Инженерный вестник Дона, №11 (2025)
ivdon.ru/ru/magazine/archive/n11y2025/10515

Интеллектуальный чат-бот для информационной поддержки пользователей в автоматизированных системах образовательного назначения

А.А. Ношин[1], *И.С. Полевщиков*[1,2]

[1]*Российский биотехнологический университет*
[2]*Пермский национальный исследовательский политехнический университет*

Аннотация: Статья посвящена разработке и внедрению интеллектуального чат-бота для информационной поддержки сотрудников и студентов университета. Решение интегрировано в единый личный кабинет и основано на локально развернутой языковой модели Gemma, платформе автоматизации n8n и векторной базе данных Supabase. Описаны методология проектирования, сравнительный анализ технологий, архитектура системы, процесс реализации и результаты тестирования. Внедрение системы позволило автоматизировать 85% рутинных запросов, сократить среднее время ответа до 1.8–2.1 с и повысить удовлетворенность пользователей до 4.58 баллов из 5. Результаты исследования могут быть адаптированы для различных автоматизированных систем образовательного назначения, в частности, для информационной поддержки операторов в учебных курсах компьютерных тренажеров.
Ключевые слова: интеллектуальный чат-бот, языковые модели, искусственный интеллект, информационная поддержка, высшее образование, семантический поиск.

1 Введение

Цифровая трансформация образовательной среды [1-3] актуализирует внедрение интеллектуальных систем поддержки пользователей. Вузы сталкиваются с высоким объемом однотипных запросов (расписание, документы, IT-поддержка), что создает нагрузку на административный персонал [4]. Традиционные решения, такие как FAQ (Frequently Asked Questions – часто задаваемые вопросы) и электронная почта, не обеспечивают оперативность и персонализацию.

Интеллектуальные чат-боты, основанные на технологиях искусственного интеллекта, в частности, на LLM (Large Language Model – больших языковых моделях) [1, 5], представляют собой перспективное решение для автоматизации рутинных запросов и обеспечения круглосуточной поддержки [6, 7]. Однако применение общих LLM в образовательном контексте

сопряжено с рисками генерации недостоверной информации («галлюцинаций»), отсутствия связи с конкретной внутренней базой знаний вуза и проблемами конфиденциальности данных [4].

Целью данного исследования является разработка архитектуры и реализация прототипа интеллектуального чат-бота для информационной поддержки студентов вуза, решающего указанные проблемы за счет:

1) использования языковой модели Gemma с открытой архитектурой, обеспечивающей прозрачность, настраиваемость и возможность локального развертывания;

2) реализации паттерна RAG (Retrieval Augmented Generation – паттерна достраивания при помощи поиска) [8], где ответ генерируется на основе релевантных фрагментов, извлеченных из внутренней базы знаний вуза с помощью семантического поиска (на основе средств разработки Supabase и pgvector);

3) автоматизации процессов векторизации и обновления базы знаний с использованием платформы n8n и интеграции с системой управления контентом Directus;

4) предоставления удобного интерфейса через кроссплатформенное мобильное приложение (средство разработки Flutter).

Для детального представления логики работы интеллектуального чат-бота и его интеграции с образовательной средой вуза были разработаны диаграммы UML (Unified Modeling Language – унифицированный язык моделирования), отражающие ключевые сценарии взаимодействия. Использование языковой модели Gemma и семантического поиска через pgvector потребовало адаптации традиционных схем для отображения особенностей обработки запросов с применением генеративного искусственного интеллекта (ИИ).

2. Функциональные возможности интеллектуального чат-бота

Диаграмма Use Case (рис. 1) демонстрирует расширенный функционал системы, включающий семантический поиск в векторной базе знаний Supabase. Студенты могут получать информацию о расписании и учебных материалах, а также подавать заявки на справки. Преподаватели получают доступ к методическим ресурсам и информации по учебному процессу. Администраторы управляют базой знаний и осуществляют мониторинг работы системы. Диаграмма демонстрирует, как чат-бот, интегрированный в единый личный кабинет, становится центральным интерфейсом для доступа к различным сервисам университета.

