任务中都显着优于传统的RAG模型。（Acharya等人，2024）介绍了一种可扩展的多语言检索模型，旨在解决支持多种语言所面临的挑战，特别是像印度语这样的低资源语言。通过利用NLLB编码器和E5多语言检索器的蒸馏方法，NLLB-E5实现了跨语言的零触发检索，无需多语言训练数据。对Hindi-BEIR等基准测试的评估显示了其强大的性能，强调特定任务的挑战，并推进多语言信息获取，以实现全球包容性。

1. **检索增强生成（RAG）当前的挑战和局限性：**

本节旨在强调RAG当前的挑战和局限性，考虑到系统的当前景观，这将塑造该领域未来的研究方向。

**可扩展性和效率**：RAG模型面临的主要挑战之一是可扩展性。由于检索组件依赖于外部数据库，因此处理庞大且动态增长的数据集需要高效的检索算法。高计算成本和内存要求也使RAG模型难以在实时或资源受限的环境中部署（Shi et al 2023），（Asai et al 2023 b）。

**检索质量和相关性**：确保检索文档的质量和相关性仍然是一个重要的问题。检索模型有时会返回不相关或过时的信息，这会对生成的输出的准确性产生负面影响。提高检索精度，特别是对于长格式内容生成，仍然是一个活跃的研究领域（Mallen et al 2022），（Shi et al 2023）。

**偏见与公平：**与其他机器学习模型类似，RAG系统可能会因检索数据集中存在的偏差而表现出偏差。基于检索的模型可能会放大检索知识中的有害偏差，导致一代人的偏差输出。开发用于检索和生成的偏差缓解技术是一个持续的挑战。

**连贯性**：RAG模型通常难以将检索到的知识整合到连贯的、上下文相关的文本中。检索到的段落与生成模型的输出之间的对齐并不总是无缝的，导致最终响应中的不一致或事实幻觉（Ji等人

**可解释性和透明度**：与许多人工智能系统一样，RAG模型通常被视为黑箱，在检索如何影响生成方面透明度有限。提高这些模型的可解释性对于培养信任至关重要，特别是在关键应用程序中（Roller et al 2020）

6.检索增强生成（Retrieval-Augmented Generation，RAG）是自然语言处理及相关领域的一个重要研究方向，它将检索机制和生成机制结合在一起。本节探讨了未来研究的关键领域，强调了在RAG系统中创新和改进的潜力。

6.1**增强多模式集成**：在RAG模型中集成文本、图像、音频和视频数据仍然是一个不断发展的挑战。未来的研究应该集中在改进多模态融合技术，以实现**不同数据类型之间的无缝交互**。这包括开发跨模态信息对齐和合成的高级方法。

等人，2022），（Yasunaga等人，2022年），（Zhu et. al. 2024）已经探索了多模态学习，但需要进一步的创新来增强多模态输出的连贯性和情境性。跨模态检索研究旨在提高RAG系统跨不同模态检索相关信息的能力。例如，将基于文本的查询与图像或视频内容检索相结合，可以增强视觉问答和多媒体搜索等应用，这是RAG研究的又一个方向。

6.2**扩展和效率**：随着RAG模型在越来越大规模的应用程序中的部署，可扩展性成为一个关键的问题。研究应该集中在开发方法，以有效地扩展检索和生成过程，而不损害性能。技术，如分布式计算和有效的索引方法是必不可少的处理大型数据集。提高效率的RAG模型包括优化检索和生成组件，以减少计算资源和延迟。

6.3**个性化和适应性**：未来的RAG模型应侧重于个性化检索过程，以迎合**个人用户的偏好**和上下文。这涉及到开发技术，以适应基于用户历史、行为和偏好的检索策略。通过更深入地理解查询的上下文和情感，增强RAG模型的上下文适应性（Gupta et. al.

2024）和文档的存储库对于提高所生成的响应的相关性至关重要。

研究应探索基于互动情境的动态调整检索和生成过程的方法，包括将用户反馈和情境线索纳入RAG管道。

6.4**道德和隐私考虑：**消除偏见（Shrestha等人，2024年），（Gupta等人，2024），是未来研究的关键领域。随着RAG系统在不同应用中的部署，**确保检索和生成内容的公平性和减少偏见是至关重要的**。未来的RAG研究应该关注隐私-在检索和生成过程中保护敏感信息的保存技术。这包括开发安全数据处理方法和隐私感知检索策略。模型的可解释性也是一个关键领域，作为改进RAG的正在进行的研究的一部分。

6.5**跨语言和低资源语言**：扩展RAG技术以支持多种语言（Chirkova等人，2024），尤其是**低资源语言**，是一个很有前途的研究方向。未来的研究应致力于提高跨语言检索和生成能力，以提供跨语言的准确和相关的结果。增强RAG模型，以有效地支持低-资源语言涉及开发检索和生成具有有限训练数据的内容的方法。

研究应侧重于迁移学习和数据扩充技术，以提高在代表性不足的语言中的表现。

6.6**高级检索机制**：未来的RAG研究应探索适应不断变化的查询模式和内容要求的动态检索机制，包括开发能够根据新信息和不断变化的用户需求动态更新检索策略的模型，研究将各种检索策略（如密集检索和稀疏检索）结合起来的混合检索方法，可以提高RAG系统的有效性。研究应该探索如何整合不同的检索方法，以实现不同任务的最佳性能。

6.7**与新兴技术的集成**：将RAG模型与**脑-机接口（BCI）集成**可能会在人机交互和辅助技术方面带来新的应用。

研究应探索RAG系统如何利用BCI数据来增强用户体验并生成情境感知响应。RAG与AR和VR技术的集成为创造沉浸式和交互式体验提供了机会。未来的研究应探索RAG模型如何通过提供情境相关信息和交互来增强AR和VR应用。

7.结论

检索增强生成（RAG）经历了重大的发展，广泛的研究致力于提高检索效率和增强连贯生成以最大限度地减少幻觉。从早期的迭代到最近的进步，RAG一直有助于将外部知识整合到大型语言模型（LLM）中，从而提高准确性和可靠性。特别是，最近的领域特定工作已经展示了RAG在诸如法律的、医学和低资源语言应用等专业领域的潜力，突出了其适应性和范围。然而，尽管取得了这些进展，本文还是确定了仍然没有解决的明显差距。挑战诸如模糊或非结构化信息的集成，领域特定上下文的有效处理，以及复杂检索任务的高计算开销仍然存在。这些限制限制了RAG系统的更广泛的适用性，特别是在多样化和动态的现实世界环境中。本文概述了未来的研究方向-从改进检索机制到增强上下文管理和确保可扩展性-通过解决这些差距，下一代RAG模型有可能驱动更可靠，高效和领域适应性更强的LLM系统，进一步推动检索增强AI应用的可能性。