Е. Д. Чикирева, Е. В. Белоглазова, И. Н. Глухих
Тюменский государственный университет, г. Тюмень
УДК 004.891

НЕЙРОАССИСТЕНТ В ОБЛАСТИ КОНСУЛЬТАЦИИ ПО ОБРАЗОВАТЕЛЬНОМУ ПРОЦЕССУ

Аннотация. В статье представлена разработка нейроассистента в области консультации по образовательному процессу ТюмГУ с применением RAG-технологий.

Ключевые слова: нейроассистент, LLM, большая языковая модель, нейронные сети, виртуальный помощник, вопросно-ответная система, RAG, поисковая расширенная генерация.

Введение. Почти каждое предприятие нуждается в специалистах, которые будут отвечать на организационные вопросы сотрудников или клиентов. В настоящее время все чаще для автоматизации рутинных процессов на предприятиях и в обычной жизни используются нейросетевые технологии [1, 2].

Исследования применения ответно-вопросных систем, основанных на нейронных сетях, по базам знаний предприятий показывают перспективы их внедрения [3, 4].

В связи с этим было принято решение об исследовании возможности применения в ТюмГУ нейроассистента, который будет отвечать на вопросы студентов по организации образовательного процесса.

Проблема исследования. На данный момент актуальна проблема консультации студентов по организации образовательного процесса, в частности в Тюменском государственном университете. В данной статье описан процесс создания и тестирования нейроассистента в области консультации по образовательному процессу.

Материалы и методы. В исследовании использованы нормативные документы ТюмГУ, находящиеся в общем доступе. Разрабатываемый нейроассистент должен обладать актуальными знаниями об устройстве образовательного процесса Тюменского государственного университета, а также воспринимать вопросы различных вольных формулировок и отвечать на них в «человечной» форме. Для решения этих задач из существующих методов наиболее подходящим является RAG (Retrieval-Augmented Generation) — поисковая расширенная генерация. Его суть заключается в том, что вместе с запросом пользователя в Large Language Model — большая языковая модель (LLM) дополнительно передается документ, предположительно содержащий ответ на заданный пользователем запрос. На основе информации из этого документа модель уже будет генерировать ответ.

Таким образом, в отличии от простого использования общей языковой модели, при применении RAG языковая модель используется лишь для обработки естественного языка и выдачи ответа на нем же, а не для поиска информации в своей общей базе знаний.

Результаты. При выборе языковой модели возник ряд трудностей: невозможнось запуска офлайн-моделей таких, как Llama из-за отсутствия соответствующего оборудования и проблемы цензурирования бесплатных онлайн-моделей (GigaChat от Sber). Наиболее оптимальным выбором стала YaGPT: не блокирует нужные нам темы вопросов и выдает наиболее релевантные ответы. Главным ее минусом стал ограниченный бесплатный период. На рис. 1 представлена схема с принципом работы нейроассистента.

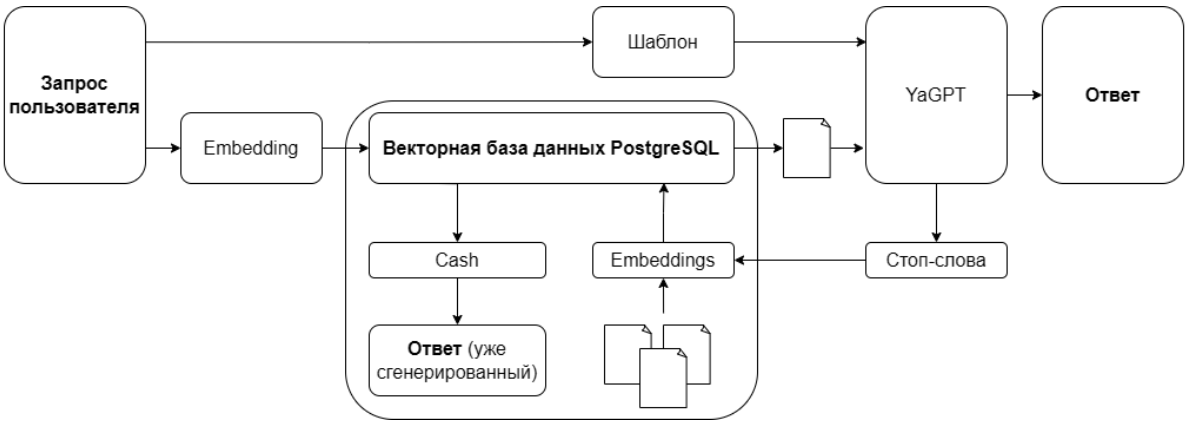

Рис. 1. Схема работы нейроассистента

Нейроассистент создавался на базе Telegram-бота. Сначала бот получает вопрос от пользователя и формирует из него эмбеддинги — числовой вектор размерностью 256 значений. При помощи него ассистент уже будет искать наиболее соответствующий теме вопроса нормативный документ. В качестве эмбеддера был использован соответствующий сервис от Яндекса.

