基于LLM的税务知识智能问答系统研究

黎栋梁[1*] 文颖[1] 伍照生[2] 郑显斌[2]

1.华东理工大学 信息科学与工程学院，中国·上海 200237

2.大象慧云信息技术有限公司，中国·贵州 贵阳 550081

摘要：针对税务政策更新快、专业性强、人工咨询效率偏低等问题，提出一种基于大语言模型的税务知识智能问答系统设计方案。系统采用“知识库检索增强＋工具调用执行＋多智能体协同”的总体思路，在税务政策文档、法规条款、办税指引和常见问题数据基础上构建税务知识库，通过分段切分、向量化编码和重排序机制完成政策知识获取；在回答生成阶段，引入工具调用完成法规检索、条件校验和规则计算，在复杂任务场景下进一步采用多智能体协同完成任务拆解、政策比对、结果审校与答复生成。研究给出了系统总体架构、检索增强方法、工具执行机制与多智能体流程设计，并结合典型税务业务对系统运行过程进行了分析。结果表明，该方案能够提升税务咨询的响应效率、结果依据性和业务处理稳定性，可为企业用户和财税人员提供可扩展的智能税务服务支持。

关键词：大语言模型；税务知识库；检索增强生成；工具调用；多智能体

Research on Intelligent Tax QA System Based on LLM

Li Dongliang[1*], Wen Ying[1], Wu Zhaosheng[2], Zheng Xianbin[2]

1. School of Information Science and Engineering, East China University of Science and Technology, China Shanghai 200237

2. Ele-Cloud, China Guizhou Guiyang 550081

Abstract: In response to challenges such as the rapid updates of tax policies, high domain specialization, and low efficiency of manual consultation, this paper proposes the design of an intelligent tax knowledge question answering system based on large language models. The system adopts an overall framework integrating knowledge base retrieval augmentation, tool invocation for execution, and multi-agent collaboration. A tax knowledge base is constructed from tax policy documents, regulatory provisions, tax service guidelines, and frequently asked questions. Policy knowledge is acquired through document segmentation, vector encoding, and re-ranking mechanisms.In the answer generation stage, tool invocation is introduced to perform regulatory retrieval, condition validation, and rule-based computation. For complex task scenarios, a multi-agent collaborative approach is further employed to accomplish task decomposition, policy comparison, result verification, and response generation. This study presents the overall system architecture, retrieval-augmented methods, tool execution mechanisms, and multi-agent workflow design, and analyzes the system operation process through typical tax-related business cases.Experimental results demonstrate that the proposed approach improves the response efficiency, evidential reliability, and operational stability of tax consultation, providing scalable intelligent tax service support for enterprise users and financial and tax professionals.

Keywords: Large language models (LLMs); Tax knowledge base; Retrieval-augmented generation (RAG); Tool invocation; Multi-agent systems

0 引言

税务政策具有更新频繁、条目繁多、适用条件复杂等特征，企业在日常经营中经常需要处理税种识别、优惠适配、发票管理和申报流程等问题。传统税务咨询方式通常依赖人工客服、关键词检索或静态知识库，难以兼顾响应速度、答案一致性和政策依据的完整呈现。随着税务服务向数字化、智能化方向发展，构建能够理解自然语言、联动政策知识并输出可追溯结果的智能问答系统，已成为提升税务服务效率的重要路径[1]。

近年来，大语言模型在自然语言理解、文本生成和上下文推理方面表现出较强能力，为知识问答系统带来了新的技术范式。相较于基于规则或模板的传统方法，大语言模型能够更自然地处理复杂表达，但在专业场景中也容易出现时效性不足、依据缺失和推理幻觉等问题。检索增强生成技术通过将外部知识检索结果注入模型上下文，可显著提升答案的事实一致性和可解释性，因此成为领域问答系统的重要实现方式[2-3]。

基于上述背景，本文围绕税务知识服务场景，设计一

种面向政策咨询与业务执行的智能问答系统。文章重点从系统总体架构、检索增强知识获取、工具调用驱动的任务执行以及多智能体协同处理流程四个方面展开论述。

1 系统总体架构设计

1.1 设计目标与分层结构

系统设计以“可检索、可执行、可追溯”为核心目标：其一，面向税务政策咨询场景实现对法规、办税指引和优惠条件的统一检索；其二，面向计算与校验类任务通过工具调用保证执行结果的结构化和可复核；其三，面向复杂业务场景通过多智能体协同降低单模型处理长链条任务时的遗漏风险。

