

УДК 004.896

DOI: 10.24412/3033-6007-2026-339-4-22

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ ПОВЫШЕНИЯ ТОЧНОСТИ RAG-СИСТЕМ НА СПЕЦИАЛИЗИРОВАННЫХ КОРПУСАХ НОРМАТИВНО-ТЕХНИЧЕСКОЙ ДОКУМЕНТАЦИИ

Агафонов А.А.^{1, a}, Пономарев А.В.^{1, b}

¹ Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербург, Россия.

E-mail: ^a agafonov.a@spcras.ru, ^b ponomarev@iias.spb.su

Аннотация. В условиях цифровизации промышленности возрастает потребность в интеллектуальных системах поддержки принятия решений, способных эффективно извлекать знания из массивов нормативно-технической документации. В статье рассматриваются интеллектуальные вопрос-ответные системы на основе технологии генерации, дополненной поиском (Retrieval-Augmented Generation, RAG), применительно к специализированному корпусу нормативно-технических документов железнодорожной отрасли. Проводится сравнительный анализ ряда методов повышения точности RAG-систем, включая стратегии обработки запросов (HyDE, Step-Back, Multi-Query, Decomposition, Pseudo-Relevance Feedback, Recursive Refinement), двухэтапное извлечение с использованием кросс-энкодера текстовых фрагментов (вопроса и нормативного пункта), а также дообучение моделей эмбединга текстовых фрагментов на синтетических парах «вопрос – нормативный пункт». Эксперименты показывают, что наиболее существенный прирост качества поиска обеспечивает предметное дообучение моделей эмбединга. Применение кросс-энкодера также дает положительный эффект, особенно для моделей с изначально менее точным ранжированием. В то же время продвинутые стратегии обработки запросов не приводят к значимому повышению качества извлечения по сравнению с базовым подходом. Сформированные корпус документов, наборы вопросов и программный код размещены в открытом доступе.

Ключевые слова: RAG, большие языковые модели, семантический поиск, модели эмбединга, кросс-энкодер, нормативно-техническая документация, железнодорожный транспорт, дообучение.

Введение

В условиях цифровой трансформации промышленности и роста объема неструктурированных текстовых данных возрастает потребность в интеллектуальных системах поддержки принятия решений (СППР), способных оперативно предоставлять лицу, принимающему решение, релевантную информацию. На железнодорожном транспорте, характеризующемся высокими требованиями к безопасности и надежности, особенно остро стоит задача

Для цитирования:

Агафонов А.А., Пономарев А.В. Сравнительный анализ методов повышения точности RAG-систем на специализированных корпусах нормативно-технической документации. *Интеллектуальный транспорт*. 2026. № 3(39). С. 4–22. DOI: 10.24412/3033-6007-2026-339-4-22.

эффективного извлечения знаний из массивов нормативно-технической документации, регламентирующей процессы эксплуатации, обслуживания и ремонта подвижного состава и инфраструктуры. Традиционные поисковые системы, основанные на точном совпадении ключевых слов, зачастую не справляются с семантической сложностью запросов, что обуславливает применение методов интеллектуального анализа текстов на основе современных подходов обработки естественного языка.

В настоящее время одним из наиболее распространенных подходов к построению текстовых СППР является технология генерации, дополненной поиском (Retrieval-Augmented Generation, RAG), впервые предложенная в работе [1]. Современные подходы к построению RAG-систем и направления их развития систематизированы в обзорной работе [2]. Применение RAG в СППР рассматривается в различных предметных областях, в частности в здравоохранении [3], сельском хозяйстве [4], электронной коммерции [5] и розничной торговле [6]. Эта технология обеспечивает сочетание механизмов семантического поиска по внешним базам знаний с генеративными способностями больших языковых моделей (БЯМ), что позволяет формировать аргументированные ответы на основе актуальных фактологических данных, а не только знаний, заложенных в веса языковой модели. Благодаря этой особенности RAG-системы становятся важным компонентом современных отраслевых СППР, обеспечивая баланс между гибкостью диалогового взаимодействия и достоверностью информации.

Однако практическая реализация RAG-систем в предметных областях со строгими требованиями к информационной безопасности и высокой степенью лингвистической специализации сталкивается с рядом вызовов, которые существенно ограничивают прямое применение стандартных решений. Среди основных проблем создания отраслевых текстовых СППР можно отметить следующие:

1. наличие различных категорий закрытой информации, которая не должна покидать контур организации, что ограничивает применение наиболее мощных моделей, доступных по API, обладающих как большим контекстным окном, позволяющим одновременно анализировать большой объем источников, так и достаточно развитыми языковыми возможностями;
2. наличие специфической отраслевой терминологии, которая может снижать эффективность работы моделей, предназначенных для широкого использования, опирающегося преимущественно на общую лексику.

Данные трудности оказываются взаимосвязаны – отраслевая терминология, вполне вероятно, более точно обрабатывалась бы передовыми БЯМ, однако необходимость предотвращения утечки данных ограничивает выбор теми инструментами, которые могут быть развернуты на собственной вычислительной инфраструктуре предприятия, которая может быть достаточно ограниченной. Таким образом, несмотря на рост эффективного контекстного окна передовых БЯМ¹ [7], технология RAG, центральным элементом которой является управление размером контекста, доступного БЯМ при формировании ответа, оказывается по-прежнему востребованной на практике.

¹Claude Blog. 1M Context Is Now Generally Available for Opus 4.6 and Sonnet 4.6 [Электронный ресурс]. URL: <https://claude.com/blog/1m-context-ga> (дата обращения: 13.07.2026).

Технология RAG предлагает множество вариантов практической реализации, различающихся методом выделения фрагментов («чанков») и отображения их в векторное пространство для семантического поиска, стратегией формирования и обработки поискового запроса, используемыми языковыми моделями [2]. Существенное внимание также уделяется разработке методов оценки качества RAG-систем, учитывающих как качество поиска релевантных документов, так и корректность формируемых ответов [8]. Несмотря на значительное количество исследований, посвященных развитию отдельных компонентов RAG, выбор эффективной конфигурации RAG для специализированных предметных областей остается открытой научно-практической задачей. В статье проводится практическое исследование ряда распространенных техник, используемых при реализации технологии RAG, на примере специализированного корпуса нормативно-технической документации в области железнодорожного транспорта. В рамках реализации данного исследования был сформирован корпус нормативно-технической документации в области железнодорожного транспорта и тестовое множество, состоящее из набора вопросов с ожидаемыми ответами, реализован базовый вариант технологии RAG, а также ряд распространенных стратегий обработки запросов. Кроме того, была проведена оценка перспективности дообучения модели отображения текста в векторное пространство низкой размерности (модели эмбединга текстовых фрагментов) для повышения точности извлечения фрагментов. Реализованные варианты оценены как по показателям, характеризующим качество извлечения релевантных фрагментов, так и по правильности формируемых ответов. Сформированный корпус, база фрагментов, а также программный код, воспроизводящий реализованные алгоритмы и результаты экспериментов, размещены в открытом доступе².

