УДК 004.89

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИИ RAG ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Тлумач Е.Д. (ИТМО)
Научный руководитель – кандидат технических наук Русак А.В.
(ИТМО)

Введение. Современные большие языковые (LLM) модели существенно поменяли подход к обработке естественного языка, помогая решать задачи перевода, генерации, суммаризации, классификации текстовых документов. Вместе с тем, у таких моделей есть ряд недостатков. Одна из ключевых проблем – ограниченный контекстный объём, модель может использовать только ту информацию, на которой она была обучена, и не способна выдавать информацию за рамками своего обучающего корпуса. Это приводит к тому, что модель может генерировать устаревшие и неточные ответы. Среди прочего трудно использовать LLM, обученные на открытых источниках, для бизнес-задач, требующих актуальной информации, так как они не способны обратиться к внутренним корпоративным документам и работать с персональными данными.

Основная часть. Решением вышеперечисленных проблем является метод Retrieval-Augmented Generation (RAG). Суть этого подхода заключается в сочетании функционала больших языковых моделей и поискового механизма для подбора релевантной информации во внешних источниках данных, что повышает точность и актуальность ответа. Метод можно разделить на два этапа: извлечение и генерация [1]. На этапе извлечения пользователь вводит запрос, который преобразуется в векторное представление и начинается поиск по внешнему хранилищу, а на этапе генерации обнаруженные фрагменты расширяют единый контекст, который затем передается в большую языковую модель для генерации ответа. Метод позволяет получать более актуальные ответы на запросы, потому что этот подход даёт возможность учитывать новые данные и избегать дообучения, а также работать с узконаправленными данными. Дообучение больших языковых моделей представляет собой довольно продолжительный по времени и ресурсоёмкий процесс, так как приходится пересчитывать большое количество весов модели. Хотя дообучение может обеспечить некоторые улучшения, метод RAG последовательно демонстрирует лучшие результаты как в обновлении имеющихся знаний, так и в интеграции совершенно новой информации [2][3].

Выводы. В ходе исследования были проведены предобработка данных, формирование векторного представления документов с помощью эмбеддинг-модели, интеграция механизма извлечения релевантной информации и генерации ответов на основе объединённого контекста. Была разработана реализация RAG-системы, которая даёт релевантные ответы на запросы с помощью поиска во внешнем источнике информации.

Список использованных источников:

1. Оболенский Д.М., Шевченко В.И. Использование метода RAG и больших языковых моделей в интеллектуальных образовательных экосистемах // Экономика. Информатика. – 2024. – URL: https://cyberleninka.ru/article/n/ispolzovanie-metoda-rag-i-bolshih-yazykovyh-modeley-v-intellektualnyh-obrazovatelnyh-ekosistemah (дата обращения: 14.02.2025).
2. Бородулин И.В. Увеличение точности больших языковых моделей с помощью расширенной поисковой генерации // Вестник науки. – 2024. – URL: https://cyberleninka.ru/article/n/uvelichenie-tochnosti-bolshih-yazykovyh-modeley-s-pomoschyu-rasshirennoy-poiskovoy-generatsii (дата обращения: 14.02.2025).

3. Овадия О., Бриф М., Мишаэли М., Элиша О. Тонкая настройка или извлечение? Сравнение методов внедрения знаний в большие языковые модели // arXiv:2312.05934 [cs.AI]. — 2024. — URL: https://arxiv.org/abs/2312.05934 (дата обращения: 14.02.2025).