一种基于大语言模型面向医保监督的数据查询服务

崔永琦[1,2] 赵卓峰[1,2]

1 北方工业大学信息学院，北京，100043；

2 大规模流数据集成与分析技术北京市重点实验室，北京，100043；

摘要：在医保监督业务中，检察人员常常需要对存储在关系数据库中的各种线索信息进行高效而准确的查询。而专用的 SQL 语法和逻辑对检察人员来说难以理解,迫切希望能够以自然语言方式来满足查询需求。大语言模型的出现使得基于自然语言的数据查询成为可能。然而，在医保欺诈监督背景下数据查询的大模型解决方案，仍然面临以下困难：一方面，大语言模型无法准确理解医保欺诈监督领域的专业关键词，导致查询结果的精度不高；另一方面，大语言模型在处理需要多轮对话或者一次性处理大量数据的复杂查询任务时，响应速度较慢，查询效率受限。针对上述问题，本文提出了一种基于大语言模型面向医保欺诈监督的数据查询服务。首先，针对大语言模型难以准确理解医保欺诈监督领域专业关键词的问题，通过动态构建 Few-Shot 组建提示词工程，达到提升查询准确率的效果。其次，针对复杂查询任务响应速度较慢的问题，通过 RAG 方法对向量数据库进行模式匹配，达到提升查询速度的效果。通过在真实数据上的实验证明了本文方法的效果与效率，同时本文方法也在医保欺诈检察业务中获得实际应用验证。

关键词：大语言模型；Text2SQL；RAG；向量数据库

DOI：10.69979/3041-0673.25.09.005

引言

医疗保障系统对城市生活至关重要，但医保药物非法流通问题依然严重。检察机关通过分析微信聊天记录提取倒卖药物线索，并存储于数据库中。为便于查询，自然语言查询可替代复杂 SQL 操作，大语言模型能将其转化为数据检索并生成结果描述，提升效率。

然而，在医保欺诈监督背景下的数据查询，大语言模型的解决方案面临以下难点。一方面，大语言模型难以准确理解医保欺诈监督领域的专业术语和关键词，这导致生成的查询结果精度不高；另一方面，处理复杂的文本到 SQL 任务时，模型的响应速度较慢，尤其是在需要多轮对话来完成模式匹配，或者一次性输入大量表的结构信息时，这会显著增加单次查询的时间，进而导致查询速度慢。

针对上述问题，本文利用大语言模型提出了一种面向医保欺诈监督的数据查询服务[1]，以提升查询的准确性和速度。本文的主要贡献如下：首先，针对大语言模型难以准确理解医保欺诈监督领域专业关键词的问题，通过动态构建 Few-Shot 组建提示词工程，达到提升查询准确率的效果。通过对数据库进行预处理，生成潜在的查询问题及其对应的 SQL 语句，并手动添加与领域相关的专业问题对，进一步提高了模型在专业领域的理解和查询的准确性。其次，针对复杂查询任务响应速度较慢的问题，我们通过 RAG 方法对向量数据库进行模式匹配，达到提升查询速度的效果。基于潜在问题检索相应的数据库信息组建提示词，减少多轮对话的需求，从而提升查询响应速度。最后，实际数据上的实验表明，本文提出的数据查询服务，在检察院的医保欺诈监督业务中取得了显著成效。

1 相关工作

1.1 检索增强生成

检索增强生成在利用知识库中存储的大量知识并使其为用户所用方面发挥着至关重要的作用[2]。通过结合知识库检索和大语言模型，可以在医保欺诈监督领域实现更加高效的查询和分析。

林哲毅[3]的研究提出了一种结合领域知识的知识蒸馏方法，通过微调大语言模型提升其在垂直领域任务中的表现。张浩然[4]等人提出了 DF-RAG 方法，通过查询重写和知识选择来优化检索增强生成模型，解决了复杂查询的局限性。

