УДК 004.891.2
ИССЛЕДОВАНИЕ ПОДХОДА RAG В ИНТЕЛЛЕКТУАЛЬНЫХ ВОПРОСНО-ОТВЕТНЫХ СИСТЕМАХ

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

Научный руководитель – кандидат физико-математических наук, доцент Фёдоров Д.А.
(ИТМО)

Введение.

В сфере жилищно-коммунального хозяйства (ЖКХ) операторы ежедневно обрабатывают множество запросов от граждан, что требует значительных временных и трудовых ресурсов. Для оптимизации этого процесса предлагается разработать интеллектуальную вопросно-ответную систему, использующую подход Retrieval-Augmented Generation (RAG). Данный метод сочетает механизмы поиска информации с генеративными возможностями больших языковых моделей, позволяя формировать ответы на основе исторических данных и нормативных документов [1]

Основная часть.

Предлагаемая система состоит из двух основных компонентов: модуля поиска и генеративной модели. На первом этапе модуль поиска анализирует поступающий вопрос и извлекает релевантные данные из базы вопросов и ответов, а также из нормативных документов. Затем генеративная модель, опираясь на полученную информацию, формирует развернутый и точный ответ.

Для повышения эффективности системы рассматриваются следующие аспекты.

Оптимизация процесса поиска: использование современных алгоритмов ранжирования и векторных представлений текста для более точного извлечения релевантной информации [2]. (добавить про чанк\ корпус текстов)

Обработка нормативных документов: разработка методов интеграции информации из нормативных актов в процесс генерации ответов, что особенно важно в сфере ЖКХ, где многие вопросы требуют ссылок на законодательные нормы.

Оценка качества системы: применение метрик, таких как точность (precision), полнота (recall) и F-мера, для объективной оценки работы системы и ее сравнения с традиционными методами.

Выводы.

Разработка интеллектуальной вопросно-ответной системы на основе подхода RAG позволит снизить нагрузку на операторов в сфере ЖКХ, обеспечивая быстрые и точные ответы на запросы граждан. Интеграция исторических данных и нормативных документов в процесс генерации ответов повысит их релевантность и достоверность. Дальнейшие исследования будут направлены на улучшение алгоритмов векторного поиска и генерации, а также на расширение базы знаний системы. Предлагаемые методы позволят снизить галлюцинации языковой модели.

Список использованных источников:

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Küttler, H., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.
2. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333-389.
3. Li, S., Stenzel, L., Eickhoff, C., & Bahrainian, S. A. (2025). Enhancing Retrieval-Augmented Generation: A Study of Best Practices. arXiv preprint arXiv:2501.07391.