

ПРОЕКТИРОВАНИЕ АГЕНТНЫХ ДИАЛОГОВЫХ СИСТЕМ НА ПРИМЕРЕ ГОСУДАРСТВЕННЫХ БАЗ ДАННЫХ С ПРИМЕНЕНИЕМ ТЕХНОЛОГИИ RAG И LLM В РАМКАХ ЗАДАЧИ TEXT2SQL

Леманов А. А. (Университет ИТМО), Скворцов Д. А. (Университет ИТМО),
Ореховский И. А. (Университет ИТМО)

Научный руководитель – кандидат технических наук, доцент Федоров Д. А.
(Университет ИТМО)

Введение. В эпоху цифровой трансформации государственных услуг, одной из ключевых задач является автоматизация взаимодействия с базами данных, что позволяет значительно сократить временные и ресурсные затраты на обработку запросов. Особенно актуальной эта задача становится в таких областях, как жилищно-коммунальная инспекция (ЖКИ), где операторы ежедневно сталкиваются с большим количеством запросов от граждан. Эти запросы часто формулируются на естественном языке и требуют преобразования в структурированные SQL-запросы для извлечения данных из государственных баз. Задача text2sql, заключающаяся в автоматическом преобразовании текстовых запросов в SQL-код, становится важным инструментом для повышения эффективности работы с данными.

Основная часть. Для решения задачи text2sql применяются LLM, такие как LLaMa и Mistral, которые отличаются высокой гибкостью и способны обрабатывать сложные запросы. Однако их релевантность SQL запроса может снижаться при работе с неполными или недостаточно структурированными данными.

Основные задачи исследования включают:

- 1) Анализ проблем, возникающих при решении задачи text2sql с учетом специфики данных
- 2) Изучение преимуществ использования LLM для генерации SQL-запросов
- 3) Изучение и применение различных подходов RAG для улучшения эффективности SQL запросов
- 4) Апробация подхода на государственных данных, с использованием таких метрик как semantic similarity, precision, recall, F-мера, factual correctness и context recall
- 5) Реализация подходов проектирования агентных диалоговых систем с использованием разработанной технологии на примере ЖКИ

Заключение. Использование LLM в сочетании с технологией RAG для решения задачи text2sql представляет собой перспективное направление, которое позволяет повысить релевантность генерации SQL-запросов, минимизировать ошибки и улучшить адаптацию к контексту запросов. Это открывает новые возможности для эффективной работы с базами данных, анализа больших объемов информации и повышения качества обслуживания, в таких областях, как ЖКИ.

Список использованных источников:

1. SiriusBI: Building End-to-End Business Intelligence Enhanced by Large Language Models // URL: <https://arxiv.org/html/2411.06102v1> // (дата обращения: 17.02.2025).
2. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
// URL: <https://arxiv.org/abs/2005.11401> // (дата обращения: 17.02.2025).
3. Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation
// URL: <https://arxiv.org/abs/2308.15363> // (дата обращения: 17.02.2025).
4. PAUQ: Text-to-SQL in Russian
// URL: <https://aclanthology.org/2022.findings-emnlp.175.pdf> // (дата обращения: 17.02.2025).
5. Shifted PAUQ: Distribution shift in text-to-SQL
// URL: <https://aclanthology.org/2023.genbench-1.18.pdf> // (дата обращения: 17.02.2025).