RAG 能够通过结合外部知识提升复杂任务的生成质量和执行效率，但针对特定的医保欺诈监督领域的应用

研究尚未深入。

1.2 提示词工程

提示词工程通过优化提示语，引导模型生成更准确的输出。在医保欺诈监督中，可帮助生成符合需求的 SQL 查询，提升查询效率，尤其在复杂数据任务中效果显著。

王娟[5]的研究通过提示词工程和“思维链”策略扩展数据集，并结合 LoRA 算法优化模型，提升了油气田领域的 NL2SQL 任务性能。翁玉鹏[6]的研究通过提示词调优和 PCA-LoRA 微调技术，优化了开源 LLMs 在 NL2SQL 任务中的准确性。

综上所述，提示词设计能够显著提升模型在复杂查询任务中的表现，但这些方法大多依赖静态的提示词模板，缺乏动态调整和优化的能力，而且提示词设计往往缺乏灵活性，无法针对医保欺诈监督数据查询场景进行动态调整。

2 本文方法

2.1 方法概述

本文的方法框架如图 1 所示，主要包括检索增强生成和提示词构建设计两个部分。输入为用户问题，输出为查询结果。首先，在检索增强生成阶段，利用大语言模型针对数据库中的所有表生成问题样本和 SQL 语句，再结合相关表名，构建向量数据库[7]，添加针对医保欺诈监督的案例，以增强向量数据库的检索效果，确保更精确地匹配复杂查询。其次，在提示词构建设计阶段，通过动态构建 Few-Shot 方法，结合检索出的样本和相关信息，组建提示词工程。将这些信息输入大模型，生成 SQL 语句查询数据库返回给大模型，返回文字描述的结果。

