Разработка модуля управления документами для AI-ассистента на основе LLM и RAG

К. И. Еремина
Санкт-Петербургский политехнический университет Петра Великого
ksenerem03@gmail.com

П. Д. Дробинцев
Санкт-Петербургский политехнический университет Петра Великого
drob@ics2.ecd.spbstu.ru

Е. В. Скуднева
Санкт-Петербургский политехнический университет Петра Великого
skudneva_ev@spbstu.ru

Н. В. Воинов
Санкт-Петербургский политехнический университет Петра Великого
voinov@ics2.ecd.spbstu.ru

В. С. Тутыгин
Санкт-Петербургский политехнический университет Петра Великого
tutygin_vs@spbstu.ru

Д. Ф. Дробинцев
Санкт-Петербургский политехнический университет Петра Великого
drobintsev_df@spbstu.ru

Аннотация. **Работа посвящена повышению производительности работы AI-ассистента на основе LLM и RAG технологий путем разработки и внедрения специализированного модуля для управления документами из внутренней базы знаний, используемыми при ответах на запросы пользователей. В рамках работы предложен подход, реализующий хранение информации в векторном представлении, поиск и удаление повторяющейся информации, выявление противоречащей информации, возможность выбора приоритетной информации. После внедрения разработанного модуля ожидается увеличение эффективности работы AI-ассистента с точки зрения сокращения времени обработки и объема хранения данных, а также повышения скорости поиска запрашиваемой информации.**

Ключевые слова: управление документами, AI-ассистент, LLM, RAG, база знаний

I. Введение

Искусственный интеллект проник во все сферы нашей жизни, начиная от помощников для технической поддержки клиентов и заканчивая помощью врачам в проведении клинических исследований [1,2]. Современные информационные системы используют AI-ассистентов на базе больших языковых моделей (Large Language Models, LLM) и дополненной генерации с поиском (Retrieval Augmented Generation, RAG) для автоматизации рутинных задач, например, для качественного сопровождения продукта на стороне заказчика, уменьшения времени на обучение новых сотрудников и поддержки базы знаний документов в актуальном состоянии и доступной для всех сотрудников [3-5].

Однако всегда актуальна проблема корректности полученного ответа, поэтому для многих ассистентов добавляются функции, демонстрирующие пользователю процесс размышлений и ссылки на источники информации, что позволяет самостоятельно контролировать качество. Также большинство предприятий и компаний имеют внутренние документы, которые невозможно найти в открытом доступе, но они играют важную роль при формировании ответа. Без информации из этих источников корректность ответа снижается.

Таким образом, можно отметить, что технологии весьма полезны, но обладают существенным недостатком, который заключается в отсутствии возможности управления документами, информация из которых используется для ответа, самостоятельно пользователем. Это влечет недостоверную информацию в ответе и противоречивые ответы на один и тот же вопрос. Поэтому актуальным является решение задачи по управлению документами и созданию персональной базы знаний для каждого заказчика, использующего AI-ассистент. Для решения данной проблемы в статье предлагается специализированный инструмент по управлению документами, позволяющий заказчикам эффективно работать с базой знаний для поддержки ее в актуальном виде. В ходе работы были изучены существующие методы управления документами; проведен сравнительный анализ современных инструментов, с помощью которых реализованы методики управления документами в информационных системах; предложен подход, реализующий хранение информации в векторном представлении, поиск и удаление повторяющейся информации, поиск противоречащей друг другу информации, возможность выбрать приоритетную информацию. Данный подход был реализован в виде отдельного модуля в рамках AI-ассистента.

II. Технологии LLM и RAG

LLM представляют собой модели глубоких нейронных сетей, позволяющие обрабатывать естественные языки. Благодаря достижениям в области глубокого обучения, которое является подмножеством машинного обучения и искусственного интеллекта, ориентированного на нейронные сети, LLM обучаются на огромных объемах текстовых данных [6,7].

