Meta-RAG:基于元数据驱动的电力领域检索增强生成框架

王合庆,魏杰,景红雨,宋晖,徐波
(东华大学计算机科学与技术学院,上海 201600)

摘 要: 大语言模型(LLM)在对话、推理和知识保留能力方面展现了显著优势,但在处理电力领域知识密集型任务时仍面临事实准确性不足、知识更新难以及高质量领域数据集匮乏的问题。针对这些挑战,引入一种改进的检索增强生成(RAG)策略,该策略融合了混合检索策略和经过微调的生成模型,提供了更高效的知识捕获和更新能力。基于对现有方法的深入分析,针对电力领域的知识问答(QA)任务,提出了元数据驱动的 RAG 框架 Meta-RAG,该框架包含数据准备、模型微调和检索推理 3 个阶段。数据准备阶段包括文档转换、元信息抽取与增强及文档解析模块,在此阶段,借助元信息的提取与增强确保了电力规范文档的高效索引和结构化处理,并且构建了电力领域的 EleQA(Electricity Question Answering)数据集,这是一个包含 19 560 个问答对的电力规范问答数据集。在模型微调阶段,通过多问题生成、思维链提示生成和监督指令微调数据集构建模块,优化了模型在特定电力问答任务上的推理能力。在检索推理阶段则采用混合编码和重排序策略,结合检索和生成模块,进一步提高了答案的准确性和合理性。通过一系列实验,Meta-RAG 的有效性得到验证。与 Self-RAG、Corrective-RAG、Adaptive-RAG、RA-ISF 等基线模型相比,Meta-RAG 具有更高的回答准确率和检索命中率,其中,基于 Qwen1.5-14B-Chat 模型的 Meta-RAG 达到了整体准确率 0.804 3,高于其他方法。消融实验和文档召回实验结果表明文档检索对框架性能影响最大,失去检索能力整体准确率下降了 0.292 8。

关键词: EleQA 数据集;元信息抽取;知识问答;电力领域;检索增强生成;模型微调;文档转换
源代码链接: https://gitee.com/coldz/EleQA
中图分类号: TP18 **文献标志码:** A **DOI:** 10.19678/j.issn.1000-3428.0070415

Meta-RAG: A Metadata-Driven Retrieval-Augmented Generation Framework for the Power Industry

WANG Heqing, WEI Jie, JING Hongyu, SONG Hui, XU Bo
(School of Computer Science and Technology, Donghua University, Shanghai 201600, China)

【Abstract】 Large Language Models (LLMs) have made significant progress in dialogue, reasoning, and knowledge retention. However, they still face challenges in terms of factual accuracy, knowledge updates, and a lack of high-quality domain datasets for handling knowledge-intensive tasks in the electricity sector. This study aims to address these challenges by introducing an improved Retrieval-Augmented Generation (RAG) strategy. This strategy combines hybrid retrieval with a fine-tuned generative model for efficient knowledge capturing and updating. The Metadata-driven RAG framework (Meta-RAG) is proposed for knowledge Question Answering (QA) tasks in the electricity domain. This includes data preparation, model fine-tuning, and reasoning retrieval stages. The data-preparation stage involves document conversion, metadata extraction and enhancement, and document parsing. These processes ensure efficient indexing and structured processing of power regulation documents. The Electricity Question Answering (EleQA) dataset, consisting of 19 560 QA pairs, is constructed specifically for this sector. The model fine-tuning stage uses multi-question generation, chain-of-thought prompting, and supervised instruction fine-tuning to optimize the reasoning abilities in specific tasks. The retrieval reasoning stage employs mixed encoding and re-ranking strategies, combining retrieval and generation modules to improve answer accuracy and relevance. Experiments validate the effectiveness of Meta-RAG. Compared to baseline models such as Self-RAG, Corrective-RAG, Adaptive-RAG, and RA-ISF, Meta-RAG shows higher answer accuracy and retrieval hit rates. Meta-RAG with the Qwen1.5-14B-Chat model achieves an overall accuracy of 0.804 3, surpassing the other methods. Ablation and document recall experiments indicate that document retrieval significantly impacts the framework performance, with a 0.292 8 drop in accuracy when the retrieval capability is lost.

【Key words】 EleQA dataset; meta-information extraction; knowledge Question Answering (QA); power industry; Retrieval-Augmented Generation (RAG); model fine-tuning; document conversion

作者简介: 王合庆,男,硕士研究生,主研方向为自然语言处理;魏杰、景红雨,硕士研究生;宋晖,教授;徐波(通信作者),副教授。
收稿日期: 2024-09-27 **修回日期:** 2024-10-23 **E-mail:** xubo@dhu.edu.cn

0 引言

电力规范问答(QA)系统涵盖电力安全操作规程、设备维护标准和应急处理指南等重要法规,电力从业人员须遵循这些法规以确保安全和合规。然而,这些法规文件的长度和复杂性使得信息提取变得困难,耗时的搜索过程给从业者和培训人员带来挑战,并可能导致误解和操作失误。问答系统旨在解决此问题。

大语言模型(LLM)在对话、推理和记忆方面表现出色,但在知识密集型任务中面临准确性和更新的问题。为提升大语言模型的能力,检索增强生成(RAG)方法由 LEWIS 等[1]于 2020 年提出,其通过结合预训练的检索器和生成器,以更模块化的方式捕获知识。传统 RAG 方法主要利用稠密段落检索(DPR)[2]等方法,其与生成模型结合,在端到端优化上取得进展[3]。然而,大语言模型仍面临幻觉和知识更新等挑战[4-5]。研究表明,将 RAG 用于大模型的上下文学习,可以有效缓解这些问题。随着改进,近年来出现了许多新的 RAG 策略。Self-RAG[6]是一种先进的 RAG 框架,其通过自我反思机制提升语言模型在文本生成任务中的表现,能判断是否需要进行文档检索,并对生成的文本段进行质量评估。在文档相关性上,Corrective-RAG[7]引入检索评估和知识精炼算法,通过轻量级评估机制衡量检索文档的相关性,并据此触发相应的知识检索策略。Adaptive-RAG[8]是一种自适应的 RAG 框架,其能根据问题的复杂性选择最合适的策略。RA-ISF[9]通过迭代自反馈方法,显著提升了 RAG 模型在开放域问答任务中的性能。

