СОДЕРЖАНИЕ

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

LLM (Large Language Model, большая языковая модель) – большая языковая модель, обученная на текстовых данных, используемая для генерации ответов на запросы пользователя.

Эмбеддинг-модель (энкодер, эмбеддер) – модель, преобразующая входные данные (текстовые) в векторное представление (эмбеддинги).

Промпт – запрос в большую языковую модель.

Пайплайн (конвейер) – последовательность действий или процессов преобразований.

Бенчмарк – набор данных для вычисления метрик работоспособности системы в различных сценариях.

RAG (Retrieval Augmented Generation)– подход, при котором большая языковая модель генерирует ответы на основе извлеченной из внешних источников информации

Retrieve (получать, извлекать) и Generate (генерировать) – основные компоненты RAG-пайплайна.

Чанк (фрагмент текста) – извлеченный на этапе Retrieve фрагмент информации из внешнего источника.

ВВЕДЕНИЕ

В современном мире темп технологического прогресса и количество научных публикаций растут с каждым днём, и эффективное освоение научной литературы становится фактором успешной работы в научной и инженерной сферах. Особенно это актуально для специалистов в области информационных технологий. Доступ к научным публикациям зачастую ограничивается не только объёмом информации, но и её формой представления: большинство современных научных статей публикуется на английском языке и содержат много терминов и абстракций.

Задачи, связанные с систематическим изучением больших массивов научной информации, требуют от исследователей значительных временных затрат на поиск, фильтрацию, чтение и осмысление данных. В условиях информационной перегрузки традиционные методы работы с текстами (ручной перевод, чтение и поиск литературы) становятся всё менее эффективными. Это делает актуальным использование искусственного интеллекта для упрощения ключевых этапов взаимодействия человека с научной информацией.

Одним из перспективных направлений является подход Retrieval-Augmented-Generation (RAG), совмещающий механизмы информационного поиска с возможностями больших языковых моделей. RAG позволяет не только находить необходимую информацию в больших корпусах документов, но и формировать на её основе осмысленные и понятные ответы, включая пояснения терминов и перевод [1].

Интересно, что идея разговора с книгой не новая: так, еще в фантастическом романе “Алмазный век” писателя Нила Стивенсона есть описание “Букваря благородных девиц” – уникальной интерактивной книги, с которой можно было разговаривать, задавать ей вопросы [2]. Такой “диалог” с книгой отражает потребность человека в том, чтобы не только прочитать, но и обсудить материал. Современные технологии, в частности RAG, позволяют воплотить в реальность образ книги из романа в виде обучающего материала, с которым можно взаимодействовать в формате беседы.

Целью выпускной квалификационной работы является разработка системы, использующей подход RAG для помощи пользователям в изучении учебных материалов, в частности научных статей в формате PDF. Система должна обеспечивать автоматический поиск и извлечение релевантной информации из предоставленного документа, перевод с английского языка, а также отвечать на вопросы пользователя по документу. Предполагается, что подобный инструмент позволит повысить эффективность изучения научной литературы студентами, снизить языковые и когнитивные барьеры, а также сократить время на обработку и усвоение информации.

Для достижения поставленной цели были сформулированы следующие задачи:

1. Изучить современные подходы к построению систем на основе RAG, определить их применимость к поставленной задаче.
2. Разработать интуитивно понятный и удобный пользовательский интерфейс для взаимодействия с системой.
3. Спроектировать и реализовать сервис на основе RAG и обеспечить взаимодействие с пользователем в реальном времени.
4. Реализовать алгоритм оценки качества системы с точки зрения полноты и точности предоставляемых ответов.

1 ТЕОРЕТИЧЕСКИЕ ОСНОВЫ RETRIEVAL AUGMENTED GENERATION

С развитием технологий обработки естественного языка большие языковые модели показали выдающиеся результаты в различных задачах, включая генерацию текста, машинный перевод, вопросно-ответные системы и обобщение информации. Большие языковые модели обладают способностью усваивать огромные объёмы информации из обучающих данных, в результате функционируя как неявная база знаний. Однако такая архитектура имеет существенные ограничения [1,3].

Во-первых, языковые модели обучаются на статических наборах данных и не могут обновлять свои знания без дорогостоящего дообучения. Во-вторых, модели склонны к галлюцинациям (генерации некорректной или вымышленной информации). Эти проблемы особенно критичны в задачах, где требуется точность, достоверность и актуальность ответов, например, при работе с научными текстами.

Для преодоления этих ограничений в 2022 году был предложен подход Retrieval Augmented Generation (RAG). Основная идея RAG заключается в том, чтобы обогатить запрос пользователя дополнительным контекстом, и только потом передать его на вход генеративной модели. Таким образом, RAG позволяет обеспечить более достоверные, обоснованные и контекстуально релевантные ответы [1].

На рисунке 1 представлена обобщённая архитектура RAG-пайплайна. Она включает в себя три ключевых этапа: разделение корпуса данных на фрагменты (Chunking), поиск релевантных фрагментов (Retrieve) и генерацию ответа на их основе (Generate). На первом этапе система извлекает из внешнего хранилища фрагменты, наиболее подходящие под запрос. Затем фрагменты вместе с пользовательским запросом подаются на вход большой языковой модели, которая генерирует финальный ответ, используя как внутренние знания, так и предоставленный контекст.

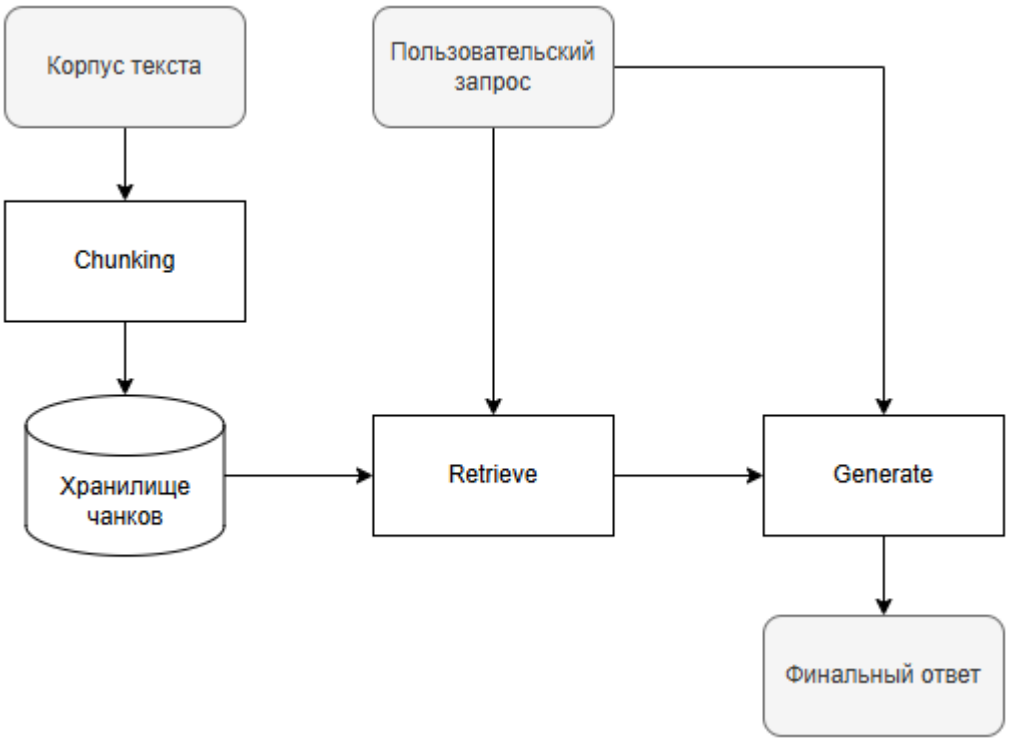


Рисунок 1 - Устройство RAG-пайплайна

1.1 Ключевые этапы RAG-пайплайна

Подход RAG модульный и состоит из нескольких этапов, каждый из которых играет важную роль в обеспечении качества итогового ответа.

1.1.1 Chunking

Этап Chunking (разбиение текста на фрагменты) представляет собой подготовку корпуса текста для поиска. Исходные документы разбиваются на небольшие логически связанные чанки. Задача этого этапа получить фрагменты, которые содержат достаточное количество информации для понимания контекста. Кроме того, они не должны быть слишком длинными, чтобы поместиться в контекст большой языковой модели [3, 4]. Существует два основных подхода к чанкингу:

1. Эвристический подход, основанный на структурных признаках текста. В качестве границ чанков используются элементы форматирования: абзацы, заголовки, списки, таблицы, а также знаки препинания. Такой подход прост в реализации и даёт хорошие результаты при наличии чётко структурированных документов.
2. Семантический подход, ориентированный на сохранение смысловой

целостности фрагментов. Здесь применяются методы на основе векторных представлений текста: текст сначала представляется в виде последовательности векторов, а затем разбивается так, чтобы внутри чанков максимизировалось семантическое сходство, а между ними минимизировалось. [3]

Выбор подхода зависит от специфики задач и характеристик документов: для технических текстов зачастую достаточно эвристик, в то время как для художественных или слабоформализованных текстов предпочтительнее использовать семантический подход.

1.1.2 Retrieval

На этапе Retrieval происходит поиск релевантных к запросу фрагментов текста. Существуют два основных подхода к построению Retrieval-этапа:

1. Традиционные методы информационного поиска, такие как триграммный поиск, TF-IDF и BM25 [5]. Эти методы обеспечивают быструю обработку запросов, но они не учитывают семантику текста, из-за чего показывают низкую точность при наличии различных формулировок или опечаток.
2. Методы на основе эмбеддингов. В этом случае как запрос, так и текстовые фрагменты кодируются в векторном пространстве с помощью эмбеддинг модели. Сходство между векторами измеряется с помощью косинусного расстояния или других метрик. Такой подход учитывает смысл текста, а также он устойчив к опечаткам и синонимам.

На практике для Retrieval этапа часто применяются гибридные схемы, объединяющие преимущества обоих подходов. Например, можно сначала выполнить “грубый поиск”, отобрав 100 кандидатов с помощью BM25, а затем провести их переранжирование с использованием эмбеддинг-моделей. Важно также учитывать особенности корпуса данных: логические блоки, заголовки разного уровня и другие структурные элементы позволяют точнее учитывать логические связи и контекст, повышая качество поиска [6,7].

2 ПРОДВИНУТЫЕ ПОДХОДЫ

С момента появления RAG было предложено множество подходов для повышения его эффективности. Улучшения затрагивают практически все стадии пайплайна. Далее рассмотрим некоторые усовершенствования, сгруппированные по соответствующим этапам.

2.1 Chunking

Главная задача на этапе разделения текста добиться оптимального баланса между объемом фрагментов и сохранением их смысловой целостности. Это особенно важно в доменно-специфичных системах, где плотность информации может значительно варьироваться. Оптимально, если чанк охватывает не отдельные предложения, а их логически связанные группы, так как нарушение границ предложений или смысловых блоков может привести к потере критически важной информации [3]. В связи с этим при разбиении текста важно учитывать не только его содержимое, но и структурные элементы текста. Так, в хорошо оформленных текстах, например в научной или учебной литературе первоначальное разбиение текста целесообразно проводить по заголовкам разного уровня, отражающим иерархию и логику изложения. Однако даже в хорошо структурированных текстах встречаются такие блоки текста, которые превышают по длине контекст большой языковой модели. В таких случаях возникает необходимость в механизме разбиения текста в несколько этапов, где сначала текст делится по очевидным структурным элементам, а затем слишком длинные фрагменты дополнительно обрабатываются с помощью более гибких нейросетевых методов.

2.1.1 Similarity Chunking.

Один из таки методов использует эмбеддинг-модели для группировки связанных по смыслу предложений. В зависимости от имплементаций алгоритм может меняться, но в общем виде он реализуется в два этапа. Первым этапом текст делят на предложения и вычисляют их векторные представления. Затем между соседними парами предложений вычисляют косинусные расстояния, в результате чего получается распределение значений. Наконец, текст разделяется

по пороговому значению, например, выбирается 95-й процентиль по получившемуся распределению [6] (красная линия на рисунке 2).