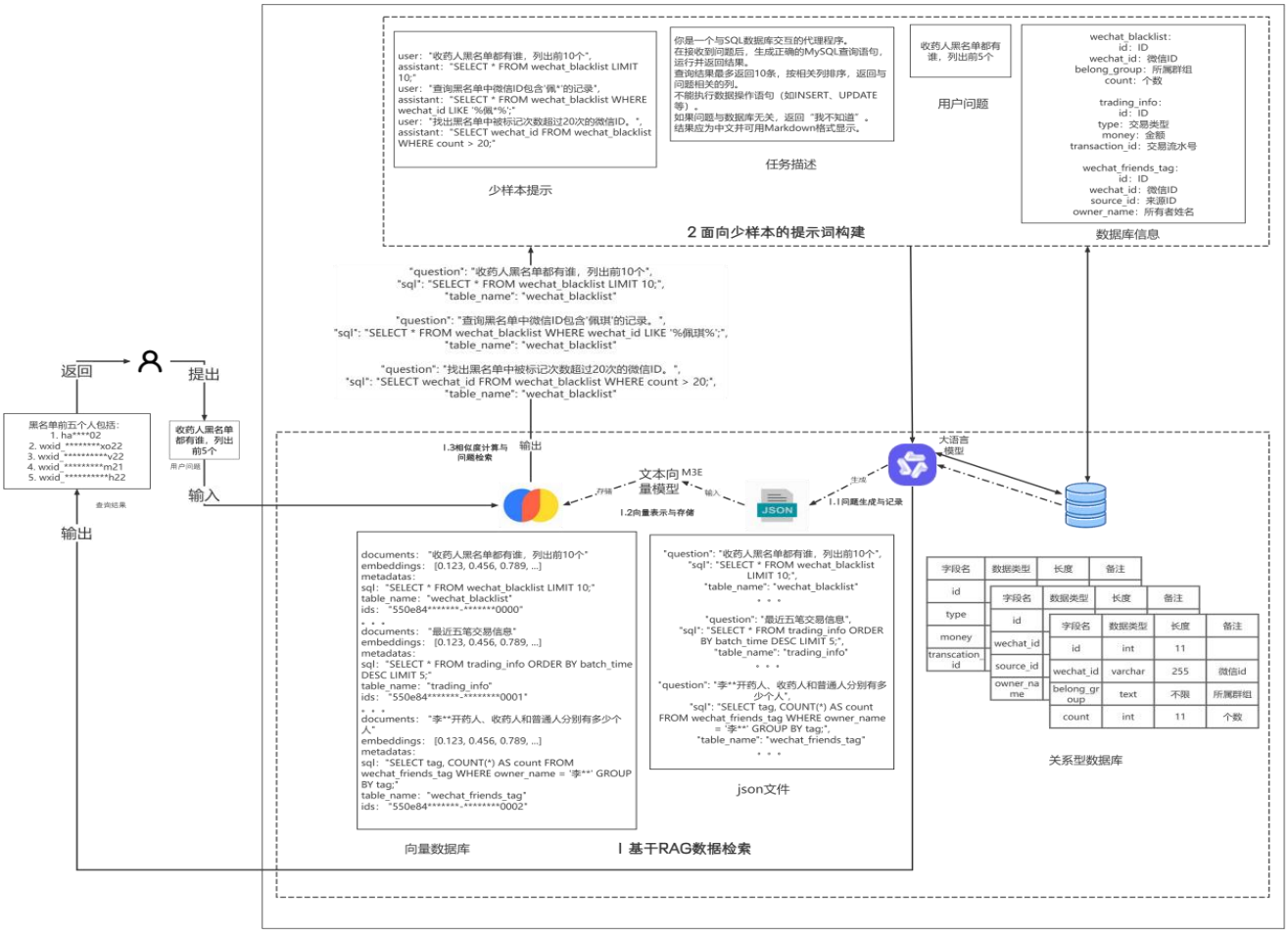

图 1 医保欺诈监督数据查询服务流程图

2.2 基于检索增强的医保欺诈监督数据检索

检索增强生成是通过结合检索和生成技术来从向量库中提取与用户问题最相关的问题等，并生成对应的

SQL 查询语句的过程。该模块包括三个部分：问题生成与记录、向量表示与存储、和相似度计算与问题检索。

针对问题生成与记录，如图 1 中 1.1 所示通过对数据库中的每张表进行详细分析，包括 31 张医保欺诈分析信息的数据表，部分表如表 1 数据库表及注释概览所示，基于其内容和结构构建查询问题集合，部分表结构如图 1 关系型数据库所示。为便于后续查询，我们记录每个问题对应的 SQL 语句及表名，构建问题、SQL 与表名组成的问题集。

表 1 数据库表及注释概览

表名	注释	表名	注释
call_records	通话记录表	wechat_baseinfo	微信基础信息表
drug_info	药品信息表	wechat_friends_tag	微信好友标签表
wechat_blacklist	微信黑名单	trading_info	交易表

针对向量表示与存储，如图 1 中 1.2 所示利用预训练的 embedding 模型将问题集中的问题转换为高维向量表示。将问题集中的问题中作为 documents，SQL 和对应的表名作为 metadatas，加上问题的高维向量存入一个向量数据库中，具体如图 1 向量数据库所示。

针对相似度计算与问题检索，如图 1 中 1.3 所示，通过将用户输入的问题转化为高维向量表示，并与向量数据库中的问题向量进行相似度计算来完成匹配[8]。相似度计算能够筛选出与用户问题最相关的前 N 个问题。这些问题作为检索结果，确保了系统能够快速、准确地定位用户的查询需求，并生成对应的 SQL 语句，从而有效减少多轮查询的需求，提升整体查询效率和系统响应速度[9]。

2.3 动态少样本构建医保欺诈监督数据查询的提示词构建

在医保欺诈监督数据查询服务中，由于检察人员不熟悉 SQL 语法，难以直接操作关系型数据库进行数据查询。为此，引入 Text2SQL 技术，将自然语言查询转换为数据检索操作，并结合大语言模型生成结果描述。但由于大语言模型在处理复杂查询时响应较慢，影响查询效率。为此，我们通过 RAG 方法对向量数据库进行模式匹配，动态构建提示词，提升查询速度。

本文的提示词构建是通过设计和优化提示词来指导大语言模型生成更为准确的查询语句的过程。该模块包括四个部分：任务描述、少样本提示[10]、用户问题分析以及数据库信息匹配。完整的提示词如图 1 中 2 面向少样本的提示词构建中所示，其中，任务描述和少样本提示是提示词工程的核心部分，主要用于明确模型的生成规则与目标，并通过提供结构化示例帮助模型更好地生成高质量的 SQL 查询语句。

