铁路计算机应用
RAILWAY COMPUTER APPLICATION

基于文本大模型的财务本地智库系统的设计与实现

翟鲁杰，张　鹏，于　健，刘洪军

Financial local think-tank system based on large language model

ZHAI Lujie, ZHANG Peng, YU Jian, and LIU HongJun

引用本文:
翟鲁杰, 张鹏, 于健, 等. 基于文本大模型的财务本地智库系统的设计与实现[J]. 铁路计算机应用, 2025, 34(4): 82–86.
ZHAI Lujie, ZHANG Peng, YU Jian, et al. Financial local think-tank system based on large language model[J]. Railway Computer Application, 2025, 34(4): 82-86.

在线阅读 View online: http://tljsjyy.xml-journal.net/2025/I4/82

文章编号：1005-8451（2025）04-0082-05

基于文本大模型的财务本地智库系统的设计与实现

翟鲁杰，张　鹏，于　健，刘洪军
（中国铁路济南局集团有限公司　财务共享服务中心，济南　250001）

摘　要：针对财务共享服务中心及财务共享服务管理信息系统在推广应用过程中用户频繁培训及重复提问的问题，设计了一套基于文本大模型的财务本地智库系统（简称：本地智库系统）。通过内网本地部署国产开源文本大模型，结合 LangChain 框架，搭建了本地智库系统的整体架构，并详细阐述了该系统的功能及关键技术。该系统依托向量数据库中上传的文档资源，经过后期微调优化处理，提供了生成式对话问答服务，辅助财务共享服务中心工作人员开展线上和线下培训工作。实验验证表明，该系统能够准确回答用户提出的问题，减轻了培训和推广人员的工作负担，在铁路其他专业培训工作方面也具有应用前景。

关键词：文本大模型；培训；本地化部署；生成式对话；智能问答系统
中图分类号：F530.68 : TP39　**文献标识码：**A　**DOI：**10.3969/j.issn.1005-8451.2025.04.15

Financial local think-tank system based on large language model

ZHAI Lujie, ZHANG Peng, YU Jian, LIU HongJun
(Financial Shared Service Center, China Railway Jinan Group Co. Ltd., Jinan　250001, China)

Abstract: In response to the frequent training and repeated questioning of users during the promotion and application of the Financial Shared Service Center and financial shared service management information system, this paper designed a financial local think-tank system (referred to as the local think-tank system) based on a large language model, built the overall architecture of a local think-tank system by deploying a domestic open-source large language model locally on the intranet, combined with the LangChain framework, and elaborated on the system's functions and key technologies in detail. The system was relied on the document resources uploaded from the vector database and fine tuned and optimized in the later stage, could provide a generative dialogue and question answering service to assist the staff of the Financial Shared Service Center in conducting online and offline training work. Experimental verification shows that the system can accurately answer the questions raised by users, reduce the workload of training and promotion personnel, and also has application prospects in other railway professional training work.

Keywords: large language model; training; localized deployment; generative dialogue; intelligent question-answering system

近年来，知识管理和信息检索在各行各业中扮演着越来越重要的角色。传统知识库存在信息孤岛现象且更新不及时，难以满足快速变化的需求。利用先进的自然语言处理（NLP，Natural Language Processing）技术，可以显著提升知识库的灵活性和智能化水平。

随着中国国家铁路集团有限公司（简称：国铁集团）财务共享中心模式的逐步实施，中国铁路济南局集团有限公司（简称：济南局集团公司）开始积极推进财务共享中心建设及财务共享服务管理信息系统的应用。在财务共享中心模式与财务共享服务管理信息系统分批推广的过程中，许多用户因初次接触财务共享概念，短时间内难以适应新的工作模式；尽管接受了一次或多次线上或线下的培训，部分用户仍对业务操作和系统使用问题存在疑惑。为了满足数智化发展的新要求，提升财务共享中心的建设水平，减轻共享中心工作人员与系统用户的工作压力，济南局集团公司财务部与济南局集团公

收稿日期：2024-12-30
作者简介：翟鲁杰，工程师；张 鹏，高级会计师。

司信息技术所联合，基于前期工作中遇到的问题，在传统问题库的基础上，提出了构建基于文本大模型的财务本地智库系统（简称：本地智库系统）的构想。

相关科研人员对开源大模型及其框架进行了深入研究，许洁[1]、杨明澔[2]、申强[3]等人对文档向量化进行了研究，提出了可行的技术方案；丁志坤[4]、贾春燕[5]等人对智能知识问答系统的本地部署方案进行了研究；张海龙等人[6]分析了 LangChain 框架中各个功能的用法和优势；陈涵等人[7]阐述了基于 DevOps 理念的铁路软件开发平台设计；陈俊臻等人[8]阐述了如何搭建问答系统并使用 LLaMa-Factory 对大模型进行微调与训练。