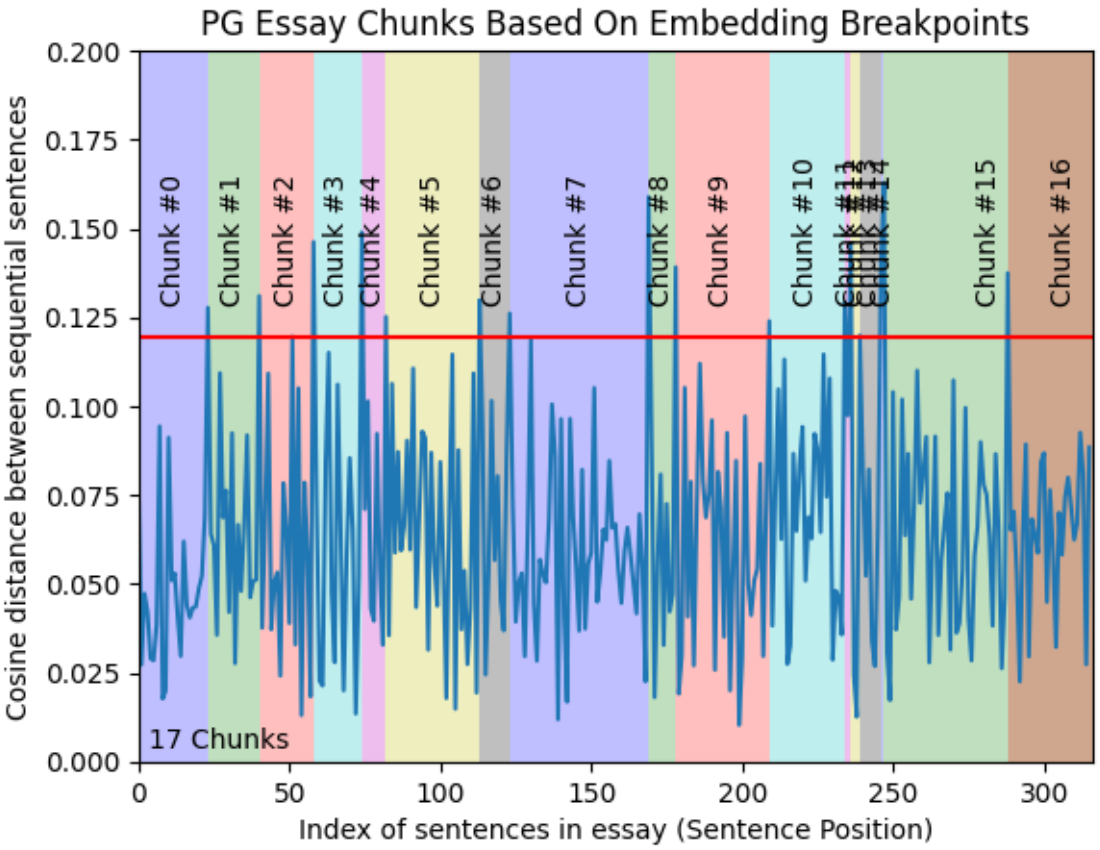


Рисунок 2 – Распределение косинусных расстояний

2.1.2 Perplexity Chunking.

Цель данного метода - сформировать набор чанков (X_1, X_2, ..., X_k), где каждый чанк представляет собой логически связанное объединение исходных предложений [7]. На начальном этапе текст разделяется на набор предложений (x_1, x_2, ..., x_n). Для объединения исходных предложений в чанки модель вычисляет перплексию (PPL) [7] для каждого предложения x_i на основе предшествующих предложений:

$$PPL_M(x_i) = \frac{\sum_{k=1}^{K} PPL_M\left(t_k^i \middle| t_{<k}^i,\ t_{<i}\right)}{K} \quad (1)$$

где K - общее количество токенов в x_i, t_i^k - k-й токен в x_i, а $t_{<i}$ обозначает все токены, предшествующие x_i. Для определения границ чанков алгоритм

анализирует следующую последовательность:

$$PPL_{seq} = (PPL_M(x_1), PPL_M(x_2), \ldots, PPL_M(x_n)\ ) \quad (2)$$

В поисках минимальных значений. Минимумы рассматриваются как потенциальные границы фрагментов, если верно хотя бы одно из следующих утверждений:

1. Значения PPL по обе стороны точки выше, чем в самой точке, и разница хотя бы с одной стороны превышает заданный порог θ:

$$\{i \mid min(PPL_M(x_{i-1}), PPL_M(x_{i+1})) - PPL_M(x_i)\ >\ \theta\} \quad (3)$$

2. Разница между левой точкой и текущей больше θ, а значение справа равно значению в текущей точке:

$$\{i \mid PPL_M(x_{i-1})\ -\ PPL_M(x_i)\ >\ \theta\ and\ PPL_M(x_{i+1})\ =\ PPL_M(x_i)\} \quad (4)$$

2.2 Rewriting

Цель этапа Rewriting заключается в адаптации пользовательского запроса к структуре и стилю данных, хранящихся в индексе. Это особенно важно, если пользователь использует в своем запросе общие или неоднозначные формулировки, которые не соответствуют документам в целевом корпусе [8].

2.2.1 HyDE

Метод HyDE (Hypothetical Document Expansion) предполагает генерацию гипотетического документа на основе запроса. Языковая модель получает похожую инструкцию: “Напиши документ, который мог бы быть ответом на данный вопрос”. Сгенерированный документ не обязательно достоверен, он может содержать фактологические ошибки, так как его задача лишь имитировать релевантный текст. Ожидается, что в результате кодирования документа с помощью эмбеддинг модели удалятся лишние или недостоверные детали и в

результате полученный вектор станет семантически ближе к релевантным фрагментам, чем вектор оригинального запроса пользователя [9].

2.3 Reranking

После первоначального извлечения релевантных документов на этапе Retrieval (например, с помощью алгоритма BM25 [5]), часто применяются алгоритмы переранжирования (этап Reranking), уточняют порядок релевантности результатов. Рассмотрим два наиболее распространенных подхода:

1. Bi-Encoder [10, 11]. Запрос и каждый документ обрабатываются раздельно с помощью эмбеддинг моделей (энкодеров на рисунке 3). На выходе получаются векторы, между которыми вычисляется мера векторной близости. На рисунке 3 показано, как фрагменты текста преобразуются в векторы для вычисления их семантической схожести.

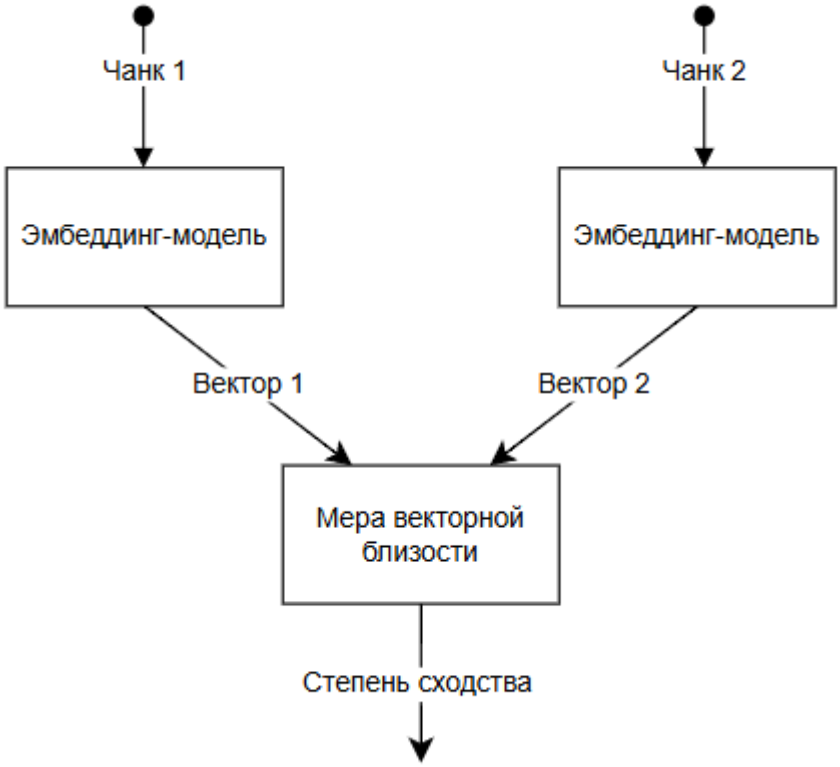


Рисунок 3 — Bi-Encoder

2. Cross-Encoder [10, 11]. Запрос и документ объединяются в одну строку и подаются на вход эмбеддинг модели. Полученный вектор подается на вход предобученного классификатора, который возвращает значение сходства

для фрагментов (см. рисунок 4).

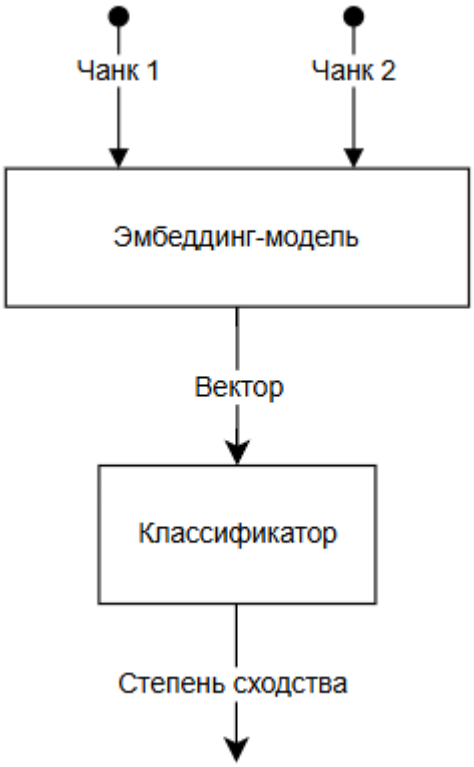


Рисунок 4 — Cross-Encoder

Cross-Encoder показывает себя точнее Bi-Encoder-подхода в задаче реранжирования [12]. Однако за счет своей архитектуры Cross-Encoder существенно более вычислительно затратный: для его работы необходимо попарное вычисление степени сходства между индексированными чанками и запросом, сложность достигает $O(n^2)$. Bi-Encoder в свою очередь позволяет предварительно индексировать все чанки, а затем проводить их сравнения с вектором запроса, таким образом сложность достигает O(n).

2.4 Agentic RAG

Agentic RAG представляет собой усовершенствованный подход к построению RAG. В отличие от традиционного RAG, где модель генерирует ответы на основе единственного шага извлечения данных, Agentic RAG позволяет большой языковой модели самостоятельно формулировать уточняющие запросы, критически оценивать полученные результаты и, при необходимости, повторно обращаться к источникам [12]. Более того, такие системы предполагают использование произвольных инструментов – внешних

функций, к которым имеет доступ большая языковая модель. Такие инструменты работают в пользовательском окружении для реализации разных задач, например веб-поиска, математических вычислений, доступа к базе данных и многого другого [13].

Один из наиболее известных подходов к построению агентных систем описан в статье ReAct. Он сочетает рассуждение (Reasoning) и действие (Acting). В отличие от простого вызова инструментов, ReAct заставляет модель вести цепочку размышлений, которая помогает ей выбирать нужные инструменты и приходить к финальному ответу. В ReAct цепочка рассуждений работает циклически до достижения финального ответа, в рамках одного цикла выполняются следующие шаги:

1. Thought (рассуждение): модель генерирует шаг рассуждения и решает, какое действие необходимо вызвать и с какой целью.

2. Action (действие): модель на основе шага рассуждения самостоятельно вызывает инструмент, подставляя необходимые параметры, либо инициирует генерацию финального ответа.

3. Observation (наблюдение, вывод) Получив результаты инструмента, модель делает выводы, и цикл повторяется (см. рисунок 5) [14].

