Чернышов Д.Д.

РТУ МИРЭА, Москва, Россия

Графсков Ф.С.

РТУ МИРЭА, Москва, Россия

Юдина Е.Е.

Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, Московская область, Россия

ГИБРИДНАЯ RAG-СИСТЕМА СПРАВОЧНОГО ОБСЛУЖИВАНИЯ СТУДЕНТОВ: АРХИТЕКТУРА И ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА НА ГЕТЕРОГЕННОМ КОРПУСЕ УНИВЕРСИТЕТСКИХ ДОКУМЕНТОВ

Аннотация. Представлена гибридная интеллектуальная справочная система для студентов, объединяющая структурную обработку расписаний и справочных сущностей с retrieval-augmented generation по неструктурированным университетским документам. Отличительная особенность решения состоит в intent-маршрутизации запросов: факты, для которых существенны группа, дата, подразделение или тип процедуры, извлекаются из нормализованного структурного слоя, тогда как содержательные и обзорные запросы обслуживаются семантическим поиском по векторному индексу. Корпус экспериментального стенда включает 75 исходных документов, 221 текстовый фрагмент, 336 структурированных событий расписания и справочник из 204 кафедр. На фиксированном наборе из 150 запросов сравнивались три режима: языковая модель без внешнего контекста, детерминированный rule-based baseline и предложенный RAG-контур. Для RAG доля полностью корректных ответов составила 66,0 %, частично корректных — 20,67 %, некорректных — 13,33 %; следовательно, 86,67 % ответов содержали применимый релевантный результат. Rule-based baseline обеспечил 36,0 % полностью корректных ответов, а генерация без локального контекста не сформировала полностью корректных ответов на

доменно-специфической выборке. Recall@10 достиг 0,6667. При пяти параллельных пользователях p95 времени отклика составил 2,2426 с. Полученные результаты подтверждают, что совместное использование структурного маршрута, семантического извлечения и контекстно-ограниченной генерации повышает фактическую корректность ответов и обеспечивает прослеживаемость источников в образовательной информационной среде.

Ключевые слова: retrieval-augmented generation, интеллектуальная справочная система, образовательные технологии, семантический поиск, векторная база данных, большая языковая модель, intent-маршрутизация, фактическая корректность.

Abstract. The paper presents a hybrid intelligent information system for students that combines structured processing of schedules and reference entities with retrieval-augmented generation over unstructured university documents. The system uses intent routing: facts dependent on a student group, date, department, or administrative procedure are retrieved from a normalized structured layer, while content-oriented and exploratory questions are processed through semantic search in a vector index. The experimental corpus contains 75 source documents, 221 text chunks, 336 structured schedule events, and 204 department entities. Three modes were compared on a fixed set of 150 queries: a language model without external context, a deterministic rule-based baseline, and the proposed RAG pipeline. The RAG system produced 66.0% fully correct, 20.67% partially correct, and 13.33% incorrect answers; thus, 86.67% of responses contained an applicable relevant result. The rule-based baseline achieved 36.0% fully correct answers, whereas the context-free language model produced no fully correct answers on the domain-specific test set. Recall@10 reached 0.6667. With five concurrent users, the p95 response time was 2.2426 s. The results demonstrate that combining structured routing, semantic retrieval, and context-constrained generation improves factual correctness and source traceability in an educational information environment.

Keywords: retrieval-augmented generation, student information service, educational technology, semantic search, vector database, large language model, intent routing, factual accuracy.

1. Введение

Цифровая инфраструктура университета содержит значительный объем справочной информации: расписания, сведения о кафедрах и институтах, локальные нормативные документы, инструкции, контактные данные и объявления. Формальная доступность этих материалов не означает их операционной доступности для студента. Источники различаются по формату, структуре, периодичности обновления и области применимости; один вопрос может требовать одновременного учета учебной группы, даты, подразделения и действующей редакции документа. Поэтому задача справочного обслуживания определяется не только поиском текста, но и выбором факта, применимого в конкретной ситуации.