Однако можно заметить, что у LLM существует проблема – знания, согласно которым языковые модели дают ответ, ограничены данными, на которых обучались модели. В связи с этим был придуман архитектурный подход, который решает данную проблему.

Дополненная поиском генерация (RAG) является методом работы с языковыми моделями, который используется для расширения их возможностей за счет интеграции механизмов поиска информации. Этот подход позволяет системе обращаться к актуальным источникам во время формирования ответа вместо дорогостоящего переобучения модели на новых данных [8,9].

Таким образом, подход, согласно которому происходит соединение языковой модели с базой знаний для обогащения контекстом, решает проблему с быстрой актуализацией данных для ответа в обход переобучению модели. Однако нерешенной остается проблема удобного управления данными в базе знаний. Именно на решение данной проблемы нацелена эта работа, предлагая инструмент для управления знаниями, которые используются для обогащения контекстом большой языковой модели, тем самым повышая качество ответов AI-ассистента.

III. Предлагаемый подход

Рассматриваемый в работе AI-ассистент основан на технологиях LLM и RAG и использует в ответах документы из внутренней базы знаний. Эта база знаний будет формироваться на основе документации конкретной компании, что повысит качество ответов ассистента.

A. Общая архитектура AI-ассистента

Общий пайплайн работы AI-ассистента представлен на рис. 1. Этап, на котором реализуется поиск релевантного контекста, выделен лиловым цветом. На этом этапе система обращается к базе знаний за данными, релевантными запросу пользователя.

UML-схема со всеми компонентами системы представлена на рис. 2. Модуль анализа качества ответов состоит из модуля контроля качества, который оценивает входные данные, релевантные документы и ответ LLM, и модуля визуализации статистики, который будет демонстрировать наиболее актуальные темы вопросов пользователей, темы, на которые система не может дать ответ, и аналогичную статистику.