Question	Musician and satirist Allie Goertz wrote a song about the "The Simpsons" character Milhouse, who Matt Groening named after who?
Thought 1	The question simplifies to "The Simpsons" character Milhouse is named after who. I only need to search Milhouse and find who it is named after.
Action 1	Search[Milhouse]
Observation 1	Milhouse Mussolini Van Houten is a recurring character in the Fox animated television series The Simpsons voiced by Pamela Hayden and created by Matt Groening.
Thought 2	The paragraph does not tell who Milhouse is named after, maybe I can look up "named after".
Action 2	Lookup[named after]
Observation 2	(Result 1 / 1) Milhouse was named after U.S. president Richard Nixon, whose middle name was Milhous.
Thought 3	Milhouse was named after U.S. president Richard Nixon, so the answer is Richard Nixon.
Action 3	Finish[Richard Nixon]

Рисунок 5 – пример цепочки рассуждений ReAct

Для реализации агентных систем такого рода можно использовать библиотеку smolagent [13]. Она предоставляет возможность задавать инструменты для модели, а внутри реализует алгоритм, подобный ReAct: он начинается с пользовательского запроса, а далее в цикле происходит цепочка

размышлений-действий (см. рисунок 6).

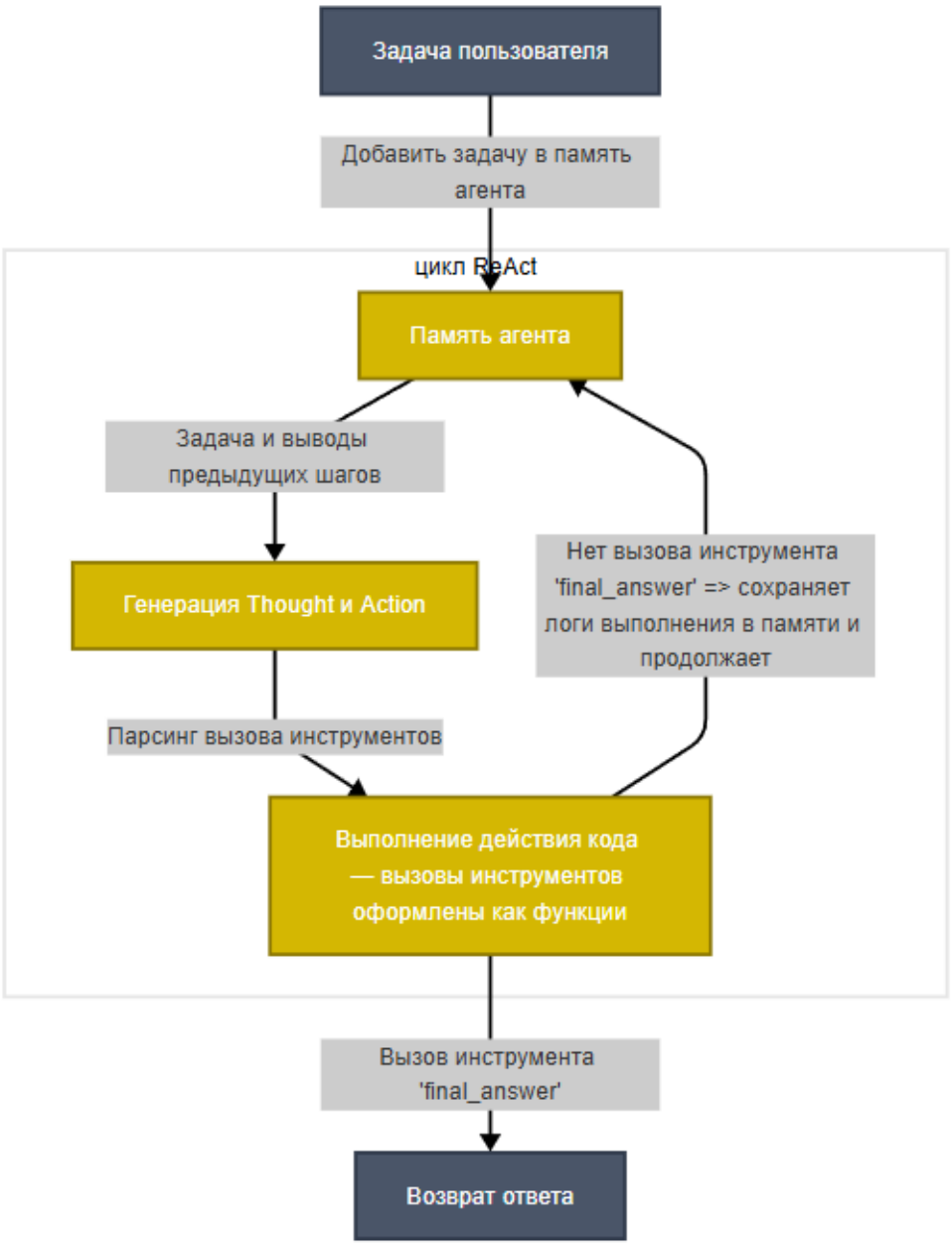


Рисунок 6 – Цикл выполнения smolagents

3 ТОЧКИ ОТКАЗА RAG И МЕТРИКИ ДЛЯ ОЦЕНКИ

RAG-системы обладают значительным потенциалом в задачах генерации ответов с опорой на внешние источники знаний. Однако стоит отметить, что при проектировании и эксплуатации подобных систем возникает ряд характерных точек отказа, способных существенно повлиять на качество ответов. В статье “Seven failure points when engineering a retrieval augmented generation system” авторы выделяют семь наиболее распространённых проблем, с которыми сталкиваются разработчики:

1. Отсутствие необходимого контента. Даже при корректной работе всех компонентов, системе может не удаваться извлечь релевантную информацию, если она отсутствует в базе знаний. Это может быть обусловлено неполнотой корпуса документов.

2. Пропуск высокоранжированных результатов. Алгоритмы извлечения или ранжирования могут не отобрать релевантные фрагменты, даже если они присутствуют в индексе. Такое поведение часто связано с недостаточной точностью эмбеддинг моделей или ошибками в алгоритмах поиска или реранжирования.

3. Ошибки при добавлении чанков в запрос к модели. Релевантные документы быть найдены, но не попали в финальный промпт из-за ограничений контекста модели или неправильной предобработки перед добавлением в финальный промпт.

4. Потеря информации во время разделения текста на чанки: необходимая для ответа информация содержится в корпусе текста, но в процессе индексации получаются противоречивые или бессмысленные чанки.

5. Неправильный формат ответа модели. Эта ошибка возникает в случае, если модель игнорирует инструкцию возвращать ответ в определенном формате, например в виде таблицы или словаря.

6. Неправильная степень специфичности. Ответы, формируемые моделью, могут быть либо чрезмерно обобщёнными, либо излишне детализированными. Это снижает их полезность и не соответствует ожиданиям

пользователя.

7. Неполные ответы: система может предоставлять ответы, содержащие лишь часть необходимой информации, даже если информация для полного ответа доступна в найденных документах.

Кроме того, RAG-системы подвержены деградации качества со временем и требуют периодической переоценки ключевых компонентов. Особенно это критично при изменении пользовательских предпочтений или обновлении источников данных. [3]

Ранее было показано, что при проектировании и внедрении RAG-систем возникает множество потенциальных точек отказа, поэтому для их выявления нужны надежные методы оценки. Поскольку RAG объединяет в себе несколько компонентов, методы оценки должны быть комплексными, охватывающими как отдельные этапы пайплайна, так и работу всей системы в целом. В обзорной статье “Evaluation of Retrieval-Augmented Generation: A Survey” авторы систематизируют метрики для оценки RAG, исходя из необходимых для рассчета данных, и этапов пайплайна, на которых применяются эти метрики. На этапе поиска (Retrieve):

1. Релевантность (найденные чанки ↔ запрос пользователя) – определяет релевантность найденных чанков к запросу пользователя.

2. Точность (найденные чанки ↔ эталонный набор чанков) – определяет, насколько точно система отбирает чанки относительно эталонных чанков, определенных в используемом датасете.

На этапе генерации (Generate):

1. Релевантность (ответ ↔ запрос пользователя) – определяет степень соответствия ответа системы пользовательскому запросу по теме, и требованиям пользователя

2. Достоверность (ответ ↔ найденные чанки) – вычисляет степень достоверности ответа относительно найденных ответов

3. Правильность (ответ ↔ эталонный ответ) – вычисляет степень соответствия ответа эталонному ответу, представленному в датасете [15].

Библиотеки и датасеты предоставляют воспроизводимые, масштабируемые и автоматизированные способы оценки. Библиотека RAGAS разработана специально для оценки качества работы RAG-систем. Она предоставляет как средства для генерации тестового датасета, так и набор метрик для комплексной оценки системы. Метрики RAGAS основаны на подходе LLM-as-a-Judge, в рамках которого большая языковая модель используется в качестве судьи для оценки качества работы системы. Далее рассмотрим ряд метрик из библиотеки RAGAS, используя для них введенные ранее обозначения [16, 17].

3.1.1 Релевантность поиска

Метрика релевантности поиска (Context Precision) рассчитывается, как доля релевантных к запросу чанков. В приведенных ниже формулах рассчитывается данная метрика:

$$Context\ Precision@K\ = \frac{\sum_{k=1}^{K}(Precision@k\ \times\ u_k)}{K} \tag{5}$$

$$Precision@k\ = \frac{true\ positives@k}{(true\ positives@k\ +\ false\ positives@k)} \tag{6}$$

где K – количество найденных чанков, $u_k \in \{0, 1\}$ – значение релевантности для найденного чанка на позиции k, определенное с помощью большой языковой модели. $Precision@k$ является долей релевантных документов среди первых k чанков, таким образом итоговая метрика увеличивается, когда релевантные чанки находятся ближе к началу списка.

3.1.2 Достоверность генерации

Метрика достоверности генерации (Faithfulness) измеряет, насколько фактологически точен ответ относительно извлеченного контекста. Значения варьируются от 0 до 1, где 1 означает, что утверждения в ответе подтверждаются контекстом. Алгоритм расчета метрики начинается с разбиения ответа системы на утверждения с помощью языковой модели. Далее каждое утверждение

проверяется с помощью большой языковой модели на релевантность к извлеченному контексту. Итоговая метрика рассчитывается следующим образом:

$$\text{Faithfulness Score} = \frac{\text{Количество релевантных утверждений}}{\text{Общее количество утверждений}} \quad (7)$$

3.1.3 Релевантность генерации

Метрика релевантности ответа (Response Relevancy) оценивает, насколько ответ соответствует пользовательскому запросу. Высокий балл означает, что ответ полноценно отвечает на запрос без избыточной или нерелевантной информации. Алгоритм расчета метрики начинается с генерации 3-х вопросов на основе ответа системы. Таким образом большая языковая модель старается восстановить запрос пользователя по ответу. Затем измеряется степень схожести между векторами восстановленных запросов и векторным представлением оригинального запроса. Финальная метрика высчитывается, как среднее между полученными на прошлом этапе значениями.

3.1.4 Правильность генерации

Метрика правильности генерации (Answer Correctness) оценивает соответствие ответа эталонному. Балл варьируется от 0 до 1, где 1 означает полное совпадение. Метрика рассчитывается, как взвешенное среднее между фактологическим соответствием и семантическим сходством ответов. Для оценки фактологического соответствия из ответов с помощью большой языковой модели выделяют и сравнивают утверждения, чтобы оценить фактологическое соответствие по следующей формуле:

$$F1\ Score = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (8)$$

где TP – количество утверждений, присутствующих и в эталонном, и в сгенерированном ответах. FP – количество утверждений, которые есть в сгенерированном ответе, но отсутствуют в эталонном. FN – количество

утверждений, которые есть в эталонном ответе, но отсутствуют в сгенерированном. Для расчёта семантического сходства рассчитывают меру косинусной близости между векторами сгенерированного и эталонного ответов. Ответ рассчитывается, как взвешенное среднее между фактологическим соответствием и семантическим сходством.