近年来,伴随多种新技术的探索,电力智能问答系统已有一定的技术进步。张峰等[10]设计了结合知识库与自然语言处理的新型平台,减少了客服资源消耗。周帆等[11]提出了基于知识图谱的电网模型问答系统来实现问答。覃祥坤[12]通过知识图谱和深度学习解决电力企业的检索与问答问题。ZHANG 等[13]采用“比较-聚合”框架提升故障诊断准确性。MENG 等[14]结合 TF-IDF(Term Frequency-Inverse Document Frequency)与余弦相似度提高了系统准确性。研究表明,通过余弦相似度与皮尔逊相关系数能有效评估大语言模型的答案准确性[15],基于链路推理的方法可提高设备知识问答效率[16]。

在电力领域,利用大语言模型开展知识问答研究已有一些探索性尝试。SCER(Self-consistency, Extract and Rectify)框架[17]将自洽性方法引入电力领域,提升了大语言模型在问答任务中的表现。HUANG 等[18]验证了大语言模型在电力系统中的潜力。CHENG 等[19]介绍了适用于电力调度的大语言模型 GAIA。MAJUMDER 等[20]探讨了大语言模型在电力行业的应用与限制,强调优化数据收集和 RAG 以提升安全场景的响应质量。在使用大语言模型辅助电力相关工作时,人们对大语言模型自身和电力数据的安全性有了更多的考虑。RUAN 等[21]分析了将大语言模型应用于现代电力系统的潜在威胁,并强调需要研究和开发应对措施。电力领域需要权衡安全性和成本,在选择底座模型时,尽管 GPT-4 等商业大模型在自然语言处理方面取得了显著进展,但高昂的训练费用限制了其在推理阶段的经济性,同时也增加了安全审查的复杂性。相较之下,开源大模型通过集成 RAG 模块,不仅能够高效整合外部知识,降低训练成本,还缓解了由于知识源有限而产生的幻觉问题。同时,开源大模型提供了更大的可定制性,可以进行全面的安全审查,因此使用开源大模型进行知识问答比使用闭源商业大模型更加安全。

尽管 RAG 方法已在部分领域取得了显著成效,但在电力领域,仍然面临多个重要挑战。首先,该领域缺乏专用的高质量数据集,难以支撑模型有效地训练和推理。其次,目前没有针对电力领域需求优化的 RAG 框架,从而限制了模型的实际应用效果。最后,对于如何更高效地整合文档信息以提升模型回答准确性,也尚未有成熟的解决方案。

在改进的 RAG 方法中,元信息的有效性至关重要。元信息提供了结构化上下文帮助模型更准确地理解和索引文档,这对于长而复杂的电力规范文档尤为重要。现有的方法虽然各有优势,但仍存在局限,如 Self-RAG[6]由于自我反思机制增加了计算复杂度,Corrective-RAG[7]在复杂查询时可能引入不相关信息,Adaptive-RAG[8]的动态策略选择算法复杂性较高,RA-ISF[9]增加了模型的计算开销。

针对上述挑战,本文提出了一个新颖的元数据驱动的 RAG 框架 Meta-RAG。这个框架不仅能够评估和思考文档的有效性,还涵盖了从问题集和文档集的构建,到大模型微调的完整流程。通过提取文档的元信息,Meta-RAG 能够更高效地整合和评估文档内容,提高检索的精准度。通过设

计对比实验，包括与 Self-RAG[6]、Corrective-RAG[7]、Adaptive-RAG[8] 以及 RA-ISF[9] 的对比，证明了 Meta-RAG 在提升准确性和适应电力领域特殊需求方面的优势。

本文的贡献主要体现在以下几个方面：

1）提出了一个适用于电力领域的元数据驱动的 RAG 框架 Meta-RAG。该框架不仅解决了大语言模型在处理知识密集型任务过程中的事实准确性和知识更新问题，还具备易于集成和高效运行的特点，显著提高了系统的实用性。

2）构建了 EleQA(Electricity Question Answering) 数据集。这是一个专为电力领域知识问答任务设计的高质量、多样化数据集，包括 32 610 条电力规范条例和 19 560 个问答对，其中，电力规范条例覆盖火电技术、水电站设备检修管理、电力电容电感测试和高空救援等多个应用场景，在题型上涵盖单选、填空和判断 3 种题型，数量分别为 6 150、6 590 和 6 820 道，单选题的正确选项出现在 A、B、C、D 4 个位置的比例相当，判断题的答案为“正确”“错误”的比例相当，确保了题型的均衡性和合理性。

3）提出了一套有效的推理结果评估策略。对单选题和判断题使用同一种评估策略，并且单独为填空题设计一种评估策略，实现对答案与推理结果的比对达到关键词级别，进一步提高评判的准确率。

结合 Meta-RAG 框架以及 EleQA 数据集，形成了一个完整的 benchmark，为后续研究提供了坚实的基础和标准化的评估平台。

1 基于元数据驱动的 RAG 框架

针对电力领域中的知识问答任务，本文设计了一个创新性框架 Meta-RAG。这个框架结合了 RAG 策略和大语言模型，显著提升了对电力相关查询的回答质量和准确性，为处理电力领域的知识密集型任务提供了有效解决方案。具体来说，当用户提出一个与电力领域相关的查询时，系统会依据文档库中的电力规范文档片段，结合大语言模型生成准确且有依据的答案，并给出合理性解释。

Meta-RAG 框架如图 1 所示（彩色效果见《计算机工程》官网 HTML 版，下同）。该框架由 3 个阶段组成：数据准备阶段，模型微调阶段，检索推理阶段。

1）数据准备阶段：通过文档转换模块，将电力规范文档自动化处理为仅含主体内容的 Markdown 格式，元信息抽取和增强模块相继提取并丰富章节间的层次信息，随后文档解析模块将其细分为独立且语义完整的文本块，确保文档的结构化和可读性，为后续应用奠定基础。

2）模型微调阶段：利用 LoRA（Low-Rank Adaptation）方法[22]引入低秩矩阵进行参数优化，提高微调效率并减少参数量，采用了受 RAFT（Retrieval-Augmented Fine-Tuning）[23]启发的多阶段干扰文档引入策略，逐步增强模型对相关信息的筛选和推理能力，适应特定任务需要。