基于上述研究，本文设计了一套本地智库系统。该系统通过对用户输入问题的分析，快速定位典型问题答案，或根据文本向量库进行大模型推理后给出准确答复，及时为用户答疑解惑，有力支持了济南局集团公司财务共享工作的开展，提升财务和业务人员的工作效率。

1 系统架构

本地智库系统采用分层设计和模块化设计思想，其架构如图 1 所示。

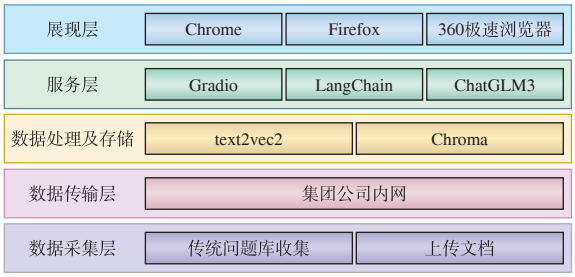


图1 本地智库系统架构

1.1 数据采集层

数据采集层的主要任务是通过上传文档和将传统智库中的数据转化为数据集，用以对模型进行训练，进而实现数据的采集。其数据来源主要包括相关知识文档和传统智库。

1.2 数据传输层

数据传输层利用 HTTP、TCP/IP 等传输协议，将数据采集层的采集的数据经铁路办公网传输至数据处理及存储层。

1.3 数据处理及存储层

数据处理的主要任务是对文档进行向量化，主要包括文本提取、文本拆分、文本标注及文本向量化。存储主要是将向量化后的结果保存到本地向量库中。本地智库系统使用的向量化方式为词嵌入（Word Embedding）。

1.4 服务层

服务层基于自然语言处理、深度学习、前馈神经网络、提示工程等机器学习方法与技术，结合财务共享用户需要和财务共享知识库，进行数据分析。同时，从交互界面、数据接口、数据检索、限制上传文档的类型和大小、为文本大模型设置提示词和角色，以及其他常规系统功能等方面，提供数据服务。

1.5 展现层

展现层结合服务层的数据服务分析结果，为用户提供应用界面、响应展示、辅助信息、本地智库系统状态及多语言支持等功能。本地智库系统采用 B/S 模式部署，用户仅通过浏览器即可访问该系统。

2 系统功能

2.1 知识库文档上传

为用户提供文档上传入口，上传后的文档经过文本拆分与向量化后被存储到本地向量库中供大模型调用。

2.2 生成式问答

为用户提供基于文本大模型的问答服务。用户可通过前端 Web 页面输入问题，文本大模型在接收到问题后，对问题进行理解、分析，并结合提示词、向量库中的内容以及自身的推理能力，给出答案。

3 关键技术

3.1 文档向量化

文档向量化是构建本地知识库的关键前置技术，依赖于先进的词嵌入模型和向量库实现文本的分解与向量化处理。文本拆分和重叠参数的选择，对硬件资源的占用和向量化的效果具有直接影响。若文本拆分的片段过长，或相邻文本块间的重叠字符数

量过多，将导致内存和处理器资源的高负载，甚至引发程序崩溃；相反，若文本片段较短、重叠字符较少，则文本间的语义连续性可能被破坏，从而降低向量化结果的准确性。本文采用 Transformer 架构中的词嵌入技术，结合 Word2Vec词嵌入模型捕捉词汇间的语义关联，实现文本向量化。在向量化过程中，平衡了硬件资源限制与文本处理的精度之间的关系，有效解决了准确度、语义连续性与系统性能之间的矛盾，确保了文本向量化的高效与准确。

3.2 构建知识库

构建知识库是本地智库系统的核心基础工作。为确保知识库内容全面、数据准确且易于维护，需要综合处理数据收集、清洗、标准化、标注、表示及更新等多方面的需求。本文通过加强内部管理、培训常态化、增加人力资源等措施，解决多来源异构数据的清洗、整合、解析及统一标注等问题，去除冗余数据、修复错误、填补缺失值并消除数据冲突，从而全面提高数据质量。使用机器人流程自动化（RPA，Robotic Process Automation）技术，降低数据采集、清洗及整合过程中的人力资源投入和人工成本。基于深度文档理解的检索增强生成（RAG，Retrieval-Augmented Generation）开源引擎 RAGFlow，更加便捷地创建和更新知识库，有效提高文本大模型回答问题的效率。通过组织数据采集人员和数据清洗人员学习相关的数据隐私法律法规、对采集数据中的敏感数据进行脱敏处理，确保数据隐私的合法合规。通过内网部署、加密存储、加密传输、数据备份及访问控制等方式，保障了数据和知识库的安全性。

