

DOI:10.3969/j.issn.1671-0673. 2026. 02. 003

基于嵌入模型微调的信号领域知识问答优化

田文浩, 屈 丹, 司念文, 刘励超

(信息工程大学, 河南 郑州 450001)

摘要: 针对检索增强生成(RAG)在信号处理这类术语密集、逻辑严谨的专业领域中面临的领域知识覆盖不足和术语语义理解偏差等问题, 提出一种基于嵌入模型微调的信号领域知识问答优化方法。首先, 构建高质量的信号领域数据集, 并采用对比学习策略对检索器中的嵌入模型进行领域适配微调; 其次, 基于微调后的嵌入模型搭建 RAG 系统, 并在检索环节引入混合检索与重排序机制, 在生成环节优化提示词设计, 以提升对信号领域知识的理解与生成质量。实验结果表明, 经微调后的嵌入模型在信号领域检索任务中多项性能指标显著提升, 优化后的 RAG 系统在忠实度、答案相关性和上下文精确率等方面均优于基线方法。

关键词: 检索增强生成; 信号; 大语言模型; 模型微调; 问答系统

中图分类号: TP391.41

文献标识码: A

文章编号: 1671-0673(2026)02-0145-08

Optimization of Signal Domain Knowledge QA Based on Embedding Model Fine-Tuning

TIAN Wenhao, QU Dan, SI Nianwen, LIU Lichao

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To address the challenges of insufficient domain knowledge coverage and semantic misunderstanding of technical terms when applying retrieval-augmented generation (RAG) to terminology-intensive and logically strict fields such as signal processing, a QA optimization method for signal processing knowledge based on embedding model fine-tuning is proposed. First, a high-quality dataset in the signal domain is constructed, and a contrastive learning strategy is adopted to fine-tune the embedding model in the retriever for better domain adaptation. Subsequently, a RAG system is built using the fine-tuned embedding model. Hybrid retrieval and re-ranking mechanisms are introduced during the retrieval phase, while prompt optimization is applied in the generation phase to enhance the comprehension and quality of outputs related to signal domain knowledge. Experimental results demonstrate that the fine-tuned embedding model significantly improves multiple performance metrics in signal-domain retrieval tasks. The optimized RAG system also outperforms baseline methods in terms of faithfulness, answer relevancy, and contexts precision.

Key words: retrieval-augmented generation; signal; large language model; model fine-tuning; question answering system

近年来, 大语言模型 (Large Language Models, LLMs) 在知识表示与推理方面展现出突破性能力。以 GPT^[1] 系列为代表的通用大模型通过预训练-微

调范式, 已在开放域知识问答、立场检测等任务中验证了其语义理解优势^[2]。然而, 直接将通用大模型应用于专业领域面临知识长尾分布、术语语义

收稿日期: 2025-10-29; 修回日期: 2025-11-25

基金项目: 河南省自然科学基金 (252300420990); 河南省科技攻关项目 (252102211040)

作者简介: 田文浩 (2003—), 男, 硕士生, 主要研究方向为人工智能、大语言模型。E-mail: tianwh224@163.com

鸿沟等领域化挑战。检索增强生成(Retrieval-Augmented Generation, RAG)^[3]提供了解决思路。通过构建领域知识库与向量数据库的协同索引机制,结合大模型的理解生成能力与检索系统的知识管理能力,可显著提升生成内容的准确性、时效性与专业性,对于推动大模型在垂直领域应用具有重要意义。

RAG经历了从基础到复杂、从单一到模块化的范式转变,可分为朴素RAG、高级RAG、模块化RAG等3个发展阶段^[4]。在检索方面,重排序^[5]的引入可滤除干扰文档,查询重写、转换和扩展^[6-8]等方法优化了查询表达,混合检索策略^[9]结合向量相似性与关键词匹配以提高检索精度。在生成方面,检索和生成的迭代协同^[10]通过多轮交互减少幻觉问题,自我反思机制^[11]则显著提升了RAG系统的生成质量和事实准确性。上述进展提升了RAG在领域化场景的应用效果,推动了电信^[12]、电力^[13]和制造^[14]等垂直领域的应用。这些应用主要通过结构化数据处理和生成模型微调等方法增强大模型领域知识适应性,提升了问答效果。然而,这些方法成本普遍较高,且检索器采用通用嵌入模型,未充分考虑检索器嵌入模型对于系统性能的影响。尤其在信号处理这类数理逻辑强、特殊术语符号密集的专业领域中,通用嵌入模型对于密集术语语义表征能力弱、区分度低,易导致检索性能下降,进而制约RAG专业领域知识覆盖率和术语符号理解能力。