3）检索推理阶段：分为离线和在线部分，离线阶段通过稀疏和稠密编码器将文档转化为向量存储，在线阶段结合用户输入实时检索，并通过结构化推理输出最终答案，提升检索和推理效率。

在数据准备阶段，首先将原始的电力规范文档输入文档转换模块，转化为 Markdown 文档。这些 Markdown 文档随后进入元信息抽取模块，生成相应的元信息。接着，元信息与Markdown文档一起

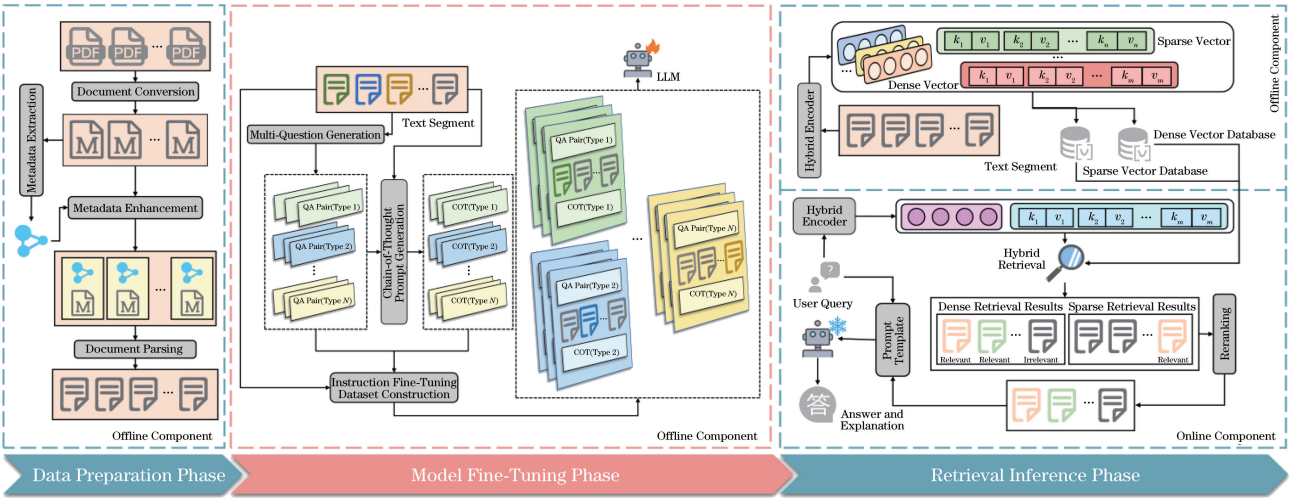

图 1 Meta-RAG 框架

Fig. 1 Meta-RAG framework

输入元信息增强模块以获得增强后的 Markdown 文档。这些增强后的文档再经过文档解析模块处理,转化为便于程序读取和处理的文本段。

在模型微调阶段,文本段被输入到多问题生成模块,生成大量不同类型的问答对。这些问答对连同文本段一起进入思维链提示生成模块,生成从问题到答案的合理性解释。最后,这些解释、问答对及相关和无关文档被输入到指令微调数据集构建模块,形成用于大语言模型训练和评估的数据集。

在检索推理阶段,文本段首先输入混合编码模块,生成稠密向量和稀疏向量,并分别存入相应的向量库中。当用户查询时,查询问题被输入到混合编码器,生成查询稠密向量和查询稀疏向量。将这些查询向量与向量库中的文档向量进行相似度匹配,获取相关文档集。这些相关文档接下来进入重排序模块,生成递减相关性的文档列表。最后,相关文档集与查询问题一起写入提示模板,作为大语言模型的输入,生成含详细解释的最终回答。

2 方法

2.1 数据准备阶段

在数据准备阶段,通过一系列模块的部署,整体提升了电力规范文档的实用性和信息质量。首先,文档转换模块通过自动化的预处理和格式转换,仅保留正文内容,并将其转化为 Markdown 格式。接着,元信息抽取模块高效提取章节标题等关键信息,增强了文本段内容的理解。然后,元信息增强模块将文本段与父级信息结合,丰富了层次化理解。最后,文档解析模块通过分解和提取文档要素,使每个文本段成为独立且语义完整的文本块。这些模块的共同作用确保了文档的结构化和可读性,为后续的处理和应用奠定了基础。

2.1.1 文档转换模块

为了有效提升电力规范文档的信息质量,文档转换模块被设计用于实现自动化预处理和格式转换过程。总体来说,该模块通过识别和剔除无关内容,仅保留正文信息,从而提高文档的实用性和清晰度。具体来说,现有 PDF 工具在分辨正文和非正文内容方面存在不足,该模块引入一种文档正文识别算法,有效过滤题目、引言、目录、附录、参考文献等非正文部分;同时建立文档元数据体系,实现全流程批量自动化,避免了人工干预;完成内容处理后,将正文内容转化为 Markdown 格式,利用 PDF 到 Markdown 的转换算法,确保标题等级与原文一致。在转换过程中,算法考虑了页码和表格数据对标题识别的干扰,通过检查标题的合法性,保证同级和层级标题的有序性。这一模块有效提高了文档处理的精确度和效率,为后续的文档利用奠定了基础。本文中,文档转换模块所处理的 PDF 文件,即电力规范文件和电力故障文件,全部为来源于国家能源局或全国标准信息公共服务平台的公开试行文件。这些文件都是可搜索的 PDF,可以直接进行文本搜索和提取。

2.1.2 元信息抽取模块

元信息抽取模块在文本处理中扮演着关键角色,旨在通过提取章节标题,提升对文本段内容的理解。元信息提供了对文档内容的准确性和简洁性的支持,特别是在问答过程中,其辅助作用尤为明显。具体来说,得益于之前的文档格式转换,将内容转为 Markdown 格式后,该模块能够高效地从中提取元信息。通过这一过程,能够确保文档在后续使用中更加易于理解和访问。