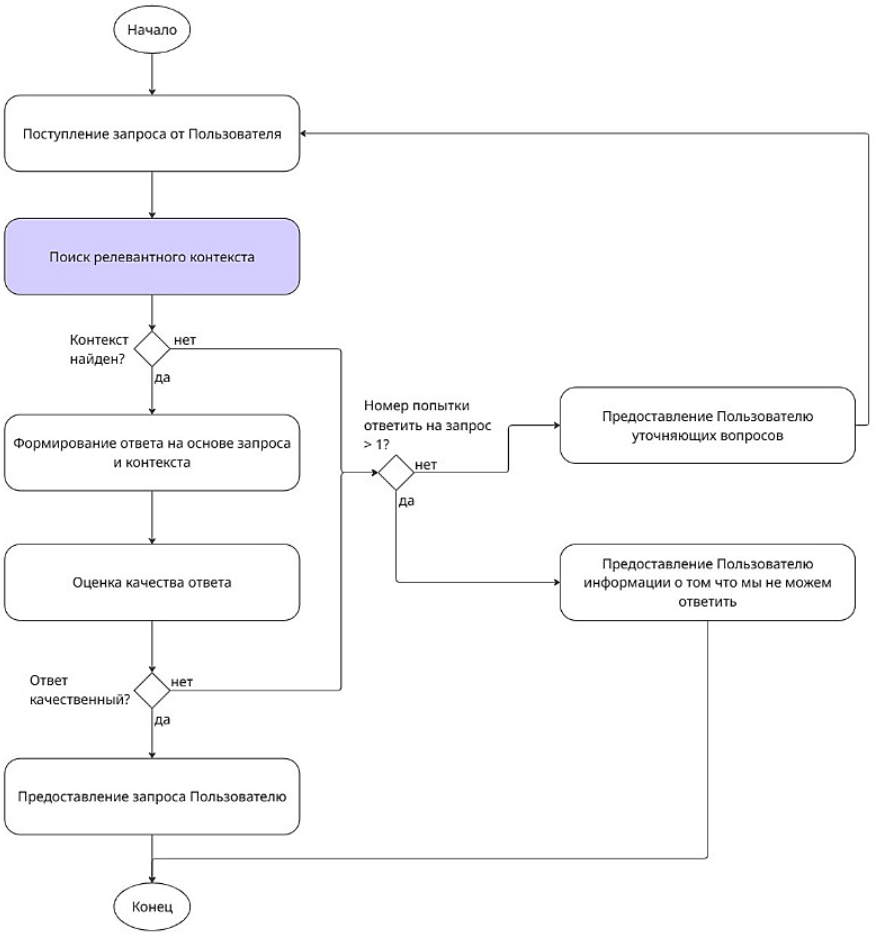

Рис. 1. Общий пайплайн работы AI-ассистента

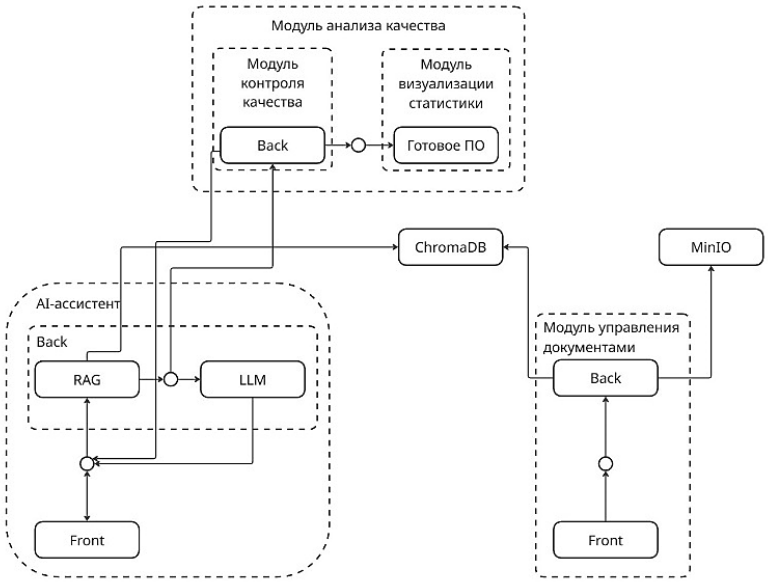

Рис. 2. UML-схема работы AI-ассистента

B. Модуль управления документами

Разработанный модуль отвечает за автоматизацию процессов, связанных с загрузкой, обновлением и просмотром документов в базе знаний. Модуль управления документами предназначен для решения следующих задач:

- Создание актуальной базы знаний, которой способен управлять специалист со стороны заказчика.
- Предоставление актуальной информации из первоисточника.
- Избавление от дублирования информации в документах.
- Повышение качества ответов за счет выдачи приоритетов высказываниям при противоречивом контексте.