Для хранения текстов нормативных документов, а также их векторных представлений, была создана векторная база данных на языке PostgreSQL. После того как создан эмбеддинг вопроса, на основе косинусного расстояния находятся 3 наиболее релевантных ему документа. По очереди тексты этих документов вместе с изначальным вопросом пользователя передаются в следующий шаблон:

«Ты ассистент в деканате ТюмГУ и помогаешь студентам разобраться в устройстве образовательного процесса. Отвечать на вопрос в тегах question нужно только на основе предоставленного документа в тегах info. Отвечай на вопросы очень кратко. [info] 2.1. Промежуточная аттестация — это часть образовательного процесса, направленная на оценивание результатов обучения по учебным предметам… (продолжение документа) [/info] [question] Чем зачет отличается от дифферецированного зачета? [/question]».

После ассистент дожидается ответа от языковой модели, и, если в нем содержатся ключевые слова, указывающие на отсутствие решения проблемы пользователя в текущем документе, ассистент переходит к следующему документу и повторяет для него все то же самое. Наиболее часто используемые ключевые слова — «простите», «к сожалению», «извините», «прошу прощения», «не указан», «невозможно точно», «нельзя точно».

Если ни в одном документе не было найдено решения проблемы пользователя, то в качестве ответа выдается первый сгенерированный моделью ответ. Иначе пользователю предоставляется полученное из документа решение. Примеры ответов бота показаны на рис. 2 и 4.

Из-за того что в качестве вопросов модели передаются тексты целых документов, довольно быстро расходуются токены. Для снижения вычислительной нагрузки было принято решение о создании дополнительной таблицы в базе данных, которая будет содержать заданные ранее вопросы, их эмбеддинги и ответы модели на них. Таким образом, перед тем как отправить новый вопрос пользователя в языковую модель, ассистент сначала находит наибо-

лее похожий на него ранее заданный вопрос, и если степень их соответствия (косинусное соответствие эмбеддингов) не ниже определенного значения, то пользователю выдается ответ из базы. Пример такого ответа представлен на рис. 3.

С учетом того, что модель выдает не всегда абсолютно верные ответы, в чат-бот была добавлена функция обратной связи от пользователей. На основе этой информации ассистент будет определять, можно ли выдавать данный ответ и для других пользователей, а также поможет совершенствовать проект. Примеры ответов со сбором обратной связи представлены на рис. 5 и 6.