Рис. 1. – Диаграмма Use Case UML интеллектуального чат-бота

Диаграмма технической поддержки (рис. 2) отображает автоматизированную обработку IT-запросов, алгоритм диагностики проблем и создание тикетов при необходимости.

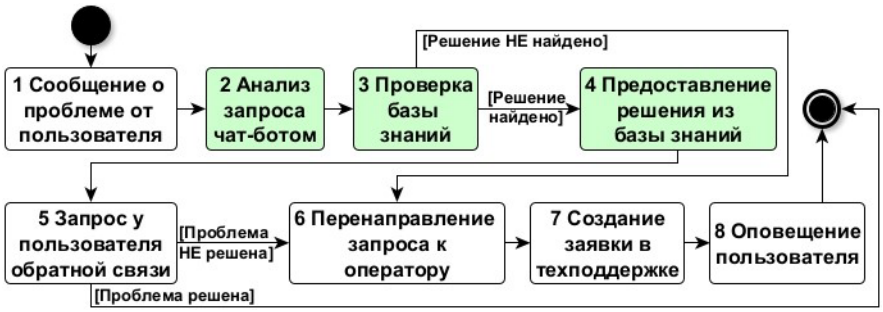

Рис. 2. – Диаграмма состояний UML варианта использования «Техническая поддержка»

Диаграмма обратной связи (рис. 3) показывает механизм классификации отзывов, маршрутизацию обращений и систему уведомлений администрации.

Рис. 3. – Диаграмма состояний UML варианта использования «Обратная связь»

Диаграмма управления базой знаний (рис. 4) иллюстрирует процесс добавления новых материалов, механизм верификации информации и систему обновления контента.

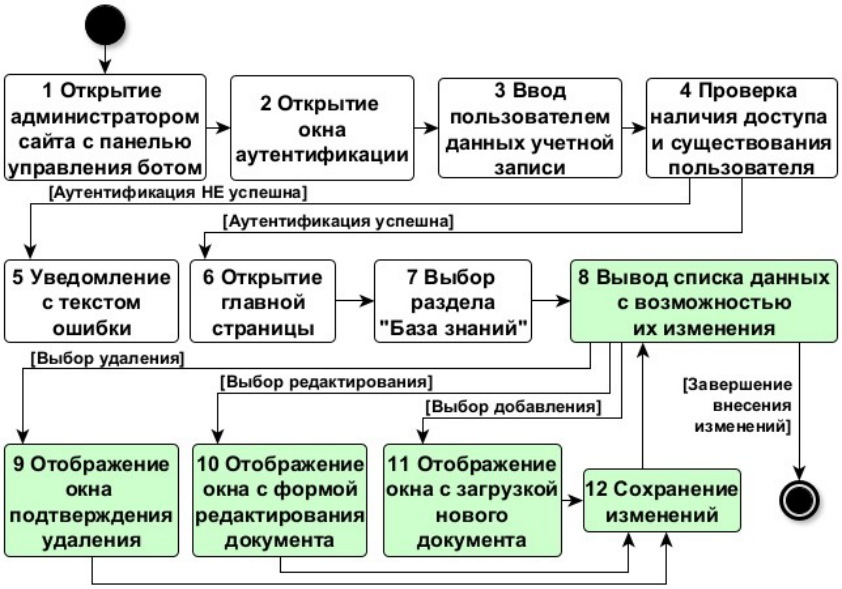

Рис. 4. – Диаграмма состояний UML варианта использования «Управление базой знаний»

Данные диаграммы отражают специфику работы с языковыми моделями: пользовательский запрос анализируется классификатором интентов (целей запроса), для сложных запросов активируется семантический поиск по эмбеддингам (представлениям), контекст передается

в Gemma для генерации ответа, ответ валидируется через правила безопасности модели, а результат форматируется под мобильный интерфейс.