通过以上各项技术难点的处理和优化，构建出一个覆盖广泛、质量高、具备动态更新能力的智能知识库，为后续的文本大模型应用和智能服务提供有力支持。

3.3 上下文管理

上下文管理技术是确保文本大模型能够进行连贯、持续对话的关键技术。随着对话的深入，模型可能会丧失之前的上下文，导致对话的连贯性和准确性下降。为解决这一问题，本文采用了 LangChain 框架中的 Memory 模块，记录人机对话历史，并在每次对话结束后手动更新记忆内容，以确保重要信息得以保存。

在长对话场景中，文本大模型需要能够有效地记住关键信息，并根据对话的进展适时“遗忘”不再相关的内容，避免过时信息影响后续对话的流畅性和准确性。本文采用关闭窗口即清空上下文的策略，以确保在每次新的对话开始时，文本大模型能够从一个清晰的起点进行处理，从而提升对话的精准性和效率。

4 实验与结果分析

4.1 环境准备

本文采用 Ubuntu 操作系统作为开发环境。项目构建使用了 Anaconda 脚手架。Anaconda 集成了多个开发工具，可以便捷地管理 Python 相关包，尤其适用于科学计算和数据分析领域。

使用 Anaconda 创建了 Python 虚拟环境，并安装了核心工具。环境的具体配置如下。

（1）LangChain：LangChain 是一个开源框架，旨在帮助开发者构建与语言模型交互的应用。其模块化设计允许开发者根据需求自由组合不同模块（如数据加载、文本生成和知识存储等）。LangChain 支持多种模型集成，包括 OpenAI、Hugging Face 以及本地模型，并提供灵活的上下文管理功能，以提高对话的相关性和连贯性。

（2）ChatGLM3：ChatGLM3 是一款基于大规模预训练的生成式语言模型，具备高效的自然语言理解能力，能够处理复杂的语言结构并生成高质量的对话结果。它通过对大量文本数据的训练积累了丰富的知识，适用于多种应用场景。此外，ChatGLM3 支持在本地环境中部署，便于满足数据隐私和安全的需求。

（3）text2vec2：text2vec2 实现了 Word2Vec、RankBM25、BERT、Sentence-BERT、CoSENT 等多种文本表征、文本相似度计算模型，为文本表示学

习和自然语言处理任务提供支持。

以上环境配置为本文的开发和实验提供了强大的技术支持和灵活的开发工具，有助于确保项目的顺利开展。

4.2　开发接口

设计并实现用户人机交互接口、上传接口，确保用户能够方便地访问知识库。使用 LangChain 快速搭建本地智库系统后端框架，用于进行文本提取、文本拆分、词嵌入、为大模型提供提示词、记录人机对话以及其他常规系统功能，使用 Gradio 快速搭建系统前端，用于用户与本地智库进行交互。

4.3　数据准备

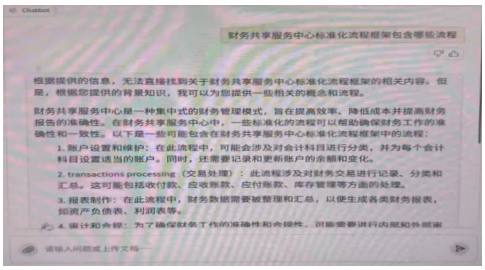
选择所需要的知识文档上传至本地智库系统，对文档进行拆分处理。利用 text2vec2 词嵌入模型对拆分后的文档进行向量化，并将生成的向量保存到本地的向量数据库中。同时，从传统问题库收集典型问题及答案，对数据进行清洗，去除无关信息，并统一数据格式。

4.4　模型集成

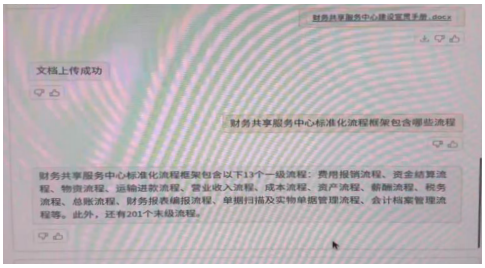
ChatGLM3 采用 API（Application Programming Interface）方式对外提供服务，作为独立服务单独运行。在模型配置文件中，可以配置服务的监听端口。启动服务后，它会提供一个外部可访问的地址和端口，本地智库系统前端通过 HTTP 协议对其进行访问并进行数据交互。

