УДК 004.891.2:336.02
DOI 10.26118/2782-4586.2026.72.76.051

Лыков Артем Валерьевич
Кубанский государственный технологический университет
Мурлин Алексей Георгиевич
Кубанский государственный технологический университет

Повышение надёжности рекомендаций искусственного интеллекта в корпоративных финансах с помощью Retrieval-Augmented Generation

Аннотация. Рассматривается применение архитектуры RAG (Retrieval-Augmented Generation, генерация, дополненная поиском) для задач корпоративного финансового анализа. Предлагается прототип системы, объединяющей механизм поиска по документам и генеративную модель обработки естественного языка. В качестве источников используется демонстрационный корпус финансовых документов, включающий годовые отчёты компаний, регуляторные материалы и отраслевые ориентиры. Реализован гибридный механизм поиска, сочетающий поиск по ключевым словам и векторное представление текста с доменной приоритизацией документов по компании и отчётному периоду. Описана архитектура прототипа, включающая этапы индексации документов, извлечения релевантных фрагментов и генерации аналитического ответа на основе найденного контекста. Показано, что использование поискового компонента повышает интерпретируемость результатов и обеспечивает согласованность аналитических выводов с источниками данных. Ограничения прототипа связаны с демонстрационным объёмом корпуса документов и применением эвристической процедуры проверки фактологической опоры ответа.

Ключевые слова: искусственный интеллект, большие языковые модели, Retrieval-Augmented Generation, генерация, дополненная поиском, корпоративные финансы, финансовый анализ, анализ отчётности, кредитный риск, долговая нагрузка, рентабельность капитала, интерпретируемость моделей, гибридный поиск, векторное представление текста, информационный поиск.

Lykov Artem Valerievich
Kuban State Technological University
Murlin Alexey Georgievich
Kuban State Technological University

Improving the reliability of AI recommendations in corporate finance with the help of Retrieval-Augmented Generation

Abstract. The paper examines the application of the RAG (Retrieval-Augmented Generation) architecture to corporate financial analysis tasks. A prototype system combining document retrieval mechanisms with a generative natural language processing model is described. The system operates on a demonstration corpus of financial documents including corporate annual reports, regulatory materials, and sector benchmarks. A hybrid retrieval mechanism is implemented that combines keyword-based search and vector representations of text with domain-specific prioritization of documents by company and reporting period. The architecture of the prototype includes document indexing, retrieval of relevant fragments, and generation of analytical responses based on the retrieved context. The results indicate that integrating a retrieval component improves interpretability and ensures that generated analytical conclusions are grounded in source documents. The limitations of the prototype are related to the small size of the document corpus and the use of heuristic procedures for verifying factual grounding of generated responses.

Key words: artificial intelligence, large language models, Retrieval-Augmented Generation, corporate finance, financial analysis, financial reporting analysis, credit risk, leverage analysis, return on equity, model interpretability, hybrid retrieval, vector embeddings, information retrieval.

Введение

Цифровизация корпоративных финансов сопровождается активным внедрением систем искусственного интеллекта для автоматизации анализа отчётности, оценки кредитных рисков и подготовки управленческих рекомендаций. Большие языковые модели демонстрируют высокую эффективность при обработке текстовой информации, однако их применение в финансовом контуре ограничено проблемой фактических искажений. Модель способна формировать правдоподобный, но неверный вывод, что в условиях финансового анализа недопустимо.

В корпоративной среде ошибка в интерпретации коэффициентов рентабельности, неверный расчёт ожидаемых потерь или некорректная ссылка на регуляторные требования может привести к искажению управленческого решения. Это особенно критично при работе с нормативными документами и стандартами регулирования банковской деятельности, включая Положение Банка России № 590-П и требования МСФО 9 (IFRS 9, International Financial Reporting Standard 9, «Финансовые инструменты») [2–4].