Генеративные языковые модели упрощают естественноязыковое взаимодействие, однако их параметрические знания не обеспечивают актуальность локальных университетских данных. Подход retrieval-augmented generation (RAG) связывает генерацию с внешним корпусом и позволяет указывать происхождение фактов [1]. Эффективность такого контура определяется не только возможностями генератора, но и качеством извлечения: нерелевантный фрагмент становится содержательной основой уверенно сформулированного ответа. Современные исследования поэтому рассматривают retrieval, фильтрацию, оценку контекста и генерацию как единую систему [6–10].

Для образовательной организации дополнительную сложность создает гетерогенность источников. Расписание является структурным событием с датой, временем, группой и аудиторией; положение или инструкция представляют собой протяженный текст; карточка кафедры сочетает название, адрес, телефон и электронную почту. Принудительное сведение всех классов

данных к одинаковым текстовым фрагментам упрощает индексирование, но ослабляет точность ответов на запросы, в которых критичны точные поля. В представленной работе используется гибридная стратегия: структурируемые сущности обслуживаются детерминированным intent-контуром, а свободные тексты — семантическим retrieval с последующей генерацией.

Цель исследования — разработать и экспериментально оценить интеллектуальную справочную систему, обеспечивающую контекстно релевантный доступ к неоднородным университетским данным. Научно-практический вклад работы включает: 1) двухконтурную модель представления структурных событий и неструктурированных документов; 2) intent-маршрутизацию с учетом учебной группы и типа справочной сущности; 3) воспроизводимую оценку retrieval, фактической корректности ответов и производительности на едином наборе из 150 запросов. В отличие от универсального чат-бота, система формирует ответ в границах локальной базы знаний и сохраняет связь между репликой, найденным контекстом и источником.

2. Связанные исследования

Базовая архитектура RAG была предложена как объединение параметрической памяти генеративной модели и непараметрической памяти внешнего индекса [1]. Методы плотного поиска показали, что семантические представления позволяют извлекать фрагменты, лексически отличающиеся от формулировки запроса [2]. Для построения таких представлений широко применяются трансформерные модели [4, 5] и архитектура Sentence-BERT, обеспечивающая эффективное сравнение предложений по косинусной близости [3]. Эти результаты сформировали основу современных доменных RAG-систем.

Обзоры RAG выделяют стадии предварительной обработки, retrieval, постобработки контекста и генерации [6]. Self-RAG вводит адаптивное извлечение и самопроверку сформированного текста [8], а Corrective RAG — оценку качества найденных документов и корректирующие действия при

неуверенном retrieval [9]. Отдельное направление связано с диагностикой: RAGAS разделяет оценку релевантности контекста, верности ответа и его соответствия запросу [7]. Для прикладной системы это означает, что итоговая доля корректных ответов должна анализироваться совместно с метриками retrieval и характеристиками нагрузки.

Увеличение контекстного окна само по себе не устраняет проблему отбора фактов. Эксперименты с длинным контекстом демонстрируют зависимость качества от позиции релевантного фрагмента и наличие эффекта «lost in the middle» [10]. Поэтому в университетском справочном сервисе предпочтительнее не передавать генератору весь корпус, а формировать компактный контекст из тематически и процедурно применимых фрагментов. Метаданные типа документа, даты актуализации, подразделения и учебной группы становятся частью алгоритма, а не только описанием источника.

Образовательные RAG-системы уже применяются для навигации по ресурсам университетов [11], учебного вопросно-ответного взаимодействия [13] и работы с изменяющимся авторитетным знанием [12]. Исследования отмечают одновременно ценность привязки ответа к проверенному материалу и необходимость учитывать человеческое восприятие полезности [13, 14]. В отличие от систем, ориентированных на один учебник или один курс, рассматриваемое решение работает с административно неоднородным корпусом, где часть запросов требует не генерации объяснения, а точного извлечения события или реквизита.

3. Материалы и методы

3.1. Архитектура гибридной справочной системы

Система реализована как совокупность независимых сервисов. Клиентский контур принимает запрос через веб-интерфейс или мессенджер и передает его сервису извлечения и оркестрации. Оркестратор выполняет модерацию, нормализацию, определение намерения, учет профиля пользователя и выбор маршрута обработки. Сервис языковой генерации

получает только сформированный контекст и инструкцию ответа. Векторное хранилище содержит эмбеддинги документных фрагментов и связанные метаданные. Такое разделение позволяет независимо обновлять корпус, retrieval и генеративную модель.