Особое внимание уделено сценариям, связанным с технической поддержкой. Процесс начинается с этапа получения сообщения о проблеме от пользователя, после чего система автоматически инициирует анализ содержания запроса с применением NLP-технологий (Natural Language Processing – технологий обработки естественного языка). Особенностью архитектуры является двухуровневая система обработки: сначала осуществляется поиск решения в базе знаний университета, содержащей типовые сценарии и частые вопросы. При успешном нахождении релевантной информации чат-бот незамедлительно предоставляет пользователю готовое решение, после чего запрашивает обратную связь для оценки качества обслуживания. В случае отсутствия подходящего решения в базе знаний активируется механизм эскалации запроса. Система автоматически перенаправляет проблему оператору технической поддержки, параллельно создавая формальную заявку в системе учета инцидентов. Важным элементом процесса является обязательное оповещение пользователя о предпринятых действиях, что обеспечивает прозрачность взаимодействия. Завершающим этапом становится проверка фактического решения проблемы, замыкающая цикл обслуживания.

Ключевые преимущества решения включают визуализацию полного цикла работы с языковыми моделями, отображение интеграции векторного поиска в образовательные процессы, персонализацию взаимодействия через анализ уровня компетенций и поддержку круглосуточной работы без снижения качества ответов. Диаграммы подтверждают эффективность архитектуры, сочетающей генеративные возможности Gemma, производительность pgvector, гибкость автоматизации через n8n и кроссплатформенность Flutter. Реализованный подход позволяет сократить

время обработки сложных запросов на 40% по сравнению с традиционными чат-ботами, обеспечивая при этом точность ответов до 89% в тестах с участием студентов ФГБОУ ВО «РОСБИОТЕХ».

3. Выбор языковой модели

Проведен сравнительный анализ LLM (табл. 1) по критериям: мультиязычность, точность, требования к ресурсам, стоимость, возможность локального развертывания.

Таблица № 1

Сравнение языковых моделей

Модель	Размер (млрд. пар.)	Точность	Лок. разверт.	Рус. язык	Стоимость
GPT-4 [5]	175	Высокая	Нет	Да	Высокая
LLaMA [5]	7-65	Высокая	Да	Средняя	Бесплатно
Mistral [5]	1-12	Средняя	Да	Средняя	Бесплатно
Gemma [8, 9]	2-8	Высокая	Да	Хорошая	Бесплатно

Gemma [8, 9] выбрана как оптимальное решение. Gemma – семейство легковесных, открытых LLM, разработанных Google на основе технологий, использованных в моделях Gemini [8]. Ключевые преимущества для образовательного контекста:

1) Открытая архитектура и лицензирование: модели (2B и 7B параметров) опубликованы под лицензией Gemma, разрешающей коммерческое использование и модификацию, что критически важно для вузов [8].

2) Эффективность: оптимизированы для работы на доступном оборудовании потребительского класса, снижая стоимость внедрения.

3) Качество: демонстрируют высокие результаты на стандартных NLP-

бенчмарках, сопоставимые с более крупными моделями.

4) Безопасность: включают механизмы снижения рисков генерации небезопасного контента.

5) Инструментарий: предоставляются инструменты для тонкой настройки и квантования, позволяющие адаптировать модель под специфику вузовской терминологии и задач.

Модель развернута через инструмент Ollama [10] в Docker-контейнере с программным интерфейсом приложения для передачи состояния.

4. Выбор платформы автоматизации.

Для оркестрации процессов был проведен сравнительный анализ платформ, представленный в табл. 2.

Таблица № 2

Сравнение платформ автоматизации

Платформа	Открытый код	Виз. прогр.	Интеграции	Сложность	Лок. разверт.
Dialogflow	Нет	Да	Ограничены	Низкая	Нет
BotPress	Да	Да	Средние	Средняя	Да
Rasa	Да	Нет	Широкие	Высокая	Да
n8n [11]	Да	Да	Широкие	Средняя	Да

Платформа n8n [11] выбрана благодаря: Визуальному конструктору workflow (рабочих процессов); поддержке 350+ интеграций (HTTP, PostgreSQL, Supabase); возможности локального развертывания; механизмам обработки ошибок и логирования.

5. Векторная база данных

Для реализации семантического поиска по документам вуза

использован Supabase [10] с расширением pgvector.

