УДК 004.8

РАЗРАБОТКА ИНТЕЛЛЕКТУАЛЬНОГО СЕРВИСА АВТОМАТИЗАЦИИ АНАЛИТИКИ КАДРОВОГО СОСТАВА С ИСПОЛЬЗОВАНИЕМ LLM И ТЕХНОЛОГИИ TEXT-TO-SQL

Апасов М.В., **Булыгин С.А., Федоров Д.А.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Федоров Д.А. (ИТМО)

Введение. В государственных и крупных коммерческих организациях актуальной задачей является оперативное получение кадровой аналитики. Зачастую данные хранятся в реляционных базах данных, доступ к которым требует знания SQL и обученных специалистов, ведь конечные потребители этой информации, клиенты и руководители, не имеют нужных навыков.

Перспективным направлением решения данной проблемы являются text-to-sql системы для генерации запросов на основе естественного языка. При этом, именно LLM показывают прогресс в решении NLP-задач, к классу которых и относится text-to-sql.

В данный момент на рынке существует возможность использования больших LLM от крупных компаний через API [1], локальные модели, а также специализированные алгоритмы, но каждое из этих решений имеет недостатки: безопасность чувствительных данных, невозможность тонкой настройки, требования к интернет-соединению и оборудованию, необходимость платить за токены, трудности внедрения, а также проблемы с контекстом [2].

Основная часть. Для решения трудностей в данном исследовании предлагается использовать локальную LLM с малым количеством параметров, что позволит обеспечить безопасность данных, простоту внедрения, снизит требования к системе, а также устранит необходимость платить за токены, повышая доступность системы для рядового пользователя и предприятий без доступа к дорогим и мощным датацентрам. При этом, подобное решение снижает точность генерации и не решает проблемы с предоставлением контекста.

Для покрытия возникающих трудностей уже существует несколько инновационных решений, позволяющих улучшить работу LLM в text-to-sql:

1) Дообучение модели для снижения ошибок генерации в определенной области и базе;
2) Применение технологии RAG для предоставления точного контекста без обучения;
3) Вызов функций и MCP для тонкой настройки работы моделей.

Зачастую, в исследованиях эти методы используются по-отдельности, что в случае с малыми моделями решает одну конкретную проблему, но приводит к потеря гибкости, скорости, галлюцинациям, а сбор данных для конкретной задачи может быть слишком затратным.

Именно поэтому в данном исследовании предлагается использовать метод комбинирования сразу нескольких стратегий развертывания и оптимизации, дополняющих друг друга:

1) Была найдена и дообучена на общих SQL запросах, а также квантизирована до 4 бит модель - Llama-3-sqlcoder-8b, наилучшим образом прошедшая практические тесты;
2) Выполнена разработка RAG, состоящая из модели эмбеддинга, векторной БД [3] и скрипта для автоматической векторизации баз данных;
3) Написаны две функции для вызова со стороны LLM, позволяющие получить доступ к данным из БД даже при проблемах в работе RAG для повышения

стабильности;

4) При помощи фреймворков LangChain, LangGraph и FastAPI был собран усиленный RAG технологией, дообучением и вызовом функций ИИ-агент, доступ к которому обеспечивается через веб-интерфейс или API.

Выводы. В результате исследования был получен прототип приложения в формате API для генерации SQL-команд по запросу на естественном языке с использованием локальной LLM и State-of-Art методов развертывания и оптимизации. Также проведены тесты, показывающие существенное сокращение времени генерации в 4 раза (с 155-170 сек. до 40-50 сек.) на системе без видеоадаптера (CPU N100 и 16GB ОЗУ в одноканальном режиме) в сравнении с крупными LLM при сохранении точности ответов на уровне облачных решений.

Подобное приложение имеет ценность как для конечного пользователя, так и для бизнес-сегмента, ведь позволяет упростить доступ к информации в базах данных при сохранении безопасности данных, низких требований и простоты интеграции.

Список использованных источников:

1. Чунгулова Г.К., Оразалиева Э.Н. Возможности и проблемы больших языковых моделей в образовании на примере ChatGPT // Наука и реальность. – 2024. – №4. – С. 85–90.
2. Современные подходы «из текста в SQL»: RAG, CoT и другие хитрости [Электронный ресурс] // braintools.ru. URL: https://www.braintools.ru/article/17032.
3. Выбираем векторную БД для AI-агентов и RAG: большой обзор баз данных и поиск смысла [Электронный ресурс] // habr.com. URL: https://habr.com/ru/articles/961088/