2.1.3 元信息增强模块

元信息增强模块为文档的层次化理解提供了重要支持,其主要目的是提高元信息的丰富性和关联性。总体而言,通过结合各层级的文本段信息,该模块可以生成更为详尽的元信息描述。具体操作中,每个文本段都有相应的元信息,小层级内容通常包含其父级的更多信息,该模块通过将深层级文本段的元信息与相关父级元信息拼接,生成更丰富的描述。这种方式确保了每个文本段获得增强的上下文关联,使文档的整体理解更加准确和全面。

2.1.4 文档解析模块

文档解析模块旨在系统化地分解和提取文档信息,以提升文本的可读性和实用性。该模块通过解析带有增强元信息的文本段,提取出元信息、文本内容、标题、标题层级及相关文档索引等关键要素。具体来说,这种拆分方法使每个文本段成为独立的文本块,与传统的 LangChain 拆分相比,保留了更完整的语义。此方法不仅增强了文档的结构化程度和层次逻辑,也提高了信息的可读性,使其在实际应用中更加有价值。

2.2 模型微调阶段

模型微调阶段通过优化技术和数据策略提升模型在特定任务上的表现。使用 LoRA[22] 方法进行监督指令微调,通过引入低秩矩阵优化预训练模型的参数,提高效率并减少了参数量;此外,从 RAFT[23] 微调方法中获得启发,采用多阶段训练策略逐步增加干扰文档的复杂程度和数量,以增强模型的筛选和推理能力。

具体而言，LoRA 通过引入低秩矩阵的方式对预训练模型进行参数调整，从而提高效率并减少参数量。设预训练模型的权重矩阵为 $\boldsymbol{W}$，LoRA 微调通过引入两个低秩矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 进行权重更新，如式(1)所示：

$$\boldsymbol{W}_{\text{new}}=\boldsymbol{W}+\boldsymbol{AB} \tag{1}$$

式中：$\boldsymbol{A}\in\mathbb{R}^{d\times r}$ 和 $\boldsymbol{B}\in\mathbb{R}^{r\times k}$ 为低秩矩阵，r 为低秩维度，通常满足 $r\ll d$ 和 $r\ll k$。这种方法得到新的权重 $\boldsymbol{W}_{\text{new}}$，通过仅适应 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 参数，而保持原有的预训练权重 $\boldsymbol{W}$ 不变，使参数数量大幅减少。通过这种低秩表示，LoRA 能够有效捕获模型微调所需要的关键信息，从而增强模型在特定任务上的表现。

从 RAFT[23] 微调方法中获得启发，本文在指令微调的数据集中引入了干扰文档，并随机放置相关文档的位置。不同于 RAFT，本文通过多阶段训练流程，逐步增加干扰文档的复杂程度和数量，从而增强模型的筛选和推理能力，让大语言模型意识到在回答用户问题时需要首先识别和利用有用文档，再结合文档逐步推理。

2.2.1　多问题生成模块

多问题生成模块是构建电力知识问答数据集的重要环节，该模块通过生成多样化的问题集合，为大语言模型的微调数据集奠定基础，生成的问题按比例分配至微调数据集。具体流程包括以下几个步骤：

1)提示编写：不同题型的提示内容有所不同，每种题型都有其特定的出题要求，并提供少量具体例子，以帮助大语言模型更好地理解这些要求。

2)问题生成：在提示中写入文档的文本部分，利用大语言模型生成简洁明确的答案。

3)问答对质量过滤：初步生成的问答对需要经过质量过滤。为确保问答对的格式和内容质量，需要进行人工评估，具体流程为：安排 3 名志愿者根据准确性、相关性、完整性、流畅度对所有批次的候选问答对进行评分，每项分值 25 分，总分为 100 分；志愿者对其负责的问答进行逐项评分，并记录结果；将各志愿者的评分汇总后，计算出每个问答对的平均得分，选择每段文本中平均得分最高的问答；如果出现并列得分，志愿者对这些问答进行投票，选出得票最高者。整个过程耗时约两周。此外，在填空题生成提示中，特别强调了题目应具备“创造性”而非简单使用原文随机缺失关键词留下空缺的“挖空”，以确保具有更高的回答难度。

$$\max_{\theta}E_{(q,\text{cot},P_{\text{gold}},P_{\text{interf}})\sim D_{\text{train}}}\left[\ln(p_\theta(y\mid q,\text{cot},P_{\text{gold}}))+\ln\left(\frac{p_\theta(P_{\text{gold}}\mid q)}{p_\theta(P_{\text{gold}}\mid q)+p_\theta(P_{\text{interf}}\mid q)}\right)\right] \tag{2}$$

式中：θ 表示模型的参数集合；E 是期望值操作符，应用于训练数据集 D_{train} 中的样本；q 代表问题；cot 是附加提示或逻辑推理过程；P_{gold} 为包含正确答案的黄金文本段落；P_{interf} 是干扰文本段落；y 是正确答案；$p_\theta(y\mid q,\text{cot},P_{\text{gold}})$ 表示在给定问题、提示和黄金段落的情况下生成正确答案的概率；$p_\theta(P_{\text{gold}}\mid q)$ 和 $p_\theta(P_{\text{interf}}\mid q)$ 分别是模型判定段落为相关段落的概率。

2.3　检索推理阶段

检索推理阶段旨在通过结构化流程提升检索和推理效率，包括离线和在线两个部分。离线部分主要负责将生成的元数据文档集通过稀疏和稠密编码器嵌入对应的向量，并分别存储在相应的向量库中；在线部分主要是结合用户输入进行实时检索和推理，从而生成最终的回答。

2.3.1　混合编码器

混合编码器的目的是通过结合稀疏和稠密编码方法，为文本提供丰富且精细的语义表示。稀疏编码器和稠密编码器各自发挥不同作用，共同提升检

2.2.2　思维链提示生成模块

思维链提示生成模块旨在提升回答的质量和可信度。总体来说，该模块通过显式展示推理过程，帮助模型更好地理解问题的逻辑关系，以提供精准且有依据的答案。具体而言，多问题生成模块首先生成高质量的问答对，其中包含明确的问题和答案，这些问答对成为思维链提示生成的输入，输出的内容则涵盖从问题理解到答案生成的各个推理步骤。这不仅增强了模型的推理能力，还提高了回答的准确性和合理性。