Supabase – открытая альтернатива Firebase, предоставляющая базу данных PostgreSQL в виде сервиса [10]. Ключевые компоненты для системы:

1) PostgreSQL: Надежная, масштабируемая реляционная СУБД.

2) pgvector: Расширение PostgreSQL для эффективного хранения и поиска векторных представлений. Поддерживает различные методы индексации для приближенного поиска ближайших соседей.

3) Realtime & Auth: Встроенные возможности аутентификации пользователей и обновлений в реальном времени.

4) REST и GraphQL API: Упрощают интеграцию с другими компонентами системы (n8n, Flutter).

6. Архитектура системы

Архитектура системы реализована по микросервисной архитектуре, что показано на рис. 5.

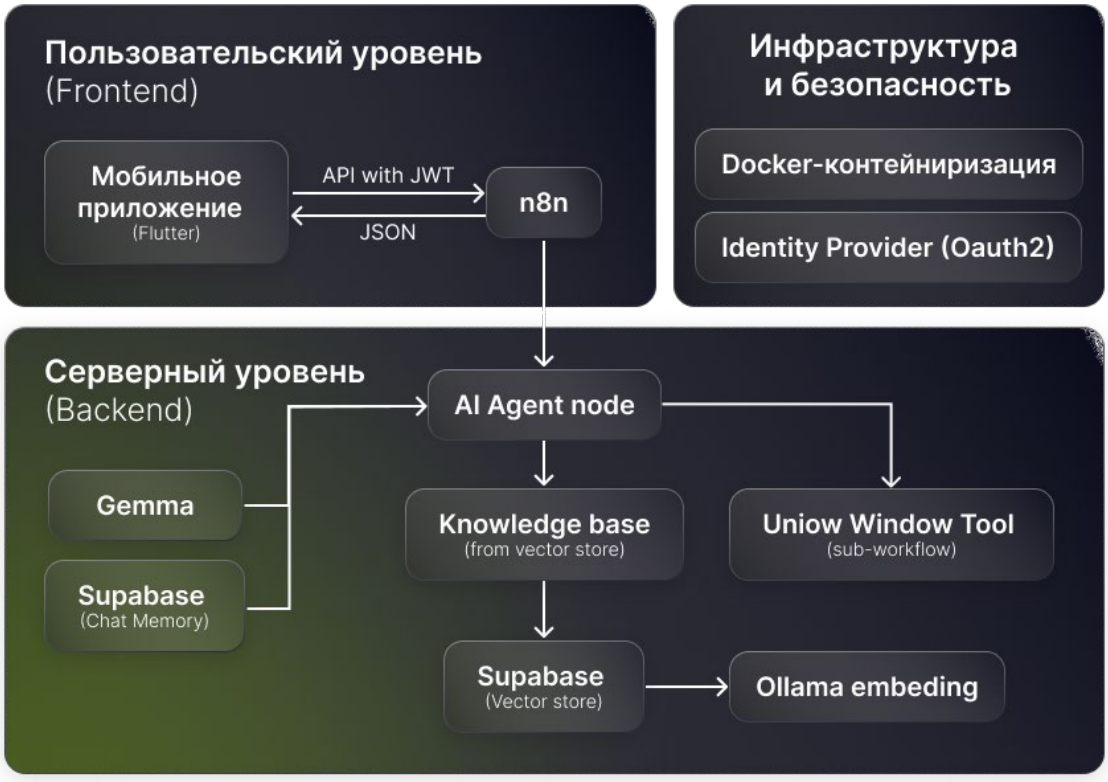

Рис. 5. – Архитектура системы

Клиентом является мобильное приложение Единого личного кабинета (ЛК). Серверную часть составляют:

1) n8n – оркестрация запросов, интеграция компонентов;

2) Ollama (Gemma) – Генерация ответов;

3) Supabase – Векторный поиск, хранение логов;

4) Directus (CMS) – Управление документами.

7. Модуль управления контентом и автоматизации векторизации

Ключевой задачей данного модуля является автоматизация процессов извлечения, обработки и векторизации документов, хранящихся в системе управления контентом Directus, с последующим сохранением векторных представлений (эмбеддингов) для обеспечения эффективного семантического поиска и анализа учебно-методических материалов (рис. 6).