针对以上问题,本文以信号领域知识为对象,构建了一个信号领域知识数据集,在该数据集上对文本嵌入模型进行微调,基于微调后的嵌入模型实现信号领域知识RAG问答系统,并在检索和生成策略上对系统进行领域适配优化,从而提高RAG系统在信号领域的性能表现。

本文的主要工作包括:1)构建高质量信号领域数据集,利用大模型辅助提取原始数据,通过多次检查剔除低质量数据,最终形成两个数据量分别为5 075条和7 846条的数据集,用于微调文本嵌入模型。2)基于信号领域数据集微调文本嵌入模型,并使用对比学习策略进一步提升文本区分度。3)基于信号文本嵌入模型搭建面向信号领域的RAG问答系统,并对其检索和生成策略进行适配优化,在忠实度、答案相关性、上下文精确率等指标上均优于基线系统,从而验证了本文“领域数据集构建-嵌入模型微调-检索生成策略协同优化”的领域RAG优化框架。

1 所提方法

1.1 信号领域数据集

数据集所涉及的原始文献资料主要来源于信号领域的经典教材、技术报告、权威网站和相关协议标准。原始文献资料将依次经过预处理、文本对提取、质量检查、负采样、数据分集,最终形成可以直接使用的数据集,用于嵌入模型的微调和测试。数据集构建的具体流程如图1所示。

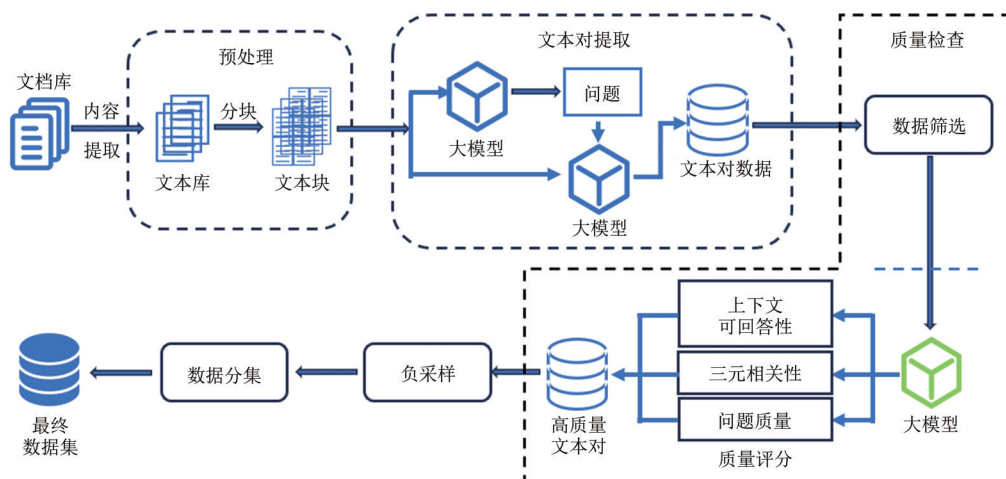


图1 数据集构建流程

1.1.1 预处理和文本对提取

预处理包括内容提取和分块。主要是将原始文献资料中的文本、表格等内容提取出来并进行分块,形成文本语料。

文本对提取主要利用大模型从文本语料中分别提取出文本问答对数据。该步骤中利用大模型做“问题生成-回答生成”的两阶段提取。提取出的文本对的内容格式如下:

{“question”:[], “answer”:[], “context”:[]}]}

其中:question是利用大模型从输入语料中提取的高价值问题,涵盖多个方面;answer是大模型根据问题和输入语料给出的回答;context是回答问题所使用的上下文,来源于输入语料,支持问题的回答。