2.2.3　指令微调数据集构建模块

指令微调数据集构建模块旨在为模型提供全面的训练样本，以增强其信息检索和问答能力。该模块通过整合多方面的内容，包括问答对、答案推理过程、黄金文本段落及干扰文本段落，形成完整的训练数据集。

具体来说，干扰文本段落来自整个文档集，在排除与当前问答对相关的文档后随机检索得到；然后，将黄金文本段落与干扰文本段落随机排列，以保证模型在训练时能够更好地分辨和利用相关信息。在训练过程中，模型不仅要学会生成正确的答案，还要学会从黄金文本段落和干扰文本段落中检索和区分相关信息，这通过优化以下目标函数来实现，如式(2)所示：

索效果。稀疏编码器迅速提取与用户查询匹配的关键词，适用于电力法规和标准的快速检索；稠密编码器则深入分析上下文和语义，捕捉更深层次的文档关联。最终，将两种结果经过排序整合，确保提供准确且高度相关的答案，这一过程提升了 Meta-RAG 在电力领域复杂查询中的效率和准确性。

Meta-RAG 核心采用 bge-m3 语义向量模型，其以多语言支持和多功能性而著称。在稀疏编码过程中，bge-m3 评估序列中每个 token 的情境化嵌入，并通过线性变换和 ReLU 激活生成稀疏向量，这不仅捕捉了单词间的语义关系，还突出了文本中的关键内容，丰富了词汇信息；同时，在稠密编码过程中，bge-m3 利用多层 Transformer 架构，精细化地表示每个 token，最终通过 CLS token 的嵌入表达整体输入。

2.3.2 混合检索模块

混合检索模块旨在综合利用稀疏和稠密检索，以提供最相关的搜索结果。总体上，该模块在单个数据集合中实现了多向量字段的高效搜索。具体操作中，首先将数据集合加载到内存，并为每个向量字段创建索引，以优化搜索性能；在搜索执行时，分别对稀疏和稠密向量字段发起请求，并根据预设的策略合并结果。采用的相似度计算方法如式(3)所示：

$$d_{\text{cosine}}(\boldsymbol{v}_q,\boldsymbol{v}_d)=1-\frac{\boldsymbol{v}_q\cdot\boldsymbol{v}_d}{\|\boldsymbol{v}_q\|\|\boldsymbol{v}_d\|} \tag{3}$$

式中：$\boldsymbol{v}_q$ 是查询向量，代表一个问题或查询的特征；$\boldsymbol{v}_d$ 是文档向量。

在混合检索模块中，对稀疏检索和稠密检索的结果给出相同的权重，混合检索计算方式如式(4)所示：

$$S_{\text{hybrid}}=\lambda\cdot S_{\text{sparse}}+(1-\lambda)\cdot S_{\text{dense}} \tag{4}$$

式中：S_{hybrid} 表示混合相似度得分，用于将稀疏和稠密表示结合起来；λ 是一个权重参数，介于 0 和 1 之间，用于调节稀疏和稠密得分在最终结果中的影响；S_{sparse} 是基于稀疏表示的相似度得分；S_{dense} 是基于稠密表示的相似度得分。

2.3.3 重排序模块

重排序模块的目的是优化检索结果的准确性和相关性。整体上，通过双重排序策略提供更优质的输出。

首先，采用倒数排名融合(RRF)策略进行初步多文档排序。RRF 利用式(5)综合多个独立检索器的排名结果，为文档计算综合得分，确保公平性且提升排序准确性。

$$S_{\text{RRF}}(d)=\sum_{i=1}^{k}\frac{1}{r_i(d)+c} \tag{5}$$

式中：$S_{\text{RRF}}(d)$代表文档 d 的综合得分；k 是参与融合的检索器数量；$r_i(d)$是第 i 个检索器中文档d 的排名；c 是一个常数，本文取 $c=60$，以防止高排名的检索器得分过大造成不公平。

接着，使用 bge-reranker-v2-m3 模型进行单文档重排序。此模型运用分层自蒸馏策略提升推理效率，在多个检索基准测试中表现卓越。通过式(6)，模型将来自混合检索模块的候选文档进行精细排序，将与查询最相关的文档排列在前，形成文档集合。通过结合稀疏和稠密检索模型，该模块在有限资源环境中依然能高效运行，显著提升检索系统性能和输出质量。

$$\text{RelevantDocs}(q)=\{d_i\mid\text{rank}(d_i,q)\leqslant k,\forall d_i\in D\} \tag{6}$$

式中：RelevantDocs(q)表示与查询 q 相关的文档集合；d_i 是候选文档，从文档集合 D 中选择出来；rank(d_i,q)是文档 d_i 相对于查询 q 的排名；k 是预定义的阈值，表示选择排名在前 k 名的文档；D 是所有可能的文档集合。

2.3.4 大语言模型提示模块

大语言模型提示模块旨在通过结构化的框架，引导模型进行有效的推理和解题。总体上，该模块提供了一种系统化的方法，以提高大语言模型的回答准确性和逻辑性。具体而言，该模块包括以下步骤：

1)确定最相关的文档：首先识别和选择能够最大化满足问题需求的相关文档，以确保信息的准确和全面。

2)提取关键文档片段：在相关文档中识别最具信息性和指导性的片段，找到可以直接回答问题的部分，提升信息的有效性。

3)详细推理说明：结合选定的文档和片段进行深入的推理分析，解释如何综合利用这些信息得出准确答案。此步骤强调逻辑推理和实质性分析，增强了答案的说服力和逻辑性。

4)提供明确答案：根据推理过程选择最合适的答案，并以清晰简洁的形式呈现给用户。

3 实验结果与分析

3.1 基准测试

EleQA 数据集是专为电力领域知识问答而打造的高质量数据集，包含 32 610 条电力规范条款和 19 560 个问答对，其覆盖的领域非常广泛，包括水电站设备检修管理、电力电容电感测试、电力建设、风电场电量评估、电力变压器选用、核电厂设备维修、发电厂锅炉机组除尘器检修、调相机检修、输电