Одним из направлений снижения фактических искажений является использование подхода RAG [1, 8], предполагающего дополнение генерации поиском релевантных документов. В этом случае языковая модель формирует ответ не только на основе внутренних параметров, но и с опорой на внешние источники данных. Для финансовых задач такой подход представляется перспективным, поскольку позволяет привязывать вывод к конкретным показателям отчётности и регуляторным нормам.

Цель исследования заключается в разработке и апробации прототипа системы на основе архитектуры RAG, ориентированной на задачи корпоративного финансового анализа. Подход предполагает использование гибридного механизма поиска, учитывающего доменную специфику финансовых данных и структуру аналитических запросов, а также программную реализацию прототипа с последующей проверкой его функционирования на типовых сценариях финансовой аналитики.

Научная новизна состоит в адаптации архитектуры RAG к задачам корпоративных финансов с учётом структуры финансовых показателей, регуляторных требований и особенностей аналитических запросов. Предложена схема гибридного поиска с приоритизацией документов по компании и актуальности источника, а также реализован механизм эвристической проверки фактологической опоры формируемого ответа на retrieved-документы.

В связи с этим актуальна практическая значимость, которая определяется возможностью повышения воспроизводимости и интерпретируемости результатов генерации при использовании языковых моделей в задачах корпоративного финансового анализа. Предложенный прототип может рассматриваться как основа для дальнейшего масштабирования и интеграции подобных решений в корпоративные информационные системы.

Теоретические основы исследования

Галлюцинации больших языковых моделей в финансовых задачах

Галлюцинацией больших языковых моделей является генерация правдоподобной, но фактически некорректной информации [10, 11], не подтверждаемой источниками или контекстом запроса. В финансовых задачах подобные искажения представляют повышенный риск, поскольку результаты генерации могут использоваться в процессе принятия управленческих решений.

В корпоративных финансах галлюцинации проявляются в нескольких типовых формах. К ним относятся фактические ошибки, выражающиеся в неверных значениях

финансовых коэффициентов, исторические искажения данных по периодам отчётности, некорректная интерпретация регуляторных требований, а также нереалистичные прогнозные допущения при моделировании денежных потоков. Даже небольшое отклонение в показателях рентабельности или долговой нагрузки способно привести к существенным искажениям оценки компании [5, 6].

В финансовой предметной области корректность ответа может быть проверена конкретными документами: годовыми отчётами, регуляторными актами, внутренним методикам. Это создаёт предпосылки для интеграции генеративных моделей с механизмами поиска и опоры на внешние источники данных.

Снижение вероятности подобных искажений требует использования методов, обеспечивающих опору генеративной модели на внешние источники данных. Одним из наиболее распространённых решений является архитектура RAG, предполагающая объединение механизмов поиска и генерации текста.

Retrieval-Augmented Generation: архитектура и принципы работы

Подход RAG представляет собой архитектурную схему, в которой генерация ответа дополняется предварительным поиском релевантных документов. В отличие от использования языковой модели в изолированном режиме, RAG предполагает, что формирование ответа осуществляется с учётом извлечённого контекста [8, 9].

В общем виде система включает два функциональных контура: модуль поиска и модуль генерации. Поисковый модуль формирует набор релевантных документов на основе текстового запроса. Для этого могут применяться как разреженные методы поиска по ключевым словам, так и векторные представления текста, что соответствует современным направлениям развития поисковых архитектур [9, 12, 14]. Полученные документы ранжируются по степени релевантности, после чего их содержимое используется в качестве дополнительного контекста при формировании ответа.

В отличие от полной донастройки языковой модели на специализированном датасете, retrieval-подход позволяет работать с внешними документами без изменения параметров модели. Это упрощает внедрение и обеспечивает привязку результата к конкретным источникам.

Метрики оценки качества RAG-систем

Оценка качества RAG-систем в финансовых задачах должна учитывать не только лингвистическую корректность ответа, но и его фактическую обоснованность. В контексте прототипа учитываются показатели retrieval-качества, фактологической опоры ответа и корректности извлечения финансовых метрик.