Рисунок 1 — Общая архитектура интеллектуальной справочной системы

Intent-маршрутизация выделяет запросы о расписании, кафедрах, контактах, местоположении и учебной группе. Если запрос соответствует структурному сценарию и содержит необходимые слоты, ответ строится по нормализованным данным; при отсутствии группы или другого обязательного атрибута система инициирует уточнение. Для вопросов, не сводимых к точной выборке полей, оркестратор выполняет векторный поиск, фильтрует кандидаты по метаданным и формирует контекст генерации. Гибридность здесь означает не простое сосуществование правил и LLM, а выбор вычислительного пути в зависимости от семантики и институциональной цены ошибки.

Генеративный сервис использует локальную модель Qwen2.5-7B-Instruct в формате GGUF с квантованием Q3_K_M [17]. Температура установлена равной 0,1, размер контекстного окна — 4096 токенов, максимальная длина ответа — 1024 токена. Низкая стохастичность выбрана для воспроизводимости

и минимизации вариативности формулировок. Сервис не рассматривается как самостоятельный источник фактов: его функция состоит в преобразовании отобранного контекста в связный ответ с сохранением ссылок на исходные материалы.

3.2. Формирование корпуса и индекса

Исходные данные получены из открытых университетских источников и включают HTML-страницы, PDF-документы и табличные файлы XLSX. Конвейер загрузки фиксирует адрес источника и момент актуализации. Для расписаний используется отдельный структурный парсер: из HTML-каталога извлекаются ссылки по институтам, а из PDF/XLSX — дата, время, группа, подразделение и аудитория. Записи проходят форматную проверку и сохраняются как события. Свободные тексты разбиваются на фрагменты с перекрытием 150 символов, что уменьшает потерю смысловой связи на границах.

Векторизация выполняется русскоязычной моделью sbert_large_nlu_ru [18], а хранение и выборка представлений — в ChromaDB [19]. Каждый фрагмент связан с идентификатором документа, типом источника и ссылкой на происхождение. Переиндексация выполняется после обновления файлового слоя; тем самым состояние корпуса и состояние индекса синхронизируются в одной процедуре. Профиль стенда приведен в таблице 1.

Таблица 1 — Профиль экспериментального корпуса

Показатель	Значение	Назначение в системе
Исходные документы	75	Открытые материалы университета в PDF, XLSX и HTML
Текстовые фрагменты	221	Семантический retrieval по неструктурированным материалам

Показатель	Значение	Назначение в системе
Структурные события расписания	336	Точные ответы по дате, времени, группе и аудитории
Справочные сущности	204 кафедры; 7 кампусов; 2 филиала	Intent-обработка контактов и местоположений
Перекрытие фрагментов	150 символов	Сохранение локального контекста на границах
Дата контрольной индексации	19.05.2026	Фиксация состояния корпуса для эксперимента

3.3. Экспериментальный протокол

Оценка проведена на фиксированном наборе из 150 естественноязыковых запросов. Набор охватывает расписания и экзаменационные события, сведения о кафедрах и контактах, навигационные вопросы, регламентные процедуры, перефразированные запросы и формулировки с неполными слотами. Для каждого запроса зафиксированы ожидаемый ответ и основной эталонный фрагмент либо структурная запись. Такой формат позволяет раздельно оценить наличие релевантного основания и фактическое качество итоговой реплики.

Сравнивались три режима. В режиме LLM без RAG та же генеративная модель получает только пользовательский запрос и системную инструкцию, не имея доступа к локальному корпусу. Rule-based baseline использует нормализацию, регулярные выражения и детерминированные правила для расписаний, кафедр, телефонов и электронной почты. Гибридный RAG применяет intent-маршрутизацию, структурную выборку и семантический поиск с последующей контекстно-ограниченной генерацией. Во всех режимах сохраняется один и тот же набор запросов.

