基于改进混合检索的大语言模型税务问答研究

岳 帅[1] 王 业[2] 周泽冰[1]
YUE Shuai WANG Ye ZHOU Zebing

摘 要 针对税务问答任务中检索效果不足的问题，文章提出了一种基于改进混合检索的大语言模型问答方法。首先，构建覆盖多类别、多标签的高质量税务政策问答数据集，涵盖增值税、企业所得税、个人所得税、发票管理等；然后，设计融合稀疏检索（BM25）与密集检索（DPR）的改进型混合检索算法，通过引入轻量级神经网络动态预测稀疏与密集检索的最优权重比例，并结合交叉编码器重排序，有效提升相关文档的检索准确性；最后，基于 ChatGLM3-6B-base 大语言模型，采用 LORA 微调方法，对模型进行针对性优化，增强其对税务政策语义的理解能力。实验证明，所提出的数据集与改进检索方法能够显著提升税务问答系统的检索召回率和准确率，同时，所提出的框架为大语言模型在税务政策问答任务的应用提供了理论与实践支撑。

关键词 税务问答；稀疏检索与密集检索融合；混合检索算法；大语言模型；LORA 微调

doi：10.3969/j.issn.1672-9528.2026.02.008

0 引言

随着税务政策体系[1]的日益复杂和智能化服务需求的不断增长，建设高效、准确的税务问答系统[2]已成为当前的重要研究方向。然而，现有系统在税务领域仍存在检索效果有限、文档相关性不足等问题，难以为用户提供及时、精准的答案支持。

为此，本文聚焦税务问答场景，开展了税务问答数据集构建与混合检索算法改进研究。首先，针对税务政策信息来源分散、结构复杂等问题，自主构建涵盖多类别税务知识的高质量税务问答数据集。其次，设计融合稀疏检索（BM25）[3]与密集检索（DPR）[4]的改进型混合检索算法，并结合交叉编码器重排序机制[5]，来提升检索结果的召回率与相关性。此外，基于 ChatGLM3-6B-base 大语言模型[6]，针对税务政策问答任务进行了 LORA 微调[7]，以进一步提高模型对税务领域专业问题的语义理解与匹配能力。本文的主要贡献如下：

（1）构建了多类别、结构化的税务问答数据集，涵盖增值税、企业所得税、个人所得税、发票管理等，满足税务问答任务的数据需求。

（2）提出了一种基于 BM25 与 DPR 融合的混合检索算法，结合交叉编码器重排序，显著提升了检索结果的相关性与准确性。

（3）针对税务领域，基于 ChatGLM3-6B-base 大语言模型，进行了 LORA 微调，有效提升了模型在税务场景下的问题匹配与答案检索能力。

1 相关理论

1.1 混合检索算法

传统的稀疏检索方法最早由 Robertson 等[3]提出，其核心基于词频与倒排索引机制（如 BM25），能够高效检索与查询关键词高度匹配的文档，具有计算速度快、可解释性强的优势。然而，稀疏检索方法高度依赖词面匹配，缺乏对语义层面相关性的建模能力，容易出现语义检索失败或遗漏潜在相关文档的问题。

随着深度学习的发展，Karpukhin 等[4]提出了密集检索方法（DPR），通过双塔编码器结构将查询与文档分别映射到低维语义向量空间，利用向量相似度进行检索，有效捕捉文本的深层语义关联，从而弥补了稀疏检索方法的不足。然而，单一使用稀疏或密集检索方法在复杂问答任务中依然存在召回率不足、结果冗余噪声较多的问题。

针对上述问题，Mala 等[8]提出了混合检索策略，旨在融合 BM25 与 DPR 的检索结果，充分结合稀疏检索的关键词匹配优势与密集检索的语义理解能力，实现信息互补与召回率提升。然而，现有混合检索方法仍存在以下不足：

（1）部分研究对稀疏与密集检索结果的融合策略较为简单，无法充分发挥二者的协同优势；

（2）当前混合检索系统多数缺乏针对税务等垂直领域

1. 新疆农业大学计算机与信息工程学院 新疆乌鲁木齐 830052
2. 新疆农业大学网络与信息技术中心 新疆乌鲁木齐 830052

的场景适配与优化，导致领域召回效果不足；

（3）多数混合检索框架尚未充分引入交叉编码器重排序机制，检索结果的相关性与排序精度仍有较大提升空间。

1.2 大语言模型

自 2022 年 ChatGPT 发布以来 [9]，大语言模型（LLM）在各垂直领域的应用研究迅速发展，逐步成为智能化任务的重要基础支撑。目前，专业领域的大语言模型设计通常遵循以下开发路径：选取通用大模型作为基础，结合参数高效微调技术与任务定制化策略，针对特定场景进行适配与优化。

以司法领域的大语言模型 Lawyer-LLaMA 为例，Huang 等 [10] 基于法律判决文书、法律条文、司法解释、普法文章与国家司法考试数据，对 Chinese-LLaMA 模型进行持续的领域预训练，并引入基于 RoBERTa 的法律条文检索模块，有效缓解了大模型在法律应用中的幻觉问题，为垂直领域大模型设计提供了可借鉴路径。

受到司法领域大模型设计思路的启发，针对当前国内缺乏面向税务领域的大语言模型现状 [11]，本文自主构建了高质量的税务问答数据集，并基于 LORA 微调技术对通用大语言模型进行任务适配，旨在填补税务智能问答与要素识别领域的技术空白。

2 研究与方法

2.1 税务政策数据集的构建

2.1.1 税务政策数据采集与处理

构建面向税务问答任务的智能系统，需要高质量、结构化的税务知识数据。为实现精确的政策理解与要素识别，系统必须依赖覆盖多维度、多层级的税务政策数据，包括政策条款、适用范围、执行细则、历史修订版本等内容。税务数据的完整性与规范性直接影响到下游模型的训练质量与系统性能。