Показатели поискового качества, отражающие число найденных документов и степень их релевантности запросу. В рамках реализованной системы применяется нормированная оценка согласованности поискового ранжирования, рассчитываемая на основе распределения сходства между найденными документами.

Показатель фактологической опоры ответа на поисковые документы. В прототипе реализована эвристическая проверка наличия извлечённых метрик в исходных документах, что позволяет оценить, опирается ли ответ на фактический контекст.

Учитывается количество извлечённых финансовых показателей и их соответствие содержанию документов. Данный критерий отражает способность системы корректно интерпретировать финансовые данные.

Для полноформатных исследований могут применяться стандартные метрики качества ответов, такие как точность совпадения, F1-score (F1-мера, гармоническое среднее точности и полноты) и специализированные показатели согласованности с источником. Эти метрики рассматриваются как направление дальнейшего расширения эксперимента при масштабировании корпуса и набора запросов.

Методика экспериментального исследования

Корпус финансовых документов и структура данных

Для исследования использован демонстрационный корпус финансовых документов, сформированный для проверки работоспособности RAG в задачах корпоративной аналитики. Корпус включает десять текстовых документов, отражающих типовые источники финансовой информации: годовые отчёты компаний, материалы инвесторских презентаций, регуляторные документы и отраслевые бенчмарки.

Каждый документ содержит текстовую часть и набор метаданных, включающих наименование компании, год публикации и тип источника. Часть финансовых показателей представлена в структурированном виде, что позволяет проверить корректность их извлечения и интерпретации.

Назначение корпуса состоит в проверке логики работы поискового контура, корректности извлечения метрик и устойчивости системы к фактическим искажениям. Исходный код экспериментальной реализации размещён в открытом репозитории [15].

Для апробации использованы типовые запросы, отражающие задачи корпоративных финансов:

- анализ динамики показателей рентабельности;
- расчёт кредитного резерва на основе PD (Probability of Default, вероятность дефолта) и LGD (Loss Given Default, доля потерь при дефолте);
- сопоставление EBITDA (Earnings Before Interest, Taxes, Depreciation and Amortization, прибыль до вычета процентов, налогов и амортизации) с отраслевым ориентиром;
- оценка долговой нагрузки и рисков рефинансирования.

Такой набор сценариев позволяет протестировать систему как в рамках микроаналитики отдельной компании, так и при работе с нормативными требованиями.

Архитектура и программная реализация прототипа

Прототип системы реализован на языке Python. Архитектура построена по принципу разделения на поисковый и генеративный контуры.

Поисковый модуль реализует гибридный механизм, сочетающий:

- поиск по ключевым словам (разрежённый подход);
- упрощённые векторные представления текста, основанные на TF-IDF (Term Frequency – Inverse Document Frequency, частота терма – обратная частота документа).

Результаты двух механизмов объединяются, после чего формируется нормированная оценка релевантности документов. Дополнительно применяется приоритизация по совпадению компании и году публикации, что повышает доменную согласованность ответа.

По итогам ранжирования отбираются top-k документов при фиксированном пороге релевантности. В текущей конфигурации используются следующие параметры:

- k = 5;
- порог релевантности 0,75;
- нормированная оценка согласованности retrieval в диапазоне от 0 до 1.

Отобранные документы используются в качестве контекста для формирования аналитического ответа. В прототипе генерация строится на основе шаблонной логики и извлечённых финансовых метрик, что позволяет контролировать фактическую корректность результата.

Дополнительно реализован механизм проверки фактологической опоры ответа. Он анализирует, присутствуют ли ключевые показатели, указанные в ответе, в retrieved-документах. Результат проверки фиксируется в виде логического индикатора.

Экспериментальные сценарии

Апробация системы проведена на наборе демонстрационных сценариев, отражающих практические задачи корпоративной аналитики.