如图 1 所示，系统整体可划分为用户交互层、应用编排层、智能服务层、工具执行层以及知识与数据层。用户交互层负责接收问题与展示结果；应用编排层负责问题解析、会话状态维护和权限控制；智能服务层承担检索增强、提示构造和答案生成；工具执行层封装政策查询、条件校验与规则计算等可调用能力；知识与数据层则统一存放税务知识库、向量索引和问答记录。

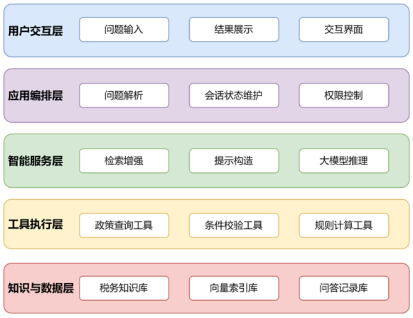

图1 税务智能问答系统总体架构

1.2 运行流程设计

系统运行时，用户首先以自然语言提交问题，系统对问题进行实体抽取和意图识别，判断其属于政策咨询、流程问答、税收优惠判断还是规则计算任务。随后，系统进入检索增强或工具执行链路：对于知识型问题，优先从税务知识库召回相关政策片段；对于计算型与流程型问题，则在检索到政策依据后进一步调用结构化工具执行；若任务涉及多条件判定或多阶段办理，则由多智能体进行任务拆解和协同处理。

2 检索增强的税务知识获取方法

2.1 知识库构建与预处理

税务知识库的数据来源包括国家税务总局公开政策文件、地方税收政策通知、办税服务指南、案例解读以及常见问题库。针对 PDF、Word 和网页等异构文档，系统首先进行文本抽取与清洗，去除页眉页脚、无效编号和重复段落；随后依据政策主题、税种类别、适用主体和时间属性对文本进行语义切分，并保留标题、来源、发布日期和适用范围等元数据，以便在问答阶段实现精准召回与结果追溯。

完成文本切分后，系统利用嵌入模型将知识片段编码为向量并写入向量数据库。与传统倒排检索相比，向量检索能够更好地处理税务咨询中的口语化表达、同义表达和条件组合表达，为复杂问句的初始召回提供支撑。

2.2 语义召回与重排序

在检索阶段，系统将用户问题向量化后与知识库中的候选片段计算相似度。常用的语义匹配形式可表示为：

$$\mathrm{sim}(q,d)=(q,d)/(\|q\|\cdot\|d\|)$$

式中 q 表示用户问题向量，d 表示知识片段向量，sim(q,d) 表示二者的语义相关度。系统先基于相似度完成向量召回，再结合政策发布时间、来源可信度和主题匹配度进行重排序，从而提升返回片段的有效性[2]。

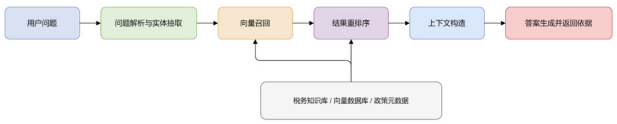

图2 RAG税务知识检索流程

如图 2 所示，检索增强流程包括问题解析、向量召回、结果重排序、上下文构造和答案生成五个环节。与直接让模型独立作答相比，RAG 方法将税务政策依据嵌入回答链路，有助于缓解模型幻觉并增强回答的可验证性。

3 工具调用驱动的税务任务执行机制

3.1 工具抽象与任务路由

仅依赖语言模型生成文本，难以稳定完成税率判断、条件校验、材料清单生成和办理链路检查等执行型任务。为此，系统在问答引擎之外封装政策检索工具、条件判别工具、税额计算工具和结果追溯工具，并通过任务路由模块决定何时由模型直接回答、何时调用外部工具完成结构化执行[4]。

在路由逻辑上，系统先根据问题意图识别结果判断任务类型：若用户主要寻求法规解释，则进入“检索增强—生成回答”链路；若任务要求计算、比对或校验，则进入“知识检索—工具执行—结果生成”链路。该设计能够将模型的语言理解优势与工具的精确执行能力结合起来。

3.2 结果校验与可追溯输出

以税额试算类问题为例，系统可将税基、适用税率和扣除项参数结构化后交由工具进行规则计算，其形式可抽象表示为：

$$\mathrm{Tax\_due}=\mathrm{Base}\times\mathrm{Rate}-\mathrm{Deduction} \quad (2)$$