Статья организована следующим образом. В разделе 1 рассматриваются общая схема технологии RAG и исследуемые способы повышения качества поиска, включая стратегии обработки запросов, переранжирование с использованием кросс-энкодера и предметное дообучение моделей эмбединга. В разделе 2 описаны сформированный корпус нормативно-технической документации и наборы вопросов, использованные в экспериментах. В разделе 3 приведены методика экспериментального исследования и результаты сравнительной оценки рассмотренных подходов. В заключении сформулированы основные выводы работы.

1. Технология RAG

Общая схема технологии RAG изображена на рисунке 1. Весь комплекс операций, осуществляемых в рамках данной технологии, можно разделить на две фазы: фаза подготовки и фаза выполнения. На фазе подготовки производится идентификация корпуса документов, содержащих факты о предметной области, которые должны быть использованы при ответах на запросы пользователя. Документы разбиваются на фрагменты, которые обрабатываются моделью эмбединга текстовых фрагментов для отображения в семантическое векторное пространство таким образом, что близкие по смыслу документы соответствуют близким (по значению определенной меры –

²Репозиторий проекта rag-intrans-paper [Электронный ресурс]. URL: <https://gitverse.ru/avponomarev/rag-intrans-paper> (дата обращения: 28.08.2026).

например, косинусного сходства) векторам. Пары, состоящие из фрагмента и соответствующего ему вектора, помещаются в векторное хранилище, обеспечивающее возможность эффективного поиска по заданному вектору. На фазе выполнения поступивший от пользователя запрос также преобразуется в векторное представление и в векторном хранилище производится поиск фрагментов, обладающих схожими векторными представлениями. Исходный запрос пользователя и найденные в векторном хранилище фрагменты включаются в промпт к БЯМ, содержащий инструкции по формированию ответа на запрос на основе представленной информации.

При практической реализации этой общей схемы необходимо принять ряд инженерных решений, определяющих:

1. Стратегию разбиения корпуса на фрагменты. Эта стратегия является результатом поиска компромисса между богатством контекста (чем больше фрагмент, тем больше информации он содержит, что может быть полезно как для его семантической интерпретации, так и для построения ответа) и «сфокусированностью» фрагмента (облегчающей поиск конкретных фактов, релевантных запросу, и позволяющей сократить объем контекста БЯМ, формирующей итоговый ответ).
2. Модель эмбединга. Существует множество моделей, характеризующихся различным качеством и требованиями к вычислительным ресурсам.
3. Стратегию обработки запроса. В качестве поискового образа не обязательно должен выступать сам запрос, для повышения качества извлечения он может быть предварительно обработан (например, переформулирован в другом стиле).
4. БЯМ, осуществляющую формирование окончательного ответа.

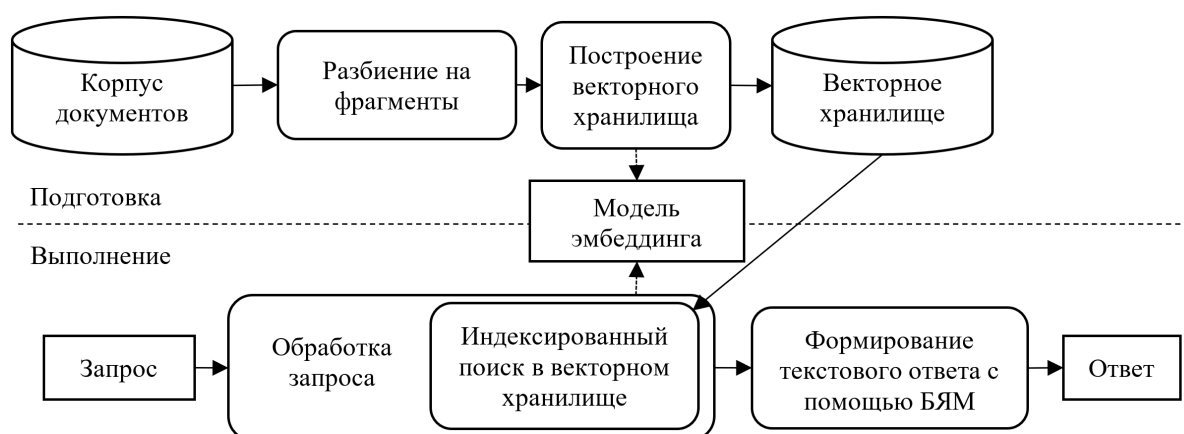


Рис. 1. Общая схема технологии RAG

В статье представлены результаты экспериментального исследования, нацеленного на сравнительную оценку эффективности решений, касающихся нескольких ключевых компонентов этой схемы, а именно:

- стратегии обработки запросов;
- стратегии извлечения результатов из векторного хранилища (одноэтапная или двухэтапная);

- модели эмбединга и возможности ее дообучения для учета предметной терминологии;
- БЯМ, используемой для формирования текстовых ответов.

1.1. Стратегии обработки запросов

Прямое использование исходного запроса пользователя в качестве поискового образа не всегда является оптимальным: запросы могут быть сформулированы нестрого, содержать избыточные детали или, напротив, быть слишком краткими для точного семантического сопоставления. Кроме того, сложные многосоставные вопросы могут требовать извлечения информации из нескольких источников, каждый из которых релевантен лишь отдельному аспекту запроса. В связи с этим в практике построения RAG-систем широкое распространение получили методы предварительной обработки запросов, направленные на трансформацию исходного пользовательского вопроса в форму, более удобную для семантического поиска. В настоящей работе исследован ряд таких стратегий, различающихся как по принципу действия, так и по вычислительной сложности.

HyDE (Hypothetical Document Embeddings) [9] – вместо прямого использования пользовательского запроса сначала с помощью большой языковой модели генерируется гипотетический документ-ответ на заданный вопрос, после чего этот синтезированный текст используется в качестве поискового запроса для извлечения из векторного хранилища. Предполагается, что гипотетический ответ, даже если он содержит неточности, лучше отражает семантическую структуру релевантных документов, чем исходный вопрос.

Step-Back (шаг назад) [10] – метод, основанный на идее абстрагирования. Стратегия заключается в том, чтобы сформулировать более общую версию исходного вопроса, сохраняющую его суть, но абстрагирующуюся от частных деталей. Затем выполняется поиск как по исходному, так и по обобщенному запросу, а их результаты объединяются.

Multi-Query (мультизапросы) [11] – стратегия, заключающаяся в генерации нескольких альтернативных формулировок одного и того же вопроса с помощью языковой модели. Каждая из сгенерированных формулировок используется для отдельного поиска, после чего результаты всех запросов объединяются. Данный подход призван компенсировать возможные неудачные формулировки исходного вопроса и повышать полноту извлечения.