Первый сценарий связан с анализом изменения показателя рентабельности собственного капитала за два последовательных года и выявлением факторов динамики. Второй сценарий посвящён расчёту ожидаемых кредитных потерь на основе заданных

значений PD и LGD с учётом нормативных требований. Третий сценарий предполагает сопоставление показателя EBITDA компании с отраслевым ориентиром. Четвёртый сценарий направлен на оценку долговой нагрузки и рисков рефинансирования.

Подобный формат апробации позволяет продемонстрировать работу ключевых компонентов системы и выявить типовые ситуации, в которых retrieval-подход повышает согласованность ответа с источниками.

В условиях демонстрационного прототипа основной акцент сделан на оценке согласованности ответа с retrieved-контекстом, а не на статистическом сопоставлении с большим эталонным набором ответов.

Оценка качества включает следующие показатели:

1. Количество найденных релевантных документов и значение нормированной поисковой оценки.
2. Число корректно извлечённых финансовых метрик.
3. Результат проверки фактологической опоры ответа.
4. Время формирования ответа, характеризующее применимость системы в операционном режиме.

Классические метрики совпадения текста, такие как Exact Match (метрика точного совпадения ответа) или F1-score, в рамках настоящего прототипа не применяются ввиду отсутствия масштабного эталонного набора ответов. Их использование рассматривается как направление дальнейшего развития при расширении корпуса и формализации процедуры валидации.

Архитектура корпоративной RAG-системы

Каркас системы включает десять финансовых документов, охватывающих годовые отчёты компаний, квартальную отчётность, регуляторные нормы и отраслевые бенчмарки. Документы представлены в виде структурированной доменной модели, включающей классы FinancialDocument, FinancialMetric и RAGAnalysisReport.

Общая архитектура системы (представлена на рисунке 1) включает поисковый и аналитический контуры. Поисковый модуль реализует гибридный механизм ранжирования документов, после чего отобранный контекст используется для формирования аналитического ответа с последующей проверкой фактологической опоры.

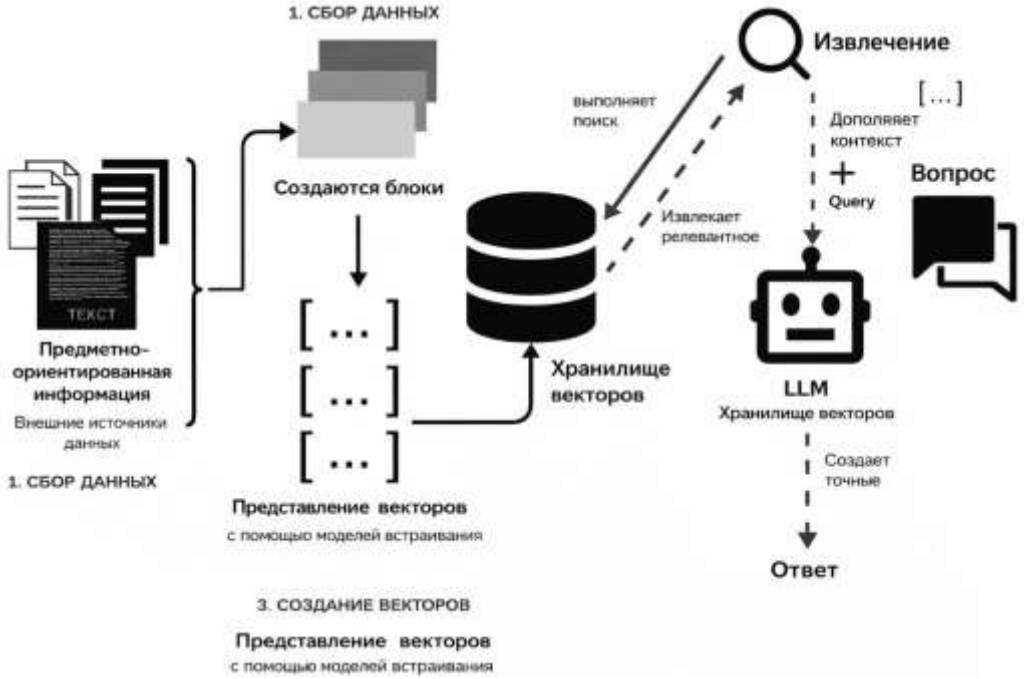

Рисунок 1 – Архитектура корпоративной RAG-системы

Каждая метрика связана с конкретным источником, годом публикации и компанией, что обеспечивает прослеживаемость результатов анализа и снижает вероятность фактических искажений [10, 13]. Состав корпуса представлен в таблице 1.