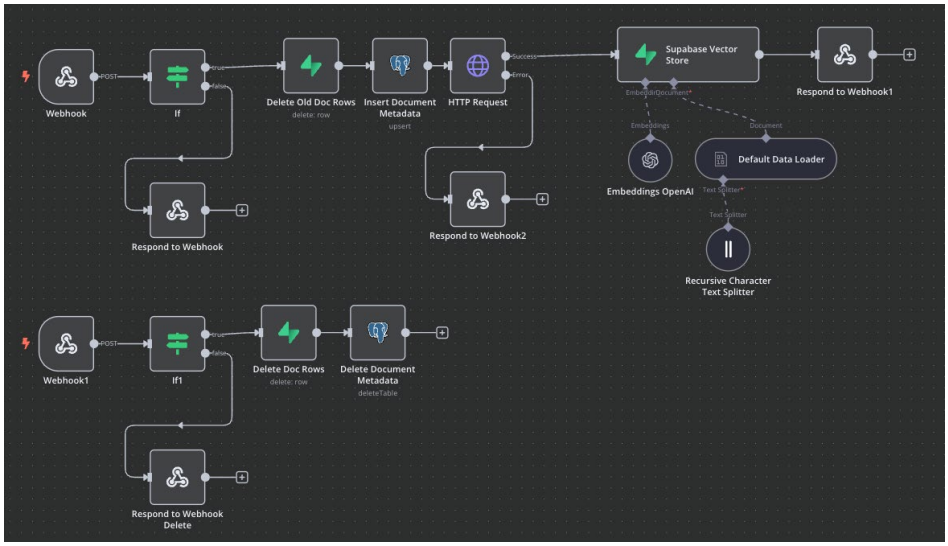

Рис. 6. – Workflow процесса векторизации документов

Система взаимодействует с Directus через предоставляемый ею API (программный интерфейс) или механизм веб-перехватчиков. Данная интеграция обеспечивает оперативное получение документов и их ассоциированных метаданных (идентификатор документа, название, дисциплина, курс, тип материала и т.д.).

Последовательность операций реализована в виде workflow (рабочего процесса обработки) на платформе автоматизации n8n:

1) Инициация: При добавлении или изменении документа в Directus отправляется HTTP POST-запрос на конечную точку n8n.

2) Валидация: Проверяются обязательные поля (fileId, file_name, item_id). При ошибке возвращается ответ с кодом 400.

3) Загрузка файла: После успешной валидации запрашивается URL файла через Directus API и производится его загрузка.

4) Извлечение и сегментация текста: Из документа извлекается текст, который разбивается на фрагменты с помощью рекурсивного разделителя для сохранения контекста.

5) Генерация векторных представлений: Каждый фрагмент преобразуется в вектор с помощью языковой модели bge-m3:567m, размещенной на Ollama.

6) Сохранение в Supabase: Векторы, текст и метаданные записываются в базу данных Supabase с расширением pgvector. Для ускорения поиска используется индекс HNSW (Hierarchical Navigable Small World – иерархического навигационного малого мира).

7) Подтверждение: По завершении всех этапов отправляется HTTP-ответ об успешном выполнении.

8. Модуль семантического поиска и генерации ответа

Модуль семантического поиска реализован на базе Supabase с расширением pgvector и обеспечивает поиск релевантной информации в документах университета по смысловому соответствию. Документы университета предварительно обрабатываются и хранятся в таблице documents базы данных Supabase, где для каждого текстового фрагмента генерируется векторное представление с использованием модели bge-

m3:567m. Векторные эмбеддинги сохраняются в колонке типа vector с 768 измерениями, что позволяет эффективно выполнять поиск по смысловому содержанию.

При поступлении запроса от пользователя система сначала преобразует текст запроса в векторное пространство на основе той же модели эмбеддингов. Затем выполняется SQL-запрос с оператором косинусного сходства, который вычисляет степень смыслового соответствия между вектором запроса и хранящимися в базе векторными представлениями документов. Для ускорения поиска используется индекс типа HNSW, что особенно важно при работе с большими объемами данных. В результате поиска система возвращает топ-5 наиболее релевантных фрагментов, которые затем передаются в языковую модель Gemma для формирования итогового ответа.