1.1.2 质量检查

质量检查包括数据筛选和质量评分,以提高文本对质量。

数据筛选环节包含字符串匹配、N-gram 相似度检验、语义相似度检验以及数值实体检验等步骤,如图2所示。

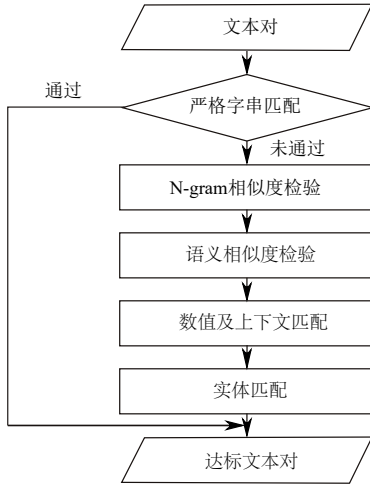


图2 数据筛选流程

首先,对文本对进行严格字符串匹配,即判断文本对是否严格存在于输入语料中,可大幅减少后续环节开销。对于未通过严格字符串匹配的文本对,其后续筛选流程可以用式(1)表达:

$$F_{\text{filter}} = E(S(N(c_p, C, \theta_N), C, \theta_S), C) \quad (1)$$

式中: c_p 表示文本对 p 中的context; C 表示输入语料

块; θ_N 表示N-gram相似度阈值; θ_S 表示语义相似度阈值; F_{filter} 表示整个后续筛选操作,包含3个操作; $N(\cdot)$ 表示N-gram检验,是第1个操作,对 c_p 和 C 逐一计算Jaccard相似度,相似度最大值大于 θ_N 的 p 通过检验; $S(\cdot)$ 表示语义相似度检验,是第2个操作,使用语言模型对 c_p 和 C 进行语义相似度计算,相似度大于 θ_S 即通过检验; $E(\cdot)$ 表示数值与实体检验,是第3个操作, c_p 中的数值与实体存在于 C 中则通过检验。

质量评分阶段,利用大模型进行文本对评分,从高层次语义和逻辑上进行筛选。该阶段检查聚焦3方面指标:1)上下文可回答性,判断能否根据上下文得到问题答案;2)三元相关性,包括question和answer、question和context、answer和context的相关性;3)问题质量评分,衡量问题和目标领域的贴合程度、实际价值。

1.1.3 负采样和数据分集

经过质量检查的数据仍不可直接用于嵌入模型微调,需要进行负采样,以同时包含正负样本。数据格式如下:

{“query”:[], “pos”:[], “neg”:[], “prompt”:[]}

其中:query是查询,由第1.1.2节得到的文本对中的question直接转换得到;pos是查询的相关文本,由文本对中的context直接转换得到;neg是与查询无关或者相关性很低的文本,每条数据均包含了10条无关文本;prompt是用于检索的指令,设置为“为这个句子生成表示以用于检索相关文章:”。

数据分集将数据集分为训练集和测试集,分别用于训练和检索测试,数据样例如表1所示(无关文本仅部分展示)。

表1 数据样例

字段	内容
query	若某无噪无损信道的输入符号数 $r=8$ 且等概率分布,求其信道容量 C (需给出推导过程)
pos	无噪无损信道...其信道容量 $C=\max I(X;Y)=\max H(Y)=\max H(X)=\log_2 8=\log_2 2^3=3$ (4.14)···当信道输入等概率分布时,输出也为等概率分布,信道达到信道容量 “定理5.11.3生成随机变量 X 的最优算法所需的均匀比特数的期望值在 $H(X)$ 和 $H(X)+2$ 之间: $H(X) \leq E T < H(X)+2/\ln(5-115)$ ” “当要求滤波器为线性相位响应特性时,则通带内群延迟特性就应是常数”
neg	“3.21 如题3.21图所示的复合系统由3个子系统组成,它们的单位序列响应分别为 $h(k)=\delta(k)$, $h(k)=\delta(k-N)$, N 为常数, $h(k)=\varepsilon(k)$,求复合系统的单位序列响应。” ...
prompt	为这个句子生成表示以用于检索相关文章:

1.2 基于对比学习的嵌入模型微调

首先明确本文所涉及的信号领域检索任务:给定一个查询 $q \in Q$ 和一个领域知识文档集合 $D =$

$\{d_1, d_2, \dots, d_N\}$,任务目标是学习一个 $g: Q \times D \rightarrow [0,1]$,使对于任意 q ,其相关文档 d^* 的得分 $g(q, d^*)$ 明显高于无关文档得分。而嵌入模型是决定文档得分的

关键一环。

在 RAG 系统的嵌入模型选择上,本文选择 bge-large-zh-v1.5^[15]和 bge-m3^[16]模型。前者对中文文本表示能力优秀,参数规模较小;而后者则是一个性能强大的多语言模型,模型较复杂,参数规模大。两者同属于智源研究院的 BGE 系列双编码器架构模型,尽管精度低于同系列交叉编码器模型,但是效率高,适用于初步检索。该系列模型在 C-MTP^[15]数据集上进行纯文本预训练、通用微调、特定任务微调 3 个阶段训练。