Таблица 1 – Состав корпуса документов

Тип документа	Примеры	Ключевые метрики
Годовые отчёты	XYZ Corp 2024, XYZ Corp 2023	ROE (Return on Equity, рентабельность собственного капитала), Debt/Equity (отношение долга к собственному капиталу), EBITDA
Квартальная отчётность	ABC Inc Q4 (четвёртый квартал) 2024	EBITDA, FCF (Free Cash Flow, свободный денежный поток)
Регуляторные документы	Положение ЦБ РФ №590-П	PD, LGD, ECL (Expected Credit Loss, ожидаемые кредитные потери)
Аналитические модели	DCF Valuation (Discounted Cash Flow, модель дисконтированных денежных потоков), Risk Scoring (оценка риска)	WACC (Weighted Average Cost of Capital, средневзвешенная стоимость капитала), PD, LGD
Отраслевые бенчмарки	Sector Benchmarks Technology 2024	ROE, Debt/Equity

Поисковый контур

В системе реализован гибридный механизм поиска, сочетающий:

- поиск по ключевым словам;
- векторное представление текста на основе TF-IDF;
- косинусную меру сходства для оценки релевантности.

Итоговая оценка документа формируется как агрегированная величина, нормированная в диапазоне от 0 до 1.

Дополнительно применяется доменно-ориентированная приоритизация:

- совпадение компании из запроса повышает итоговую оценку;
- документы последнего отчётного года получают дополнительный вес;
- годовые отчёты имеют более высокий приоритет по сравнению с вспомогательными источниками.

Подобная схема позволяет учитывать контекст запроса и снижает вероятность выбора нерелевантных документов.

Последовательность обработки запроса

Обработка запроса включает следующие этапы:

1. Лексический и векторный поиск релевантных документов.
2. Ранжирование кандидатов и отбор top-k документов (k = 5) при пороге релевантности 0,75.
3. Извлечение финансовых показателей из структурированных данных и текстовых фрагментов.
4. Формирование аналитического ответа.
5. Проверка фактологической опоры.

Пример обработки запроса представлен на рисунке 2 и включает этапы поиска, ранжирования, извлечения метрик, формирования аналитического вывода и проверки фактологической согласованности.

Рисунок 2 – Пример последовательности обработки запроса

Извлечение метрик реализовано с использованием заранее заданных структурированных значений и регулярных выражений для текстовых документов. Поддерживается распознавание показателей ROE, Debt/Equity, PD, LGD, EBITDA и WACC. Если числовые параметры указаны непосредственно в запросе пользователя, они имеют приоритет над значениями из документов.

Аналитический модуль и оценка согласованности

Аналитический блок выполняет интерпретацию извлечённых показателей на основе фиксированных правил. Для динамических показателей рассчитывается изменение год к году. Порог существенного отклонения по ROE установлен на уровне 10%.

Кредитный риск оценивается через расчёт ожидаемых потерь по формуле:

EL (Expected Loss, ожидаемые потери) = PD × LGD. Это соответствует регуляторной логике расчёта ожидаемых потерь [2, 5, 7].

Долговая нагрузка интерпретируется с использованием граничного значения Debt/Equity = 2,0. Рекомендации формируются по типовым сценариям, привязанным к выявленным отклонениям показателей.

Предложенная схема повышает интерпретируемость результата и снижает вероятность фактических искажений.

