УДК 658.8:004.9

РАЗРАБОТКА ГОЛОСОВОГО ПОМОЩНИКА ДЛЯ ПРИЕМНОЙ КОМИССИИ УНИВЕРСИТЕТА НА ОСНОВЕ BIG DATA И RAG-АРХИТЕКТУРЫ

А.В. Казаков
Аспирант ТУ им. А.А. Леонова (филиал) МИИГАиК
kazalexit@gmail.com

А.Н. Казаков
Окончил магистратуру – ВШЭ, Специалитет – МИФИ.. Область научных интересов: голосовые технологии (ASR/TTS), ИИ применительно к тексту и голосу, обработка естественного языка (NLP), интеллектуальные системы поддержки принятия решений, большие языковые модели.

Аннотация. В статье рассматривается процесс разработки интеллектуального голосового помощника для автоматизации работы приемной комиссии Технологического университета им. А.А. Леонова, (филиал) МИИГАиК

Ключевые слова: голосовой помощник, приемная комиссия, RAG, Yandex SpeechKit, подготовка данных, векторизация, NLP, автоматизация образования.

Введение и постановка задачи. Современные вызовы цифровизации образования требуют внедрения интеллектуальных систем, способных обрабатывать большие объемы входящих запросов абитуриентов.

Приемная комиссия в период кампании сталкивается с пиковыми нагрузками, когда преподаватели не успевают обрабатывать звонки, что приводит к потере потенциальных студентов или недовольству родителей абитуриентов. Решением является разработка голосового помощника, способного круглосуточно консультировать по правилам приема, срокам подачи документов, перечню экзаменов и проходным баллам.

Целью данного проекта является создание такого помощника для Технологического университета им. А.А. Леонова на основе технологии RAG и сервисов Yandex Cloud.

На момент написания статьи проект находится на начальном, но наиболее важном этапе – этапе подготовки и структурирования данных, от качества которых напрямую зависит точность ответов итоговой системы.

Математический аппарат анализа и фильтрации признаков.Архитектура решения базируется на технологии Retrieval-Augmented Generation (RAG) с интеграцией сервисов Yandex SpeechKit (ASR/TTS) и телефонии (рисунок 1).

Основной фокус работы сделан на критически важном этапе подготовки данных: сборе, очистке, структурировании и векторизации нормативно-справочной информации университета.

Приводятся требования к датасету, методика обработки неструктурированных текстов (приказы, правила, программы) и подходы к эмбеддингу для обеспечения семантического поиска.

Проект находится на начальной стадии, что обуславливает детальное описание препроцессинга как фундамента качественной работы всей системы.

Исследование архитектуры проектируемого решения. Проектируемая система голосового помощника имеет модульную архитектуру, включающую следующие ключевые компоненты: телефонный шлюз (Asterisk) для приема звонков, модуль распознавания речи (ASR Yandex SpeechKit), оркестратор диалогов (LangGraph), RAG-сервис с векторной базой данных, базу знаний университета, модуль синтеза речи (TTS Yandex SpeechKit).

Принцип работы заключается в том, что текст запроса, полученный из голоса, используется для поиска по векторной базе знаний.

Найденные релевантные фрагменты документов (контекст) вместе с исходным запросом передаются языковой модели (LLM) для генерации точного ответа, что минимизирует риск галлюцинаций нейросети.