本研究所使用的税务政策数据来源于国家税务总局官网、地方税务局公开发布平台，共收集了约 22 147 份税务文档，涵盖增值税、企业所得税、个人所得税、发票管理等多个政策类别。原始字段包括政策标题、发布时间、政策正文、条款编号、适用主体、适用范围、税率描述、政策依据等信息。为保证模型训练与分析的准确性，本文对原始数据进行了如下预处理流程：

数据清洗：由于公开获取的数据存在格式不统一等问题，本文采用以下策略进行清洗处理：对文本字段（如适用范围、政策依据等）缺失内容，视重要性决定是否删除或用“暂无”填充；对结构重复、内容冗余的政策记录进行比对筛查，剔除重复项，确保数据唯一性与一致性。

异常值处理：基于经验规则和领域知识对关键字段（如税率、申报周期、适用主体类型）进行分布分析，对于出现明显偏离常规值的数据，进行人工审核并修正或剔除，避免影响训练结果。

数据标准化与格式转换：为方便后续模型输入与知识建模，本文采用以下策略进行数据标准化与格式转换：对结构化字段进行统一标准化，如将“适用于小微企业”“面向中小型企业”统一编码为同一类别；文本内容进行分句切分，便于用于问答模型中的句级输入处理；所有类别型字段采用 One-Hot 编码进行数值化；在高维类别数据（如“适用地区”）中，尝试采用嵌入表示（Embedding）进行压缩表示，以捕捉类别之间的语义关系；对数值型字段（如税率、金额阈值等）进行归一化处理，提升模型训练的稳定性。

通过上述预处理操作，本文构建的税务政策数据集在结构清晰性、标签规范性与语义完整性方面达到了较高标准，为后续混合检索算法构建、文档重排序，以及大语言模型的微调提供了高质量的数据支撑基础。

2.1.2 税务政策数据标签化处理

在税务政策数据预处理完成后，为进一步提升数据结构的标准化程度与模型处理的效率，本文对税务政策知识进行了系统性的标签化处理。标签化过程的目标是将原始文本中的关键信息和结构化字段转换为模型可直接处理的数值化标签或向量化表示，便于后续检索与大语言模型的输入设计。

首先，对于政策的顶层分类，分为增值税、企业所得税、个人所得税和发票管理。采用 One-Hot 编码将类别型数据转换为离散的数值特征，确保不同政策类别在模型输入中具备独立的向量表示，有助于提升检索与分类的准确性。其次，针对每类税务政策所涉及的核心要素（如适用范围、纳税人类型、税率标准、申报方式等），本文通过人工标注与规则抽取相结合的方法，建立了全面的要素标签体系。标签的处理方式包括：对少量标签的类别型数据采用 One-Hot 编码；对多标签样本，使用多标签二进制编码，以支持一对多的标签标注形式；对部分连续型要素（如税率、金额阈值），采用归一化处理，便于模型输入。然后，对于政策正文、条款内容、适用说明等非结构化文本，采用 TF-IDF 和 BERT 向量化等文本编码方法进行处理。

通过上述标签化处理，本文构建的税务政策数据集实现了结构化标签与文本向量的全面融合，不仅提升了检索效率，也为后续的大模型训练与个性化问答系统设计提供了高质量、标准化的数据基础。

2.1.3 税务政策数据集问答对生成

在完成税务政策数据采集、预处理及标签化处理后，本文进一步设计了系统化的问答对。问答的设计不仅要求涵盖政策内容全面，同时要求问题具有多样性和合理性，需要为后续检索与大语言模型微调提供有效支撑。

在问题模板设计上，针对税务政策的多类别标签特性，本文参考实际税务咨询场景与税务从业者提问习惯，设计了多种标准化问题模板，适配四大类税务政策：增值税、企业所得税、个人所得税、发票管理。问题示例如下：

（1）关于“税务政策类别”的“标签”，具体规定是什么？

（2）企业在“税务政策类别”的“标签”方面应该如何办理？

（3）纳税人在“税务政策类别”中，关于“标签”有哪些注意事项？

（4）“税务政策类别”涉及的“标签”适用范围是什么？

通过模板自动化替换税务类别与标签信息，生成个性化、多样化的问题内容，确保问句在结构上通用且在语义上针对性强。

在已标签化的税务政策原文中，本文采用基于标签锚点的答案抽取方法。具体步骤如下：

（1）标签定位：根据人工标注的标签位置，确定答案在政策原文中的起止位置。

（2）答案抽取：提取与标签严格对应的政策条文或描述，确保答案具备政策权威性与准确性。

（3）问答对匹配：将自动生成的问题与抽取得到的答案进行一一匹配，生成结构化问答对。

对于存在标签缺失或描述模糊的样本，本文设置了人工审核环节，确保生成问答对的准确性与可用性。

最终，本文共构建了约 12 000 组高质量税务政策问答对，涵盖增值税、企业所得税、个人所得税、发票管理四类主要税种及其下属多个标签，单个政策样本最多对应 5 个标签，最少对应 1 个标签，平均每个样本包含 2.6 个标签。具体信息如表 1 所示。

表 1 税务政策问答对统计信息

税务类别	问题内容	答案内容	标签名称
增值税	企业适用的增值税税率是多少？	一般纳税人适用 13%、9% 等税率，小规模纳税人适用3%税率。	增值税税率
企业所得税	企业所得税的申报周期是什么？	企业所得税实行按季度预缴，年度汇算清缴。	申报周期
个人所得税	个人所得税的专项附加扣除有哪些？	包括子女教育、继续教育、大病医疗、住房贷款利息、住房租金、赡养老人六项扣除。	专项附加扣除
发票管理	发票开具的时间和内容要求是什么？	销售商品、提供服务或者发生应税行为时，应当在规定时间内开具发票，且需按规定内容开具。	发票开具时间 发票开具内容

2.2 改进的混合检索算法

2.2.1 问题描述

针对本文构建的税务政策问答数据集，所有回答构成集合 $D = \{D_1, D_2, \cdots, D_n\}$，其中每个回答 D_i 对应一条完整的税务政策条款或权威政策解释。用户查询问题构成集合 $Q = \{q_1, q_2, \cdots, q_n\}$，其中每个问题 q_i 为用户输入的自然语言查询。本文将问题与回答之间的匹配关系形式化为最优检索任务，目标是在给定问题 q_i 的情况下，通过改进的混合检索算法，从候选回答集合 D 中检索出与之最相关的回答，并对检索结果进行合理排序，以最大化召回率与相关性准确率，提升整体检索效果。

