УДК 004.891

ИСПОЛЬЗОВАНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ГЕНЕРАЦИИ ОТВЕТОВ НА ВОПРОСЫ С ИСПОЛЬЗОВАНИЕМ ИЗВЛЕЧЕННЫХ ИМЕНОВАННЫХ СУЩНОСТЕЙ

Тихонов С.Н. (ИТМО)

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

Введение. В современном мире цифровых технологий, использование больших языковых моделей (LLM) становится всё более востребованным и эффективным. Данное исследование проводится в рамках разработки чат-бота для муниципального округа Васильевский “Чат-бот Василий”. Оно поднимает проблему генерации ответов с использованием контекста, получить который с помощью RAG проблематично, по причине того, что смысловая составляющая запроса сильно отличается от релевантной информации, необходимой для генерации ответа. Для решения данной проблемы предлагается использовать LLM, которые будут обрабатывать запросы пользователей на естественном языке, определять наличие именованных сущностей в запросе, и генерации ответа с использованием данных, связанных с этими именованными сущностями, например можно будет попросить список проектов, выполняющихся в диапазоне определенных дат, после чего LLM определит даты в запросе и обратится к БД [1], в которой хранятся данные (в том числе даты) о выполняемых проектах, затем данные о проектах в нужном диапазоне времени из бд будут переданы в контекст LLM и на их основе будет сгенерирован ответ со списком проектов. “Чат-бот Василий” разрабатывается как бот для мессенджера Telegram, и позволит обеспечить удобный и доступный канал связи между гражданами и муниципальными службами.

Основная часть.

1) Целью данной работы является разработка архитектуры решения по генерации ответов с использованием релевантной информации, недоступной к получению с помощью RAG. Данное решение будет интегрировано в чат-бота, который будет информировать жителей муниципального округа Васильевский о проводимых и плановых работах по благоустройству, а также обеспечивать связь между депутатами и жителями.
2) Архитектура будет основана на 2 LLM, первая будет распознавать и извлекать именованные сущности, после чего на их основе будет делаться запрос к БД, а вторая будет получать исходный запрос пользователя и данные из БД и на их основе генерировать ответ..
3) Функциональные возможности
   - Обработка запросов пользователей;
   - Извлечение именованных сущностей;
   - Получение информации из БД;
   - Генерация ответа с использованием релевантных данных.
4) Технологии и методы
   - Использование современных LLM для извлечения именованных сущностей и генерации ответов;
   - Промпт инжиниринг [2];
   - Интеграция с базами данных и источниками информации;

Выводы. Было разработано решение для автоматического распознавания и извлечения именованных сущностей, получение информации из бд на их основе и

генерации ответов с использованием полученной информации.. Использование LLM значительно улучшило качество релевантной информации и генерации ответов на естественном языке

Список использованных источников:

1. Официальный сайт с документами муниципального округа Васильевский. – URL: https://msmov.spb.ru/документы.html (дата обращения 24.02.2025).
2. Искусство общения с LLM: Гайд по техникам Prompt Engineering – URL: https://habr.com/ru/articles/827546/ (дата обращения 24.02.2025).
3. Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, Hsiao-Wuen Hon. Unified Language Model Pre-training for Natural Language Understanding and Generation. - URL: https://arxiv.org/abs/1905.03197
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - URL: https://arxiv.org/abs/1810.04805