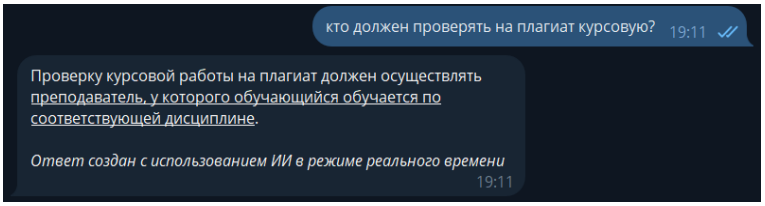

Рис. 2. Пример ответа на прямой вопрос по содержанию документа

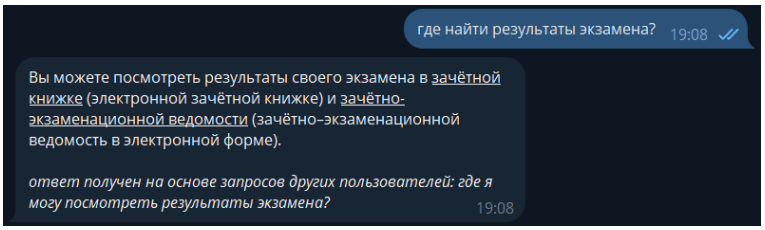

Рис. 3. Пример ответа на основе ранее заданных вопросов

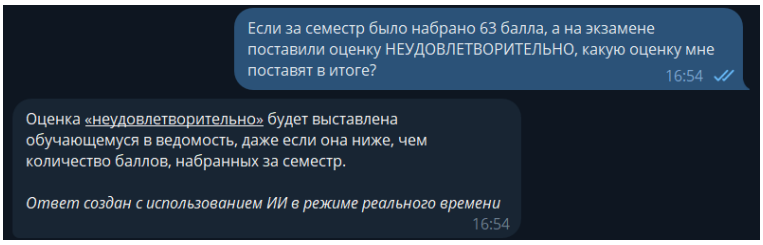

Рис. 4. Пример ответа, не содержащегося в явном виде в документе

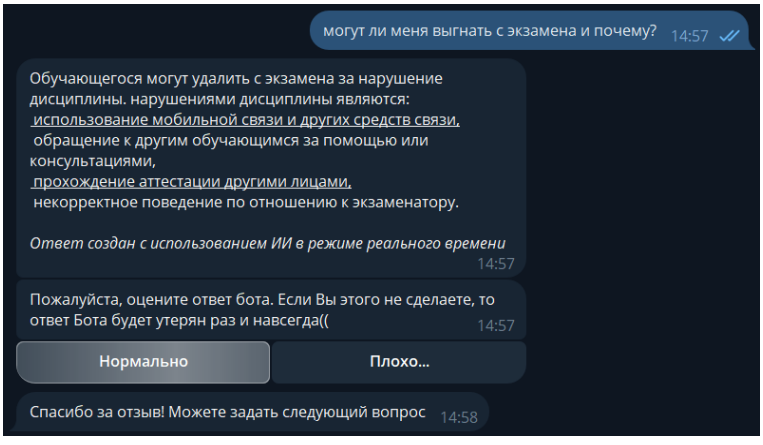

Рис. 5. Пример ответа, удовлетворяющего пользователя

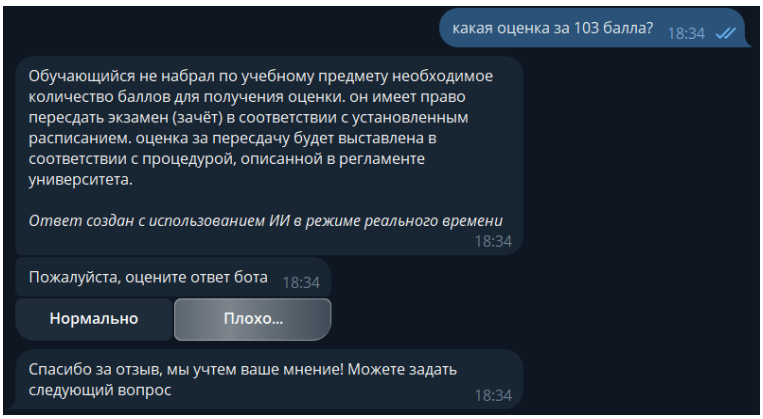

Рис. 6. Пример ответа, неудовлетворяющего пользователя

Заключение. По итогу был разработан чат-бот, отвечающий на вопросы студентов по организации образовательного процесса, в частности по правилам проведения зачетов, экзаменов и выставления баллов. В боте разработана возможность сравнения нового вопроса с предыдущими и при должном уровне соответствия выдачи ранее сгенерированного ответа. Также был реализован сбор обратной связи для более корректной работы бота и дальнейших его улучшений.

СПИСОК ЛИТЕРАТУРЫ

1. Головнин О.К. Нейроассистент для составления индивидуального плана тренировок / О.К. Головнин, А.В. Маркелов // Перспективные информационные технологии (ПИТ 2019). Текст: труды Международной научно-технической конференции / редкол.: С.А. Прохоров (гл. ред.) [и др.]. — Самара: Издательство Самарского научного центра РАН, 2019. — С. 240-242.
2. Атаева О.М. Модель поиска схожих документов в семантической библиотеке / О.М. Атаева, В.А. Серебрякова, Н.П. Тучкова. — Текст: электронный // Научный сервис в сети интернет. — 2021. — № 23. — С. 54-64.
3. Федоров В.О. Большие языковые модели с поисковой расширенной генерацией: обзор и перспективы / В.О. Федоров, Р.А. Поляков. — Текст: непосредственный // Оригинальные исследования. — 2023. — Т. 13, № 12. — С. 43-47.
4. Бородулин И.В. Увеличение точности больших языковых моделей с помощью расширенной поисковой генерации / И.В. Бородулин. — Текст: непосредственный // Вестник науки. — 2024. — Т. 3, № 3 (72). — С. 400-405.