2.2.2 改进的混合检索算法

针对传统混合检索算法中 BM25 与 DPR 权重比例静态设定、难以适配不同查询的问题 [12]，本文设计了一种基于轻量级神经网络的动态权重预测混合检索算法（DyHR），实现针对查询的自适应检索策略调整，从而提升税务问答系统的检索准确性。

对于输入查询 Q_i 和候选回答集合 $D = \{D_1, D_2, \cdots, D_n\}$，首先进行分词和预处理，得到问题序列和回答序列，用公式表示为：

$$Q_i = \{w_1, w_2, \cdots, w_m\}, \quad D_j = \{t_1, t_2, \cdots, t_n\} \tag{1}$$

式中：w_k 和 t_l 分别表示查询和文档中的第 k 个和第 l 个词。随后，基于 BM25 计算问题 Q_i 与回答 D_j 之间的稀疏检索得分，用公式表示为：

$$\begin{aligned}&\mathrm{Score}_{\mathrm{BM25}}(Q_i, D_j)\\ &= \sum_{w \in Q_i} \mathrm{IDF}(w) \cdot \frac{f(w, D_j) \cdot (k_1 + 1)}{f(w, D_j) + k_1 \cdot \left(1 - b + b \cdot \dfrac{|D_j|}{\mathrm{avgdl}}\right)}\end{aligned} \tag{2}$$

式中：$f(w, D_j)$ 表示词 w 在回答 D_j 中的出现次数；$|D_j|$ 表示回答 D_j 的长度；avgdl 表示回答集合的平均长度；k_1、b 表示 BM25 超参数；IDF(w) 表示词 w 的逆文档频率。

针对 DPR 密集检索过程，本文采用带有正负样本对的数据构造方式。对于问题 $Q_i \in Q$，其正样本为数据集中最佳匹配的回答 D_i^+，而负样本则通过 BM25 获取，选择得分最高但非最佳匹配的回答 D_i^- 作为困难负样本，以增强模型对相似干扰项的区分能力。

密集检索训练数据采用分批次训练，每一批次样本的 B_i 都由问题、正样本和困难负样本组成，用公式表示为：

$$B_i = \{Q_{B_i}, P_{B_i}, N_{B_i}\} \tag{3}$$

式中：Q_{B_i}表示第 i 个批次的问题；P_{B_i}表示第 i 个批次的正样本。

文本表征过程采用 BERT 网络进行编码，输入问题与回

答经过分词、填充与 ID 映射后，依次通过嵌入层与编码层，最终提取 CLS 向量作为问题与回答的语义表征，具体过程用公式表示为：

$$h_Q = \mathrm{BERT}(Q_i)_{|\mathrm{CLS}|}, h_D = \mathrm{BERT}(D_i)_{|\mathrm{CLS}|} \quad (4)$$

查询与回答的语义相似度通过余弦相似度计算，用公式表示为：

$$\mathrm{Score}_{\mathrm{DPR}}(Q_i, D_j) = \frac{h_Q \cdot h_D}{\|h_Q\| \|h_D\|} \quad (5)$$

BERT 参数的优化采用 InfoNCE 损失函数来进行优化，损失函数用公式表示为：

$$\mathcal{L} = -\log \frac{\exp(\mathrm{Score}_{\mathrm{DPR}}(Q_i, P_{Bi})/\tau)}{\sum_{k=1}^{K} \exp(\mathrm{Score}_{\mathrm{DPR}}(Q_i, D_k)/\tau)} \quad (6)$$

式中：P_{Bi}为问题 Q_i 的正样本回答；τ 为温度系数；K 为当前批次样本数。

为解决 BM25 与 DPR 权重静态设定的问题，本文设计了轻量级神经网络动态权重预测模块。对于每一个输入查询，提取查询特征向量 $\boldsymbol{F}_Q$，具体包括：查询长度 L_Q、BM25 得分统计特征（Top-5 均值、方差）、DPR 得分统计特征（Top-5 均值、方差）、税务关键词命中数量 K_{Tax}。查询特征向量用公式表示为：

$$\boldsymbol{F}_{Q_i} = \left[L_{Q_i}, \mu_{\mathrm{Top5}}^{\mathrm{BM25}}, \sigma_{\mathrm{Top5}}^{\mathrm{BM25}}, \mu_{\mathrm{Top5}}^{\mathrm{DPR}}, \mu_{\mathrm{Top5}}^{\mathrm{DPR}}, K_{\mathrm{Tax}}\right] \quad (7)$$

轻量级神经网络结构用公式表示为：

$$\alpha_i = \sigma\left(W_2 \cdot \mathrm{ReLU}\left(W_1 \cdot \boldsymbol{F}_{Q_i} + b_1\right) + b_2\right) \quad (8)$$

式中：σ 表示 Sigmoid 激活函数，W_1、W_2、b_1、b_2 表示神经网络参数。最终，BM25 与 DPR 的混合检索得分计算用公式表示为：

$$\begin{aligned} &\mathrm{Score}_{\mathrm{Hybrid}}(Q_i, D_j) \\ &= \alpha_i \cdot \mathrm{Score}_{\mathrm{BM25}}(Q_i, D_j) + (1-\alpha_i) \cdot \mathrm{Score}_{\mathrm{DPR}}(Q_i, D_j) \end{aligned} \quad (9)$$

动态权重预测模块根据查询复杂度与检索统计特征自适应调整稀疏与稠密检索的权重分配，从而提升检索结果的匹配准确性。

2.2.3 改进的交叉编码器重排序

2.2.3.1 轻量级交叉编码器设计

针对传统交叉编码器计算开销大、难以满足实时检索需求的问题，本文设计了一种基于轻量级蒸馏与分阶段重排序的改进交叉编码器训练与推理框架（LCE-Ranker），旨在提升税务问答场景下的检索效率与排序精度。具体做法为采用知识蒸馏技术，以完整交叉编码器作为教师模型，训练轻量级学生模，从而显著降低推理计算成本。