本文基于自主构建的数据集,对嵌入模型进行微调,以适配信号领域。微调过程采取的方法与 BGE 系列模型训练的第 3 阶段方法基本一致。在对比学习策略基础上,使用指令微调,明确微调任务目标,从而提高模型在信号领域的检索任务性能表现。微调时的目标函数如式(2)所示:

$$\min \sum_{(p,q)} -\log_2 \frac{e^{\langle e_p, e_q \rangle / \tau}}{e^{\langle e_p, e_q \rangle / \tau} + \sum_{q' \in Q'} e^{\langle e_p, e_{q'} \rangle / \tau}} \quad (2)$$

式中: p 和 q 是一对文本; Q' 是一个负样本集合; $q' \in Q'$ 是一个负样本; τ 是温度; e_p 、 e_q 分别是 p 和 q 的向量表示; $e_{q'}$ 是 q' 的向量表示; $\langle e_p, e_q \rangle$ 表示向量内积。

该目标函数所优化的损失本质上是 InfoNCE 损失变体。在训练过程中,通过优化损失,最大化正样本对的互信息估计,模型学习到信号领域的嵌入空间,使得语义相近的信号文本在嵌入空间中距离小,而语义无关文本距离大,从而提高文本区分度,进而提高检索性能。

1.3 检索与生成策略优化

为了进一步提高 RAG 在信号领域的问答准确性、专业性及可靠性,本文将采取相应措施对系统展开优化。优化将聚焦 RAG 中基础但关键的两部分:检索模块和生成模块。

在检索模块,本文从 3 个方面进行优化研究,分别是不同嵌入模型、混合检索权重和重排序,探究其各自对检索性能的影响。

嵌入模型表示是文本向量化表示的主要方式之一。不同嵌入模型获得的文本向量不尽相同,从而在检索时向量相似度不同,影响检索效果。本文除了使用前文提到的两个嵌入模型,还将使用基于这二者微调得到的模型,以研究不同嵌入模型对检索性能的影响。使用场合包括知识库构建和检索。

混合检索结合稀疏检索关键词相似度和稠密

检索向量相似度,计算混合相似度,其计算过程可用式(3)表示:

$$S_h(q,d) = \alpha \cdot S_k(q,d) + (1 - \alpha) \cdot S_v(q,d) \quad (3)$$

式中: q 和 d 是查询和文档; S_h 表示混合相似度; S_k 表示关键词相似度; S_v 表示向量相似度; α 是权重超参数。

相关研究表明^[16],混合检索能够取得比单一稀疏检索或者稠密检索更好的效果,因此本文仅采用混合检索,并寻找最优权重。

重排序是对检索结果进行二次排序,直接对检索结果序列进行优化。其过程是,对初步检索结果 $R = \{d_1, d_2, \dots, d_k\}$,用重排序模型 $S_{re}(q,d)$ 打分,根据得分 $S_{re}(q,d_i)$ 降序重排。从而将相关性更高的结果前置,大幅提高检索效果。本文选择的重排序模型 bge-reranker-v2-m3^[15]为交叉编码器,输入查询文档组合并直接输出得分,速度慢但精度高,适合精排。

在生成模块,本文将基于提示工程,改进提示词,对模型输出的角色、领域、表达和推理行为进行约束,还提供少量特殊样本供模型参考,从而提高模型输出的准确性。

2 实验与分析

2.1 实验设置

本文的 RAG 系统基于开源引擎 RAGFlow^[17]搭建。RAGFlow 是 2024 年由 InfiniFlow 提出的一款开源的端到端检索增强生成引擎,深度融合文档解析、混合检索与动态优化技术,具有良好的问答准确性与可解释性。在生成模型的选择上,本文选择 DeepSeek-R1-Distill-Qwen-7B 模型。该模型通过知识蒸馏从 Qwen^[18]从 DeepSeek-R1^[19]中提取而来,在数理逻辑推理方面表现突出,能更好适配信号领域对数理公式推导的需求,同时 7B 的参数量取得了规模、效率和性能的平衡。