线高空救援等多个场景，在题型上覆盖单选、填空和判断 3 种主要题型。数据集通过严格的文档解析和元信息抽取处理，确保数据的高质量与领域特异性。本文提出了一种评估策略，采用准确率为核心评估指标，并结合多任务评估方法区分不同题型的模型响应。此外，基准模型选择包括 Self-RAG[6]、Corrective-RAG[7]、Adaptive-RAG[8]等在内的主流 RAG 模型进行对比，验证模型在不同任务中的表现，为电力领域的智能问答系统提供标准化的评估基准。

3.1.1 数据

EleQA 数据集各类题型的数量及答案分布如表 1 所示。本文在构建时考虑到了单选题答案和判断题答案的随机分布，以帮助减少模型对某一种答案模式的偏见，按比例选出不同题型用来作为训练语料，这有助于训练出更加鲁棒的模型。各类题型样本示例如表 2 所示。

表 1 各类题型分布情况

Table 1 Distribution of question types

Question Type	Quantity	Answer Distribution
Single Choice	6 150	A/B/C/D:1 490/1 530/1 570/1 560
Fill-in-the-Blank	6 590	—
True/False	6 820	True/False:3 160/3 640

表 2 各类题型样本示例

Table 2 Sample examples of various question types

Question Type	Question	Answer	Text Passage
Multiple Choice	在哪些地区，瓷绝缘子胶合剂外露表面通常需要涂密封胶？A. 高温地区 B. 干旱地区 C. 高湿地区 D. 易结冰且冰期较长的地区	D	密封胶易结冰且冰期较长的地区，瓷绝缘子胶合剂外露表面宜涂密封胶
Fill-in-the-Blank	在进行检测时，如果设备处于_____状态，且设备表面没有进行外部工作，可以进行检测操作	带电运行	被检测设备是带电运行设备，且设备上无各种外部作业
True/False	所有文件都应该混合立卷，不分载体类型	错误	不同载体的文件应分别立卷

3.1.2 评估指标

本文对生成内容采用的评估指标为准确率。在针对 EleQA 数据集的评估中，本文设计了一套评估公式，以准确衡量大语言模型在不同题型上的表现。具体而言，EleQA 数据集包括单选题、填空题和判断题，这些题型只接受一个预定义的正确答案。由于填空题的特殊性，如果回答中包含正确答案或者核心关键词，视为正确。现有的评估技术很少能够对不同类型的回答进行区分处理。例如，根据答案是否存在以二进制方式衡量大语言模型响应的正确性[24]，这种评估策略并非没有挑战。例如，预定义答案为“纽约市”时，如果回答为“纽约”，按照严格标准会视为回答错误。为了更全面地评估大语言模型生成回答的准确性，设计了一套针对不同题型进行分别判定的评估方法。从大语言模型的响应中区分答案部分和解释部分，这对选择题来说至关重要，关于错误选项分析的这段文字如果评估时不排除，给出错误选项也可能得分，造成准确率虚高，因此需要对响应结果进行后续处理；在判断题的判定中，由于大语言模型响应具有随机性，通过微调和精心设计的提示（如利用链式思维提示），确实可以大幅提高模型按照指定格式输出的能力，但使用明确、重复和具体的指示并不能完全消除偏差，因此在模型响应中仍有很低的概率使用类似“对”和“不正确”的近义词来表示“正确”和“错误”的含义，类似的表示通过后续处理，将响应中相关近义词转换为“正确”或者“错误”，进一步减少误判。

对模型在填空题上的性能评估公式如式(7)所示：

$$\mathrm{Eval}(\theta,D_{\mathrm{test}})=\frac{1}{|D_{\mathrm{test}}|}\sum_{(q,A)\in D_{\mathrm{test}}}\mathbb{I}(A\subseteq\hat{A}) \quad (7)$$

式中：$\mathbb{I}$ 是指示函数，如果 $A\subseteq\hat{A}$ 为真，指示函数结果为 1，否则为 0；A 是预定义答案拆分得到的关键词集合，将模型输出中的答案部分拆分，得到的关键词集合 $\hat{A}$；q 是测试数据集 D_{test} 中的查询；θ 是模型参数。对单选题、判断题的评估公式如式(8)所示：

$$\mathrm{Eval}(\theta,D_{\mathrm{test}})=\frac{1}{|D_{\mathrm{test}}|}\sum_{(q,\alpha)\in D_{\mathrm{test}}}\mathbb{I}(\hat{\alpha}=\alpha) \quad (8)$$

式中：α 是预定义答案；$\hat{\alpha}$ 为模型输出中的答案部分；$\mathbb{I}$ 是指示函数，指示函数的标准定义为如果 $\hat{\alpha}=\alpha$ 为真，指示函数结果为 1，否则为 0。

检索器的评估指标为召回文档命中率(Recall@k)[25]。F1 值是衡量自动索赔验证系统性能的常用指标，尽管对分类任务的评估具有价值，但不足以评估非分类组件的性能。Recall@k 衡量的

是检索系统在前 k 个结果中返回相关文档的能力。在知识密集型任务(如问答系统)中,确保检索到的文档集中包含相关文档至关重要,这样能够提供准确且信息丰富的回答。高召回率确保系统能够捕捉到生成正确答案所需的足够信息。

3.2 参数设置

在模型训练阶段,设置了以下关键参数:每轮设备训练批次大小为 1,学习率(Learning Rate)为 1×10^{-4},训练轮次为 3,梯度累计步数为 8,学习率调度器类型为 cosine。训练使用的显卡为 2 块 NVIDIA RTX A6000 48 GB,涉及的深度学习库包括:PyTorch 2.3.1+cu121;Transformers 4.46.0.dev0;Datasets 2.20.0;Accelerate 0.33.0;PEFT 0.11.1;TRL 0.9.6。

在推理阶段,实验使用的显卡型号为 NVIDIA RTX A6000 48 GB x2,使用 vllm 加速推理,Python 版本为 3.11.0,涉及的深度学习框架包括:PyTorch 2.3.1;torchvision 0.18.1;Transformers 4.42.4;sentence-transformers 2.2.2。此外,还包括 NVIDIA CUDA 工具集的多个组件,用于支持在 NVIDIA GPU 上的加速计算。