Несмотря на то, что RAGAS и подобные библиотеки предоставляют масштабируемые и воспроизводимые способы оценки RAG, метрики, основанные на подходе LLM-as-Judge, имеют ряд ограничений и потенциальных проблем. Это связано с тем, что RAGAS стремится компенсировать эффект различия между большими языковыми моделями за счет упрощения форматов вопросов до бинарных с ответами да/нет, однако это не решает проблему возможной вариативности ответов полностью. Также, модели, выступающие в роли судьи, могут галлюцинировать. В связи с этим использование больших языковых моделей в качестве единственного средства оценки не позволяет гарантировать надежность результатов. LLM-as-Judge метрики скорее представляют собой вспомогательный инструмент, позволяющий предварительно оценить систему. LLM-as-Judge лишь имитирует человеческую оценку, и не может быть ее заменой, окончательная валидация системы по-прежнему требует экспертной оценки человеком. Совмещение автоматической оценки с помощью метрик и датасетов с проверкой человеком позволяет повысить корректность и воспроизводимость результатов оценки [16, 17].

4 ПРОЕКТИРОВАНИЕ СИСТЕМЫ

Прежде чем начинать разработку проекта, был проведен этап проектирования архитектуры. На данном этапе необходимо было уточнить требования к функционалу приложения и инструментам разработки.

Для проектирования были использованы средства визуального моделирования, в частности UML-диаграммы. Они позволили формализовать и структурировать требования к системе на раннем этапе, а также визуализировать архитектуру. Это позволило выделить ключевые компоненты и их взаимодействие, что упростило дальнейшую разработку приложения.

В дополнение был разработан прототип пользовательского интерфейса. Он стал важным инструментом в определении логики взаимодействия пользователя с системой. Благодаря ему удалось сформировать представление о сценариях использования системы.

4.1 Дизайн-документ

В качестве основы для проектирования использовался документ ML System Design Doc, предложенный командой Reliable ML [18]. Он был сильно адаптирован под специфику создаваемого приложения и стал основным ориентиром при проектировании клиент-серверной архитектуры. Документ помог разбить разработку системы на этапы, а также сформулировать задачи, которые необходимо выполнить для успешной реализации проекта.

4.2 Постановка задачи

Необходимо создать прототип системы с минимально необходимым функционалом (Minimum Viable Product), который в дальнейшем мог бы быть масштабирован и дополнен новыми функциями. Разработка системы разделена на несколько этапов:

1. Разработка серверной части. Реализация API с поддержкой авторизации, загрузки пользовательских файлов и построения индексации. В этот же этап входила разработка RAG-модуля для обработки пользовательских запросов, а также набор модульных тестов для отдельных классов и функций.

2. Реализация клиентского приложения. Разработка интерфейса на

основе решений ChatPDF [19] и ChatUI [20] с поддержкой диалога с большой языковой моделью и возможностью загрузки PDF-документов.

3. Интеграция и тестирование. Проверка корректности взаимодействия между клиентом и сервером, функциональное и нагрузочное тестирования.

4. Пилотный запуск. Развертывание системы на ограниченной пользовательской выборке, сбор обратной связи, выявление точек роста и недоработок.

4.3 Архитектура серверной части

Для серверной части необходимо реализовать эндпоинты и модули для их функционирования. При проектировании компонентов серверной части применялась UML-нотация. Диаграмма компонентов позволила визуализировать основные части системы и отразить их взаимодействие друг с другом и с базой данных (см. приложение А).

4.3.1 Хранение данных

Для хранения информации о пользователях на ранних этапах проекта было решено использовать SQLite [22] с библиотекой SQLAlchemy [22]. Такой выбор обусловлен необходимостью быстрого прототипирования. SQLite идеально подходит для проекта на начальной стадии, так как не требует развертывания отдельного сервиса. Использование SQLAlchemy позволяет эффективно взаимодействовать с базой данных и упрощает расширение системы при необходимости. В будущем можно будет перейти на более сложное решение, если нагрузка на базу данных увеличится.

Для хранения файлов на этапе пилотной версии было принято решение использовать локальное файловое хранилище. Каждый документ сохраняется в виде обычного файла на сервере, что позволяет сэкономить время на интеграции с облачными хранилищами и подходит для пилотной версии с ограниченной нагрузкой.

Для реализации информационного поиска по фрагментам текста было принято решение использовать модуль FTS5 для SQLite. Это решение подходит под требования проекта и позволяет быстро реализовать полнотекстовый поиск.

В будущем, если потребности в поиске будут увеличиваться, можно рассмотреть переход на более мощные решения.

4.4 Клиент

В качестве основы для реализации клиентской решено использовать ChatUI [20]. Он представляет собой современный чат-интерфейс для общения с LLM. Дополнительно интерфейс приложения должен содержать компоненты для просмотра и загрузки PDF-файлов.

4.5 Пилотный запуск

Пилотный запуск направлен на апробацию системы в условиях, приближенных к реальному использованию. Основной целью пилотной версии является сбор обратной связи от пользователей, чтобы выявить недочеты, узкие места в логике работы, а также сформулировать направления для дальнейшего развития приложения.

4.6 Требования к работе системы

Системные требования рассчитаны с учетом ограниченного количества пользователей пилотной версии приложения (до 10 человек). Сперва была рассчитана пропускная способность. Исходя из предположения об 1 запросе на пользователя каждые 15 секунд, расчетная нагрузка составит:

$$10 \text{ пользователей} \times \frac{1 \text{ запрос}}{15 \text{ секунд}} = 0.67\ RPS \quad (9)$$

где RPS – количество запросов за секунду. Для обеспечения устойчивой работы системы установлена целевая пропускная способность в 1.5 RPS. Это позволит учесть различные сценарии, такие как задержки в сети или повышение активности пользователей. Ожидаемое время отклика API — не более 1 секунды, что обеспечит работу с системой в режиме реального времени. Кроме того, были рассчитаны требования для вычислительных ресурсов:

1. GPU: для запуска эмбеддинг модели deepvk/USER-bge-m3 (359 млн параметров) требуется ~1.44 ГБ VRAM. При пакетной обработке по 32

фрагмента системе потребуется до 4 ГБ VRAM.

2. RAM: достаточно 2–4 ГБ для хранения данных и работы API.

Для хранения файлов пользователей достаточно дискового пространства в 200 мегабайт на каждого пользователя. Кроме хранения загруженных файлов необходимо хранить еще и индексированные чанки для каждого документа, а также базу данных для информации о пользователях и метаданных файлов. Таким образом, для поддержания системы из 10 пользователей понадобится:

$$200\text{МБ} \times 2 \times 10 + 500\text{МБ} \cong 4.5\text{ГБ} \quad (10)$$

4.6.1 Механизмы безопасности

Для обеспечения безопасности доступа к системе в рамках пилотного проекта используется метод предопределенного ключа API, который ограничивает доступ только для заранее определенной группы пользователей. Этот ключ передается в заголовке каждого запроса, и позволяет убедиться, что доступ к системе получают только авторизованные лица.

4.7 UML-диаграммы

Диаграмма вариантов использования позволила выделить ключевые сценарии взаимодействия двух типов пользователей: Пользователя (задаёт вопросы, читает ответы, просматривает документы) и Администратора, который кроме базового функционала обладает правами для настройки системы (см. рисунок 7).

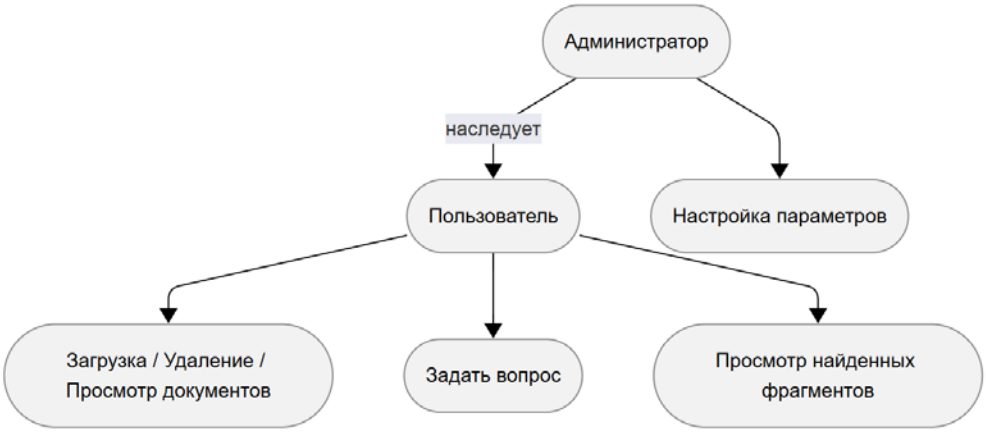


Рисунок 7 – Диаграмма вариантов использования сервиса

Следующим шагом была разработана диаграмма последовательностей. Она иллюстрирует, как объекты системы взаимодействуют друг с другом для выполнения разных сценариев работ, к тому же по ней можно проследить, как данные перемещаются внутри системы (см. приложение Б). В дополнение к созданным диаграммам была составлена диаграмма активностей. Она представляет анализ активностей пользователей, который помог понять, как именно пользователи взаимодействуют с системой в рамках различных сценариев. (см. рисунок 8)

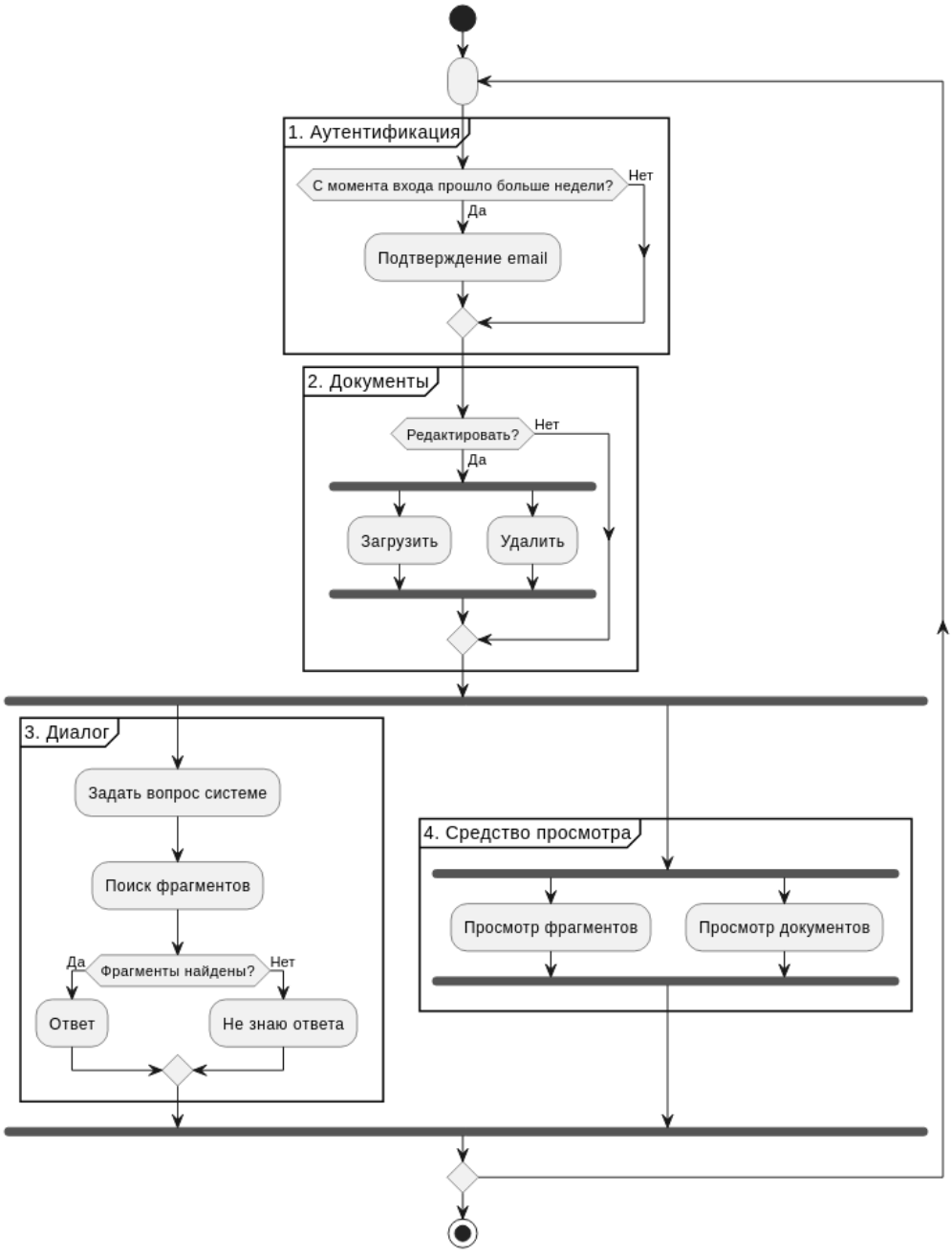


Рисунок 8 – Диаграмма активностей

5 РЕАЛИЗАЦИЯ ПРОТОТИПА

После завершения проектирования был реализован прототип системы. В данной главе приводится описание реализованной системы, охватывающее как серверную, так и клиентскую часть, структуру хранения данных, а также RAG-модуль, играющий ключевую роль в системе.