对于输入问题 Q_i 与候选回答 D_j 输入格式为：

$$\mathrm{Input} = [\mathrm{CLS}]Q_i[\mathrm{SEP}]D_j[\mathrm{SEP}] \quad (10)$$

学生模型输出相关性得分：

$$\mathrm{Score}_{\mathrm{LCE}}(Q_i, D_j) = \mathrm{TinyBERT}_{\mathrm{CLS}}(Q_i, D_j) \quad (11)$$

蒸馏损失函数采用软标签与硬标签联合优化，其公式为：

$$\mathcal{L}_{\mathrm{distill}} = (1-\lambda) \cdot \mathcal{L}_{\mathrm{hard}} + \lambda \cdot \mathcal{L}_{\mathrm{soft}} \quad (12)$$

式中：$\mathcal{L}_{\mathrm{hard}}$表示真实标签的排序损失；$\mathcal{L}_{\mathrm{soft}}$表示为学生模型输出与教师模型输出之间的 KL 散度；λ 表示平衡系数。

2.2.3.2 训练样本构建与优化流程

对于一个问题 Q_i，有四个阶段：

（1）初步检索阶段：对于一个问题 Q_i，基于动态权重预测的混合检索算法，获取 Top-N 候选文档集合：$D_{Q_i} = \{D_1, D_2, \cdots, D_n\}$。若 Top-$N$ 集合中未包含问题的最佳匹配回答 D_i^+，则将集合中一条文档替换为 D_i^+，确保训练样本包含正确答案。

（2）样本标志阶段：为每个候选文档标注标签，最佳匹配文档标签为 1，其他文档标签为 0。

（3）输入序列构造阶段：将问题与文档拼接形成输入序列，输入轻量级交叉编码器进行训练，输入序列格式为：

$$x = [\mathrm{CLS}]q_{i1}, \cdots, q_{ik}[\mathrm{SEP}]d_{j1}, \cdots, d_{jl}[\mathrm{PAD}] \quad (13)$$

式中：[CLS] 表示起始标志；[SEP] 表示问题与回答的分隔符；[PAD] 表示填充符号。

（4）轻量级交叉编码器训练阶段：输入序列通过 TinyBERT，通过嵌入层和编码层，得到各时间步的隐藏状态：

$$h_x = \mathrm{TinyBERT}(x), \quad h_x \in \mathbb{R}^{L_x \times d} \quad (14)$$

式中：L_x 表示输出序列长度；d 表示隐藏状态维度。选取 [CLS] 位置的隐藏状态，通过 tanh 激活函数与线性层映射为池化特征，其公式为：

$$p = \tanh\left(h_x(0)\right)W_1 + b_1, W_1 \in \mathbb{R}^{d \times 2}, b_1 \in \mathbb{R}^d \quad (15)$$

最后，通过线性层输出分类结果，其公式为：

$$\hat{y} = \mathrm{softmax}(pW_2 + b_2), W_2 \in \mathbb{R}^{d \times 2}, b_2 \in \mathbb{R}^d \quad (16)$$

分类损失函数采用交叉熵损失，其公式为：

$$L_{\mathrm{ce}} = -\frac{1}{N}\sum_{i=1}^{N} y \log \hat{y} \quad (17)$$

此外为提升轻量级交叉编码器的判别能力，训练阶段引入知识蒸馏损失，联合优化真实标签与教师模型输出的软标签，其公式为：

$$L_{\mathrm{distil}} = (1-\boldsymbol{\lambda}) \cdot \mathrm{KL}(\hat{y}_{\mathrm{student}} \,\|\, \hat{y}_{\mathrm{teacher}}) \quad (18)$$

式中：$\boldsymbol{\lambda}$ 表示蒸馏损失权重。

改进后的交叉编码器（LCE-Ranker）具有以下优势：

（1）轻量级交叉编码器显著降低推理计算复杂度，适配在线税务问答场景。

（2）知识蒸馏技术有效保留教师模型排序能力，提升轻量模型判别效果。

（3）分阶段重排序结构兼顾召回效率与排序精度，与前端动态混合检索算法高度契合。

2.3 LLM 微调策略

针对税务政策问答检索任务，本文选用由智谱清言与清华大学 KEG 实验室于 2023 年发布的中文大语言模型 ChatGLM3-6B-base 作为基础模型。在构建的税务政策问答数据集上，本文设计并比较了多种微调策略，包括全参数微调、低秩适配等，以评估不同微调方法在税务场景下的适应能力。和效果表现。

首先，对于输入查询与候选回答，首先将二者拼接为单一序列，输入格式如下：“[CLS] 查询内容 [SEP] 候选回答内容 [SEP]”，拼接后的输入序列 $\boldsymbol{X}$ 经过嵌入层（Embedding），映射为连续稠密的向量，其计算公式为：

$$\boldsymbol{X}_{\text{emb}} = \text{Embedding}(\boldsymbol{X};\theta) \tag{19}$$

式中：θ 表示模型参数；$\boldsymbol{X}_{\text{emb}}$ 表示输入的词向量。

随后，输入依次通过堆叠的 28 层 GLMBlock 进行特征提取。每一层 GLMBlock 的计算过程包括 LayerNorm、前馈神经网络（FFN）和多头自注意力机制，计算公式分别为：

$$\boldsymbol{X}_{\text{LN}} = \text{LN}(\boldsymbol{X}_{\text{emb}},\theta_i) \tag{20}$$

$$\theta_i = \boldsymbol{X}_{\text{LN}} + f^{\text{Attention}}(\boldsymbol{X}_{\text{LN}},\theta_i) \tag{21}$$

$$\text{GLMBlock}_i(\boldsymbol{X}_{\text{emb}}) = \text{LN}\left(\boldsymbol{X}_{\theta_i}\right) + f^{\text{FFN}}\left(\text{LN}\left(\boldsymbol{X}_{\theta_i}\right);\theta_i\right) \tag{22}$$