2.1.1 数据集

在数据集构建工作中,阈值 θ_N 和 θ_S 最终分别设置为 0.4 和 0.7,以控制数据与原始文献文本的差异。在语义相似度检验环节,为进一步提高检验效果,降低语义不连贯的影响,实验对相邻语料块进行合并和重新切割,加入待计算相似度的语料块中。而负采样中,负采样策略采取随机采样,从同批次数据中随机抽取并选择无关的样本作为负样本。分集时测试集比例为 10%。

在质量评分环节,为平衡数据集规模与质量

并支持后续实验对数据集规模和作用的比较,实验设置了6.0和8.0两个评分阈值来筛选文本对,形成Dataset-L和Dataset-S两个数据集。低阈值筛选出的数据质量较低,但数量较多,质量基本达标。高阈值筛选出的数据质量较高,但数量较少。两个数据集的情况对比如表2所示。

表2 数据集情况

数据集名称	规模	评分阈值	质量
Dataset-S	5 075	8.0	较高
Dataset-L	7 846	6.0	较低

2.1.2 评价指标

1)检索性能指标。在微调模型时,两个基础模型将基于两个数据集分别进行微调,并在各自数据集上进行检索测试。

在第1.2节明确的任务形式下,检索测试涉及的各项指标如下。

精确率(Precision)表示检索出的文档中相关文档的占比,召回率(Recall)表示检索出的相关文档在所有相关文档中的占比。计算公式分别为

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}$$

式中: N_{TP} 表示检索出的相关文档数量; N_{FP} 表示检索出的无关文档数量; N_{FN} 表示未检索出的相关文档数量。

平均精确率(Average Precision, AP)表示在不同召回率点上的精确率均值。计算公式为

$$\bar{P} = \frac{1}{M} \sum_{j=1}^N P_j \cdot r_j$$

式中: M 表示集合 D 中查询 q 的相关文档数量; P_j 表示前 j 个文档的精确率; r_j 表示第 j 个文档的相关度,若不相关则为0,相关为1。

平均精确率均值(Mean Average Precision, MAP)是在平均精确率的基础上,多个查询的平均精确率的均值。给定一组包含 S 个问题的查询 $\{q_i | i = 1, \dots, S\}$,计算公式为

$$I_{MAP} = \frac{1}{S} \sum_{i=1}^S \bar{P}(q_i) \quad (4)$$

累积排序增益(Discounted Cumulative Gain, DCG)考虑了位置带来的影响,前 j 个文档的DCG指标的计算公式为

$$I_{DCG_j} = \sum_{k=1}^j \frac{2^{r_k} - 1}{\log_2(1 + k)}$$

式中, r_k 表示第 k 个文档的相关度,有多个等级(如0、1、2)。

归一化累积排序增益(Normalized Discounted Cumulative Gain, NDCG)在DCG的基础上进行归一化,用DCG值除以前 j 个位置的最大DCG值,公式为

$$I_{NDCG_j} = \frac{I_{DCG_j}}{\max I_{DCG_j}} \quad (5)$$

平均排名倒数(Mean Reciprocal Rank, MRR)表示相关文档在检索结果中排名倒数的均值。在一组查询 $\{q_i | i = 1, \dots, S\}$ 中, d_i 表示每个查询中检索到的第1个相关文档在返回的检索结果中的位置。计算公式为

$$I_{MRR} = \frac{1}{S} \sum_{i=1}^S \frac{1}{d_i} \quad (6)$$

2)问答系统性能指标。在系统优化时,系统性能评估没有采用传统的依赖参考和标注的检索和生成评价指标,而是采用了支持无参考的开源RAG自动评估框架RAGAs^[20]中的3项指标:上下文精确率、忠实度、答案相关性,均在不提供参考的情况下进行评估。

其中,上下文精确率(Contexts Precision, CP)是检索到的文档的平均精确率,评估检索结果准确性。前 K 个检索结果的上下文精确率的计算公式如式(7):

$$I_{CP@K} = \frac{1}{K_{cor}} \sum_{k=1}^K P_k \cdot r_k \quad (7)$$

式中: P_k 表示前 k 个检索结果的精确率; r_k 在第 k 个文档相关时取1,否则取0; K_{cor} 表示前 K 个检索结果中相关文档的数量。

忠实度(Faithfulness)衡量生成的回答是否忠实于检索到的上下文,通过根据检索到的文档中推断出来的陈述数量 n 在回答中所有陈述数量 N 中的占比计算得到。在RAGAs中首先利用大模型将回答分解成若干陈述,而后再利用大模型对每个陈述进行推断。忠实度计算公式如式(8)所示:

$$I_{Faithfulness} = \frac{n}{N} \quad (8)$$

答案相关性(Answer Relevancy, AR)利用大模型评估回答和输入问题的相关性。给定问题 Q ,通过RAG系统得到了回答和检索结果,RAGAs利用大模型在回答和检索结果的基础上逆向生成若干问题 Q_i ,其中 $i = 1, \dots, N$ 。通过计算原始问题和生成问题的相关性评估回答相关性。其计算公式如式(9)所示:

$$I_{AR} = \frac{1}{N} \sum_{i=1}^N \cos(Q, Q_i) \tag{9}$$

2.2 检索和问答结果

本文在 Dataset-S 和 Dataset-L 上对 bge-large-zh-v1.5 和 bge-m3 分别进行微调,并对微调前后模型的检索性能进行测试。微调时,训练轮次为 4,学

习率为 1×10^{-5} ,温度 τ 为 0.02,批次大小设置为 8,实验验证该设置在训练效率、稳定性和效果上表现良好且取得平衡。各模型的检索性能具体表现见表 3。实验结果表明,经过微调后,所有模型在测试集上的性能指标均较大幅提升,尤其在规模较大的 Dataset-L 上微调后整体性能提升更大。

表 3 微调前后模型在数据集上的检索性能(前 10 个检索结果)

数据集	模型	NDCG	MAP	Recall	Precision	MRR
Dataset-S	bge-m3	0.667	0.610	0.847	0.085	0.612
	bge-m3-S	0.753	0.702	0.909	0.091	0.699
	bge-large-zh-v1.5	0.629	0.574	0.797	0.080	0.574
	bge-large-zh-v1.5-S	0.749	0.696	0.913	0.091	0.695
Dataset-L	bge-m3	0.596	0.537	0.780	0.078	0.535
	bge-m3-L	0.753	0.702	0.909	0.091	0.699
	bge-large-zh-v1.5	0.557	0.498	0.739	0.074	0.501
	bge-large-zh-v1.5-L	0.699	0.643	0.875	0.088	0.642

在嵌入模型微调基础上,本文结合检索和生成协同优化对系统进行综合优化,内容包括:嵌入模型微调(bge-m3 在 Dataset-L 上、bge-large-zh-v1.5 在 Dataset-S 上进行)、混合检索权重 α 调为 0.6、采

用重排序、改进提示词。优化前后系统性能如表 4 所示。可见,系统经过综合优化后,3 项问答性能指标都有了显著提升,相较于未优化时,系统综合性能分别提升了约 6.7% 和 14.0%。

表 4 综合优化后 RAG 系统性能

模型	优化策略	Faithfulness	AR	CP	平均得分	综合提升/%
bge-m3	无	0.456	0.537	0.473	0.489	6.7
	综合优化	0.486	0.553	0.528	0.522	
bge-large-zh-v1.5	无	0.436	0.536	0.418	0.463	14.0
	综合优化	0.475	0.558	0.552	0.528	

2.3 实验分析

本文对各环节所采取的方法进行了进一步实

验分析,以研究这些方法的具体影响,实验结果见表 5。

表 5 不同优化方法下 RAG 系统性能

方法	Faithfulness	AR	CP
bge-m3	0.456	0.537	0.473
bge-m3-S	0.523	0.553	0.466
bge-m3-L	0.499	0.541	0.476
bge-m3+重排序	0.508	0.536	0.530
bge-m3+改进提示词	0.495	0.556	0.472
bge-large-zh-v1.5	0.436	0.536	0.418
bge-large-zh-v1.5-S	0.479	0.542	0.479
bge-large-zh-v1.5+重排序	0.476	0.539	0.475
bge-m3-L+权重优化+重排序+改进提示词	0.486	0.553	0.528
bge-large-zh-v1.5-S+权重优化+重排序+改进提示词	0.475	0.558	0.552

在对嵌入模型微调结果(表 3)进行分析时,从数据规模和模型大小两个维度出发。从数据规模看,模型在规模较大的 Dataset-L 上微调后整体性

能提升普遍更大。从模型大小看,较小的 bge-large-zh-v1.5 微调后的提升比 bge-m3 大,尤其是在规模较小的 Dataset-S 上。这一定程度上表明,在数据质量