Прототип реализован с учетом требований и проектных ограничений, описанных в предыдущей главе. В процессе реализации были учтены актуальные практики, ориентированные на промышленные стандарты разработки. Проект прошел несколько итераций разработки и тестирования, в процессе которых некоторые проектные решения были скорректированы. Итоговая архитектура и используемые технологии незначительно отличаются от тех, что были описаны в главе про проектирование системы. В результате были приняты следующие архитектурные и программные решения:

1. Retrieval этап реализован на основе векторного поиска вместо двухступенчатого с применением BM25 [5].
2. Переход на СУБД PostgreSQL из-за необходимости организации векторного поиска.
3. Клиентская часть разработана с нуля, так как адаптация готового решения вроде ChatUI [20] под требования проекта оказалась неоправданно трудозатратной. При этом в качестве основы для стилистического оформления были частично использованы элементы ChatUI.
4. Тестирование системы оказалось нетривиальной задачей, т.к. требовалось реализовать тестирование отдельных компонентов, в частности модуля для общения с СУБД в изолированной среде с “чистой” базой данных. Решение было найдено в использовании библиотеки testcontainers [23]. Она позволяет создавать каждый раз контейнеры с новой базой данных “на лету” прямо в коде проекта.

5.1 Общая структура системы

Разработанная система состоит из двух основных частей: клиентского интерфейса и серверного приложения. Между ними осуществляется обмен

данными по протоколу HTTP, а также через потоковую передачу данных посредством однонаправленного протокола передачи данных Server-Sent Events (SSE) [24]. Ключевым элементом архитектуры является RAG-модуль, обеспечивающий возможность семантического поиска и генерации ответов на естественном языке на основе загруженных документов.

5.2 Серверная часть

Серверная часть реализована на языке Python с использованием асинхронного фреймворка FastAPI. Для работы с базой данных используется библиотека SQLAlchemy. Основной функционал серверной части включает: обработку HTTP- и SSE-запросов от клиента., взаимодействие с RAG-системой, авторизация пользователей на основе JWT-токенов, доступ к хранимым файлам и метаданным. Благодаря асинхронной архитектуре, сервер способен эффективно обрабатывать множество одновременных клиентов.

5.3 Итоговая архитектура

Для формализации архитектуры системы была выбрана C4-нотация. Она предлагает гибкий набор инструментов для проектирования программных систем, и включает в себя несколько уровней детализации:

1. Контекст (Context) – показывает внешние системы и пользователей.
2. Контейнеры (Containers) – архитектура приложения без глубокого погружения в техническую часть. Она отображает основные логические блоки и используемые технологии.
3. Компоненты (Components) – раскрывает архитектуру отдельных контейнеров.
4. Код (Code) – самый низкий уровень абстракции, чтобы показать классы и их связи.

Для описания архитектуры проекта использованы второй и третий уровни: “Контейнеры” и “Компоненты”. Они позволили описать систему в достаточной детализации без избыточных подробностей о кодовой базе проекта. На рисунке 9 представлена диаграмма контейнеров, показывающая общее взаимодействие клиента, сервера, базы данных и RAG-пайплайна.

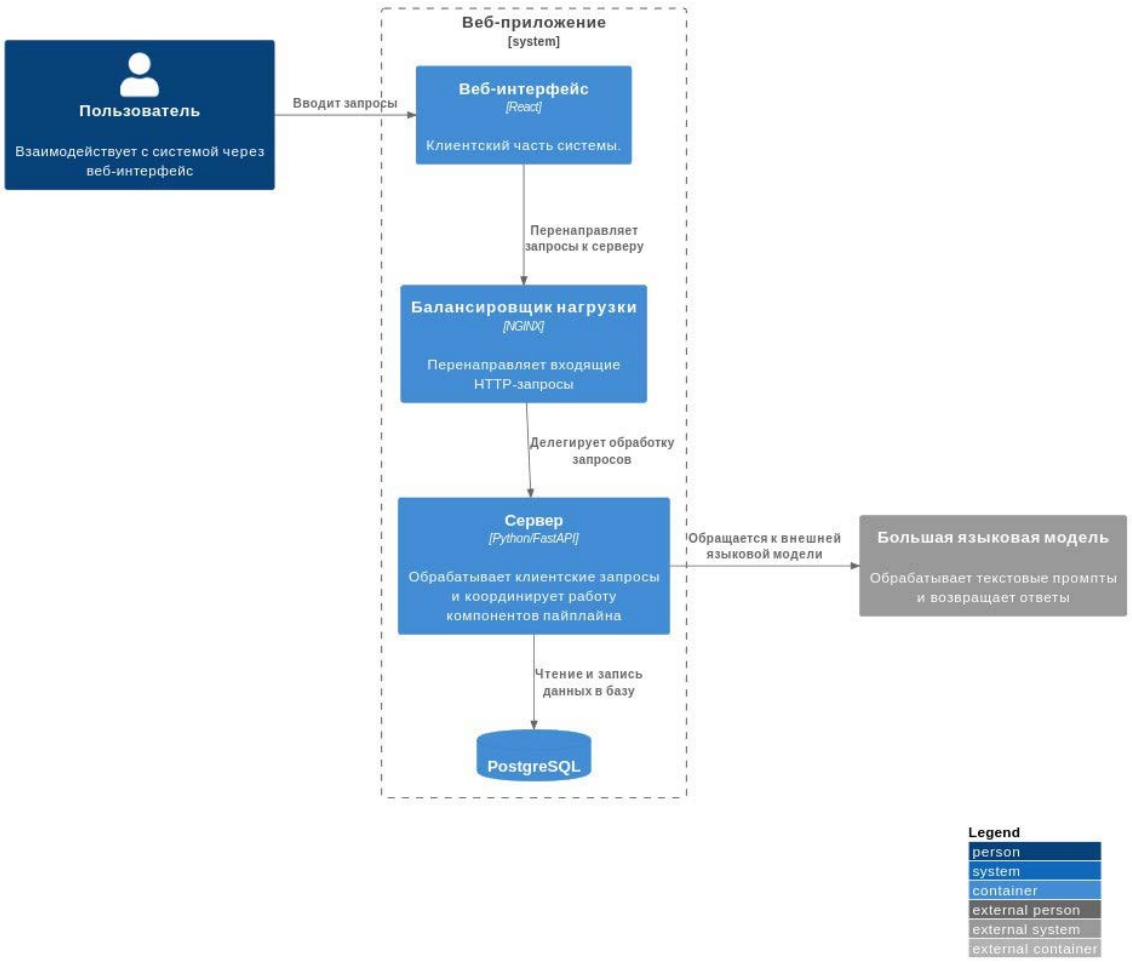


Рисунок 9 – диаграмма C4 (уровень контейнеров)

Следующий уровень детализации представляет компоненты серверной части, где можно выделить несколько ключевых модулей. Центральное место занимает API-сервис, который служит основной точкой входа в систему. Он принимает запросы от клиента, авторизует пользователей, управляет загрузкой файлов, а также инициирует взаимодействие с RAG-пайплайном. Сервисы индексации и генерации RAG для реализации функционала RAG: после загрузки файлы проходят обработку через сервис индексации, а пользовательские запросы поступают в сервис генерации, где происходит поиск релевантных чанков и потоковая генерация ответа с помощью большой языковой модели (см. рисунок 10).

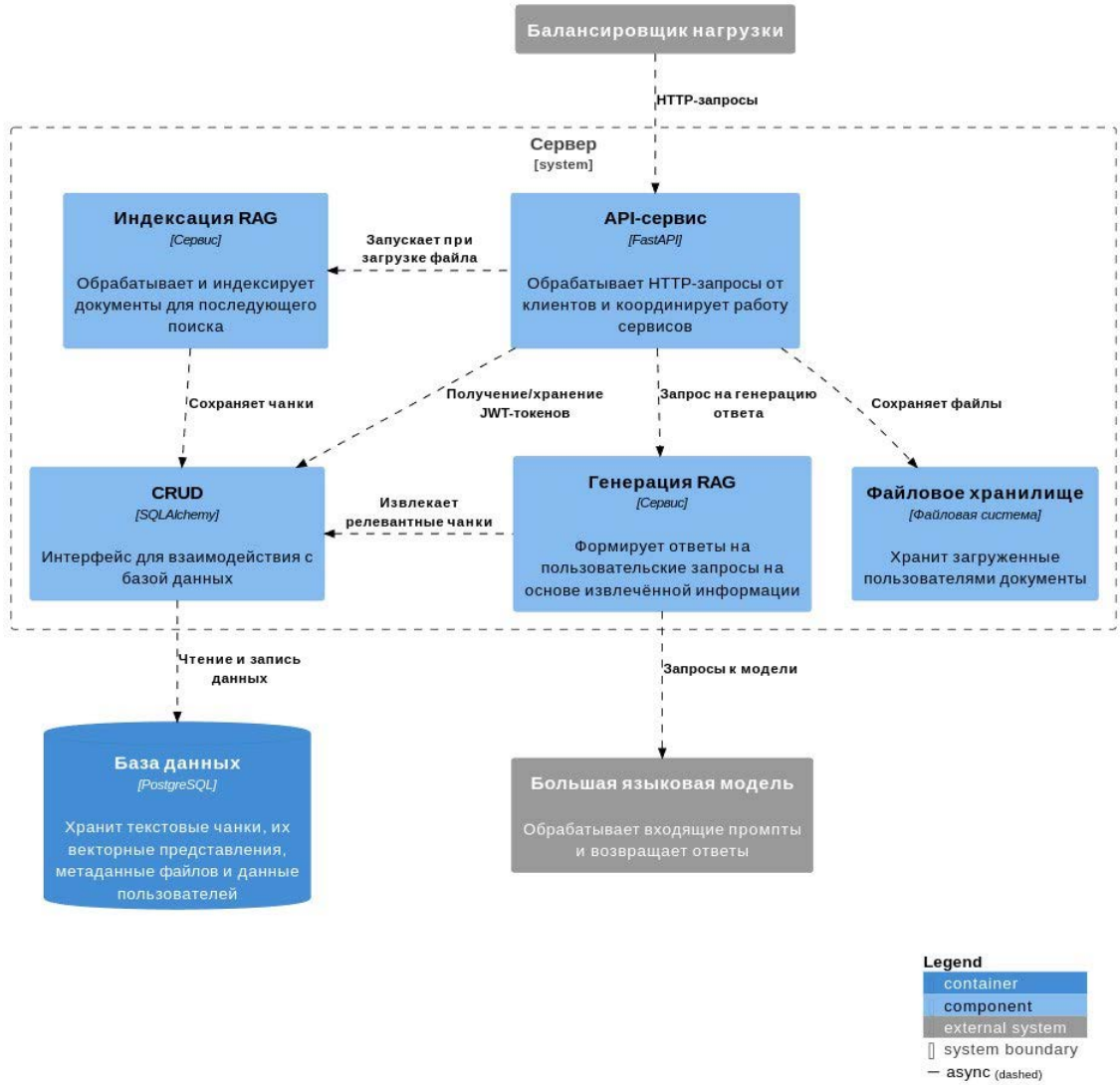


Рисунок 10 – диаграмма C4 (уровень компонентов, сервер)

В дополнение, была разработана более детализированная схема взаимодействия внутри RAG-пайплайна с этапами индексации и генерации. Работа модуля включает два этапа: Индексация и Генерация. Первоначально рассматривался классический подход с использованием двухступенчатого поиска на основе BM25 [5]. Но несмотря на его простоту и эффективность в ряде задач, он имеет ограничение при работе с мультиязычными данными. Поскольку как запросы, так фрагменты в базе данных могут быть представлены как на русском, так и на английском языках, алгоритм поиска на основе BM25 не подходит. В качестве альтернативы был реализован семантический поиск с использованием векторных представлений как для запросов, так и для чанков из базы данных. Поиск основан на Bi-Encoder подходе, и позволяет системе находить

релевантные фрагменты даже с учетом различия в формулировках и языках (см. рисунок 11).