4.5　测试与分析

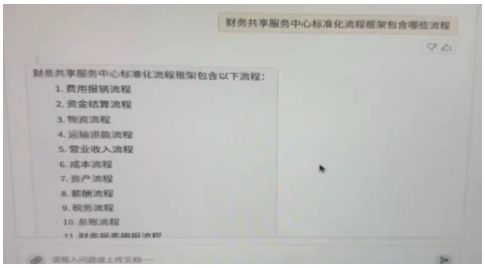
使用 LLaMa-Factory 和典型问题数据集，对文本大模型进行了多轮训练，构建了一个基于文本大模型的财务本地智能知识库。将涵盖财务会计基础知识及财务共享中心简介的相关文档导入至该知识库中，并对知识库的准确性和实用性进行测试。采用了多样化的提问策略进行测试，具体包括直接询问财务共享中心标准流程框架的详细内容、要求列出具体的流程步骤，以及探讨这些流程在实际业务操作中的应用价值等。测试结果如图 2 所示。



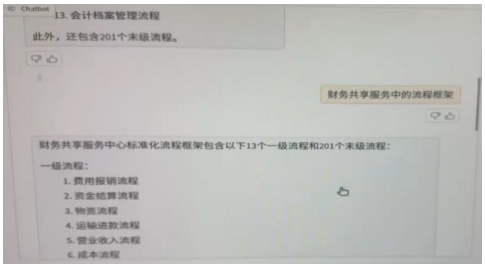
（a）未上传相关知识文档时的回答



（b）上传相关知识后的回答



（c）同一问题不同提问方式给出相同答案1



（d）同一问题不同提问方式给出相同答案2

图2　知识库准确性及实用性测试结果

无论提问形式如何变化，本地智库系统均能够准确理解问题意图，并结合上传文档中的相关信息，生成准确且符合标准的答案。这一表现证明了该系统在处理不同表述方式时的灵活性，也验证了构建的知识库在提供一致且标准化答案方面的有效性。

5　结束语

本文设计了一种结合 LangChain 和 ChatGLM3 的本地智库系统，为知识管理、信息检索和在线培训供了一种新的解决方案，初步建成了财务本地智能知识库，打破了传统知识库检索速度慢、检索条件复杂的局限，使用户能够更加便捷地获取财务领

域的专业知识，降低了用户获取知识所需要的时间成本。该系统已在济南局集团公司成功试用，为其后续在其他业务场景中的扩展应用奠定了基础。目前，该系统仅支持纯文本对话，下一步，将集中在多模态对话、优化模型性能、提升用户体验及扩展应用场景等方面进行深入研究。

参考文献

[1] 许　洁，袁小群，朱　瑞，等. 基于大模型的轻量级智能出版知识服务：理论基础与实现路径 [J]. 中国数字出版，2024，2（1）：25-35.

[2] 杨明澔，李小波，曾　倩，等. 大语言模型在油气上游业务落地的技术实践 [J]. 信息系统工程，2024（6）：61-65.

[3] 申　强. 基于 Prompt 和文本嵌入的刑事卷宗特征提取与信访风险评估模型的构建 [J]. 电脑知识与技术，2024，20（13）：34-36.

[4] 丁志坤，李金泽，刘明辉. 基于大语言模型的 BIM 正向设计问答系统研究 [J]. 土木工程与管理学报，2024，41（1）：1-7，12.

[5] 贾春燕，方伟杰，谢宇威，等. 检索增强生成技术支持下的校园问答系统研究 [J]. 通信学报，2024，45（S2）：248-254.

[6] 张海龙，黄文锋，路　翔，等. 知识增强大模型在信息系统故障分析中的应用研究 [J]. 现代计算机，2024，30（6）：87-93.

[7] 陈　涵，苗羽中，刘绍杰，等. 基于 DevOps 理念的铁路软件开发平台设计 [J]. 铁路计算机应用，2023，32（4）：53-57.

[8] 陈俊臻，王淑营，罗浩然. 融合大模型微调与图神经网络的知识图谱问答 [J]. 计算机工程与应用，2024，60（24）：166-176.

责任编辑　宣秀彬