Для retrieval рассчитывались Precision@k и Recall@k при $k \in \{3, 5, 10\}$. Пусть G_i — множество эталонных элементов для запроса i, а $R_i(k)$ — первые k результатов поиска. Средние метрики определялись следующим образом:

$$Precision@k = (1/N) \Sigma_i |G_i \cap R_i(k)| / k;$$

$$Recall@k = (1/N) \Sigma_i |G_i \cap R_i(k)| / |G_i|.$$

В принятой разметке запрос связывался с одним основным эталонным элементом, поэтому Recall@k интерпретируется как доля запросов, для которых релевантный элемент вошел в top-k. При таком устройстве теста предельное значение Precision@k равно 1/k; следовательно, его уменьшение при росте k является математическим следствием добавления кандидатов, а не признаком потери релевантного элемента.

Итоговые ответы классифицировались как полностью корректные, частично корректные или некорректные. Полностью корректный ответ должен опираться на релевантный источник и не содержать фактических искажений; частично корректный передает применимую часть ожидаемого содержания, но не полностью закрывает запрос; некорректный не позволяет выполнить требуемое информационное действие. Для доли полностью корректных ответов дополнительно рассчитаны 95%-ные интервалы Уилсона. Производительность оценивалась по p50, p95, p99 времени отклика и фактической пропускной способности при 1, 5, 10 и 20 параллельных пользователях.

4. Результаты

4.1. Качество извлечения

Таблица 2 — Метрики retrieval на тестовом наборе (N = 150)

Конфигурация	k	Precision@k	Recall@k	Число запросов
Гибридный RAG	3	0,1867	0,5600	150
Гибридный RAG	5	0,1187	0,5933	150

Конфигурация	k	Precision@k	Recall@k	Число запросов
Гибридный RAG	10	0,0667	0,6667	150

При k = 3 основной эталонный элемент найден для 56,0 % запросов. Расширение выдачи до пяти кандидатов повышает покрытие до 59,33 %, а до десяти — до 66,67 %. Равенство Precision@k ≈ Recall@k/k подтверждает, что тестовая разметка использует один основной релевантный элемент. Практически это означает, что две трети запросов получают требуемое доказательное основание уже в первых десяти результатах, а intent-маршрутизация и структурный слой дополняют векторный retrieval для запросов, где ответ определяется точными полями.

Рост Recall при увеличении k показывает, что релевантные сведения присутствуют в индексе, а основная функция последующей фильтрации состоит в выборе компактного набора без тематического шума. Такой результат согласуется с архитектурным решением: универсальный векторный top-k не используется как единственный механизм ответа, а дополняется распознаванием намерения, метаданными и структурной выборкой.

4.2. Фактическая корректность ответов

Таблица 3 — Распределение качества ответов для трех режимов

Подход	Полностью корректные	Частично корректные	Некорректные	
LLM без RAG	0 (0,00 %)	3 (2,00 %)	147 (98,00 %)	50
Rule-based	54 (36,00	0 (0,00	96 (64,00	

Подход	Полностью корректные	Частично корректные	Некорректные	
baseline	%)	%)	%)	50
Гибридный RAG	99 (66,00 %)	31 (20,67 %)	20 (13,33 %)	50