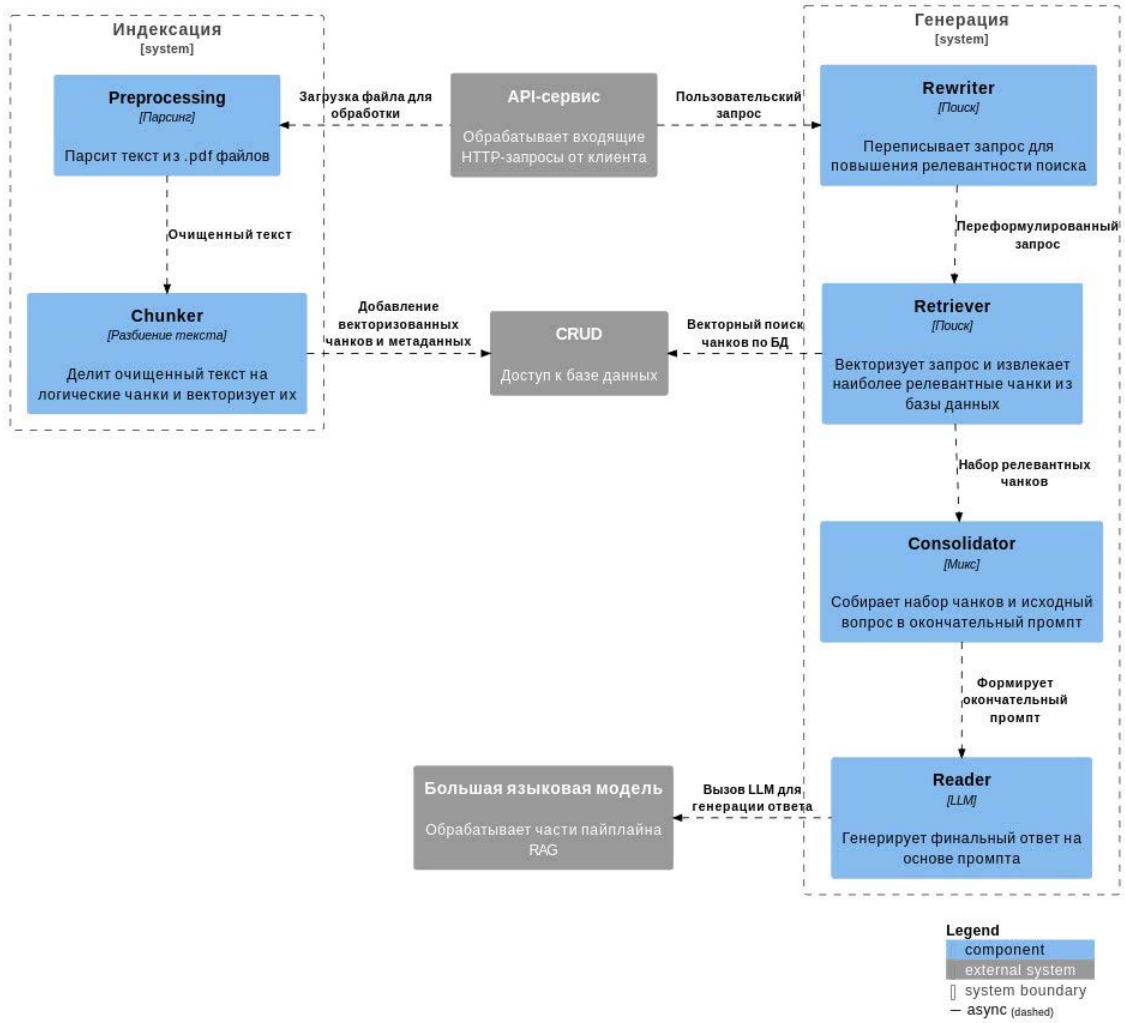


Рисунок 11 - диаграмма C4 (уровень компонентов, RAG-пайплайн)

В ходе преобразования файлов из PDF-формата в текст теряется часть структуре текста, поэтому необходимо учитывать взаимное расположение элементов на странице. Для распознавания макета документа (Layout Analysis) существуют подходы на основе визуальных трансформеров. Они точны, но вычислительно затратны, из-за чего не подходят для задач в реальном времени. В качестве компромисса между точностью и скоростью ответа была выбрана библиотека GROBID [25], использующая CRF-модели (Conditional Random Fields) для извлечения структурированного текста из PDF-файлов. Она преобразует PDF-файлы в XML-разметку, отражающую иерархию документа. Затем оттуда выделяются заголовки разных уровней, а также аннотация,

введение, заключение и др. Текст уже разделен на логические блоки по заголовкам, однако при превышении порога длины чанки дополнительно разделяются с помощью алгоритма семантического чанкинга. Последним шагом к полученным чанкам добавляются соответствующие заголовки, как часть метаинформации.

5.4 Клиентская часть

Клиент реализован с помощью библиотеки React в виде одностраничного веб-приложения (SPA). Его функционал включает интерактивный чат с возможностью ввода и отображения истории сообщений. поддержка потоковых ответов от сервера, регистрация и авторизация пользователей, загрузка и просмотр PDF-документов, навигация по загруженным файлам (см. рисунок 12).

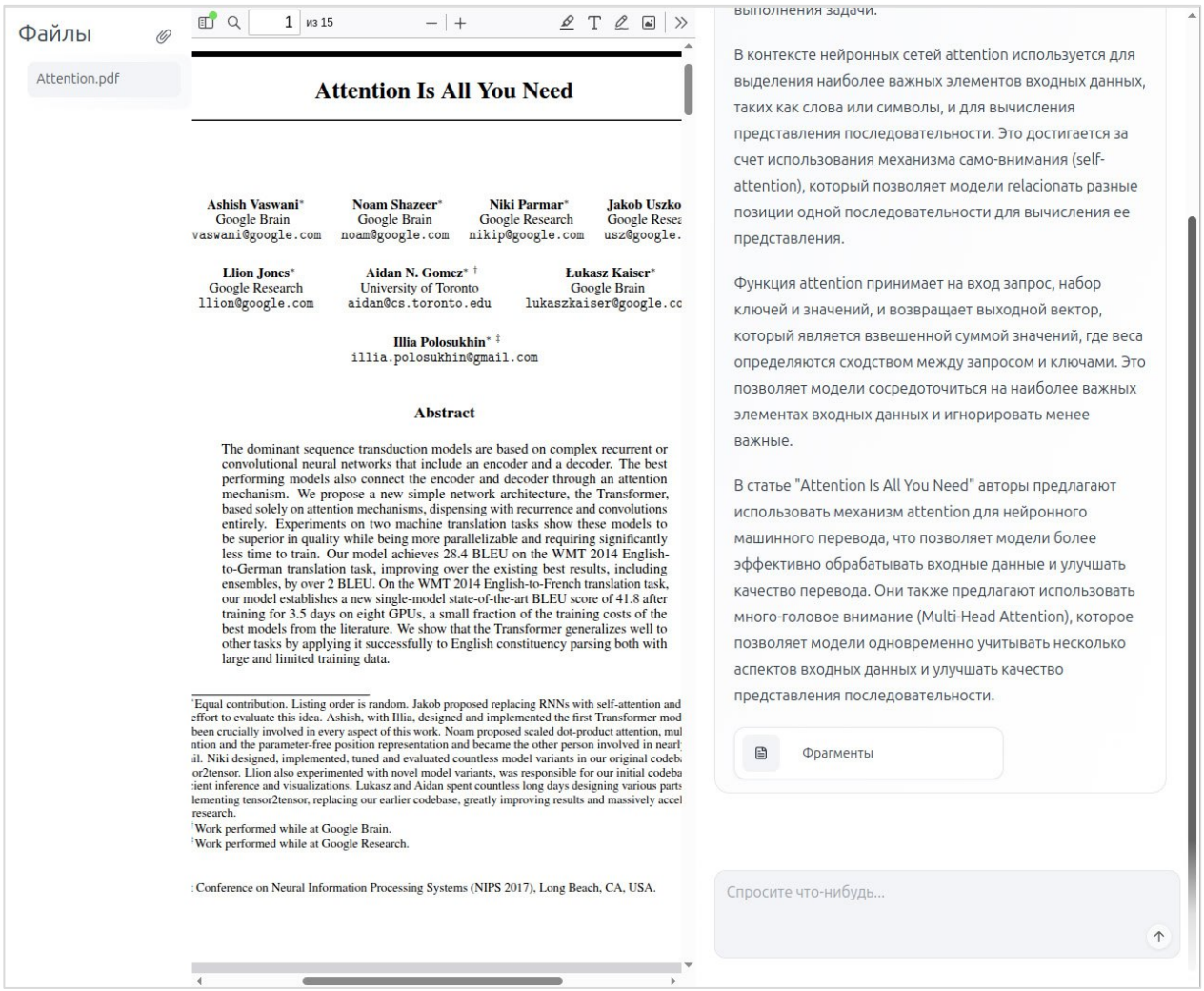


Рисунок 12 - Пользовательский интерфейс веб-приложения

6 ТЕСТИРОВАНИЕ

Для проверки корректности и надёжности функционала уже в процессе разработки составлялись модульные и интеграционные тесты с использованием библиотеки Pytest [26] и асинхронного клиента из библиотеки HTTPX [27]. Тестами покрыты основные компоненты системы, включая RAG-модуль, модуль для работы с базой данных и др. Тестирование проводится автоматически, что позволяет оперативно выявлять ошибки при изменении кода и вносить улучшения без снижения стабильности системы. Кроме того, разработан модуль, который проводит бенчмарк системы с помощью LLM-as-a-Judge [17] подхода. Он подключается к модулю RAG и выполняет его тестирование на основе датасета.

6.1 Датасет

Первоначально рассматривался датасет QASPER [28], однако в процессе разработки были выявлены его ограничения: вопросы в датасете слишком общие, к тому же каждый вопрос связан только с одной статьей. Эти ограничения делают его малопригодным для тестирования сложных Retrieval-механизмов. В итоге был выбран датасет FRAMES (Factuality, Retrieval, And reasoning MEasurement Set). Он содержит 824 вопроса, требующих информации из 2-15 статей из Wikipedia. Для каждого вопроса представлен эталонный ответ, а также список релевантных статей из Wikipedia. Датасет был разработан и выпущен 24 января 2025 года исследователями из Google с целью создания стандартизированного набора задач для оценки работы RAG-систем. В статье “ Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation ” авторы демонстрируют, что даже передовые большие языковые модели, такие как Gemini-Pro-1.5, сталкиваются со значительными трудностями при ответах на вопросы из FRAMES. При этом RAG-системы, использующие даже простые механизмы поиска, например BM25 [5], показывают лучшие результаты [29]. Это подтверждает, что добавление поискового механизма может существенно улучшить качество ответов.