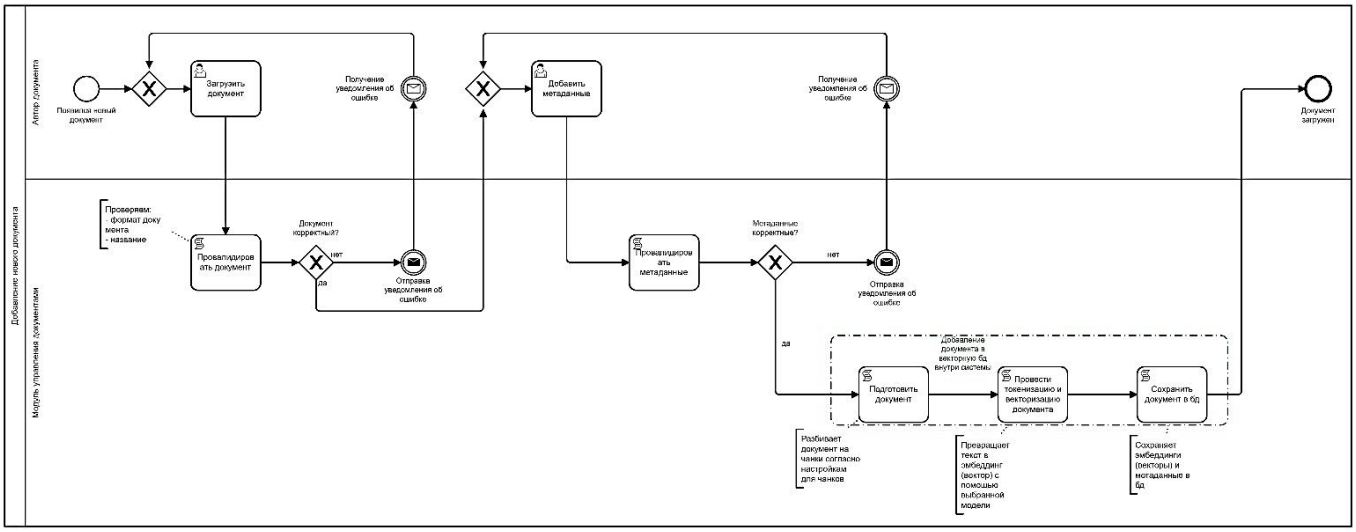

Рис. 3. BPMN-схема для добавления нового документа

Модуль управления документами обращается к двум базам данных: ChromaDB и MinIO. MinIO является объектным хранилищем и содержит оригиналы документов, информация из которых необходима для ответа LLM. В данное хранилище будут загружаться различные типы документов: PDF, Word, TXT, LaTeX, презентации и др. Обработанный текст из документов сохраняется в базу данных ChromaDB.

Основной бизнес-процесс, реализуемый данным модулем, представлен на BPMN-схеме на рис. 3. Это добавление нового документа в систему. Действия со стороны системы можно разделите на две группы: проверка введенных данных и добавление документов в векторную базу данных.

C. Инструменты для реализации

В основе AI-ассистента лежит обученная LLM-модель Llama. В качестве ретривера используется фреймворк LlamaIndex. Объектное хранилище данных – MinIO. Векторная база данных – ChromaDB. В качестве основного языка программирования выбран Python.

IV. План по оценке результатов

Результаты работы AI-ассистента после внедрения разработанного модуля планируется оценить по следующим критериям:

- время обработки информации;
- объем хранимых данных;
- время поиска.

В табл. I представлена процедура оценки времени обработки информации, затрачиваемое разработанным модулем.

ТАБЛИЦА I. ОЦЕНКА ВРЕМЕНИ ОБРАБОТКИ ИНФОРМАЦИИ

Характеристика	**Описание**
Входные данные	Модуль управления документами, 3 документа разного размера, каждый до 100 Мб
Шаги для проведения эксперимента	1. Зайти в модуль управления документами 2. Выполнить для каждого документа: 2.1. Загрузить новый документ 2.2. Зафиксировать время загрузки документа (из логов) 2.3. Обновить документ 2.4. Зафиксировать время загрузки документа (из логов) 2.5. Загрузить новый документ вручную 2.6. Зафиксировать время 2.7. Обновить документ вручную 2.8. Зафиксировать время
Метрика, ед.изм.	время обработки документа, сек; скорость обработки текста, char/сек.

В табл. II представлена процедура оценки объема данных при обработке повторной информации.

ТАБЛИЦА II. ОЦЕНКА ОБЪЕМА ДАННЫХ

Характеристика	**Описание**
Входные данные	Модуль управления документами, 3 документа с повторами информации
Шаги для проведения эксперимента	1. Зайти в модуль управления документами 2. Добавить документы в базу с помощью скрипта без поиска повторов информации 3. Зафиксировать количество элементов в базе 4. Очистить базу 5. Добавить документы с помощью скрипта с учетом поиска повторов информации 6. Зафиксировать количество элементов в базе
Метрика, ед.изм.	количество чанков, шт; размер векторного индекса, Мб

В табл. III представлена процедура оценки времени поиска информации AI-ассистентом с внедренным модулем обработки документов.