其中 Base 表示计税基础，Rate 表示适用税率，Deduction

表1 多智能体功能划分

Agent	核心职责	典型输入	典型输出
意图识别Agent	判断问题类型、抽取税种与主体要素	用户问题	任务标签与关键实体
检索规划Agent	生成检索式并选择召回策略	任务标签、关键实体	候选政策片段
政策比对Agent	比较不同条款、筛选适用条件	候选政策片段	适用条款与限制条件
工具执行Agent	调用计算、校验或表单工具	结构化参数	执行结果与中间日志
结果审校Agent	核对依据、检查冲突与风险提示	政策依据、执行结果	审校意见与最终答复
意图识别Agent	判断问题类型、抽取税种与主体要素	用户问题	任务标签与关键实体

表示可抵减或扣除项。该公式并不对应某一具体税种的完整计算模型，而是用于说明工具执行机制应将规则参数从自然语言中抽离并进行结构化求解，以减少模型在数值计算和条件组合上的误差。

工具执行完成后，系统会将计算结果与检索到的政策依据进行双重校验，并以“答案概述—适用条件—执行结果—依据条款—风险提示”的格式组织输出，从而提升结果的可读性和可复核性。对于高风险问题，系统还可自动附加“以最新官方政策为准”的提示语。

4 面向税务业务的多智能体任务处理流程

4.1 多智能体功能划分

当用户问题同时包含政策判定、条件比对、办理流程说明和结果生成等多个子任务时，单一代理往往需要处理过长的上下文和过多的中间状态，容易造成信息遗漏。多智能体方法通过将复杂任务拆解为多个具备专门职责的Agent，可在保持总体一致性的同时提升处理稳定[5,6]，见表 1。

4.2 协同执行流程

在协同执行过程中，意图识别 Agent 负责解析用户问题并输出任务标签；检索规划 Agent 根据标签选择检索策略并召回政策依据；政策比对 Agent 对召回结果进行筛选和冲突识别；工具执行 Agent 根据结构化参数完成计算、校验或材料生成；结果审校 Agent 最终整合证据链并形成答复。该流程既保留了语言模型的表达能力，也通过角色分工增强了复杂税务场景中的过程可控性。

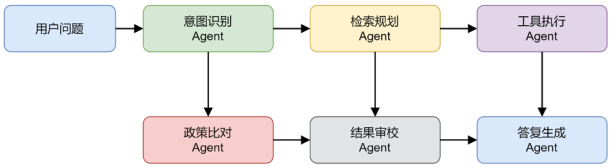

图3 多智能体协同执行流程

对于“是否满足优惠条件”“需准备哪些资料”“办理顺序如何安排”等复合问题，多智能体机制能够显著改善单轮问答在条件遗漏、步骤跳跃和依据缺失方面的不足，使系统更接近真实业务处理流程。

5 结语

本文针对税务政策咨询场景，提出了基于大语言模型的税务知识智能问答系统设计方案，并从系统总体架构、RAG 知识获取、工具调用执行和多智能体协同四个方面进行了说明。研究表明，将检索增强、工具调用与多智能体机制结合，可以在一定程度上提升税务问答系统的事实一致性、执行稳定性和结果可追溯性。

后续工作可进一步围绕政策自动更新、地域差异化知识建模、真实用户问答日志驱动的持续优化以及税务专用工具链扩展等方向展开，以推动税务智能服务系统向更高精度和更强业务适配能力发展。

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] 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.

[3] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[4] Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models[C]//The eleventh international conference on learning representations. 2022.

[5] Wu Q, Bansal G, Zhang J, et al. Autogen: Enabling next-gen LLM applications via multi-agent conversations[C]// First conference on language modeling. 2024.

[6] Li G, Hammoud H, Itani H, et al. Camel: Communicative agents for "mind" exploration of large language model society[J]. Advances in neural information processing systems, 2023, 36: 51991-52008.

基金项目：2024 年度观山湖区科技计划项目“基于大模型的税务知识中台及智能问答关键技术研究”观科合同(2024）1 号。

作者简介：* 通讯作者：黎栋梁（2001.10-），男，湖北省咸宁市，汉族，硕士研究生，华东理工大学 信息科学与工程学院在读；研究方向：电子信息工程。