УДК 004.8:004.912

РАЗРАБОТКА ЧАТ-БОТА ДЛЯ КОНСУЛЬТИРОВАНИЯ АБИТУРИЕНТОВ УНИВЕРСИТЕТА ИТМО НА ОСНОВЕ АРХИТЕКТУРЫ RAG

Фархутдинов А. А.[1]
Научный руководитель – канд. техн. наук, старший преподаватель Русак А. В.[1]
[1]Университет ИТМО
aafarkhutdinov@itmo.ru

Введение

Цифровизация образовательной среды сопровождается увеличением объёма информации о правилах приёма, направлениях подготовки и условиях поступления. Большой объём нормативных документов и их регулярное обновление затрудняют оперативное получение актуальной информации и повышают нагрузку на приёмные комиссии.

Для автоматизации консультирования применяются чат-боты на основе языковых моделей, способные обрабатывать запросы на естественном языке [1]. Однако использование исключительно генеративных моделей может приводить к формированию недостоверной информации. Перспективным решением является технология Retrieval-Augmented Generation (RAG), объединяющая методы информационного поиска и генерации текста [2]. Использование базы знаний образовательной организации позволяет повысить точность и актуальность ответов [3].

Основная часть

Разрабатываемая система представляет собой чат-бота для консультирования абитуриентов на основе архитектуры Retrieval-Augmented Generation.

Работа системы включает три основных этапа:

1. Формирование базы знаний.

Текстовые документы разбиваются на фрагменты, преобразуются в векторные представления и сохраняются в базе поиска [3].

2. Поиск релевантной информации.

Пользовательский запрос преобразуется в векторное представление, после чего осуществляется поиск наиболее близких по смыслу фрагментов базы знаний [2].

3. Генерация ответа.

Найденные фрагменты передаются в языковую модель, которая формирует итоговый ответ с учётом предоставленного контекста [1, 2]. Это позволяет снизить вероятность генерации недостоверной информации и обеспечить соответствие ответа актуальным правилам приёма.

Для реализации RAG-системы были проведены отбор и предобработка данных, получение векторного представления документов с помощью эмбеддинг-модели, реализация механизма извлечения релевантной информации и генерации ответов с помощью большой языковой модели, и тестирование на основе отобранных метрик.

Выводы

В рамках работы реализован пайплайн обработки запроса пользователя, включающий формирование базы знаний, поиск релевантной информации и генерацию ответа с использованием языковой модели. Предложенная архитектура позволяет обеспечить высокую точность ответов, актуальность предоставляемой информации и возможность обновления базы знаний без необходимости переобучения модели.

Разработанное решение может быть интегрировано в популярные платформы обмена сообщениями, такие как Telegram и ВКонтакте, что позволит обеспечить круглосуточную поддержку абитуриентов и снизить нагрузку на приёмную комиссию.

Использование чат-бота повышает доступность информации и улучшает качество взаимодействия абитуриентов с образовательной организацией.

Литература

1. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T., Child R., Ramesh A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. — 2020.
2. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. — 2020.
3. Ковалев Г., Тихомиров М., Кожевников Е., Корнилов М., Лукашевич Н. Создание русского бенчмарка для оценки моделей информационного поиска // Материалы конференции / Труды научной работы. — М.: МГУ им. М.В. Ломоносова, 2025. — 11 с.