Decomposition (декомпозиция) [12] – метод, предназначенный для обработки сложных многосоставных вопросов, ответ на которые требует информации из нескольких источников. В рамках этой стратегии исходный вопрос с помощью БЯМ разбивается на несколько более простых подвопросов, для каждого из которых выполняется независимый поиск, после чего найденные фрагменты объединяются. Такой подход особенно эффективен для многошаговых вопросов, релевантные факты для которых распределены по разным документам.

PRF (Pseudo-Relevance Feedback) [13] – сначала выполняется поиск по исходному запросу, из найденных документов извлекаются наиболее частотные ключевые слова (за исключением стоп-слов), после чего исходный запрос расширяется этими ключевыми словами и выполняется повторный

поиск. Данный подход не требует обращения к БЯМ и может быть реализован с минимальными вычислительными затратами.

PRF+LLM [14] – модификация предыдущего метода, в которой извлечение ключевых слов из найденных документов выполняется не на основе частотного анализа, а с помощью большой языковой модели. Такой подход, хотя и более вычислительно затратный, потенциально позволяет выделять семантически значимые термины, не являющиеся простыми повторяющимися словами.

Recursive Query Refinement (рекурсивная/многошаговая переформулировка) [15] – итеративная стратегия, при которой на каждом шаге на основе предыдущего запроса и найденных по нему документов языковая модель генерирует уточненную версию запроса. Результаты поиска со всех шагов объединяются. Данный подход позволяет последовательно сужать область поиска, двигаясь от общих формулировок к более точным, и особенно полезен в случаях, когда релевантная информация не находится с первой попытки.

1.2. Переранжирование фрагментов с использованием кросс-энкодера

Одним из наиболее распространенных способов повышения качества поиска в RAG-системах является использование двухэтапной схемы извлечения фрагментов [16]. На первом этапе выполняется быстрый поиск кандидатов с использованием модели эмбединга, после чего найденные фрагменты повторно ранжируются более точной, но вычислительно затратной моделью – кросс-энкодером, который получает на вход одновременно пару «запрос – фрагмент» и непосредственно оценивает степень их семантического соответствия. В рамках работы для повторного ранжирования использовалась открытая модель `qilowoq/bge-reranker-v2-m3-en-ru`³, представляющая собой русскоязычную адаптацию (усеченную по токенам до поддержки только русского и английского языков) многоязычной модели `BAAI/bge-reranker-v2-m3` [17].

Алгоритм поиска состоит из следующих этапов:

1. формирование эмбединга пользовательского запроса с использованием модели эмбединга;
2. поиск наиболее релевантных фрагментов в векторном хранилище;
3. отбор первых N фрагментов-кандидатов;
4. вычисление оценки релевантности для каждой пары «запрос – фрагмент» с помощью кросс-энкодера;
5. повторное ранжирование найденных фрагментов по полученным оценкам и выбор первых k результатов.

1.3. Дообучение модели эмбединга

Подход к обучению моделей для получения семантически значимых эмбедингов предложений представлен в работе [18], а методы обучения

³Здесь и далее указание названия модели машинного обучения в виде X/Y означает, что используется модель из репозитория HuggingFace, доступная по URL: <https://huggingface.co/X/Y>.

моделей эмбединга для задач информационного поиска рассмотрены в [19]. Несмотря на высокое качество современных моделей эмбединга общего назначения [20], они обучаются на текстовых корпусах широкой тематики и не учитывают специфику нормативно-технической документации железнодорожного транспорта. Вследствие этого семантически связанные запросы и фрагменты документов могут располагаться в пространстве эмбедингов недостаточно близко, что приводит к ухудшению качества поиска релевантного контекста.

Целью дообучения является адаптация модели эмбединга к предметной области железнодорожного транспорта таким образом, чтобы вопросы пользователей располагались в векторном пространстве ближе к соответствующим нормативным пунктам. Предполагается, что подобная адаптация позволит повысить качество компонента поиска без изменения архитектуры RAG-системы.

Особенностью предлагаемого подхода является использование для обучения только простых вопросов и вопросов средней сложности, каждый из которых в большинстве случаев имеет единственный релевантный нормативный пункт. При этом оценка качества модели производится на независимом тестовом множестве сложных вопросов, для ответа на которые требуется извлечение нескольких взаимосвязанных нормативных пунктов. Данный подход позволяет оценить способность модели к обобщению и определить, способствует ли дообучение на простых примерах улучшению поиска для более сложных пользовательских запросов.

Для исследования были выбраны три современные модели, поддерживающие русский язык и широко применяемые в задачах семантического поиска:

- **ai-forever/ru-en-RoSBERTa** [21] выбрана как относительно компактная двуязычная модель, обеспечивающая высокую скорость вычисления эмбедингов. Использование данной модели позволяет оценить, насколько эффективной может быть предметная адаптация модели небольшого размера и способен ли ее дообученный вариант конкурировать с более современными моделями эмбединга;
- **BAAI/bge-m3** [17] представляет собой одну из наиболее современных открытых моделей, специально разработанных для задач информационного поиска. Модель изначально оптимизирована для этой задачи и демонстрирует высокое качество поиска без дополнительного обучения, поэтому представляет интерес с точки зрения оценки потенциала ее дальнейшей предметной адаптации;
- **intfloat/multilingual-e5-large** [22] относится к числу современных моделей эмбединга и широко применяется при построении RAG-систем.

Дообучение выполнялось с использованием библиотеки SentenceTransformers. В качестве обучающей выборки использовались пары вида (q_i, d_i) , где q_i – вопрос пользователя, а d_i – соответствующий нормативный пункт, содержащий правильный ответ.

В рамках подхода к дообучению для каждого положительного образца остальные фрагменты текущего батча автоматически рассматриваются как отрицательные образцы. Таким образом, модель обучается увеличивать коси-

нусное сходство между вопросом и соответствующим нормативным пунктом, одновременно уменьшая сходство между вопросом и всеми остальными фрагментами батча.

Дообучение всех моделей выполнялось по одинаковой схеме. Использовались две эпохи обучения, размер батча 32, скорость обучения $2 \cdot 10^{-5}$, коэффициент затухания весов 0,01. Обучение осуществлялось на графическом процессоре RTX 3090.