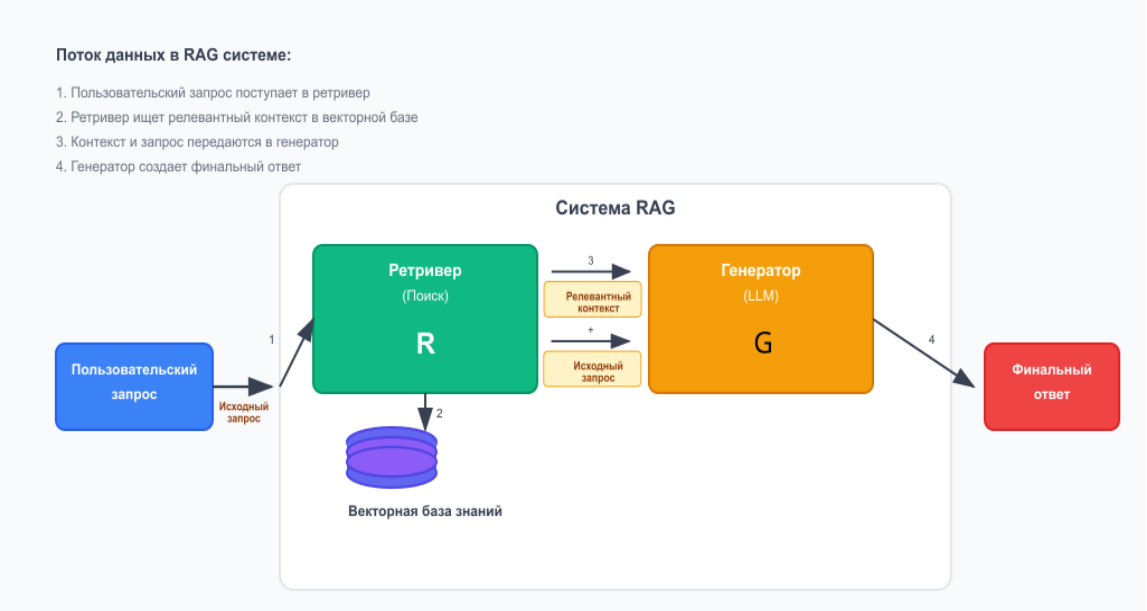

Рисунок 1. Обобщенная схема RAG

Ключевая особенность применения RAG в образовании – высокая цена ошибки.

Недопустимо сообщать абитуриенту неверную дату окончания приема документов.

Поэтому этап подготовки данных является критическим. Работа включает следующие шаги:

- Сбор и первичная обработка. Источниками данных являются официальные правила приема, приказы Министерства науки и высшего образования РФ, программы вступительных испытаний, перечни направлений подготовки, архив вопрос-ответ (FAQ) социальной сети «Вконтакте». На этом данные очищаются от графики, таблиц и нерелевантного форматирования.

- Сегментация документов. Большие документы разбиваются на небольшие фрагменты. Для нашей задачи выбран семантическая фрагментация (разделение по смысловым блокам – главам, параграфам). Размер фрагмента: 500-700 токенов с перекрытием (overlap) в 100 токенов для сохранения контекста на стыках.

- Создание оценочного датасета. Для оценки качества RAG-пайплайна формируется оценочный датасет. Он содержит пары «вопрос – эталонный ответ» с указанием ссылки на конкретный документ-источник.

Таблица 1. Параметры фрагментации документов

Тип документа	**Стратегия фрагментации**	**Размер чанка (токены)**
Правила приема	По главам/пунктам	700
Программы экзаменов	По дисциплинам	500
FAQ (архив)	По парам вопрос-ответ	300

В рамках этапа подготовки данных и проектирования задействован следующий стек технологий: язык программирования Python (библиотеки langchain, pypdf, pandas), модели эмбеддингов от Yandex (YandexGPTEmbeddings) для создания векторных представлений текста, векторное хранилище FAISS (Meta) для быстрого косинусного поиска на этапе прототипирования.

Речевые технологии представлены Yandex SpeechKit, оркестрация тестируется в LangGraph.

Завершение этапа подготовки данных позволит перейти к фазе активной разработки. Ожидается, что внедрение голосового помощника позволит снизить нагрузку на приемную комиссию в «горячий сезон» на 60-70%, обеспечить доступность информации 24/7 и сократить время ожидания ответа приемной комиссии.

Следующими этапами станут: интеграция телефонии, разработка промптов для оркестратора диалогов и нагрузочное тестирование системы.

Заключение. Разработка голосового помощника на базе RAG для приемной комиссии демонстрирует потенциал применения технологий Big Data и искусственного интеллекта в образовании.

Успех проекта напрямую зависит от качества подготовительного этапа – превращения неструктурированных документов в чистую, структурированную базу знаний. Предложенный подход к подготовке данных может быть масштабирован и адаптирован для других подразделений университета.

Список литературы

[1]. Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020;33:9459-9474. DOI: 10.48550/arXiv.2005.11401.

[2]. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS). Long Beach; 2017. DOI: 10.48550/arXiv.1706.03762.

[3]. Сошников Д.В. Технологии искусственного интеллекта в облаке: от классических моделей до генеративных нейросетей. Москва: ДМК Пресс; 2025.

[4]. Yandex Cloud. Документация Yandex SpeechKit (ASR/TTS) [Электронный ресурс]. Режим доступа: https://cloud.yandex.ru/docs/speechkit. (Дата обращения: 15.02.2026).

Авторский вклад

Казаков Алексей Владимирович – разработка концепции RAG-архитектуры, выбор методологии векторизации данных.

DEVELOPMENT OF A VOICE ASSISTANT FOR UNIVERSITY ADMISSIONS BASED ON BIG DATA AND RAG ARCHITECTURE

A.V. Kazakov

PhD Student, A.A. Leonov Technological University (Branch of MIIGAiK)

Abstract. The article discusses the development of an intelligent voice assistant to automate the work of the admission committee of the Leonov University.

Keywords: voice assistant, admission committee, RAG, Yandex SpeechKit, data preparation, vectorization, NLP, automation in education.