Экспериментальная оценка эффективности предложенной архитектуры

Апробация предложенной RAG архитектуры проведена на наборе демонстрационных сценариев, отражающих типовые задачи корпоративных финансов. Целью эксперимента являлась проверка корректности работы поискового контура, устойчивости извлечения финансовых метрик и согласованности формируемых рекомендаций с исходными документами. Результаты эксперимента представлены в таблице 2.

Таблица 2 – Результаты демонстрационной апробации прототипа

Сценарий	Найдено документов	Извлечено метрик	Проверка факт-опоры	Характер аналитического вывода
Анализ ROE (XYZ Corp 2023–2024)	2	4	Подтверждено	Негативная динамика, рост долговой нагрузки
Расчёт EL (PD, LGD)	2	2	Подтверждено	Корректный расчёт ожидаемых потерь
EBITDA vs benchmark	3	3	Подтверждено	Сопоставление с отраслевыми значениями

Долговая нагрузка	3	4	Подтверждено	Повышенный риск рефинансирования

В отличие от массового тестирования на крупном QA-датасете, исследование ориентировано на функциональную верификацию прототипа.

Для каждого сценария фиксировались следующие параметры:

– число найденных релевантных документов;
– значение нормированной поисковой оценки;
– количество извлечённых финансовых метрик;
– результат проверки фактологической опоры;
– время формирования ответа.

Во всех сценариях система корректно идентифицировала ключевые показатели, присутствующие в документах корпуса. При анализе динамики ROE были извлечены значения показателя за последовательные годы и рассчитано изменение год к году. В задаче расчёта кредитного резерва корректно использованы заданные пользователем параметры PD и LGD с вычислением ожидаемых потерь по формуле EL = PD × LGD. В сценариях сопоставления EBITDA и оценки долговой нагрузки retrieved-контекст содержал соответствующие показатели, что позволило сформировать интерпретацию без привлечения внешних данных.

Механизм проверки фактологической опоры подтвердил наличие ключевых метрик в retrieved-документах во всех случаях, где ответ содержал числовые значения из корпуса. Это свидетельствует о согласованности результата генерации с источниками.

Среднее время формирования ответа находилось в пределах, допустимых для аналитических задач, и не создаёт ограничений для интерактивного использования системы.

Качественный анализ показал, что внедрение гибридного поискового механизма повышает интерпретируемость результата за счёт явной привязки к документам корпуса, что согласуется с выводами современных исследований RAG-архитектур [8, 9, 12]. При отсутствии релевантных документов система демонстрирует снижение значения поисковой оценки, что может служить индикатором недостаточности контекста.

Следует учитывать ограничения проведённой апробации. Корпус документов носит демонстрационный характер и не отражает всего разнообразия финансовых источников. Эксперимент не включает статистическую оценку метрик точности на масштабной выборке и не охватывает нагрузочные сценарии промышленного уровня. Расширение корпуса и формализация процедуры количественной валидации рассматриваются как направление дальнейших исследований.

Полученные результаты позволяют заключить, что предложенная архитектура обеспечивает воспроизводимость анализа и снижает вероятность фактических искажений в пределах демонстрационного прототипа.

Выводы и перспективы развития

Апробация гибридной RAG архитектуры на демонстрационных сценариях корпоративной аналитики показала согласованность формируемых выводов с источниками данных корпуса. Система корректно извлекает ключевые финансовые показатели, рассчитывает производные величины, включая ожидаемые кредитные потери, и формирует интерпретации на основе заданных аналитических правил. Механизм проверки фактологической опоры подтверждает наличие используемых метрик в retrieved-документах, что повышает интерпретируемость результата.

Предложенный подход демонстрирует применимость поисковых архитектур для задач корпоративного финансового анализа и возможность повышения воспроизводимости результатов без изменения параметров языковой модели. Ограничения связаны с демонстрационным объёмом корпуса и использованием эвристических процедур проверки фактической корректности. Дальнейшее развитие связано с расширением корпуса финансовых документов и внедрением формализованных метрик оценки качества.