通过少样本提示实例的构建，提高了模型对 SQL 语法结构和生成规则的理解能力。少样本提示实例的构建同样源于“检索增强生成”阶段的输出。基于与用户查询最相似的问题及其对应的 SQL 查询，组成了 Few-Shot 示例，具体如图 1 提示词中少样本提示所示，将这些示例用作提示词输入到大语言模型中，以帮助模型更好地学习 SQL 查询的结构和生成规则，从而提高生成 SQL 的准确性。少样本提示：通过分析用户查询和相似的历史查询，生成少样本示例，并将这些示例作为提示词输入给模型，帮助其更好地学习 SQL 查询的结构与生成规则。

通过这样的少样本提示的方式，动态构建提示词工程，能有效引导模型生成 SQL 语句，避免了大语言模型在处理需要多轮对话或者一次性处理大量数据的复杂查询任务时，解决了基于大语言模型在医保欺诈监督数据查询服务中响应速度较慢的问题。

3 评估分析

3.1 环境与设置

3.1.1 数据集

本文数据集是由 31 张表所涉及的潜在问题，这些问题主要基于数据库中的具体数据查询。数据集中部分问题及其对应的表名如表 2 所示。

表 2 用户问题

表名	问题
wechat_friends_tag	找出名字为'李**'的所有好友的标签。
call_records	收药人黑名单都有谁，列出前 5 个

3.1.2 实验环境

本文实验环境如表所示。大语言模型为 Qwen 和 GLM 模型，文本嵌入模型采用 m3e。实验具体配置如表 3 所示：

表 3 实验环境

名称	配置信息
操作系统	Ubuntu 20.04
CPU	Intel® Xeon® Silver 4214R
GPU	2*NVIDIA GeForce RTX 3090

3.1.3 评价指标

为了评估所提出方法的有效性，本文主要关注准确性和响应时间两个方面的指标。准确性是执行准确度关注生成的SQL查询语句执行后产生的结果是否与参考查询语句执行产生的结果相同。执行准确度可以表示为

$$EA = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(r_{gen_i} = r_{ref_i})$$

其中，

$R_{ref} = r_{ref_1}, r_{ref_2}, \ldots, r_{ref_n}$ 表示参考SQL查询语句执行后的结果集合。

$R_{gen} = r_{gen_1}, r_{gen_2}, \ldots, r_{gen_n}$ 表示生成SQL查询语句执行后的结果集合。

（ ）是指示函数，当 成立时，（ ）= 1，否则（ ）= 0。

响应时间是记录大语言模型从输入自然语言到生成查询结果所需的时间，以评估模型的响应速度。

后续实验将重点关注以下三个问题：

（Q1）与其他方法的医保欺诈数据查询性能相比如何？

（Q2）检索增强生成对查询准确性的影响？

（Q3）N-shot 示例在医保欺诈监督数据查询中性能的影响？

3.2 实验分析

3.2.1 数据查询性能分析

针对问题Q1，也即为了评估不同方法在处理同一组问题时的响准确性，对比了本文方法与Langchain的SQLAgent和LlamaIndex在时间和准确性方面的表现。具体结果如表4所示。

表 4 不同方法的性能对比

方法	平均时间 (秒)	EA(%)
本文方法	7.8	89
SQLAgent	12.2	78.4
LlamaIndex	11.8	84.8

在实验结果的分析中可以明显看出，本文方法在查询的平均时间和精准率方面均表现出色，显著优于其他方法。具体来说，本文方法模式的平均查询时间为7.8秒，精准率达到了89%，在所有测试方法中均为最佳。造成这一差异的主要原因在于本文方法方法的独特优势。通过使用RAG方法构建Few-Shot，能够有效提升模型在少量数据上的准确性，同时保持较高的处理速度。

3.2.2 消融实验与分析

针对问题Q2，也即为了评估使用RAG以及结合医保欺诈领域人工标注对系统查询精准率的影响，我们设计了消融实验。实验对比了三种情况下的表现：不使用RAG、仅使用RAG、以及使用RAG结合人工标注的效果。结果如表5所示。