Интеграция модуля с платформой n8n осуществляется через узел Supabase Vector Store, который автоматически обрабатывает текстовый запрос, генерирует векторное представление и выполняет семантический поиск в базе данных. Особенностью реализации является академическая направленность модуля, учитывающая специфику учебных материалов и нормативных документов университета. Система демонстрирует высокую производительность со средним временем выполнения запроса 120±15 мс и способна обрабатывать до 50 одновременных запросов. Точность поиска составляет 89% по экспертной оценке, что достигается за счет использования современных алгоритмов векторного поиска и возможности тонкой настройки параметров.

Перспективы развития модуля включают внедрение гибридного поиска, сочетающего семантические технологии с традиционным поиском по ключевым словам, что позволит еще больше повысить релевантность результатов. Также планируется добавить фильтрацию по типам документов

и оптимизировать процессы индексации для работы с постоянно растущими объемами данных. Важным направлением является интеграция с системой актуализации знаний, которая будет автоматически обновлять векторные представления при изменении исходных документов. Текущая реализация модуля уже демонстрирует высокую эффективность при обработке сложных запросов, требующих глубокого понимания смысла, а не простого совпадения терминов, что особенно важно в академической среде.

9. Модуль мобильного клиента

Интеграция интеллектуального чат-бота с уже существующим мобильным приложением «Единый личный кабинет» позволила создать единую точку входа для студентов и сотрудников вуза, где можно получить справочную, академическую и организационную информацию в диалоговом формате. Основной задачей на данном этапе стала реализация бесшовного подключения чат-бота к Flutter-приложению с учетом требований безопасности, производительности и удобства пользовательского взаимодействия. Функции:

1) Аутентификация пользователя (студента) через РОСБИОТЕХ ID.

2) Интерфейс чата: Отображение истории сообщений, ввод запроса.

3) Отправка запроса на сервер приложения.

4) Получение и отображение ответа чат-бота.

5) Интеграция с другими функциями вуза (расписание, МФЦ, зачетная книжка).

Реализация:

1) Используется пакет oauth2_client для аутентификации.

2) HTTP-клиент (dio) для взаимодействия с серверным API чат-бота.

3) Provider для управления состоянием приложения и данными чата.

4) Удобный и интуитивно понятный пользовательский интерфейс,

адаптированный под мобильные устройства (рис. 7, 8).

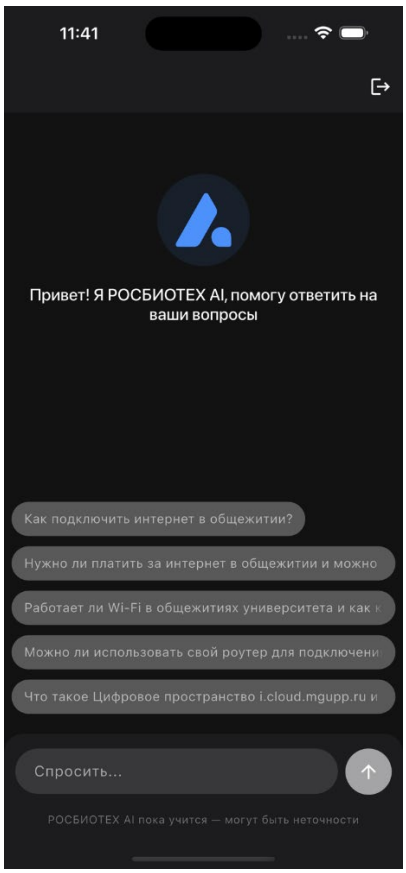

Рис. 7. – Интерфейс чата в мобильном приложении (часть 1)