2. Корпус документов и наборы вопросов

В рамках исследования был сформирован корпус нормативно-технических документов на железнодорожную тематику, который был использован тремя способами: для формирования тестового набора данных в ходе оценки различных вариантов реализации RAG, для наполнения векторной базы данных (компонент извлечения в рамках RAG осуществлял поиск непосредственно среди множества текстовых фрагментов данного корпуса) и для формирования набора данных при дообучении модели эмбединга. В корпус вошли основные публичные нормативно-технические документы железнодорожной отрасли [23]:

- Устав железнодорожного транспорта Российской Федерации⁴;
- Правила технической эксплуатации железных дорог Российской Федерации (ПТЭ)⁵;
- Инструкция по организации движения поездов и маневровой работы на железнодорожном транспорте Российской Федерации (Приложение 2 к ПТЭ);
- Приложения 1–20 к Инструкции по организации движения (содержащие текстовые данные, остальные приложения содержат преимущественно формы);
- Правила перевозок грузов, порожних грузовых вагонов железнодорожным транспортом⁶.

Актуальные на дату написания статьи версии документов были загружены с сайта РЖД по ссылкам, приведенным выше (для Инструкции по организации движения и приложений к ней ссылки не указаны, так как данные документы доступны по ссылке на ПТЭ) в форматах HTML и DOCX.

Следует отметить, что, несмотря на то что одним из важных мотивов, определяющих актуальность проводимого исследования, является наличие множества документов, предназначенных для служебного пользования (ДСП), в корпус сознательно включались только открытые документы. С одной стороны, это позволило сделать сам корпус и все его производные открытыми, что способствует обеспечению воспроизводимости исследования и облегчает его развитие, с другой стороны, основные проблемы (в частности,

⁴Устав железнодорожного транспорта Российской Федерации // Веб-портал РЖД [Электронный ресурс]. URL: <https://company.rzd.ru/ru/9353/page/105104?id=154> (дата обращения: 13.07.2026).

⁵Правила технической эксплуатации железных дорог Российской Федерации // Веб-портал РЖД [Электронный ресурс]. URL: <https://company.rzd.ru/ru/9353/page/105104?id=1827> (дата обращения: 13.07.2026).

⁶Правила перевозок грузов, порожних грузовых вагонов железнодорожным транспортом // Веб-портал РЖД [Электронный ресурс]. URL: <https://company.rzd.ru/ru/9353/page/105104?id=1758> (дата обращения: 13.07.2026).

специфическая отраслевая терминология) проявляются и на сформированном корпусе документов, поэтому авторы считают, что результаты, полученные с его помощью, могут быть легко обобщены.

Первым шагом обработки корпуса при применении технологии RAG является разбиение его на фрагменты («чанки»), которые сохраняются в векторном хранилище и в дальнейшем служат фактографической базой при формировании ответов на вопросы пользователя (см. раздел 1). Особенностью нормативной документации и, в частности, всех документов, включенных в корпус, является высокая степень структуризации – документы разбиты на разделы и пункты (или статьи – в зависимости от вида документа), каждый из которых посвящен определенной ситуации, сценарию или действию. Подобная структуризация, наряду с высокой тематической сфокусированностью текста пунктов и их относительно небольшим размером, позволяет непосредственно использовать их в качестве фрагментов. В некоторых случаях текст пунктов, в свою очередь, бывает разбит на абзацы – при обработке корпуса данное разбиение также было учтено и сохранено в структурированном представлении корпуса, это может быть полезно при проведении экспериментов с альтернативными стратегиями разбиения на фрагменты, однако во всех экспериментах, проведенных в статье, в качестве фрагмента использовался весь текст пункта (или статьи).

Для оценки различных вариантов реализации RAG был сформирован набор вопросов трех уровней сложности:

- Простые вопросы, относящиеся к одному факту, например, «Для чего предназначен базовый размер исчисления сборов и штрафов?».
- Вопросы средней сложности, относящиеся к нескольким фактам, расположенным в одном или нескольких абзацах нормативного пункта, например, «В чем различие операций, выполняемых в местах общего пользования и в местах необщего пользования?» или «По каким основаниям перевозчик имеет право отказать в согласовании заявки на перевозку груза?».
- Вопросы повышенной сложности, требующие извлечения нескольких взаимосвязанных нормативных положений, например, «Каковы обязательства перевозчика и заявителя переадресовки в случае изменения грузополучателя и необходимости информирования сторон о переадресовке, если перевозчик является владельцем инфраструктуры?».

Всего в тестовое множество входит 150 вопросов, из которых 37 простых, 13 средней сложности и 100 повышенной сложности. Для оценки дообученных моделей эмбединга использовались только 100 вопросов повышенной сложности, тогда как в остальных экспериментах использовалось полное тестовое множество из 150 вопросов.

Формирование вопросов повышенной сложности выполнялось автоматически с использованием модели GPT-4o-mini. Генерация проводилась не по отдельным пунктам, а по заранее сформированным группам тематически связанных фрагментов. Для каждого нормативного пункта вычислялось векторное представление с помощью модели *intfloat/multilingual-e5-large*, после чего для случайно выбранного опорного пункта находились ближайшие по косинусному сходству фрагменты. Из найденных кандидатов формировались

группы размером два или три пункта. В зависимости от конфигурации эксперимента применялись три режима отбора: пункты из одного раздела одного документа; пункты из одного документа, но необязательно одного раздела; пункты без ограничения по документу и разделу.

Для формирования вопросов, отражающих различные информационные потребности пользователей, использовались три пользовательские роли:

1. поездной диспетчер, рассматривающий эксплуатационные ситуации, условия, обязанности и порядок действий;
2. специалист по нормативному контролю, сопоставляющий общие правила, специальные условия, ограничения и исключения;
3. разработчик технологического процесса, анализирующий последовательность операций и распределение ответственности.

Для каждой роли задавался предпочтительный набор типов вопросов, включая вопросы вида «условие – действие», описание процедуры, сопоставление общего правила и исключения, агрегацию требований и восстановление причинно-следственной последовательности.

Для предметного дообучения моделей эмбединга был сформирован отдельный синтетический набор пар «вопрос – релевантный нормативный пункт». БЯМ (GPT-4o-mini) передавались сведения об источнике, версии документа, разделе и номере пункта, а также полный текст всех его абзацев. Вопросы формировались в двух категориях сложности: простые и средней сложности.

Генерация выполнялась с использованием структурированного вывода. Температура генерации была установлена равной 0,2, что обеспечивало достаточное разнообразие формулировок при сохранении устойчивости и воспроизводимости. В результате было сформировано 3159 вопросов, из которых 2071 относились к категории простых, 1088 – к категории вопросов средней сложности.

3. Экспериментальное исследование

3.1. Методика

Целью экспериментального исследования являлась оценка эффективности различных способов повышения точности RAG-системы, предназначенной для работы с нормативно-технической документацией железнодорожного транспорта. Экспериментальное исследование проводилось последовательно в несколько этапов.

На первом этапе выполнялось сравнение трех базовых моделей эмбединга (ai-forever/ru-en-RoSBERTa, BAAI/bge-m3 и intfloat/multilingual-e5-large) без дополнительной предметной адаптации.