基本达标时,微调时使用的数据规模对模型性能影响更大,模型越大,越需要大规模的数据。

在检索器嵌入模型选择上,本文对不同的嵌入模型展开了实验(表5),基于微调前后的 bge-large-zh-v1.5 和 bge-m3 模型构建知识库并进行检索和问答。就 CP 这个检索强相关的指标而言, bge-m3 模型微调前后 CP 变化不大,在 Dataset-L 上微调后略微提升。相比之下, bge-large-zh-v1.5 微调后提升较为明显,指标领先。

本文选择权重超参数 α 时设置了 0.8、0.7、0.6、0.5 等 4 组。由于信号领域术语规范多, α 不宜过低,实验中始终保持在 0.5 以上。最终系统性能表现如表 6 所示。结果表明, α 取 0.6 表现最优。

表 6 不同混合检索权重下 RAG 系统性能

模型	参数 α	Faithfulness	AR	CP
bge-m3	0.8	0.460	0.552	0.449
	0.7	0.456	0.537	0.473
	0.6	0.468	0.548	0.478
	0.5	0.503	0.543	0.475
bge-large-zh-v1.5	0.8	0.414	0.535	0.398
	0.7	0.436	0.536	0.418
	0.6	0.480	0.547	0.438
	0.5	0.467	0.548	0.419

本文还就重排序的影响进行了消融实验(表 5)。结果表明,重排序通过检索结果序列优化,为 RAG 提供质量更高的上下文,对于系统性能尤其是检索性能提升作用显著。

在生成方面,本文使用改进前后的提示词进行对比(表 5)。从结果中的忠实度和 AR 看,改进后的提示词确实提升了输出质量。

综合来看,RAG 系统使用微调后的嵌入模型,再结合混合检索权重优化(α 取 0.6)、重排序和改进提示词策略,系统性能最佳,相较于未优化的基线系统,各项指标综合提升最大。

3 结束语

本文针对信号处理领域专业术语密集、逻辑严谨的特点,提出了基于嵌入模型微调与检索策略优化的 RAG 增强方法。通过构建高质量信号领域数据集,采用大模型辅助提取与多重质量控制机制,形成不同规模与质量的数据子集;在此基础上,对嵌入模型进行领域适配微调,显著提升了模型在信号领域的检索性能。进一步,通过系统实

验分析了嵌入模型选择、混合检索权重调配、重排序机制以及提示词优化等策略对系统性能的影响,最终确定了一组较优配置,有效提升了 RAG 系统在信号领域问答任务中的综合表现。实验表明,本文方法在忠实度、答案相关性和上下文精确率等指标上均取得显著提升。从而验证了本文所设计的“领域数据集构建—嵌入模型微调—检索生成策略协同优化”的领域 RAG 优化框架,对于其他专业领域 RAG 应用具有一定借鉴意义。

然而,本文工作仍存在一定局限。首先,所构建的数据集在规模与构建细节方面尚有扩展与完善空间;其次,嵌入模型微调对不同模型结构与参数规模的具体影响机制仍需深入探索和验证。未来将围绕以上方面展开持续研究,以进一步推动 RAG 技术在信号领域中的有效应用。

参考文献:

- [1] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.
- [2] 卢银鹏,郭凯威,卢记仓,等.基于上下文学习和语义检索增强的零样本立场检测方法[J].信息工程大学学报, 2025, 26(2): 196-202.
- [3] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.
- [4] GAO X F, XIONG Y, GAO X Y, et al. Retrieval-augmented generation for large language models: a survey [DB/OL]. (2023-12-18) [2025-06-10]. <https://arxiv.org/abs/2312.10997>.
- [5] NOGUEIRA R, CHO K. Passage re-ranking with BERT. [DB/OL]. (2019-01-13) [2025-06-10]. <https://arxiv.org/abs/1901.04085>.
- [6] ZHENG H X S, MISHRA S, CHEN X Y, et al. Take a step back: evoking reasoning via abstraction in large language models [DB/OL]. (2023-10-09) [2025-06-10]. <https://arxiv.org/abs/2310.06117>.
- [7] 张浩然,郝文宁,靳大尉,等.DF-RAG:基于查询重写和知识选择的检索增强生成方法[J].计算机科学, 2025, 52(11): 30-39.
- [8] 王海涛,师杨坤.基于大语言模型的查询扩展方法研究[J].计算机技术与发展, 2025, 35(3): 148-155.
- [9] SAWARKAR K, MANGAL A, SOLANKI S R. Blended RAG: improving RAG (retriever-augmented generation) accuracy with semantic search and hybrid query-based retrievers [C]//Proceedings of the 2024 IEEE 7th Interna-