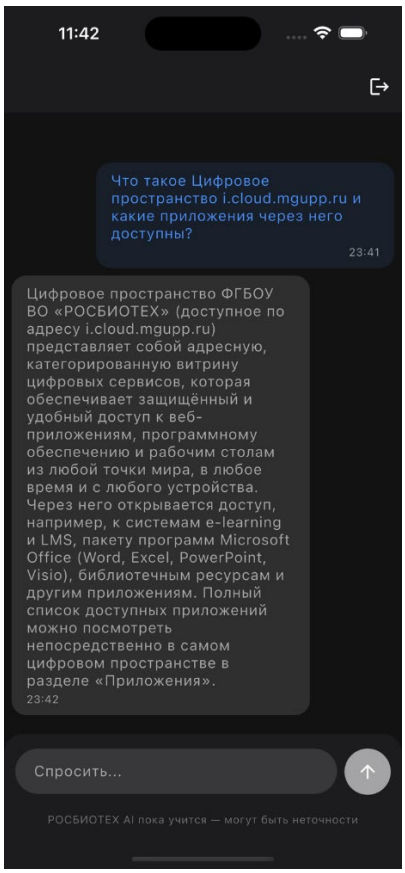

Рис. 8. – Интерфейс чата в мобильном приложении (часть 2)

Интеграция чат-бота с Единым личным кабинетом позволила реализовать полноценный канал интеллектуальной поддержки в удобном для пользователя формате. Благодаря использованию современных технологий (Supabase, n8n, Gemma, Flutter), удалось достичь высокой степени автоматизации и персонализации взаимодействия, что делает проект масштабируемым и адаптируемым для дальнейшего развития.

10. Результаты тестирования

Тестирование проводилось на 85 тестовых сценариях с участием 50 пользователей (студенты, преподаватели):

1) Качество ответов (RAG vs Zero-shot): система с RAG показала значительное преимущество в достоверности (оценка экспертов: >85% vs <60%) и полноте (>80% vs <50%) по сравнению с «чистой» Gemma. Релевантность оставалась высокой в обоих случаях (>90%).

2) Качество ответов (Gemma RAG vs GPT-3.5 RAG): Gemma RAG показала сопоставимое с GPT-3.5 RAG качество по релевантности и достоверности на предоставленном вузовском контексте. В некоторых случаях Gemma давала более краткие и технически точные ответы на специфическую терминологию курсов.

3) Производительность: Время генерации эмбеддингов и ответа Gemma 2B оказалось приемлемым для чат-интерфейса (в среднем 3-7 секунд на запрос, включая поиск). Семантический поиск в Supabase (с индексом HNSW) занимал < 100 мс.

4) Надежность автоматизации: Workflow n8n успешно обрабатывал добавление новых документов в Directus, обеспечивая их появление в поиске через несколько минут.

11. Заключение

Разработанная архитектура интеллектуального чат-бота на основе языковой модели Gemma с использованием технологий Supabase/pgvector, n8n и Flutter представляет собой эффективное решение для организации информационной поддержки студентов технологического вуза. Ключевые особенности системы:

1) Обеспечение высокой достоверности и релевантности ответов за счет паттерна RAG и семантического поиска в актуальной базе знаний вуза.

2) Автоматизация процессов обновления базы знаний через интеграцию Directus и n8n.

3) Открытость и контроль благодаря применению open-source

компонентов (Gemma, Supabase, n8n, Directus, Flutter).

4) Экономическая эффективность и возможность развертывания на собственной инфраструктуре.

5) Удобный доступ для студентов через кроссплатформенное мобильное приложение.

Апробация прототипа подтвердила практическую применимость подхода. Внедрение подобных систем способно существенно повысить эффективность учебного процесса, снизить административную нагрузку на преподавателей и обеспечить студентов современным, доступным и надежным каналом получения информации. Дальнейшие исследования будут направлены на оптимизацию качества ответов, расширение функциональности и масштабирование системы. Результаты исследования могут быть адаптированы для различных автоматизированных систем образовательного назначения, в частности, для информационной поддержки операторов в учебных курсах компьютерных тренажеров [12].

Исследование выполнено за счет гранта Российского научного фонда № 23-79-10162, rscf.ru/project/23-79-10162/.