ТАБЛИЦА III. ОЦЕНКА ВРЕМЕНИ ПОИСКА

Характеристика	**Описание**
Входные данные	Модуль управления документами, AI-ассистент, база знаний из 10 документов, набор тестовых запросов - 10 штук
Шаги для проведения эксперимента	1. Зайти в AI-ассистента 2. Для каждого запроса выполнить: 2.1. Задать вопрос 2.2. Получить ответ 2.3. Зафиксировать время выполнения (из логов)
Метрика, ед.изм.	время поиска контекста, сек

V. Заключение

Внедрение AI-ассистентов на базе LLM и RAG в современные информационные системы открывает значительные возможности. Однако текущий уровень развития таких систем связан с проблемой недостаточной прозрачности и контролируемости источников информации, что может приводить к снижению корректности ответов и появлению противоречивых результатов. Ключевая проблема, выявленная в работе, заключается в отсутствии у заказчика полноценных средств управления собственными документами, используемыми в базе знаний AI-ассистента. Это обусловливает необходимость разработки специализированного инструмента, позволяющего заказчикам самостоятельно формировать и поддерживать персональную базу знаний, устранять дублирующую и противоречивую информацию и устанавливать приоритеты для источников. В перспективе дальнейшее развитие подобных решений может привести к созданию более ответственных и адаптивных AI-ассистентов, тесно интегрированных в корпоративные процессы и поддерживающих доверие пользователей к автоматически генерируемым ответам.

Список литературы

[1] Hou T., Li M., Tan Y., Zhao H. Physician adoption of AI assistant. Manufacturing & Service Operations Management, 2024, vol. 26, no. 5, pp. 1639-1655. DOI: https://doi.org/10.1287/msom.2023.0093

[2] Senanayake S., Karunanayaka K., Ekanayake K. V. J. P. Review on AI assistant systems for programming language learning in learning environments. 2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI). 2024, pp. 1-6. DOI: https://doi.org/10.1109/SLAAI-ICAI63667.2024.10844969

[3] Wang L., Huang N., He Y., Liu D., Guo X., Sun Y., Chen G. Artificial intelligence (AI) assistant in online shopping: A randomized field experiment on a livestream selling platform. Information Systems Research, 2025, vol. 36, no. 4, pp. 2358-2374. DOI: https://doi.org/10.1287/isre.2023.0103

[4] Papageorgiou G., Sarlis V., Maragoudakis M., Tjortjis C. Hybrid multi-agent GraphRAG for e-government: Towards a trustworthy AI assistant. Applied Sciences, 2025, vol. 15, no. 11, p. 6315. DOI: https://doi.org/10.3390/app15116315.

[5] López-Galisteo A. J., Borrás-Gené O. The creation and evaluation of an AI assistant (GPT) for educational experience design. Information, 2025, vol. 16, no. 2, p. 117. DOI: https://doi.org/10.3390/info16020117

[6] Yu M., Meng F., Zhou X., Wang S., Mao J., Pan L., Chen T., Wang K., Li X., Zhang Y., An B., Wen Q. A survey on trustworthy llm agents: Threats and countermeasures. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2. 2025, pp. 6216-6226. DOI: https://doi.org/10.1145/3711896.3736561

[7] Wu X. K., Chen M., Li W., Wang R., Lu L., Liu J., Hwang K., Hao Y., Pan Y., Meng Q., Huang K., Hu L., Guizani M., Chao N., Fortino G., Lin F., Tian Y., Niyato D., Wang, F. Y. Llm fine-tuning: Concepts, opportunities, and challenges. Big Data and Cognitive Computing, 2025, vol. 9, no. 4, p. 87. DOI: https://doi.org/10.3390/bdcc9040087

[8] Şakar T., Emekci H. Maximizing RAG efficiency: A comparative analysis of RAG methods. Natural Language Processing, 2025, vol. 31, no. 1, pp. 1-25. DOI: 10.1017/nlp.2024.53

[9] Chan B. J., Chen C. T., Cheng J. H., Huang H. H. Don't do RAG: When cache-augmented generation is all you need for knowledge tasks. In Companion Proceedings of the ACM on Web Conference 2025, pp. 893-897. DOI: https://doi.org/10.1145/3701716.3715490