Для бенчмарка из датасета случайным образом были отобраны 50

вопросов, и проиндексированы связанные с ними статьи. Чтобы приблизить эксперимент к реальным условиям, в индекс также добавлены нерелевантные статьи, содержащие информацию по другим вопросам. Этот шаг позволяет проверить, насколько эффективно система извлекает нужные данные и игнорирует лишнюю информацию. Итоговая выборка включает 330 статей и 50 вопросов.

6.2 Конфигурации

Для сравнительного тестирования были подготовлены три конфигурации вопрос-ответных систем:

1. Модуль RAG из разработанной системы.
2. Версия Agentic RAG. Для его реализации была использована библиотека smolagents [13], а в качестве инструмента подключен поисковой механизм разработанного RAG-модуля. В результате получилась система, способная пошагово рассуждать и извлекать данные, если нужно. Цепочка рассуждений ограничена 6 шагами.
3. LLM без поиска (так как данная версия не использует информационный поиск, метрика Response Relevancy не может быть рассчитана).

Во всех конфигурациях использовалась большая языковая модель openai/gpt-4o-mini [30] и эмбеддинг-модель deepvk/USER-bge-m3 [31].

6.3 Анализ результатов

Стоит отметить, что большие языковые модели склонны более лояльно оценивать собственные ответы [32], поэтому для расчёта метрик LLM-as-Judge использована другая модель: google/gemini-pro [33] и библиотека RAGAS [15].

Конфигурация Agentic-RAG показывает наилучшие результаты по извлечению фрагментов благодаря многократным вызовам поиска по базе знаний. Однако это сопровождалось наибольшими затратами времени на ответ и токенами на генерацию, что ограничивает её применимость в задачах, требующих работы в реальном времени.

Конфигурация без информационного поиска показала значительно более

низкие результаты по фактологической правильности генерации.

Конфигурация с классической архитектурой RAG показывает сбалансированные результаты, и является скорее компромиссом между чистыми LLM и рассуждающими системами (см. таблица 1).

Таблица 1 – Сравнение метрик

Наименование метрики	Agentic-RAG	RAG	LLM
Релевантность поиска	**0.51**	0.36	-
Достоверность генерации	**0.68**	0.54	-
Правильность генерации	**0.62**	0.48	0.26
Среднее время работы (с)	29	6.5	3

6.4 Пилотное тестирование

В рамках разработки проекта было проведено короткое пилотное тестирование. Для этого приложение было развернуто глобально в сети Интернет. Доступ к приложению получили 6 студентов НГУ. Целью тестирования было получение обратной связи о работоспособности интерфейса, корректности генерации ответов и общего впечатления пользователей о работе приложения. Кроме того, было важно узнать, насколько потенциально полезным могло бы быть приложение такого вида в учебном процессе. Анкеты включили в себя 9 вопросов: 6 из них принимали оценку от 0 до 10 и касались разных аспектов приложения, а также общую оценку. Два последних вопроса позволяли пользователю поделиться впечатлениями (см. приложение В).

Поскольку опрос из 6 человек представляет из себя малую выборку, было решено рассчитать доверительные интервалы для оценок пользователей. Доверительный интервал используется для оценки надежности среднего значения по малой выборке и позволяет узнать диапазон, в котором с большой вероятностью лежит истинное значение для всей выборки. В результате проведенных вычислений (см. приложение Г) получились результаты с оценкой по разным аспектам приложения (см. таблица 2):

Таблица 2 – доверительные интервалы (95%) для разных аспектов

Наименование аспекта	Среднее	Нижняя оценка	Верхняя оценка
Точность ответов	4.16	2.93	5.39
Скорость работы	5.83	5.04	6.62
Удобство	4.5	3.39	5.6
Полезность	4.0	2.51	5.48
Общая оценка	4.83	3.6	6.06

В результате опроса, можно сделать следующие выводы:

1. Система корректно обрабатывает запросы, однако в некоторых ситуациях дает слишком общие ответы. Особенно это заметно в ситуациях, когда пользователь в своем запросе хочет указать на конкретный структурный элемент текста, например на определенную главу.

2. Пользователи положительно оценили концепт дизайна интерфейса. Особо отметили возможность просмотра PDF-файлов прямо в приложении, а также высокую скорость ответов системы.

ЗАКЛЮЧЕНИЕ

Современная образовательная сфера переживает трансформацию, вызванную искусственным интеллектом. Особую роль в этих изменениях играют большие языковые модели, в частности чат-боты на их основе. Большие языковые модели предоставляют быстрые и понятные ответы на вопросы, что делает их удобным инструментом при изучении материалов: пользователю гораздо проще задать вопрос модели, чем искать ответы в литературе. Несмотря на их удобство, модели склонны к галлюцинациям, особенно в специфических областях знаний, что негативно сказывается на процессе обучения.

Для преодоления этой проблемы разработчики и исследователи применяют подход Retrieval Augmented Generation – он позволяет значительно улучшить качество ответов большой языковой модели, подключив к ней внешние источники знаний. Популярность RAG-систем в образовательной сфере стремительно растет. Согласно отчету компании Яндекс “Искусственный интеллект и высшее образование: возможности, практики и будущее” [34] решения на базе искусственного интеллекта приобретают все большую популярность среди вузов России. Инструменты на базе искусственного интеллекта становятся не заменой человеку, а его интеллектуальным помощником, позволяя повышать эффективность учебного процесса. Предполагается, что RAG позволит работать образовательным организациям с заранее подобранным контентом. Необходимые материалы могут быть собраны и провалидированы самой образовательной организацией, а затем загружены в специальную библиотеку, которая подключается к большой языковой модели.

В рамках выпускной квалификационной работы была спроектирована и реализована гибкая масштабируемая система на основе RAG. Ее главная особенность заключается в способности налету индексировать данные и выдавать ответ пользователю в реальном времени. Из-за такой особенности появились высокие требования к производительности компонентов, особенно к процессу индексации. Несмотря на ограничения, удалось реализовать прототип, который демонстрирует практическую применимость подобного приложения.

Кроме того, разработанная система модульная и легко адаптируется под разные сценарии использования: от поддержки учебного процесса до работы с внутренними корпоративными документами. Она легко расширяется дополнительными инструментами и адаптируется под конкретные задачи. Проведенное тестирование показало надежность и стабильность системы. Реализованное приложение может служить основой для дальнейшей разработки прикладных и коммерческих продуктов в сфере образования.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Lewis P., Perez J., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W.-t., Rocktäschel T., Riedel S. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 9459–9474.

2 Stephenson N. The Diamond Age: Or, a Young Lady's Illustrated Primer. – New York : Spectra, 2003.

3 Barnett S., Dey D., Johnson M., MacAvaney S., McMillan L. Seven failure points when engineering a retrieval augmented generation system // Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI. – 2024. – С. 194–199.

4 Wang X., Li X., Ye D., Zhang W. Searching for best practices in retrieval-augmented generation // arXiv preprint arXiv:2407.01219. – 2024.

5 Robertson S. E., Walker S., Hancock-Beaulieu M., Gatford M., Payne A. Okapi at TREC-3 // NIST Special Publication SP. – 1995. – Т. 109. – С. 109.

6 levels of text splitting // GitHub URL: https://github.com/FullStackRetrieval-com/RetrievalTutorials/blob/main/tutorials/LevelsOfTextSplitting/5_Levels_Of_Text_Splitting.ipynb (дата обращения: 10.05.2025).

7 Zhao J., Liu L., Wu L., Yin D., Zhang D. Meta-chunking: Learning efficient text segmentation via logical perception // arXiv preprint arXiv:2410.12788. – 2024.

8 Ma X., Tan Q., Pang L., Yang Y., Zhang M. Query rewriting in retrieval-augmented large language models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. – 2023. – С. 5303–5315.

9 Gao L., Dai Z., Callan J., Liu J. Precise zero-shot dense retrieval without relevance labels // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2023. – С. 1762–1777.

10 Glass M., Pradeep R., Ferraro F., White A. Re2G: Retrieve, rerank, generate // arXiv preprint arXiv:2207.06300. – 2022.

11 Reimers N., Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks // arXiv preprint arXiv:1908.10084. – 2019.

12 Rosa G., Omidshafiei S., Simard P. In defense of cross-encoders for zero-shot retrieval // arXiv preprint arXiv:2212.06121. – 2022.

13 Smolagents // GitHub URL: https://github.com/huggingface/smolagents (дата обращения: 10.05.2025).

14 Yao S., Zhao J., Yu D., Yu Y., Zhang Y., Cao Q., et al. ReAct: Synergizing reasoning and acting in language models // International Conference on Learning Representations (ICLR). – 2023.

15 Yu H., Gao C., Ren X., Gu J. Evaluation of retrieval-augmented generation: A survey // CCF Conference on Big Data. – Singapore : Springer Nature Singapore, 2024. – С. 102–120.

16 Es S., Ren Y., Trivedi H., et al. 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. – С. 150–158.

17 Gu J., Liu Z., Ren X. A survey on LLM-as-a-Judge // arXiv preprint arXiv:2411.15594. – 2024.

18 ML System Design Doc // GitHub URL: https://github.com/IrinaGoloshchapova/ml_system_design_doc_ru/blob/main/ML_System_Design_Doc_Template.md (дата обращения: 10.05.2025).

19 ChatPDF // ChatPDF URL: https://www.chatpdf.com/ru (дата обращения: 10.05.2025).

20 ChatUI // GitHub URL: https://github.com/huggingface/chat-ui (дата обращения: 10.05.2025).

21 SQLite // Wikipedia URL: https://en.wikipedia.org/wiki/SQLite (дата обращения: 10.05.2025).

22 SQLAlchemy // GitHub URL: https://github.com/sqlalchemy/sqlalchemy (дата обращения: 10.05.2025).

23 Testcontainers Python // GitHub URL:

https://github.com/testcontainers/testcontainers-python (дата обращения: 10.05.2025).

24 Server-sent events // Wikipedia URL: https://ru.wikipedia.org/wiki/Server-sent_events (дата обращения: 10.05.2025).

25 GROBID: machine learning for parsing and structuring scientific publications // GitHub. URL: https://github.com/kermitt2/grobid (дата обращения: 15.05.2025).

26 Pytest // Wikipedia URL: https://en.wikipedia.org/wiki/Pytest (дата обращения: 10.05.2025).

27 HTTPX // GitHub URL: https://github.com/projectdiscovery/httpx (дата обращения: 10.05.2025).

28 Dasigi P., Ammar W., Swayamdipta S., Gardner M. A dataset of information-seeking questions and answers anchored in research papers // arXiv preprint arXiv:2105.03011. – 2021.

29 Krishna S., Wang Z., Zhou C., Choudhury S., Talukdar P. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation // arXiv preprint arXiv:2409.12941. – 2024.