3.3 基线模型

为了评估 Meta-RAG 框架的有效性,选择了近 1 年内表现出色的 RAG 模型作为基线模型进行对照,其中:Self-RAG[6]通过自我监督机制提升了性能,在多个任务中展现出卓越的效果,是一个非常有参考价值的对照模型;Corrective-RAG[7]采用纠错机制优化回答准确度,在电力领域的应用中可能会有较好表现,选择其作为基线模型有助于探讨 Meta-RAG 在纠错方面的潜在优势;Adaptive-RAG[8]通过适应性学习机制提高了灵活性和泛化能力,适合处理电力领域多样化的数据和任务,是评估 Meta-RAG 模型灵活性的一个良好参照;RA-ISF[9]在检索和生成回答方面结合了信息检索精度和生成模型的优势,适合用于复杂的电力信息系统,选择其作为基线能初步衡量 Meta-RAG 的综合性能。

3.4 实验结果

3.4.1 基准测试

Meta-RAG 与不同 RAG 方案在电力领域知识问答的基准测试结果如表 3 所示(表中加粗数据表示最优值,下同)。从实验结果上看:Meta-RAG 底座使用 GLM-4-9B-Chat、Baichuan2-13B-Chat 和 Qwen1.5-14B-Chat 均取得了不错的效果,其中底座使用 Qwen1.5-14B-Chat 的方案取得了最佳结果;对比方案的底座都是基于 Qwen1.5-14B-Chat 这一底座进行的。没有使用 RAG 的策略与 Naive 策略分数相差不大,这是由于简单的文本拆分用于海量的文档段时检索效率并不高,也无法保证检索到的相关文本段具有语义的完整性。

表 3 基准测试结果

Table 3 Benchmark test results

Strategy	Model	Accuracy			
		Single Choice	Fill-in-the-Blank	True/False	Overall
None	Qwen1.5-14B-Chat	0.656 1	0.044 1	0.784 3	0.495 1
Naive	Qwen1.5-14B-Chat	0.665 2	0.097 1	0.720 9	0.510 4
Self-RAG[6]	Qwen1.5-14B-Chat	0.863 3	0.495 7	0.866 8	0.739 9
Corrective-RAG[7]	Qwen1.5-14B-Chat	0.872 7	0.501 7	0.869 6	0.745 9
Adaptive-RAG[8]	Qwen1.5-14B-Chat	0.867 4	0.492 0	0.876 1	0.743 2
RA-ISF[9]	Qwen1.5-14B-Chat	0.854 7	0.489 7	0.858 8	0.732 5
Meta-RAG	GLM-4-9B-Chat	0.847 8	0.478 9	0.850 2	0.723 6
Meta-RAG	Baichuan2-13B-Chat	0.837 2	0.495 7	0.866 8	0.731 8
Meta-RAG	Qwen1.5-14B-Chat	**0.887 4**	**0.632 2**	**0.895 6**	**0.804 3**

3.4.2 消融实验

本文进行 3 个消融实验,探究 Meta-RAG 每个组件对整体的贡献,分别测试检索能力、微调策略和推理提示对结果的影响,实验结果如表 4 所示。从结果上看,失去检索能力(cot+ft)整个框架整体准确率下降了 0.292 8,不使用推理提示框架(MetaMdChunk+dense+ft)整体准确率下降了 0.029 2,无微调框架(MetaMdChunk+hybrid+cot)整体准确率下降了 0.030 1,对比可知,文档检索对整个框架性能影响最大。

表 4 消融实验结果

Table 4 Results of ablation experiment

Model	Accuracy			
	Single Choice	Fill-in-the-Blank	True/False	Overall
cot+ft	0.664 2	0.081 7	0.787 9	0.511 5
MetaMdChunk+dense+ft	0.867 0	0.568 4	0.893 4	0.775 1
MetaMdChunk+hybrid+cot	0.863 3	0.574 0	0.888 7	0.774 2
Meta-RAG	**0.887 4**	**0.632 2**	**0.895 6**	**0.804 3**

3.4.3 文档召回实验

文档召回实验计算召回的前 3 个文档中是否包含相关文档，实验结果如表 5 所示。从结果上看，无论从整体上还是各题型上，混合检索都优于单一的稠密检索和稀疏检索，其中使用 RRF + bge-reranker-v2-m3 的文档召回率最高。具体来说，RRF 策略通过融合多个检索器的优势，有效地扩大了检索结果的覆盖面，而 bge-reranker-v2-m3 则进一步优化了排序精度，从而显著提高了相关文档的召回率。这样的组合不仅在整体性能上表现出色，对各类疑难问题的处理也具有显著优势，进一步证明了融合策略在复杂检索任务中的有效性。实验结果表明，结合多种检索方法和先进的重排序技术，可以使系统的整体检索效果得到一定的提升。

表 5 文档召回实验结果(Top-3)

Table 5 Results of document retrieval experiment (Top-3)

Retrieval Strategy	Accuracy			
	Single Choice	Fill-in-the-Blank	True/False	Overall
BM25	0.498 2	0.427 3	0.603 2	0.511 3
Dense	0.546 7	0.474 0	0.589 8	0.537 4
Hybrid+RRF	0.564 3	0.489 3	0.606 0	0.553 7
Hybrid+RRF+bge-reranker-v2-m3	**0.571 6**	**0.502 6**	**0.608 6**	**0.561 4**

为了探究文档召回时 Top-k 参数的影响，采用了 Hybrid + RRF + bge-reranker-v2-m3 作为文档检索方法，实验结果如表 6 所示。根据实验结果可以看出，k 超过 5 后性能提升不再显著，且 Top-3 和 Top-5 的效果相近。然而，引入更多的无关文档可能会对模型的表现产生更大干扰。因此，本文最终选择 Top-3 作为所有检索的默认配置。

表 6 文档召回实验结果

Table 6 Results of document retrieval experiment

Top-k	Accuracy (Overall)
Top-1	0.494 4
Top-3	0.561 4
Top-5	0.564 9
Top-10	0.575 6
Top-15	0.580 5
Top-20	0.581 6

4 结束语

本文提出了一种基于元数据驱动的 Meta-RAG 框架，旨在解决大语言模型在处理知识密集型任务时的事实准确性和知识更新问题，特别是面向电力领域的知识问答任务。通过数据准备、模型微调和检索推理 3 个阶段，Meta-RAG 有效地解决了数据集缺乏、RAG 框架不适配、文档信息整合效率低等问题。实验结果表明，Meta-RAG 在电力领域知识问答任务中取得了显著成效，尤其在回答准确率和检索命中率方面优于现有基线模型。此外，本文通过 EleQA 数据集设立了电力领域的基准。未来研究可聚焦于实时更新机制、多模态数据融合及结合知识图谱优化检索，以进一步提升系统性能。

参考文献

[1] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[C]//Proceedings of Conference on Neural Information Processing Systems. Berlin, Germany: Springer, 2020: 1-10.