На втором этапе исследовалось влияние переранжирования документов с использованием кросс-энкодера qilowoq/bge-reranker-v2-m3-en-ru. Для каждой модели эмбединга оценивалось качество поиска при различном количестве фрагментов-кандидатов, подлежащих повторному ранжированию.

На третьем этапе исследовалось влияние стратегий предварительной обработки пользовательских запросов на качество поиска.

На четвертом этапе выполнялось предметное дообучение моделей эмбединга на сформированном наборе простых вопросов и вопросов средней сложности. Исследовалось влияние объема обучающей выборки (500, 1000 и 3000 вопросов) на качество поиска. Оценка проводилась на независимом тестовом множестве вопросов повышенной сложности, что позволило определить способность модели переносить приобретенные знания на более сложные пользовательские запросы.

На заключительном этапе проводилась оценка качества генерации ответов БЯМ с использованием модели эмбединга, выбранной по результатам экспериментов по оценке качества поиска.

Для количественной оценки качества поиска использовались следующие показатели:

- **Recall@10** характеризует полноту поиска и показывает долю релевантных фрагментов, найденных среди первых десяти результатов;
- **MRR@10** (Mean Reciprocal Rank) характеризует качество ранжирования и определяется как среднее значение обратного ранга первого найденного релевантного фрагмента;
- **NDCG@10** (Normalized Discounted Cumulative Gain) характеризует качество ранжирования с учетом как полноты поиска, так и позиций всех релевантных фрагментов.

Для оценки качества генерации рассматривались две БЯМ одного семейства – Qwen3 14B и Qwen3 32B [24]. В обоих случаях генерация выполнялась без режима рассуждения (Thinking), а в качестве контекста языковой модели передавались десять наиболее релевантных фрагментов, найденных с использованием модели эмбединга multilingual-e5-large, дообученной на выборке из 3000 синтетических вопросов.

Качество генерации оценивалось с использованием библиотеки Ragas⁷ [8], реализующей подход LLM-as-a-Judge, в рамках которого в качестве модели-судьи использовалась БЯМ GPT-4o-mini. Оценка выполнялась автоматически по следующим показателям:

- **Достоверность ответа** (Faithfulness) – степень, в которой утверждения, содержащиеся в ответе, подтверждаются предоставленным контекстом;
- **Релевантность ответа** (Answer Relevancy) – степень соответствия сгенерированного ответа поставленному вопросу;
- **Семантическое сходство** (Semantic Similarity) – степень семантической близости сгенерированного ответа к эталонному ответу независимо от совпадения формулировок.

3.2. Результаты

⁷Официальный сайт библиотеки Ragas [Электронный ресурс]. URL: <https://www.ragas.io/> (дата обращения: 13.07.2026).

Базовый поиск с использованием моделей эмбединга

На первом этапе была проведена оценка качества семантического поиска с использованием трех моделей эмбединга без дополнительного предметного дообучения. Результаты представлены в таблице 1.

Таблица 1. Результаты оценки базовых моделей эмбединга

Модель эмбединга	Recall@10	MRR@10	NDCG@10
ru-en-RoSBERTa	0,640	0,582	0,543
multilingual-e5-large	0,712	0,672	0,608
bge-m3	0,724	0,638	0,597

Полученные результаты показывают, что модели эмбединга multilingual-e5-large и bge-m3 существенно превосходят ru-en-RoSBERTa по всем рассматриваемым показателям качества поиска. Наибольшее значение Recall@10 продемонстрировала модель bge-m3 (0,724), что свидетельствует о ее более высокой способности находить релевантные фрагменты. Вместе с тем по показателям MRR@10 (0,672) и NDCG@10 (0,608), характеризующим качество ранжирования поисковой выдачи, лучшие результаты показала multilingual-e5-large, что означает, что данная модель чаще располагает релевантные фрагменты на первых позициях поисковой выдачи.

Результаты использования кросс-энкодера

Для оценки влияния повторного ранжирования документов была реализована двухэтапная схема поиска. На первом этапе модель эмбединга извлекала заданное число фрагментов-кандидатов (20, 50 или 100), после чего они повторно ранжировались кросс-энкодером. Полученные результаты представлены на рисунке 2. Значение 0 на горизонтальной оси соответствует базовому варианту поиска без использования переранжирования.

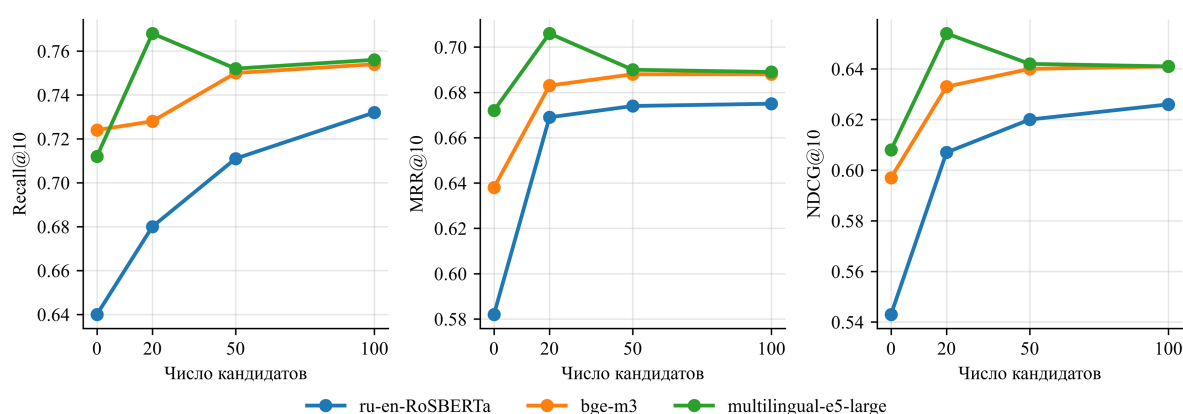


Рис. 2. Влияние переранжирования кросс-энкодером на качество семантического поиска

Полученные результаты показывают, что использование кросс-энкодера позволяет повысить качество поиска для всех рассмотренных моделей эмбединга, однако величина прироста существенно зависит от качества исходного ранжирования. Наибольший эффект наблюдается для модели ru-en-RoSBERTa. По сравнению с базовым вариантом значение Recall@10

увеличилось с 0,640 до 0,732, MRR@10 – с 0,582 до 0,675, а NDCG@10 – с 0,543 до 0,626. При этом увеличение числа фрагментов-кандидатов с 20 до 100 приводит к последовательному улучшению всех рассматриваемых показателей.

Для модели bge-m3 также наблюдается положительное влияние увеличения числа документов-кандидатов. При переходе от 20 к 100 кандидатам значение Recall@10 возрастает с 0,728 до 0,754, а показатели MRR@10 и NDCG@10 демонстрируют умеренный, но устойчивый рост.