30 GPT-4o mini // OpenAI URL: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ (дата обращения: 10.05.2025).

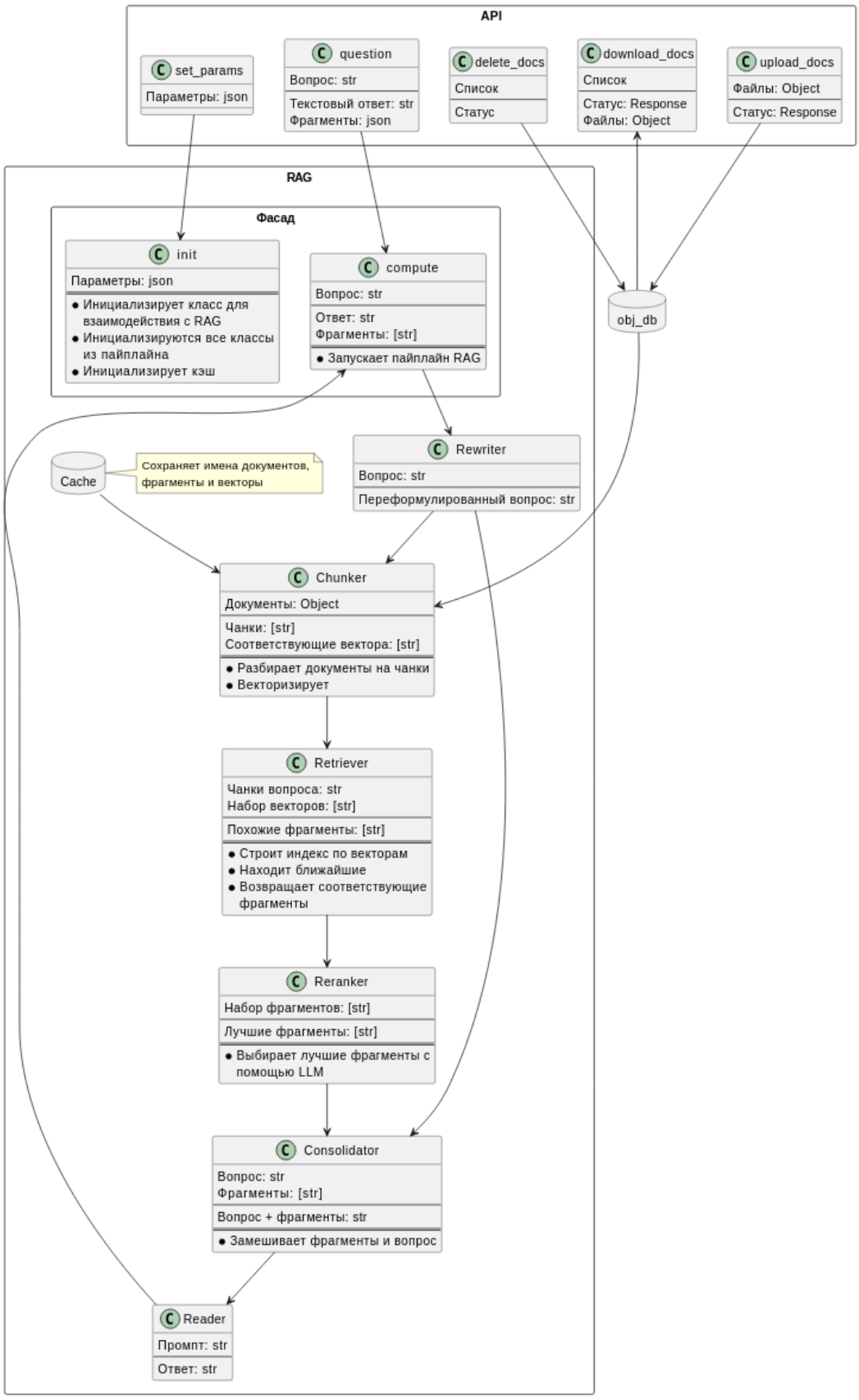
31 deepvk. USER-bge-m3 // HuggingFace URL: https://huggingface.co/deepvk/USER-bge-m3 (дата обращения: 10.05.2025).

32 Xu W., Deng Y., Wang Z., Liu J., Zhang Y. Pride and prejudice: LLM amplifies self-bias in self-refinement // arXiv preprint arXiv:2402.11436. – 2024.

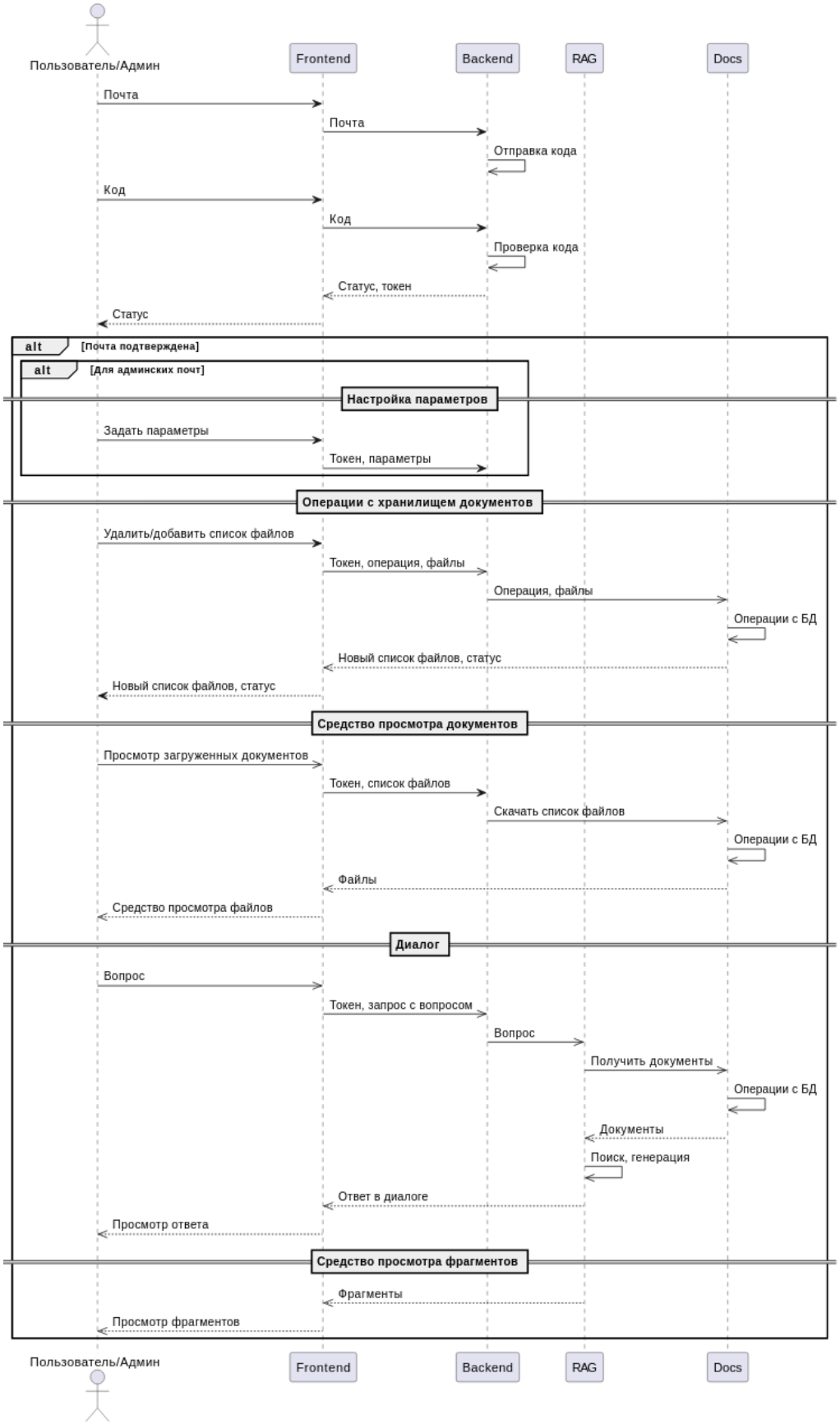
33 Gemini 2.5 Pro // Google DeepMind URL: https://deepmind.google/technologies/gemini/pro/ (дата обращения: 10.05.2025).

34 Искусственный интеллект и высшее образование: возможности, практики и будущее // Яндекс Образование URL: https://education.yandex.ru/aihighreport (дата обращения: 10.05.2025).

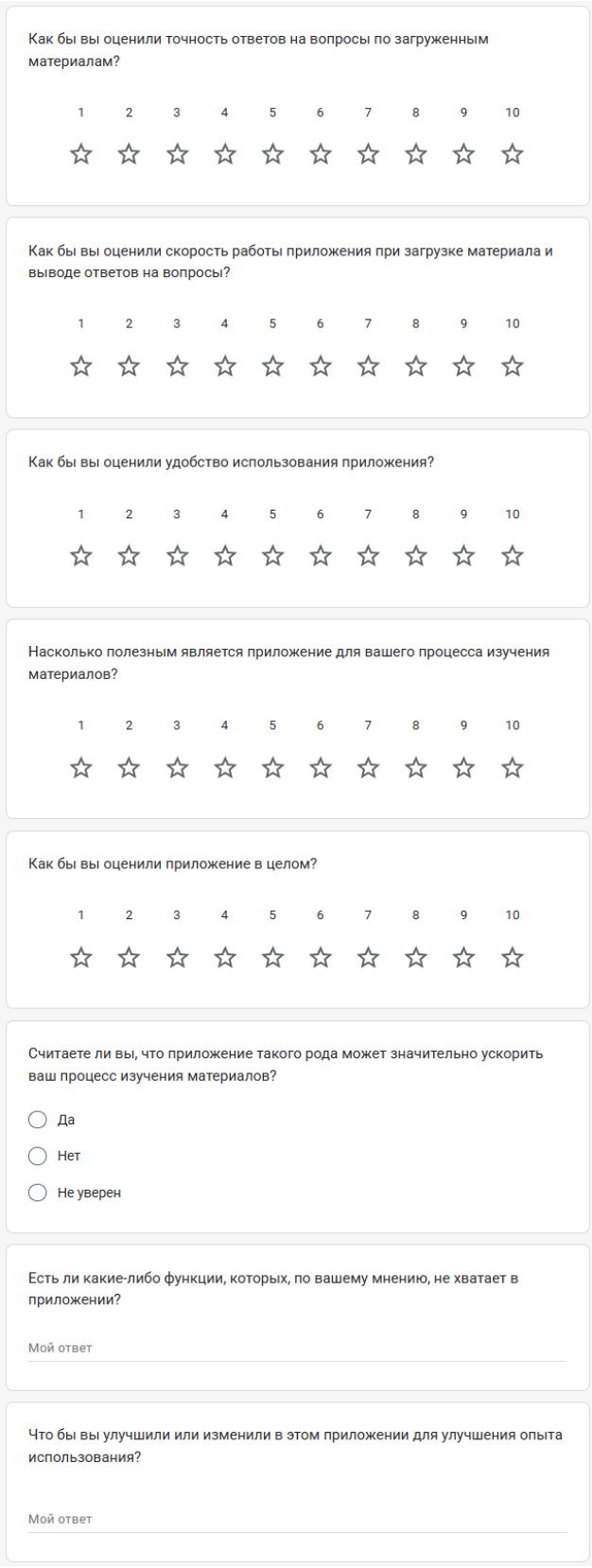
ПРИЛОЖЕНИЕ А
Диаграмма компонентов серверной части



ПРИЛОЖЕНИЕ Б
Диаграмма последовательностей



ПРИЛОЖЕНИЕ В
Анкета для опроса пользователей



ПРИЛОЖЕНИЕ Г

Программный код для подсчета доверительных интервалов

```
import numpy as np
import scipy.stats as stats
import pandas as pd

# Входные данные - результаты опроса
data = np.array([
    [3, 6, 4, 4, 5],
    [4, 5, 3, 5, 6],
    [4, 6, 5, 3, 3],
    [6, 7, 5, 4, 6],
    [3, 5, 6, 2, 4],
    [5, 6, 4, 6, 5],
])

# Параметры
confidence = 0.95  # 95% доверительный интервал
n = data.shape[0]

means = data.mean(axis=0)
std_err = stats.sem(data, axis=0)
h = std_err * stats.t.ppf((1 + confidence) / 2.0, n - 1)

intervals = pd.DataFrame({
    'Mean': means,
    'CI Lower': means - h,
    'CI Upper': means + h
})
print(intervals)
```

Выпускная квалификационная работа выполнена мной самостоятельно и с соблюдением правил профессиональной этики. Все использованные в работе материалы и заимствованные принципиальные положения (концепции) из опубликованной научной литературы и других источников имеют ссылки на них. Я несу ответственность за приведенные данные и сделанные выводы.

Я ознакомлен с программой государственной итоговой аттестации, согласно которой обнаружение плагиата, фальсификации данных и ложного цитирования является основанием для не допуска к защите выпускной квалификационной работы и выставления оценки «неудовлетворительно».

_____ _____

ФИО студента *Подпись студента*

« ____ »_____ _20 __г.
(заполняется от руки)