式中：θ_i 表示第 i 个 GLMBlock 的参数；$f^{\text{Attention}}(\cdot)$ 表示自注意力计算过程；$f^{\text{FFN}}(\cdot)$ 表示前馈网络。

经过 28 层堆叠后，得到最终的深层语义特征，具体公式为：

$$y_\theta = \text{LN}(\text{GLMBlock}(\boldsymbol{X}_{\text{emb}} \times 28);\theta) \tag{23}$$

$$y = \text{Linear}(y_\theta;\theta) \tag{24}$$

模型输出 y 作为当前输入查询与候选回答的匹配概率预测。

在微调过程中，模型训练目标是优化查询与候选回答的相关性预测能力。对于训练数据，标注 1 表示候选回答为当前查询的正确匹配；标注 0 表示候选回答与查询无关。

$$\mathcal{L} = -\sum \ln P(y_{\text{true}}|y) \tag{25}$$

式中：y_{true} 为真实标签（0 或 1）；y 为模型预测概率。

2.4 检索增强推理（RAG）

在完成 LORA 微调后，本文设计基于检索增强的推理流程 [13]，以充分发挥改进混合检索算法与微调大语言模型的协同优势，提升税务问答系统的检索准确性与推理能力。整个推理流程如图 1 所示。

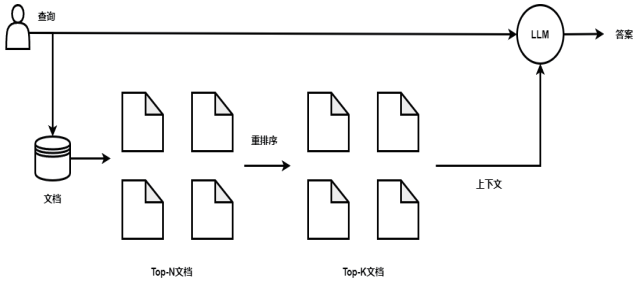

图 1 RAG 推理流程

当用户输入自然语言问题后，首先通过改进的混合检索算法，从税务政策回答库中检索 Top-N 候选回答集合。改进后的检索过程为：

（1）基于 BM25 与 DPR 的动态权重融合检索，获取候选文档初步得分；

（2）采用轻量级交叉编码器进行粗排，过滤低相关性文档；

（3）利用完整交叉编码器对 Top-N 文档进行精排，获得最终候选排序结果。

对于检索得到的 Top-K 候选回答，逐一与用户查询进行拼接，输入至微调后的大语言模型，预测查询与回答的相关性得分。模型输入格式为“[CLS] 查询内容 [SEP] 候选回答内容 [SEP]”。模型输出相关性预测概率，依据预测得分进行最终排序，选择相关性最高的回答作为当前问题的系统输出。

改进后的检索增强推理流程具有以下优势：

（1）支持动态检索策略自适应调整，适配多样化查询需求；

（2）融合轻量级重排序与大语言模型深度推理，兼顾效率与准确性；

（3）不依赖生成式回答，可输出可控且符合政策规范的标准答案；

（4）LORA 微调后的大模型更能兼顾性能与效率。

3 实验与分析

3.1 数据集

本文基于自建的税务政策问答数据集开展实验，数据集涵盖增值税、企业所得税、个人所得税、发票管理四类核心税务政策。问题内容来源于实际税务政策条文、政策解读以及典型咨询问题，答案内容基于权威政策文件进行人工抽取与审核，确保问答的准确性、专业性与规范性。本数据集共构建约 12 000 组高质量税务政策问答对，按照 8∶2 的比例划分为训练集与测试集。具体分布情况如表 2 所示。

表 2 税务政策问答数据集

税务类别	样本总量	训练集	测试集	占比 /%
增值税	4 320	3 456	864	36.0
企业所得税	2 880	2 304	576	24.0
个人所得税	3 120	2 496	624	26.0
发票管理	1 680	1 344	336	14.0

为满足大语言模型的输入限制，本文对问题与答案的文本长度进行了统计与控制。统计结果显示：问题内容长度分布在 10~50 个 tokens，平均长度约为 28 tokens；答案内容长度分布在 30~300 个 tokens，平均长度约为 120 tokens。

3.2 实验环境和参数设置

实验在 Ubuntu20.04LTS 操作系统环境下进行，硬件配置包括 IntelXeonGold6226R @2.90 GHz 处理器、NVIDIAA10080GBGPU、CUDA11.8 显卡驱动、256 GB 内存。编程语言采用 Python3.9，深度学习框架使用 PyTorch2.0.1。

在模型训练参数设置方面，本文采用 AdamW 优化器，初始学习率设为 2e-5，权重衰减为 0.01，批次大小设置为 8，训练轮数为 5，采用线性学习率衰减并设置 500 步学习率预热。随机种子设置为 42，以确保实验结果的可复现性。针对密集检索模型（DPR）与轻量级交叉编码器的训练，采用 InfoNCE 损失与交叉熵损失进行优化，蒸馏过程软标签权重 λ 设置为 0.5。上述参数配置兼顾训练效率与模型效果，经过多轮实验调整后最终确定。

3.3 评估指标

为了全面评估本文提出的改进混合检索算法、交叉编码器重排序效果以及微调后大语言模型在税务场景下的任务表现，本文选取了检索任务与分类任务中常用且具有代表性的评估指标。具体定义如下：

MRR 用于衡量检索系统返回结果的排序质量，主要关注正确答案在候选列表中的排名情况。MRR 的计算公式为：

$$\mathrm{MRR@}K = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\mathrm{rank}_i} \tag{26}$$

式中：N 表示测试样本总数；rank_i 表示第 i 个样本中正确答案在 Top-K 候选列表中的排名位置。Recall@K 衡量的是检索系统在返回的前 K 个候选结果中，是否包含正确答案的比例。用公式表示为：

$$\mathrm{Recall@}K = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\mathrm{rank}_i \le K\right) \tag{27}$$