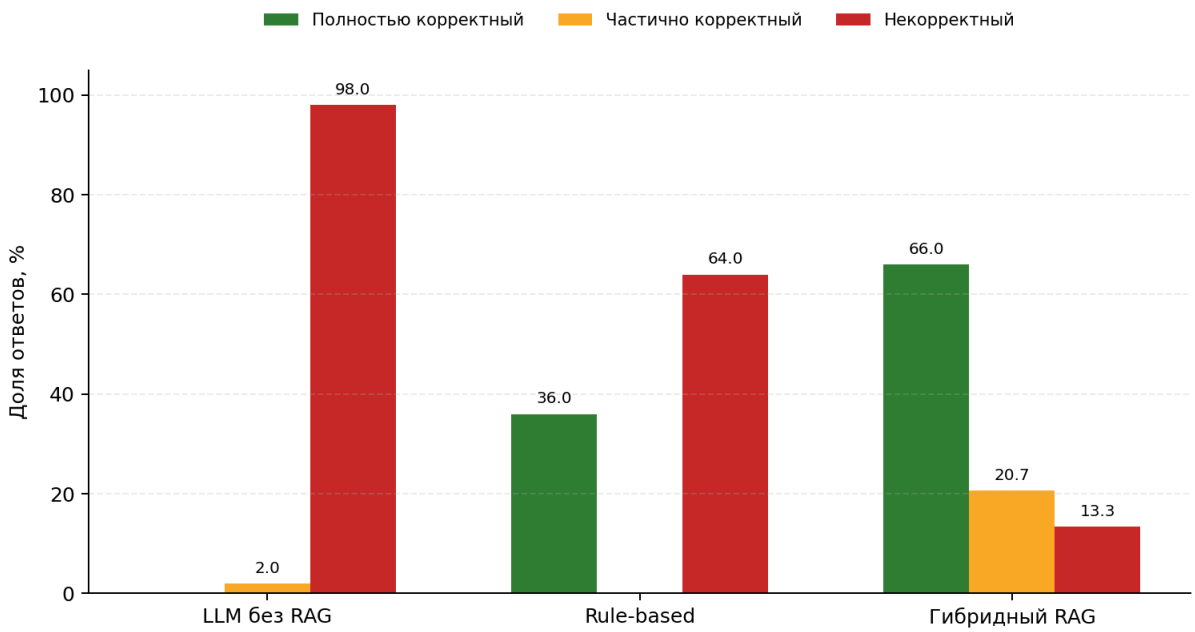

Рисунок 2 — Распределение качества ответов в сравниваемых режимах

Гибридный контур сформировал 99 полностью корректных ответов из 150, что соответствует 66,0 % (95%-ный интервал Уилсона: 58,10–73,10 %). Для rule-based baseline получено 54 полностью корректных ответа — 36,0 % (95%-ный интервал: 28,76–43,94 %). Абсолютный прирост относительно детерминированного baseline составил 30 процентных пунктов, а число полностью корректных ответов увеличилось в 1,83 раза.

Суммарная доля полностью и частично корректных ответов RAG достигла 86,67 % (95%-ный интервал: 80,30–91,20 %). Этот показатель характеризует долю запросов, для которых система возвращает применимый

релевантный результат, даже если ответ может требовать дополнительного уточнения. Доля некорректных ответов составила 13,33 %. Для LLM без RAG отсутствие полностью корректных ответов закономерно: запросы требуют локальных фактов, отсутствующих в параметрической памяти модели. Rule-based механизм надежно обрабатывает узкие шаблонные сценарии, но не покрывает лексическую вариативность и свободные содержательные вопросы.

Полученное распределение подтверждает комплементарность компонентов. Структурный маршрут обеспечивает точность там, где ответ можно вычислить по полям, а RAG расширяет покрытие для перефразированных, обзорных и междокументных запросов. Генератор при этом выполняет функцию интерфейса к доказательному контексту, а не замещает локальную базу знаний.

4.3. Производительность

Таблица 4 — Время отклика и пропускная способность

Параллельных пользователей	p50, с	p95, с	p99, с	RPS
1	0,0160	0,0276	0,0437	64,6048
5	1,4179	2,2426	2,3430	3,3177
10	1,5756	2,6500	2,6544	6,0484
20	4,6719	5,7613	6,2609	4,2333

Профиль нагрузки отражает наличие двух вычислительных путей. Одиночные запросы, разрешаемые прогретым структурным или коротким retrieval-контуром, обслуживаются за десятки миллисекунд. При пяти

параллельных пользователях медиана составляет 1,4179 с, а p95 — 2,2426 с. Наибольшая пропускная способность многопользовательского режима достигается при десяти параллельных пользователях — 6,0484 RPS при p95 2,65 с. При двадцати пользователях p95 возрастает до 5,7613 с, что соответствует насыщению генеративного компонента и очереди межсервисных вызовов.

Для пилотного университетского сервиса результаты показывают возможность интерактивного ответа при умеренной одновременной нагрузке. Микросервисное разделение позволяет масштабировать генеративный сервис независимо от клиента и retrieval, а структурный fast path снижает стоимость типовых запросов. Квантили времени отклика информативнее среднего значения, поскольку фиксируют поведение хвоста распределения в периоды массовых обращений.

5. Обсуждение

Результаты показывают, что для университетской справочной среды решающим является не сам факт подключения языковой модели, а организация доказательного контура. Контекстно-свободная LLM хорошо формирует грамматическую структуру ответа, но не располагает локальными датами, контактами и регламентами. Детерминированные правила возвращают точные значения для предусмотренных сценариев, однако требуют явного покрытия формулировок. Гибридный RAG объединяет сильные стороны обоих подходов: сохраняет точность структурной выборки и допускает семантическую вариативность естественного языка.