- tional Conference on Multimedia Information Processing and Retrieval. Piscataway, USA: IEEE, 2024:155-161.
- [10] SHAO Z H, GONG Y Y, SHEN Y L, et al. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy [C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Stroudsburg, USA: ACL, 2023:9248-9274.
- [11] ASAI A, WU Z Q, WANG Y Z, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection [DB/OL]. (2023-10-18) [2025-06-10]. <https://arxiv.org/abs/2310.11511>.
- [12] DU J J, DING J B, LI J Y, et al. SC-Telco RAG: enhancing RAG through structured contexts in telecommunications [C]//Proceedings of the 2024 IEEE Globecom Workshops. Piscataway, USA: IEEE, 2024:1-6.
- [13] 王合庆, 魏杰, 景红雨, 等. Meta-RAG: 基于元数据驱动的电力领域检索增强生成框架 [J]. 计算机工程, 2026, 52(2):383-392.
- [14] BEI Y J, FANG Z B, MAO S Y, et al. Manufacturing domain QA with integrated term enhanced RAG [C]//Proceedings of the 2024 International Joint Conference on Neural Networks. Piscataway, USA: IEEE, 2024:1-8.
- [15] XIAO S T, LIU Z, ZHANG P, et al. C-Pack: packed resources for general Chinese embeddings [DB/OL]. (2023-09-14) [2025-06-10]. <https://arxiv.org/abs/2309.07597>.
- [16] CHEN J L, XIAO S T, ZHANG P T, et al. BGE M3-embedding: multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation [C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg, USA: ACL, 2024:2318-2335.
- [17] Github-Infiniflow. RAGFlow is an open-source RAG (Retrieval-Augmented Generation) engine based on deep document understanding [EB/OL]. (2025-02-28) [2025-06-10]. <https://github.com/infiniflow/ragflow>.
- [18] BAI J Z, BAI S, CHU Y F, et al. Qwen technical report [DB/OL]. (2023-09-29) [2025-06-10]. <https://arxiv.org/abs/2309.16609>.
- [19] Deepseek-AI, GUO D Y, YANG D J, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning [DB/OL]. (2025-01-22) [2025-06-10]. <https://arxiv.org/abs/2501.12948>.
- [20] ES S, JAMES J, ESPINOSA A L, et al. RAGAs: automated evaluation of retrieval augmented generation [C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. Stroudsburg, USA: ACL, 2024: 150-158.
- (编辑: 冯 春)

(上接第 144 页)

- [13] MOU X Q, CHEN X L, GUAN J, et al. Sea clutter suppression for radar PPI images based on SCS-GAN [J]. IEEE Geoscience and Remote Sensing Letters, 2021, 18(11):1886-1890.
- [14] PEI J F, YANG Y, FANG Z H, et al. Sea clutter suppression for marine surveillance radar based on generative adversarial learning [C]//Proceedings of the IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. Piscataway, USA: IEEE, 2023:5202-5205.
- [15] CONTE E, LONGO M. Characterisation of radar clutter as a spherically invariant random process [J]. IEE Proceedings F (Communications, Radar and Signal Processing), 1987, 134(2):No.191.
- [16] MARIER L J. Correlated K-distributed clutter generation for radar detection and track [J]. IEEE Transactions on Aerospace and Electronic Systems, 1995, 31(2):568-580.
- [17] 丁斌, 夏雪, 梁雪峰, 等. 基于深度生成对抗网络的海杂波数据增强方法 [J]. 电子与信息学报, 2021, 43(7):1985-1991.
- [18] 马鑫. 基于深度生成学习的雷达杂波模拟技术研究 [D]. 成都: 电子科技大学, 2022:1-93.
- [19] 高暄皓. 微弱机动目标雷达信号相参积累算法研究 [D]. 郑州: 信息工程大学, 2023:1-119.
- [20] 关键, 刘宁波, 王国庆, 等. 雷达对海探测试验与目标特性数据获取: 海上目标双极化多海况散射特性数据集 [J]. 雷达学报, 2023, 12(2):456-469.
- [21] BLACKNELL D, TOUGH R J A. Parameter estimation for the K-distribution based on $(z \log(z))$ [J]. IEE Proceedings-Radar, Sonar and Navigation, 2001, 148(6):309-312.
- (编辑: 冯 春)