式中：$\mathbb{I}(\cdot)$表示指示函数，若第 i 个样本的正确答案排在前 K 位，取值为 1，否则为 0。本文重点考察 Recall@1、Recall@5、Recall@10、Recall@15 和 Recall@20，以全面评估检索算法在不同返回规模下的召回能力。

此外，为了全面评估微调后大语言模型在税务政策问答任务中的生成质量，本文选取了自动评价体系中应用广泛的指标，ROUGE-1、ROUGE-2、ROUGE-L 和 BLEU-4 进行评估。

3.4 核心实验与对比

3.4.1 检索重排序算法实验结果展示

本节对不同检索算法及重排序方法在自建税务政策问答数据集上的性能表现进行了系统评估，实验结果如表 3 所示。本文对比了传统稀疏检索算法 TF-IDF 和 BM25，基于 BERT 的稠密检索模型 DPR，以及本文提出的动态权重混合检索算法 Hybrid15 和结合改进重排序模块的 Hybrid15+ReRank。

表 3 各种检索重排序算法对比实验结果

算法	MRR@10	Recall@1	Recall@5	Recall@10	Recall@15	Recall@20
TF-IDF	50.42	41.87	61.20	71.58	76.01	82.36
BM25	68.25	59.14	78.87	86.43	89.15	91.07
DPR	65.73	55.62	80.25	86.79	89.32	91.02
BM25+ReRank	78.91	71.24	88.96	93.12	95.03	96.15
DyHR+LCE-Ranker	**82.67**	**74.85**	**91.54**	**95.07**	**96.38**	**97.42**

实验结果表明，传统的稀疏检索算法（TF-IDF、BM25）在 Top-K 召回指标上存在一定局限，难以有效捕捉问题与答案之间的语义关联。DPR 通过语义编码提升了部分指标，但整体性能仍低于 BM25，说明在本数据集下稠密检索的泛化能力存在不足。结合交叉编码器重排序的 BM25+ReRank 方案在各指标上取得明显提升，验证了排序优化的重要性。本文提出的改进方案 DyHR+LCE-Ranker 在 MRR@10、Recall@1、Recall@5 等核心指标上均取得最佳性能，显著优于其他对比算法。

3.4.2 检索重排序算法实验结果分析

通过与传统的 TF-IDF、BM25 和 DPR 检索算法对比，可以得出以下结论：传统稀疏检索方法在税务问答场景下检索效果有限，难以有效捕捉复杂的语义匹配关系；DPR 在 Top-K 召回率上相较 BM25 略有提升，但在整体 MRR 和 Top-1 指标上仍存在不足，说明稠密检索在本数据集上泛化能力有限。引入重排序后，BM25+ReRank 显著提升了各项指标，验证了重排序的重要性。本文提出的 DyHR+LCE-Ranker 在所有指标上均取得最优表现，充分证明了动态检索与轻量级重排序的有效性。

模型参数敏感性：实验中发现动态权重预测模块对输入特征的变化较为敏感，合理设计查询特征输入（如 BM25 与 DPR 得分统计特征）对检索效果有重要影响。此外，轻量级交叉编码器中的蒸馏权重 λ 调节对最终排序效果具有一定影响，过高或过低都会导致模型性能下降。

总之，DyHR+LCE-Ranker 在动态权重自适应检索和轻量级交叉编码器重排序方面的创新，使其能够针对不同查询灵活调整检索策略并高效优化候选排序，提高了税务问答系统的检索准确率和实时响应能力，为税务智能问答系统的高效检索与应用落地提供了可行的技术方案。

3.4.3 微调大模型实验结果展示及分析

本节对不同微调方法在 ChatGLM-6B-Base 大模型上，使用自建税务政策问答数据集上的性能表现进行了系统评估，实验结果如表 4 所示。

表 4 针对 ChatGLM-6B-Base 大模型各种微调方法系统评估

微调方法	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
—	30.12	18.45	21.05	12.34
P-Tuning V2	45.87	28.76	43.20	29.50
LoRA	70.95	60.40	68.73	58.20
全参数微调	77.30	66.85	74.90	63.10

表 4 展示了不同微调策略在自建税务政策问答数据集上的生成性能对比。实验结果表明，未经微调的 ChatGLM-6B-Base 基线模型在各项指标上表现较低，说明原始模型对税务领域的语义理解和生成能力有限。采用轻量级提示调优方法 P-TuningV2 后，模型性能有了明显提升，尤其是在 ROUGE-1 和 ROUGE-L 指标上表现较好，体现了提示学习在特定任务适应中的优势。

相比之下，基于低秩适配（LoRA）的微调策略进一步显著提升了各项指标，尤其是 ROUGE-2 和 BLEU-4，展现出更强的细粒度语义捕获和文本生成能力。全参数微调作为最全面的调整方式，在所有评测指标上均取得最佳效果，体现了完整模型参数更新带来的最大性能提升。

综上所述，不同微调策略在性能和资源效率上存在权衡，P-TuningV2 适合快速轻量调优，LoRA 兼顾性能与效率，全参数微调则适合追求最高生成质量的场景。本文的实验结果为税务问答领域的大语言模型微调方法选择提供了有价值的参考。

3.4.4 检索－问答实验及分析

检索—问答实验旨在将检索算法与微调后的大语言模型（LLM）有效结合，应用于税务政策问答场景。实验首先通过检索算法从税务知识库中提取 Top-*K* 相关文档，并将其与用户问题一同输入 LLM 生成答案。该实验主要包括两个方面：一是研究不同上下文文档数量（如 Top-1、Top-2、Top-3、Top-5、Top-10）对问答效果的影响，二是对比不同检索算法在支持 LLM 生成答案时的实际表现，全面验证本文检索—生成联动框架在税务问答任务中的有效性。