Иная картина наблюдается для multilingual-e5-large. Наилучшие результаты достигаются уже при повторном ранжировании первых 20 фрагментов-кандидатов (Recall@10 = 0,768, MRR@10 = 0,706, NDCG@10 = 0,654). Дальнейшее увеличение числа кандидатов до 50 и 100 приводит к небольшому снижению всех трех показателей. Вероятно, это связано с тем, что исходное ранжирование модели уже обладает высоким качеством, а включение менее релевантных документов увеличивает количество потенциальных ошибок повторного ранжирования.

Таким образом, эффективность использования кросс-энкодера определяется не только его собственной способностью различать релевантные фрагменты, но и качеством первоначальной поисковой выдачи. Для моделей с менее точным ранжированием (ru-en-RoSBERTa) увеличение числа кандидатов обеспечивает существенный прирост качества, тогда как для более сильной модели эмбединга (multilingual-e5-large) оптимальным оказывается повторное ранжирование сравнительно небольшого числа фрагментов.

Результаты применения различных стратегий обработки запроса

Результаты применения различных стратегий обработки запроса приведены в таблице 2. В данном эксперименте использовалась модель эмбединга multilingual-e5-large без переранжирования кросс-энкодером, а для формулировки альтернативных (или дополнительных) версий запроса – БЯМ GPT OSS 20B. Вопреки ожиданиям авторов статьи, продвинутые методы обработки запросов не привели к заметному повышению качества поиска релевантных фрагментов. Большинство методов показали примерно тот же уровень качества извлечения, что и базовая стратегия без дополнительной обработки. По результатам эксперимента наиболее перспективной выглядит стратегия StepBack, однако ее преимущество не является значительным, применение же стратегии HyDE привело к существенному падению качества результатов.

Таблица 2. Результаты использования различных стратегий обработки запроса

Стратегия обработки	Recall@10	MRR@10	NDCG@10
Без доп. обработки	0,712	0,672	0,608
HyDE	0,599	0,558	0,499
StepBack	0,711	0,688	0,611
MultiQuery	0,692	0,642	0,577
Decomposition	0,723	0,603	0,571
PRF+LLM	0,700	0,641	0,590
PRF без LLM	0,704	0,648	0,599
Recursive	0,682	0,659	0,590

Результаты дообучения моделей эмбединга

Результаты дообучения моделей эмбединга представлены на рисунке 3. На графиках показано изменение значений Recall@10, MRR@10 и NDCG@10 в зависимости от размера обучающей выборки.

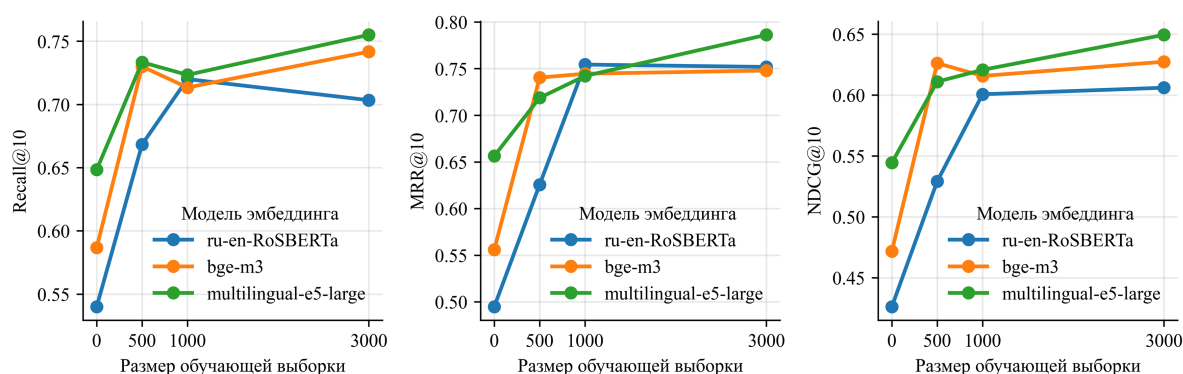


Рис. 3. Влияние дообучения моделей эмбединга на качество семантического поиска

Дообучение приводит к существенному повышению качества поиска для всех исследуемых моделей. При этом наибольший прирост наблюдается уже при использовании сравнительно небольших обучающих выборок, тогда как дальнейшее увеличение их объема обеспечивает значительно менее выраженный эффект.

Наиболее заметное влияние предметная адаптация оказала на модель ai-forever/ru-en-RoSBERTa. По сравнению с базовым вариантом значение Recall@10 увеличилось с 0,540 до 0,720 при обучении на 1000 вопросах, MRR@10 – с 0,495 до 0,754, а NDCG@10 – с 0,426 до 0,601. Дальнейшее увеличение размера обучающей выборки до 3000 вопросов не привело к существенному улучшению качества. Вероятно, это связано с тем, что для относительно компактной модели потенциал предметной адаптации оказался практически исчерпан уже при обучении на 1000 вопросах. Для модели BAAI/bge-m3 также наблюдается устойчивое улучшение всех исследуемых показателей.

Наиболее стабильный рост качества демонстрирует модель multilingual-e5-large. При обучении на 3000 вопросах модель достигает лучших результатов среди всех рассмотренных вариантов: Recall@10 составляет 0,755, MRR@10 – 0,786, а NDCG@10 – 0,649.

В целом, результаты подтверждают эффективность предложенного подхода к предметному дообучению моделей эмбединга. При этом наиболее важным является тот факт, что обучение выполнялось исключительно на относительно простых вопросах, тогда как оценка производилась на независимом наборе вопросов повышенной сложности, что говорит о способности моделей переносить знания, полученные на относительно простых примерах, на более сложные сценарии поиска, требующие извлечения нескольких взаимосвязанных фрагментов.

Результаты оценки качества генерации ответов

Полученные результаты представлены в таблице 3. В таблице приведены средние значения показателей и соответствующие им стандартные отклонения.

Таблица 3. Результаты оценки качества генерации ответов БЯМ

БЯМ	Faithfulness	Answer Relevancy	Semantic Similarity
Qwen3 14B	0,839 ± 0,216	0,913 ± 0,100	0,939 ± 0,020
Qwen3 32B	0,833 ± 0,202	0,896 ± 0,136	0,931 ± 0,021
<i>p</i> -значение	0,667	0,078	< 0,001

Модель Qwen3 14B продемонстрировала несколько более высокие значения по всем рассматриваемым показателям по сравнению с Qwen3 32B. Для оценки статистической значимости различий между моделями использовался парный критерий Уилкоксона. Статистически значимых различий по показателям Faithfulness ($p = 0,667$) и Answer Relevancy ($p = 0,078$) не выявлено. Для показателя Semantic Similarity различие оказалось статистически значимым ($p < 0,001$), однако его абсолютная величина невелика (разница составляет 0,008), поэтому данное различие не имеет существенной практической значимости.