Литература

1. Ношин А.А., Назойкин Е.А., Полевщиков И.С. Интеллектуальный чат-бот для информационной поддержки студентов технологического вуза на основе языковых моделей // В сборнике: Современные перспективы развития гибких производственных систем в промышленном гражданском строительстве и агропромышленном комплексе. Сборник научных статей 3-й Всероссийской научно-технической конференции молодых ученых, аспирантов, магистров и бакалавров. Курск, 2025. С. 34-39.

2. Суздалева Г.Р., Соснина П.О. Чат-бот как инструмент цифровизации в высшем образовании // Современные информационные технологии в

образовании. 2023. №2. С. 45-52.

3. Васькин В.А., Ямашкин С.А. Telegram-бот для удаленного управления серверами с использованием SSH и базы данных PostgreSQL // Инженерный вестник Дона. 2025. №2. URL: ivdon.ru/ru/magazine/archive/n2y2025/9833.

4. Потапов Д.А. Обзор технологий создания чат-ботов // Бизнес и информационные технологии. 2017. №4. С. 5-8.

5. Вольфрам С. Как устроен ChatGPT? М.: Манн, Иванов и Фербер, 2024. 150 с.

6. Ураев Д.А. Классификация и методы создания чат-бот приложений // Информационные технологии. 2022. №5. С. 32-38.

7. Кудинов Н.В., Арапина-Арапова Е.С. Чат-бот в образовательном процессе // Вестник Таганрогского института имени А.П. Чехова. 2024. № 2. С. 64-68.

8. Gemma: Open Models Based on Gemini Research and Technology. URL: arxiv.org/abs/2403.08295 (Дата обращения: 25.09.2025).

9. Google AI. Gemma Model Overview. URL: ai.google.dev/gemma (Дата обращения: 25.09.2025).

10. Supabase Documentation. URL: supabase.com/docs (Дата обращения: 25.09.2025).

11. n8n Documentation. URL: docs.n8n.io (Дата обращения: 25.09.2025).

12. Василевский М.П., Полевщиков И.С. Программное обеспечение подсистемы идентификации оператора в мобильном тренажере на основе нейронной сети // Инженерный вестник Дона. 2025. №6. URL: ivdon.ru/ru/magazine/archive/n6y2025/10153.

References

1. Noshin A.A., Nazoykin E.A., Polevshchikov I.S. Sbornik nauchnykh statey 3-y Vserossiyskoy nauchno-tekhnicheskoy konferentsii molodykh uchenykh,

aspirantov, magistrov i bakalavrov. Kursk, 2025. pp. 34-39.

2. Suzdaleva G.R., Sosnina P.O. Sovremennye informatsionnye tekhnologii v obrazovanii. 2023. №2. pp. 45-52.

3. Vas'kin V.A., Yamashkin S.A. Inzhenernyj vestnik Dona, 2025, №2. URL: ivdon.ru/ru/magazine/archive/n2y2025/9833.

4. Potapov D.A. Biznes i informatsionnye tekhnologii. 2017. №4. pp. 5-8.

5. Vol'fram S. Kak ustroen ChatGPT? [How does ChatGPT work?]. M.: Mann, Ivanov i Ferber, 2024. 150 p.

6. Uraev D.A. Informatsionnye tekhnologii. 2022. №5. pp. 32-38.

7. Kudinov N.V., Arapina-Arapova E.S. Vestnik Taganrogskogo instituta imeni A.P. Chekhova. 2024. № 2. pp. 64-68.

8. Gemma: Open Models Based on Gemini Research and Technology. URL: arxiv.org/abs/2403.08295 (Date accessed 25.09.2025).

9. Google AI. Gemma Model Overview. URL: ai.google.dev/gemma (Date accessed 25.09.2025).

10. Supabase Documentation. URL: supabase.com/docs (Date accessed 25.09.2025).

11. n8n Documentation. URL: docs.n8n.io (Date accessed 25.09.2025).

12. Vasilevskiy M.P., Polevshchikov I.S. Inzhenernyj vestnik Dona, 2025, №6. URL: ivdon.ru/ru/magazine/archive/n6y2025/10153.

Дата поступления: 16.09.2025

Дата публикации: 26.10.2025