表 5 使用 RAG 模式和结合人工标注对系统准确性的影响

方法	EA(%)
w/o RAG	79
only RAG	85.2
joint w/ RAG + annotated cases (ours)	88.4

实验结果表明，结合医保欺诈领域的人工标注问题能够显著提升系统的查询精准率。具体而言，仅使用RAG时，系统的查询精准率已达到85.2%，表现出较高的准确性。然而，进一步结合领域内人工标注问题后，查询精准率显著提高至88.4%。这是因为，人工标注的问题使系统能够更准确地理解和处理与医保欺诈检测相关的特定查询需求，从而在处理复杂查询时更加精准有效。

3.2.3 关键参数分析

针对问题Q3，也即为了评估N-shot示例对方法性能的影响，我们设计了关键参数实验。实验对比了在0-shot、1-shot、3-shot和5-shot设置下的表现，具体分析了系统在不同N值下使用Few-Shot示例生成时的查询准确性和平均时间。结果如表6所示。

表 6 N-shot 示例对方法性能的影响

方法	平均时间 (秒)	EA(%)
0-shot	7.5	79.4
1-shot	8.1	82.6
3-shot	8.5	89.0
5-shot	9.3	87.4

实验结果表明，随着N值的增加，系统的查询准确性总体呈现上升趋势，尤其是在3-shot设置下达到了最高值89.0%。这说明适当的Few-Shot示例数量能够显著提升系统的查询性能。然而，5-shot设置下的准确性略低于3-shot，表明过多的示例可能引入噪声，降低系统的准确性。同时，查询时间随着N值的增加而延长，

过高的 N 值还会导致计算负担的增加。因此，3-shot 设置在 Few-Shot 学习中表现最优，能够在保持较高准确性的同时，维持合理的查询时间。

4 结论

在医保欺诈监督的领域背景下，针对医保欺诈监督数据查询难以准确理解专业领域关键词和查询速度较慢的问题，本文提出了一种基于大语言模型的医保欺诈监督数据查询服务。该方法通过构建向量数据库，并结合 RAG 技术和提示词工程设计，有效提高了查询的准确性和速度。实验结果表明，本文方法在检察院的实际数据上显著提升了查询的效果，查询准确性达到 89%，平均查询速度为 7.8 秒，验证了方法的有效性。

本文方法仍存在部分局限。目前仅针对单表查询场景，未充分考虑多表查询的复杂性，这在需要整合多个数据源时可能会限制查询的全面性。这也是后续需要研究的内容。

参考文献

[1]全筱筱,熊文举,潘军杰,等.基于大语言模型的数据查询机器人在医学领域的应用[J].医学新知,2024,34(09):1057-1063.

[2]朱兵，张勇，唐波，等.基于大数据的电力信息通信预警技术探索[J]. 电子世界，2019(16)：199-200.

[3]林哲毅.基于知识增强的垂直领域大语言模型研究与应用[D].杭州电子科技大学,2024.DOI:10.27075/d.cnki.ghzdc.2024.001917.

[4]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/OL].计算机科学,1-12[2025-04-25].

[5]王娟,梁倩,王磊,等.大语言模型驱动的油气田勘探开发数据智能检索方法[J].西安工业大学学报,2024,44(06):795-802.DOI:10.16185/j.jxatu.edu.cn.2024.06.306.

[6]翁玉鹏.基于大语言模型的NL2SQL方法研究[D].西安石油大学,2024.DOI:10.27400/d.cnki.gxasc.2024.000460.

[7]路沙.向量数据库突显含金量[N].中国信息化周报,2023-08-14(022).DOI:10.28189/n.cnki.ndnjy.2023.000236.

[8]李思卓，赵辉，耿晓燕，等.关系数据库中地理空间数据存储优化研究[J].测绘与空间地理信息，2023,46(04)：158-161.

[9]王玉珏.基于提示学习的少样本分类方法研究[D].哈尔滨:黑龙江大学，2024.DOI:10.27123/d.cnki.ghlju.2024.000945.

[10]文婧.小样本语义分割方法研究[D].武汉:中南民族大学，2022.DOI:10.27710/d.cnki.gznmc.2022.000721.