[2] KARPUKHIN V, OĞUZ B, MIN S, et al. Dense passage retrieval for open-domain question answering [EB/OL]. [2024-07-27]. https://arxiv.org/abs/2004.04906.

[3] IZACARD G, LEWIS P, LOMELI M, et al. Atlas: few-shot learning with retrieval augmented language models[J]. Journal of Machine Learning Research, 2023, 24(251): 1-43.

[4] YAO J Y, NING K P, LIU Z H, et al. LLM lies: hallucinations are not bugs, but features as adversarial examples[EB/OL]. [2024-07-27]. https://arxiv.org/abs/2310.01469.

[5] BANG Y, CAHYAWIJAYA S, LEE N, et al. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity[EB/OL]. [2024-

07-27]. https://arxiv.org/abs/2302.04023.
[6] ASAI A, WU Z, WANG Y, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection[C]// Proceedings of the 12th International Conference on Learning Representations. Berlin, Germany: Springer, 2024: 1-10.
[7] YAN S Q, GU J C, ZHU Y, et al. Corrective retrieval augmented generation[EB/OL]. [2024-07-27]. https://arxiv.org/abs/2401.15884.
[8] JEONG S, BAEK J, CHO S, et al. Adaptive-RAG: learning to adapt retrieval-augmented large language models through question complexity[C]//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Stroudsburg, USA: ACL, 2024: 7036-7050.
[9] LIU Y M, PENG X Y, ZHANG X H, et al. RA-ISF: learning to answer and understand from retrieval augmentation via iterative self-feedback[EB/OL]. [2024-07-27]. https://arxiv.org/abs/2403.06840.
[10] 张峰, 杨晓艺, 刘奕湘. 电力智能问答平台架构的研究与设计[J]. 能源与环保, 2017, 39(7): 193-195.
ZHANG F, YANG X Y, LIU Y X. Research and design of power intelligent question-answering platform architecture [J]. Energy and Environmental Protection, 2017, 39(7): 193-195. (in Chinese)
[11] 周帆, 叶健辉, 肖林朋, 等. 基于知识图谱的电网模型本体智能问答系统研究[J]. 中国科技信息, 2019, 31(16): 85-86.
ZHOU F, YE J H, XIAO L P, et al. Research on intelligent question-answering system of power grid model ontology based on knowledge graph[J]. China Science and Technology Information, 2019, 31(16): 85-86. (in Chinese)
[12] 覃祥坤. 一种电力图谱问答系统设计与实现[D]. 北京:中国科学院大学人工智能学院, 2020.
QIN X K. Design and implementation of a power graph question answering system[D]. Beijing: School of Artificial Intelligence, University of Chinese Academy of Sciences, 2020. (in Chinese)
[13] ZHANG Q, JIA Q Y, WANG Y H. Question answering based assisted decision for electric power fault diagnosis[C]// Proceedings of the IEEE 5th International Conference on Cloud Computing and Big Data Analytics. Chengdu, China: IEEE Press, 2020: 194-198.
[14] MENG F Q, WANG W H, WANG J D. Research on short text similarity calculation method for power intelligent question answering[C]//Proceedings of the 13th International Conference on Computational Intelligence and Communication Networks. Lima, Peru: IEEE Press, 2021: 91-95.
[15] LI W Q, QI X M, ZHAO Q, et al. Knowledge graph-based credibility evaluation method for electric grid large language model knowledge question-answering[C]//Proceedings of the 7th International Conference on Electronic Information Technology and Computer Engineering. New York, USA: ACM, 2023: 754-759.
[16] XIN R, ZHANG P F, CHEN X, et al. Knowledge graph question-answering based on link reasoning for electrical equipment [C] // Proceedings of the 2024 International Conference on Power Electronics and Artificial Intelligence. New York, USA: ACM, 2024: 594-600.
[17] ZHAO J X, MA Z C, ZHAO H, et al. Self-Consistency, Extract and Rectify: knowledge graph enhance large language model for electric power question answering [C] // Advanced Intelligent Computing Technology and Applications. Singapore: Springer Nature Singapore, 2024: 493-504.
[18] HUANG C H, LI S Y, LIU R H, et al. Large foundation models for power systems[EB/OL]. [2024-07-27]. https://arxiv.org/abs/2312.07044.
[19] CHENG Y H, ZHAO H, ZHOU X Y, et al. A large language model for advanced power dispatch [EB/OL]. [2024-07-27]. https://arxiv.org/abs/2408.03847.
[20] MAJUMDER S, DONG L, DOUDI F, et al. Exploring the capabilities and limitations of large language models in the electric energy sector[J]. Joule, 2024, 8(6): 1544-1549.
[21] RUAN J Q, LIANG G Q, ZHAO H, et al. Applying large language models to power systems: potential security threats[J]. IEEE Transactions on Smart Grid, 2024, 15(3): 3333-3336.
[22] HU E J, SHEN Y, WALLIS P, et al. LoRA: low-rank adaptation of large language models[EB/OL]. [2024-07-27]. https://arxiv.org/abs/2106.09685.
[23] ZHANG T J, PATIL S G, JAIN N, et al. RAFT: adapting language model to domain specific RAG[EB/OL]. [2024-07-27]. https://arxiv.org/abs/2403.10131.
[24] CUCONASU F, TRAPPOLINI G, SICILIANO F, et al. The power of noise: redefining retrieval for RAG systems [EB/OL]. [2024-07-27]. https://arxiv.org/abs/2401.14887.
[25] DMONTE A, ORUCHE R, ZAMPIERI M, et al. Claim verification in the age of large language models: a survey [EB/OL]. [2024-07-27]. https://arxiv.org/abs/2408.14317.

文字编辑 金胡考
栏目编辑 赖玉玲