Заключение

В работе проведен сравнительный анализ методов повышения точности RAG-систем на корпусе нормативно-технической документации в области железнодорожного транспорта. Исследованы три группы методов: стратегии обработки запросов, двухэтапное извлечение с использованием кросс-энкодера и дообучение моделей эмбединга.

Установлено, что современные модели эмбединга существенно превосходят более компактные аналоги по всем показателям качества поиска. Применение кросс-энкодера для повторного ранжирования фрагментов позволяет повысить качество поиска, однако величина эффекта определяется качеством исходного ранжирования: для моделей с менее точной выдачей прирост значителен, тогда как для сильных моделей оптимальным является ранжирование небольшого числа кандидатов.

Стратегии предварительной обработки запросов не обеспечили значимого повышения качества извлечения по сравнению с базовым подходом, а в ряде случаев привели к ухудшению результатов.

Наибольшую эффективность среди всех исследованных методов показало предметное дообучение моделей эмбединга на синтетических парах «вопрос – нормативный пункт». Обучение на относительно простых вопросах обеспечило устойчивое улучшение поиска по сложным запросам, что свидетельствует о способности моделей к обобщению. Сравнение языковых моделей разного размера не выявило существенного преимущества более крупной модели при формировании ответов на основе предоставленного контекста.

Таким образом, в условиях ограниченной вычислительной инфраструктуры и необходимости работы с закрытыми документами, содержащими

специфическую отраслевую терминологию, наиболее практичным инструментом повышения точности RAG-систем является предметное дообучение модели эмбединга. Данный подход позволяет существенно повысить качество поиска без дополнительных вычислительных затрат на этапе выполнения и без риска утечки данных. Предложенная методология может быть обобщена на другие предметные области со схожей структурой нормативно-технической документации.

Список литературы

- [1] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus [et al.] // *Advances in Neural Information Processing Systems*. – 2020. – Vol. 33. – P. 9459–9474.
- [2] Retrieval-Augmented Generation for Large Language Models: A Survey / Y. Gao, Y. Xiong, X. Gao [et al.] // *arXiv*. – 2023. – arXiv:2312.10997. – DOI 10.48550/arXiv.2312.10997.
- [3] A Review of Large Language Models in Medical Education, Clinical Decision Support, and Healthcare Administration / J. Vrdoljak, Z. Boban, M. Vilović [et al.] // *Healthcare*. – 2025. – Vol. 13, № 6. – Art. 603. – DOI 10.3390/healthcare13060603.
- [4] A Comprehensive Survey of Retrieval-Augmented Large Language Models for Decision Making in Agriculture: Unsolved Problems and Research Opportunities / A. Vizniuk, G. Diachenko, I. Laktionov [et al.] // *Journal of Artificial Intelligence and Soft Computing Research*. – 2025. – Vol. 15, № 2. – P. 115–146. – DOI 10.2478/jaiscr-2025-0007.
- [5] Ершов, М. А. Интеллектуальная СППР для оценки конкурентоспособности в электронной коммерции: методология интеграции RAG-системы / М. А. Ершов // *Экономика строительства*. – 2025. – № 11. – С. 518–521.
- [6] Самофалов, Г. С. Интеграция больших языковых моделей (LLM) и RAG в контуры поддержки принятия решений (DSS) для задач ценообразования и управления спросом в розничной торговле / Г. С. Самофалов // *Интернаука*. – 2026. – № 16–1. – С. 38–42.
- [7] Cioplea, R. Memory in the Age of AI: How LLMs Remember, Forget, and Leak : Technical Report, ver. 3.0. – 2026. – 37 p. – DOI 10.17605/OSF.IO/9WYG3.
- [8] RAGAS: Automated Evaluation of Retrieval-Augmented Generation / S. Es, J. James, L. E. Anke, S. Schockaert // *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.
- [9] Precise Zero-Shot Dense Retrieval without Relevance Labels / L. Gao, X. Ma, J. Lin, J. Callan // *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. – 2023. – P. 1762–1777. – DOI 10.18653/v1/2023.acl-long.99.
- [10] Take a Step Back: Evoking Reasoning via Abstraction in Large Language Models / H. S. Zheng, S. Mishra, X. Chen [et al.] // *International Conference on Learning Representations (ICLR 2024)*. – 2024.
- [11] DMQR-RAG: Diverse Multi-Query Rewriting for RAG / Z. Li [et al.] // *arXiv*. – 2024. – arXiv:2411.13154. – DOI 10.48550/arXiv.2411.13154.
- [12] Ammann, P. J. L. Question Decomposition for Retrieval-Augmented Generation / P. J. L. Ammann, J. Golde, A. Akbik // *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. – 2025. – P. 497–507. – DOI 10.18653/v1/2025.acl-srw.32.
- [13] Salton, G. Improving Retrieval Performance by Relevance Feedback / G. Salton, C. Buckley // *Journal of the American Society for Information Science*. – 1990. – Vol. 41, № 4. – P. 288–297.
- [14] Mackie, I. Generative relevance feedback with large language models / I. Mackie, S. Chatterjee, J. Dalton // *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. – 2023. – P. 2026–2031.

- [15] Wang, Y. MaFeRw: Query Rewriting with Multi-Aspect Feedbacks for Retrieval-Augmented Large Language Models / Y. Wang // arXiv. – 2024. – arXiv:2408.17072. – DOI 10.48550/arXiv.2408.17072.
- [16] Nogueira, R. Passage Re-Ranking with BERT / R. Nogueira, K. Cho // arXiv. – 2019. – arXiv:1901.04085. – DOI 10.48550/arXiv.1901.04085.
- [17] M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation / J. Chen [et al.] // arXiv. – 2024. – arXiv:2402.03216. – DOI 10.48550/arXiv.2402.03216.
- [18] Reimers, N. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks / N. Reimers, I. Gurevych // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). – 2019. – P. 3982–3992. – DOI 10.18653/v1/D19-1410.
- [19] Text Embeddings by Weakly-Supervised Contrastive Pre-Training / L. Wang, N. Yang, X. Huang [et al.] // arXiv. – 2022. – arXiv:2212.03533. – DOI 10.48550/arXiv.2212.03533.
- [20] MTEB: Massive Text Embedding Benchmark / N. Muennighoff, N. Tazi, L. Magne, N. Reimers // arXiv. – 2022. – arXiv:2210.07316. – DOI 10.48550/arXiv.2210.07316.
- [21] The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design / A. Snegirev [et al.] // arXiv. – 2024. – arXiv:2408.12503. – DOI 10.48550/arXiv.2408.12503.
- [22] Multilingual E5 Text Embeddings: A Technical Report / L. Wang [et al.] // arXiv. – 2024. – arXiv:2402.05672. – DOI 10.48550/arXiv.2402.05672.
- [23] Паршин, К. А. Исследование применения признаков терминов железнодорожной отрасли при формировании классификатора / К. А. Паршин, М. С. Подгорный // Теория и практика современной науки. – 2019. – № 3. – С. 229–232.
- [24] Qwen3 Technical Report / A. Yang [et al.] // arXiv. – 2025. – arXiv:2505.09388. – DOI 10.48550/arXiv.2505.09388.