3.4.4.1 上下文文档数量统计分析

本部分旨在探究不同 Top-*K* 文档作为问题上下文对 LLM 问答效果的影响。实验选用经过 LORA 微调的 ChatGLM-6B-Base 作为问答大模型，检索阶段采用 DyHR 与 LCE-Ranker 算法。实验结果如表 5 所示。从结果可以看出，使用检索得到的多个文档作为问题上下文输入 LLM 时，整体问答效果均低于仅使用最佳匹配文档（Top-1）的场景。随着文档数量的增加，各项指标呈现下降趋势，说明检索过程中存在误差积累，这些误差会传递到生成阶段，最终影响整体性能。特别是当上下文文档数从 Top-1 增加到 Top-10 时，尽管检索的召回率有所提高，但问答性能并未得到改善，反而因无关文档的干扰，导致模型关注点分散，生成质量显著下降。实验结果证明：增加上下文文档数并非越多越好，精准检索仍然是提升检索－问答一体化效果的关键。

表 5 上下文文档数量统计分析

上下文文档数	ROUGE-1	ROUGE-2	ROUGE-L	BLEU-4
Top-1	68.20	56.70	66.50	53.40
Top-2	67.90	56.50	66.20	53.00
Top-3	67.10	55.80	65.50	52.00
Top-5	63.00	51.70	61.20	48.10
Top-10	32.50	21.00	30.50	19.50

3.4.4.2 不同检索算法比较分析

本部分探究了不同检索算法对微调后的 LLM 问答性能的影响，只选取检索到的 Top-1 文档作为问题上下文，实验结果如表 6 所示。

表 6 不同检索算法对微调后 LLM 问答性能比较分析

检索算法	ROUGE-1	ROUGE-2	ROUGE-L	BELU-4
TF-IDF	52.75	42.30	51.00	39.50
BM25	55.90	45.80	54.20	42.10
DPR	64.00	54.10	62.70	50.12
BM25+ReRank	66.80	56.30	64.80	52.20
DyHR+LCE-Ranker	78.10	67.41	73.25	65.40

实验结果表明，传统的稀疏检索算法（TF-IDF、BM25）在提供上下文时对问答性能提升有限，DPR 和 BM25+ReRank 能进一步改善生成质量，而本文提出的 DyHR+LCE-Ranker 方

案在各项指标上均取得最佳表现，验证了动态检索与轻量级重排序策略在提升 LLM 问答准确性中的有效性。

3.4.5 消融实验

为了验证模型各个模块的有效性，本节进行了详细的消融实验，通过逐步引入模型的核心创新组件，分析各部分对模型性能的贡献。具体实验设置如下：

（1）Base：BM25+ 重排序 + 未微调的 LLM：采用传统稀疏检索加经典重排序，结合未微调的通用大语言模型。

（2）Base+DyHR：通过引入动态权重混合检索，验证检索层面改进对问答效果的提升。

（3）Base+DyHR+LCE-Ranker：在上一基础上，将重排序模块升级为轻量级交叉编码器。

（4）完整（DyHR+LCE-Ranker+ 微调 LLM）：包含所有组件。

通过逐步引入模型关键组件，本文对各模块在检索－问答一体化流程中的有效性进行了消融验证。实验结果如表 7 所示。结果表明基础方案（Base：BM25+ReRank+ 未微调 LLM）在各项指标上表现相对较低，说明传统检索与未微调大语言模型的组合在复杂税务问答任务中的适应性有限。

表 7 消融实验表

模型	ROUGE-1	ROUGE-2	ROUGE-L	BELU-4
Base	66.80	56.30	64.80	52.20
Base + DyHR	70.25	59.85	68.40	55.60
Base + DyHR + LCE-Ranker	73.50	63.10	71.25	59.80
DyHR + LCE-Ranker + 微调 LLM	78.10	67.41	73.25	65.40

引入动态权重混合检索（DyHR）后，通过融合多种检索策略，动态调整不同检索结果的权重，有效提升了候选文档的相关性，解决了传统检索对语义匹配不足的问题。DyHR 模块显著增强了检索阶段对多样化问题的适应能力，提升了检索结果的准确性。

进一步加入轻量级交叉编码器（LCE-Ranker）后，通过对检索结果进行高效的细粒度语义重排序，增强了文档排序的准确性，有效缓解了检索阶段引入噪声文档对问答质量的干扰。LCE-Ranker 模块有效降低了无关文档对生成质量的干扰，提升了排序精度。

最终，结合 LORA 微调的 LLM，完整方案在所有指标上达到最优，表明模型微调能够有效增强领域适应性，提升税务问答场景下的语言理解与生成能力。微调的 LLM 进一步提升了模型对税务领域语义的深度建模与上下文理解能力，显著改善了回答的准确性与专业性。

实验数据显示，完整的模型在税务政策问答数据集上，相较于基础的 Base 模型，ROUGE-1、ROUGE-2、ROUGE-L 和 BLEU-4 分别提升了 11.3%、9.7%、8.4%、13.2%，每个模块的引入都为解决特定问题提供了有效手段，验证了本研究设计的合理性与实用价值。

4 结论

针对税务政策问答场景下检索与生成一体化效果不足的问题，本文提出了一种基于动态检索与轻量级重排序的大语言模型问答框架。首先，在传统检索基础上改进了动态权重混合检索（DyHR）模块，强化了多检索策略的动态融合能力，解决了传统稀疏检索对复杂语义匹配能力不足的问题；其次，优化了轻量级交叉编码器（LCE-Ranker）模块，缓解了检索阶段引入噪声文档对生成质量的干扰，进一步提升了检索结果的排序准确性；最后，引入了 LORA 微调的大语言模型，显著提升了模型在税务领域的语言理解和生成能力。通过与多个基线模型的性能比较以及消融实验验证，结果表明所提出的检索－问答一体化模型在自建税务政策问答数据集上的 ROUGE-1、ROUGE-2、ROUGE-L 和 BLEU-4 等指标均取得显著提升，有效验证了所提模型结构与各关键模块设计的合理性和有效性。未来的研究将进一步探索多轮对话中的检索与生成协同优化机制，提升模型在复杂税务业务场景中的泛化能力与交互体验。

参考文献：

[1] 杨磊 . 全面构建绿色税收体系的探讨 [J]. 税务研究 , 2018(6): 122-123.