Особенно значимым является разделение расписаний и свободных документов. Табличное событие нецелесообразно представлять только как текст: поля даты, группы и аудитории должны участвовать в фильтрации напрямую. Одновременно регламентный документ нельзя полностью выразить набором жестких правил без существенных затрат сопровождения.

Двухконтурное представление данных превращает гетерогенность корпуса из источника ошибок в основание для маршрутизации.

Доля 86,67 % полностью или частично корректных ответов характеризует работоспособность системы на доменно-специфической выборке и согласуется с исходной оценкой релевантности около 87 %. В статье этот показатель разделен на две категории, что делает результат интерпретируемым: 66,0 % ответов полностью закрывают запрос, еще 20,67 % передают применимую часть содержания. Такое представление позволяет отличать фактическую ошибку от ситуации, где требуется дополнительный слот или уточняющая реплика.

Метрики retrieval и качества ответа описывают разные уровни системы. Recall@10 = 0,6667 не задает верхнюю границу итоговой полезности, поскольку часть запросов обслуживается структурным intent-контуром, а генерация может использовать несколько семантически связанных фрагментов. Вместе с тем рост Recall при увеличении k указывает на перспективность переранжирования, раздельных индексов по типам документов и учета временной валидности в функции ранжирования. Эти механизмы естественно продолжают предложенную архитектуру.

Прослеживаемость ответа обеспечивается связью между репликой, выбранными фрагментами и первичными источниками. Для административной информации это имеет самостоятельную ценность: пользователь может проверить дату и область применимости основания. Обновление базы знаний выполняется без переобучения генеративной модели, поэтому система адаптируется к новым расписаниям и редакциям документов через повторный ingest и переиндексацию. Такой жизненный цикл соответствует требованиям к ответственному использованию генеративного ИИ в образовании [15, 16].

Эксперимент использует фиксированное состояние открытого корпуса и не обрабатывает персональные данные участников. Методика воспроизводима на другом образовательном контуре: требуется заменить источники, настроить структурные парсеры и сформировать локальный набор эталонных запросов.

Архитектура, метрики и схема сравнения при этом сохраняются. Для последующей эксплуатации могут быть добавлены автоматическая оценка актуальности, cross-encoder reranking, контрактное тестирование сервисов и раздельные пороги доверия для критичных и обзорных запросов.

6. Практическая реализация и применение

Прототип развертывается в контейнерной инфраструктуре [20]. Сервис оркестрации и векторный индекс допускают CPU-ориентированное исполнение, тогда как генеративный компонент использует GPU с 8–16 ГБ видеопамяти в зависимости от конфигурации. Модель и индексы подключаются как версионируемые артефакты, что сокращает время восстановления и позволяет контролировать соответствие между программным кодом, корпусом и результатами.

Пользовательский сценарий включает четыре стадии: постановку вопроса, распознавание намерения, формирование доказательного контекста и выдачу ответа со ссылками. Если запрос недостаточно определен, система запрашивает учебную группу или иной необходимый параметр. Если сообщение нарушает правила использования, срабатывает модерационный контур. Различение уточнения, отказа и технической ошибки делает поведение сервиса предсказуемым и уменьшает риск восприятия каждого сбоя как содержательного ответа.

Предложенное решение применимо в приемных комиссиях, деканатах, институтах и кампусных справочных службах. Переиспользование архитектуры не требует переноса конкретного корпуса: доменная часть локализована в источниках, парсерах и метаданных, а общая часть включает retrieval, маршрутизацию, генерацию и оценку. Это позволяет последовательно расширять сервис от справочника одного подразделения до единого информационного контура образовательной организации.

Заключение

Разработана и экспериментально проверена гибридная RAG-система справочного обслуживания студентов на гетерогенном корпусе университетских документов. Архитектура объединяет структурный слой расписаний и справочных сущностей, семантический поиск по неструктурированным материалам, intent-маршрутизацию и контекстно-ограниченную локальную генерацию. Это обеспечивает точную обработку формализуемых фактов и гибкий диалог для свободных естественноязыковых запросов.