COMPARATIVE ANALYSIS OF METHODS FOR IMPROVING RAG SYSTEM ACCURACY ON SPECIALIZED CORPORA OF REGULATORY AND TECHNICAL DOCUMENTATION

Agafonov A.A.^{1, a}, Ponomarev A.V.^{1, b}

¹ St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia.

E-mail: ^a agafonov.a@spcras.ru, ^b ponomarev@iiias.spb.su

Abstract. The digital transformation of industry drives a growing demand for intelligent decision support systems capable of efficiently extracting knowledge from regulatory and technical documentation. This paper examines Retrieval-Augmented Generation (RAG) technology applied to a specialized corpus of railway regulatory documents. A comparative analysis is conducted for several methods aimed at improving RAG accuracy, including query processing strategies (HyDE, Step-Back, Multi-Query, Decomposition, Pseudo-Relevance Feedback, Recursive Refinement), two-stage retrieval with a cross-encoder, and fine-tuning of embedding models on synthetic question-regulatory clause pairs. The experiments demonstrate that domain-specific

For citation:

Agafonov A.A., Ponomarev A.V. Comparative Analysis of Methods for Improving RAG System Accuracy on Specialized Corpora of Regulatory and Technical Documentation. *Intelligent transport*, 2026, no. 3(39), pp. 4–22. DOI: 10.24412/3033-6007-2026-339-4-22. (in Russian)

fine-tuning of embedding models yields the most significant improvement in retrieval quality. Cross-encoder reranking also provides a positive effect, particularly for models with initially less accurate ranking. At the same time, advanced query processing strategies do not lead to a substantial improvement in retrieval quality compared to the baseline approach. The constructed document corpus, question sets, and source code are made publicly available.

Keywords: RAG, large language models, semantic search, embedding models, cross-encoder, regulatory and technical documentation, railway transport, fine-tuning.

References

- [1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- [2] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., et al. (2023). *Retrieval-augmented generation for large language models: A survey* (arXiv:2312.10997). arXiv. <https://doi.org/10.48550/arXiv.2312.10997>
- [3] Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A review of large language models in medical education, clinical decision support, and healthcare administration. *Healthcare*, 13(6), Article 603. <https://doi.org/10.3390/healthcare13060603>
- [4] Vizniuk, A., Diachenko, G., Laktionov, I., Siwocha, A., Xiao, M., & Smolaq, J. (2025). A comprehensive survey of retrieval-augmented large language models for decision making in agriculture: Unsolved problems and research opportunities. *Journal of Artificial Intelligence and Soft Computing Research*, 15(2), 115–146. <https://doi.org/10.2478/jaiscr-2025-0007>
- [5] Ershov, M. A. (2025). Intelligent DSS for competitiveness assessment in e-commerce: Methodology of RAG-system integration. *Ekonomika stroitelstva*, (11), 518–521. (in Russian)
- [6] Samofalov, G. S. (2026). Integration of large language models (LLM) and RAG into decision support systems (DSS) for pricing and demand management in retail. *Internauka*, (16-1), 38–42. (in Russian)
- [7] Cioplea, R. (2026). *Memory in the age of AI: How LLMs remember, forget, and leak* (Technical Report ver. 3.0). <https://doi.org/10.17605/OSF.IO/9WXG3>
- [8] Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). RAGAS: Automated evaluation of retrieval-augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). <https://doi.org/10.18653/v1/2024.eacl-demo.16>
- [9] Gao, L., Ma, X., Lin, J., & Callan, J. (2023). Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1762–1777). <https://doi.org/10.18653/v1/2023.acl-long.99>
- [10] Zheng, H. S., Mishra, S., Chen, X., Cheng, H.-T., Chi, E. H., Le, Q. V., & Zhou, D. (2024). Take a step back: Evoking reasoning via abstraction in large language models. In *International Conference on Learning Representations (ICLR 2024)*.
- [11] Li, Z., et al. (2024). *DMQR-RAG: Diverse multi-query rewriting for RAG* (arXiv:2411.13154). arXiv. <https://doi.org/10.48550/arXiv.2411.13154>
- [12] Ammann, P. J. L., Golde, J., & Akbik, A. (2025). Question decomposition for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 497–507). <https://doi.org/10.18653/v1/2025.acl-srw.32>

- [13] Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4), 288–297.
- [14] Mackie, I., Chatterjee, S., & Dalton, J. (2023). Generative relevance feedback with large language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2026–2031).
- [15] Wang, Y. (2024). *MaFeRw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models* (arXiv:2408.17072). arXiv. <https://doi.org/10.48550/arXiv.2408.17072>
- [16] Nogueira, R., & Cho, K. (2019). *Passage re-ranking with BERT* (arXiv:1901.04085). arXiv. <https://doi.org/10.48550/arXiv.1901.04085>
- [17] Chen, J., et al. (2024). *M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation* (arXiv:2402.03216). arXiv. <https://doi.org/10.48550/arXiv.2402.03216>
- [18] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). <https://doi.org/10.18653/v1/D19-1410>
- [19] Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., et al. (2022). *Text embeddings by weakly-supervised contrastive pre-training* (arXiv:2212.03533). arXiv. <https://doi.org/10.48550/arXiv.2212.03533>
- [20] Muennighoff, N., Tazi, N., Magne, L., & Reimers, N. (2022). *MTEB: Massive text embedding benchmark* (arXiv:2210.07316). arXiv. <https://doi.org/10.48550/arXiv.2210.07316>
- [21] Snegirev, A., et al. (2024). *The Russian-focused embedders' exploration: ruMTEB benchmark and Russian embedding model design* (arXiv:2408.12503). arXiv. <https://doi.org/10.48550/arXiv.2408.12503>
- [22] Wang, L., et al. (2024). *Multilingual E5 text embeddings: A technical report* (arXiv:2402.05672). arXiv. <https://doi.org/10.48550/arXiv.2402.05672>
- [23] Parshin, K. A., & Podgornyy, M. S. (2019). A study of the application of railway industry term features in the formation of a classifier. *Teoriya i praktika sovremennoy nauki*, (3), 229–232. (in Russian)
- [24] Yang, A., et al. (2025). *Qwen3 technical report* (arXiv:2505.09388). arXiv. <https://doi.org/10.48550/arXiv.2505.09388>