[2] 陈义 , 胡志宇 , 曾玮 , 等 . 税务业务咨询问答系统 [J]. 计算机应用与软件 , 2007(2):112-115.

[3]ROBERTSON S E, ZARAGOZA H.The probabilistic relevance framework: BM25 and beyond[J].Foundations & trends in information retrieval,2009,3(4):333-389.

[4]KARPUKHIN V, OGUZ B, MIN S, et al.Dense passage retrieval for open-domain question answering[EB/OL]. (2020-09-30)[2026-01-03].https://doi.org/10.48550/arXiv.2004.04906.

[5]MACAVANEY S, YATES A, COHAN A,et al.CEDR: contextualized embeddings for document ranking[EB/OL]. (2019-08-19)[2025-12-30].https://doi.org/10.48550/arXiv.1904.07094.

[6]DU Z X, QIAN Y J, LIU X,et al.GLM:general language model pretraining with autoregressive blank infilling[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.Brussels:ACL,2022:320–335.

（下转第 45 页）

力，使分析方法在不同情感状态下均能保持高准确率，充分证明了其在在线学习情感分析任务中的鲁棒性和适应性。

3 结语

本研究提出了一种基于改进 TextCNN 与自注意力机制的在线学习情感分析方法，通过以下创新点明显提升了分析性能：（1）多模态特征融合：结合文本（课程术语、情感强度词）、面部表情（FACS 编码）及姿态动作，通过多源嵌入融合与动态注意力机制增强了特征表示能力。（2）模型结构优化：在 MFN 基础上引入自注意力重排、残差连接与层归一化，有效捕捉词语间重要关系，减少局部语言忽略问题。（3）实验验证：在 FACS 平台与公开数据集上的实验表明，模型 ACC 稳定高于 90%，混淆矩阵显示误分类率显著降低，且情感轨迹与真实状态高度一致，满足多模态融合的鲁棒性要求。

该方法为在线学习情感分析提供了高精度、低偏差的解决方案，可支持学习状态预测与教学策略优化。未来可进一步探索跨模态对齐与实时计算性能优化，以适配更复杂的在线教育场景。

参考文献：

[1] 李广宏 , 杨文静 . 沉浸式主题街区旅游形象感知及质量提升研究 : 以长安十二时辰街区为例 [J]. 北方经贸 , 2025(6): 147-154.

[2] 程刚 , 吴亚熹 , 杨岱松 , 等 . 基于 LinearSVC 和 LDA 模型的电商产品评论情感分析 [J]. 湖北理工学院学报 , 2025, 41(3): 52-57.

[3] 叶佳乐 , 普园媛 , 赵征鹏 , 等 . 混合对比学习和多视角 CLIP 的多模态图文情感分析 [J]. 计算机科学 , 2025, 52(S1): 236-242.

[4] 徐凯 , 池明得 , 李建州 , 等 . 融合 BERT 编码层的对比学习方面级情感分析 [J]. 信息技术与信息化 ,2025(5):163-167.

[5] 张亚婷 , 王鑫宇 , 郭佳祺 , 等 . 基于决策级融合的多模态图书评论情感分析 [J]. 图书情报导刊 ,2025,10(5):72-79.

[6] 王治莹 , 马若涵 , 刘翰界 . 考虑决策者不同情绪状态的洪水灾害链多阶段应急决策方法 [J]. 中国安全生产科学技术 , 2025, 21(5):209-217.

[7] 张楠 , 彭海龙 , 徐晨怡 . 国际社交媒体上中国环境议题的传播图景 : 基于推特平台的话题与情感分析 [J]. 北京科技大学学报 (社会科学版), 2025, 41(4):98-108.

【作者简介】

曾光辉（1972—），男，湖南长沙人，本科，副教授，研究方向：智能信息处理。

杨光耀（1992—），男，河南汝阳人，硕士，研究方向：人工智能算法、计算机视觉、深度学习应用等。

（收稿日期：2025-07-11 修回日期：2026-01-27）

（上接第 41 页）

[7]HU E J, SHEN Y L, WALLIS P,et al.LoRA:low-rank adaptation of large language models[EB/OL].2022[2025-12-30]. https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/.

[8]MALA C S, GEZICI G, GIANNOTTI F. Hybrid Retrieval for hallucination mitigation in large language models: a comparative analysis[EB/OL].(2025-02-28)[2025-12-30].https://doi.org/10.48550/arXiv.2504.05324.

[9]BROWN T B,MANN B ,RYDER N,et al.Language models are few-shot learners[EB/OL].(2020-07-22)[2025-12-31].https://doi.org/10.48550/arXiv.2005.14165.

[10]HUANG Q Z, TAO M X, ZHANG C, et al.Lawyer LLaMA technical report[EB/OL].(2023-10-14)[2025-12-30].https://doi.org/10.48550/arXiv.2305.15062.

[11] 任海玉 , 刘建平 , 王健 , 等 . 基于大语言模型的智能问答系统研究综述 [J]. 计算机工程与应用 ,2025,61(7):1-24.

[12]ASKARI A, ABOLGHASEMI A, PASI G, et al.Injecting the BM25 score as text improves BERT-based Re-rankers[EB/OL].(2023-01-23)[2025-12-30].https://doi.org/10.48550/arXiv.2301.09728.

[13]LEWIS P, PEREZ E, PIKTUS A, et al.Retrieval-augmented generation for knowledge-intensive NLP tasks[EB/OL].(2021-04-12)[2025-12-23].https://doi.org/10.48550/arXiv.2005.11401.

【作者简介】

岳帅（1997—），男，河南台前人，硕士研究生，研究方向：自然语言处理、大语言模型。

王业（1980—），男，新疆乌鲁木齐人，硕士，副教授，研究方向：自然语言处理、农业信息化、高校信息化。

周泽冰（2000—），男，河南驻马店人，硕士研究生，研究方向：推荐系统、自然语言处理、大语言模型。

（收稿日期：2025-07-01 修回日期：2026-01-15）