На наборе из 150 запросов гибридный RAG обеспечил 66,0 % полностью корректных и 20,67 % частично корректных ответов; суммарная доля применимых релевантных результатов составила 86,67 %. По числу полностью корректных ответов система превзошла rule-based baseline в 1,83 раза. Recall@10 достиг 0,6667, а p95 времени отклика при пяти параллельных пользователях составил 2,2426 с. Совместная оценка retrieval, фактической корректности и производительности подтверждает практическую состоятельность предложенного подхода.

Полученные результаты показывают, что интеллектуальный справочный сервис для образовательной организации целесообразно строить как управляемый контур работы с данными, а не как универсальный генератор ответов. Структурная маршрутизация, метаданные, актуализация корпуса и прослеживаемость источников формируют основу достоверности, а языковая модель обеспечивает удобную форму взаимодействия. Архитектура готова к расширению за счет переранжирования, автоматической проверки актуальности и горизонтального масштабирования генеративного сервиса.

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. Vol. 33. P. 9459–9474.

2. Karpukhin V., Oguz B., Min S. et al. Dense Passage Retrieval for Open-Domain Question Answering // Proceedings of EMNLP. 2020. P. 6769–6781. DOI: 10.18653/v1/2020.emnlp-main.550.

3. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of EMNLP-IJCNLP. 2019. P. 3982–3992. DOI: 10.18653/v1/D19-1410.

4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423.

5. Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998–6008.

6. Gao Y., Xiong Y., Gao X. et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2024. arXiv:2312.10997. DOI: 10.48550/arXiv.2312.10997.

7. Es S., James J., Espinosa-Anke L., Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024. P. 150–158. DOI: 10.18653/v1/2024.eacl-demo.16.

8. Asai A., Wu Z., Wang Y., Sil A., Hajishirzi H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection // The Twelfth International Conference on Learning Representations. 2024. URL: https://openreview.net/forum?id=hSyW5go0v8.

9. Yan S.-Q., Gu J.-C., Zhu Y., Ling Z.-H. Corrective Retrieval Augmented Generation // The Twelfth International Conference on Learning Representations. 2024. URL: https://openreview.net/forum?id=JnWJbrnaUE.

10. Liu N.F., Lin K., Hewitt J. et al. Lost in the Middle: How Language Models Use Long Contexts // Transactions of the Association for Computational Linguistics. 2024. Vol. 12. P. 157–173. DOI: 10.1162/tacl_a_00638.

11. Neupane S., Hossain E., Keith J. et al. From Questions to Insightful Answers: Building an Informed Chatbot for University Resources // 2024 IEEE Frontiers in Education Conference. 2024. DOI: 10.1109/FIE61694.2024.10892994.

12. Zheng T., Li W., Bai J., Wang W., Song Y. KnowShiftQA: How Robust are RAG Systems when Textbook Knowledge Shifts in K-12 Education? // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Vol. 2: Short Papers. 2025. P. 183–195. DOI: 10.18653/v1/2025.acl-short.16.

13. Henkel O., Levonian Z., Postle M.-E., Li C. Retrieval-Augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference // Proceedings of the 17th International Conference on Educational Data Mining. 2024. P. 315–320.

14. Kasneci E., Sessler K., Küchemann S. et al. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education // Learning and Individual Differences. 2023. Vol. 103. Art. 102274. DOI: 10.1016/j.lindif.2023.102274.

15. Miao F., Holmes W. Guidance for Generative AI in Education and Research. Paris: UNESCO, 2023. URL: https://unesdoc.unesco.org/ark:/48223/pf0000386693.

16. Newman S. Building Microservices: Designing Fine-Grained Systems. 2nd ed. Sebastopol: O’Reilly Media, 2021. 558 p.

17. Yang A., Yang B., Zhang B. et al. Qwen2.5 Technical Report. 2025. arXiv:2412.15115. DOI: 10.48550/arXiv.2412.15115.

18. AI Forever. sbert_large_nlu_ru: model card. URL: https://huggingface.co/ai-forever/sbert_large_nlu_ru (дата обращения: 25.07.2026).

19. Chroma Documentation. URL: https://docs.trychroma.com/ (дата обращения: 25.07.2026).

20. Docker Documentation. URL: https://docs.docker.com/ (дата обращения: 25.07.2026).