Список источников

1. Генерация, дополненная поиском (Retrieval-Augmented Generation) // Википедия: свободная энциклопедия. — Режим доступа: https://ru.wikipedia.org/wiki/Генерация,_дополненная_поиском?ysclid=mlwjqeon6i847408939 (дата обращения: 13.02.2026).
2. Положение Банка России от 28.06.2017 № 590-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам…» // КонсультантПлюс. — Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_220089/ (дата обращения: 13.02.2026).
3. Положение Банка России от 28 июня 2017 г. № 590-П «О порядке формирования кредитными организациями резервов на возможные потери по ссудам…» // Официальный сайт Банка России. — Режим доступа: https://www.cbr.ru/explan/590-p/ (дата обращения: 13.02.2026).
4. МСФО 9 «Финансовые инструменты» // Википедия: свободная энциклопедия. — Режим доступа: https://ru.wikipedia.org/wiki/IFRS_9 (дата обращения: 13.02.2026).
5. Рахаев В. А. Развитие методов оценки кредитного риска для формирования резервов на возможные потери по ссудам // Финансовый журнал. — 2020. — № 24. — С. 82–91. — DOI: 10.26794/2587-5671-2020-24-6-82-91.
6. Митичкин О. С. Основные принципы оценки кредитного риска в соответствии со стандартом МСФО 9 // Электронный научный журнал «Дневник науки». — 2019. — С.
7. Подходы к построению EAD-моделей на длинных временных горизонтах // Финансовый журнал•Financial Journal•№ 4•2021. — С. 91–109. — DOI: 10.31107/2075-1990-2021-4-91-109.
8. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / NeurIPS 2020 Proceedings. — 2020. — URL: https://arxiv.org/abs/2005.11401 (дата обращения: 14.02.2026).
9. Gao Y., Xiong Y., Gao X., Jia K., Pan J., Bi Y. Retrieval-Augmented Generation for Large Language Models: A Survey / arXiv preprint. — 2023. — URL: https://arxiv.org/abs/2312.10997 (дата обращения: 14.02.2026).
10. Huang L. A Survey on Hallucination in Large Language Models / arXiv preprint. — 2023. — URL: https://arxiv.org/abs/2311.05232 (дата обращения: 14.02.2026).
11. Alansari A., Luqman H. A Comprehensive Survey of Hallucination in Large Language Models: Causes, Detection, and Mitigation / arXiv preprint. — 2025. — URL: https://arxiv.org/abs/2510.06265 (дата обращения: 14.02.2026).
12. Karakurt E., Akbulut A. Retrieval Augmented Generation (RAG) and Large Language Models for Enterprise Knowledge Management and Document Automation / Applied Sciences (MDPI). — 2025. — URL: https://www.mdpi.com/2076-3417/16/1/368 (дата обращения: 14.02.2026).
13. Farquhar S., et al. Detecting Hallucinations in Large Language Models Using Entropy-Based Methods / Nature. — 2024. — URL: https://www.nature.com/articles/s41586-024-07421-0 (дата обращения: 15.02.2026).
14. Arslan M. A Survey on Retrieval-Augmented Generation and Its Applications / Procedia Computer Science. — 2024. — URL: https://www.sciencedirect.com/science/article/pii/S1877050924021860 (дата обращения: 15.02.2026).
15. Лыков А.В. Corporate RAG System Prototype [Электронный ресурс]. — Режим доступа: https://github.com/MrMixaDj32/corporate-rag-finance (дата обращения: 15.02.2026).

Сведения об авторах

Лыков Артем Валерьевич, студент направления подготовки «Программная инженерия», Кубанский государственный технологический университет, Краснодар, Россия

Мурлин Алексей Георгиевич, кандидат технических наук, доцент, Кубанский государственный технологический университет, Краснодар, Россия

Information about the authors

Lykov Artem Valerievich, student of the training program "Software Engineering", Kuban State Technological University, Krasnodar, Russia
Murlin Alexey Georgievich, Candidate of Technical Sciences, Associate Professor, Kuban State Technological University, Krasnodar, Russia