УДК 004.8
DOI: 10.26102/2310-6018/2025.50.3.027

Интеграция RAG-системы для автоматизации поиска связей показателей и мероприятий национальных проектов

И.Л. Каширина[1]✉, В.В. Кириллов[1], А.С. Албычев[1,2], Ю.В. Старичкова[1], Ш.Г. Магомедов[1], А.А. Червяков[2]

[1]МИРЭА – Российский технологический университет, Москва, Российская Федерация
[2]Федеральное Казначейство, Российская Федерация

Резюме. В условиях возрастающей сложности управления Национальными проектами, направленными на достижение Национальных целей развития РФ, актуальной задачей становится автоматизация анализа взаимосвязей между запланированными в рамках этих проектов мероприятиями и показателями, которые отражают степень достижения поставленных в проекте задач. Традиционные методы ручной обработки документов характеризуются высокой трудоемкостью, субъективностью и значительными временными затратами, что обусловливает необходимость разработки интеллектуальных систем поддержки принятия решений. В данной статье представлен подход к автоматизации анализа связей и показателей национальных проектов, который позволяет автоматически выявлять и верифицировать семантические связи «мероприятие-показатель» в документах национальных проектов, значительно повышая эффективность аналитической работы. Данный подход основан на использовании Retrieval-Augmented Generation (RAG) системы, сочетающей локально адаптированную языковую модель с технологиями векторного поиска. Работа демонстрирует, что интеграция RAG-подхода с векторным поиском и учетом онтологии проектов позволяет достичь необходимой точности и релевантности анализа. Особую ценность системе придает не только способность генерировать интерпретируемые обоснования выявленных связей, но и возможность определять ключевые мероприятия, влияющие на достижение показателей сразу нескольких национальных проектов, включая те из них, чье воздействие на реализацию данных показателей неочевидно. Предложенное решение открывает новые возможности для цифровизации государственного управления и может быть адаптировано для других задач, например, определения рисков реализации мероприятий и генерации новых мероприятий.

Ключевые слова: RAG-системы, большие языковые модели, национальные проекты, семантический поиск, автоматизация, национальные цели, искусственный интеллект в государственном управлении.

Благодарности: Работа выполнена в рамках Государственного задания на 2025 год паспорта № 6381-25 по научно-методическому и ресурсному обеспечению системы образования на тему: «Научно-методическое обеспечение работ по анализу деятельности управления общественными финансами Российской Федерации с применением искусственного интеллекта».

Для цитирования: Каширина И.Л., Кириллов В.В., Албычев А.С., Старичкова Ю.В., Магомедов Ш.Г., Червяков А.А. Интеграция RAG-системы для автоматизации поиска связей показателей и мероприятий национальных проектов. *Моделирование, оптимизация и информационные технологии.* 2025;13(3). URL: https://moitvivt.ru/ru/journal/pdf?id=2001 DOI: 10.26102/2310-6018/2025.50.3.027

Integration of the RAG system for automation of search links of indicators and activities of national projects

I.L. Kashirina[1]✉, V.V. Kirillov[1], A.S. Albychev[1,2], J.V. Starichkova[1], Sh.G. Magomedov[1], A.A. Chervyakov[2]

[1]MIREA – Russian Technological University, Moscow, the Russian Federation
[2]Federal Treasury, the Russian Federation

Abstract. In the context of the increasing complexity of managing national projects aimed at achieving the National Development Goals of the Russian Federation, an urgent task is to automate the analysis of the relationships between the activities planned within these projects and the indicators that reflect the degree of achievement of the objectives set in the project. Traditional methods of manual document processing are characterized by high labor intensity, subjectivity and significant time costs, which necessitates the development of intelligent decision support systems. This article presents an approach to automating the analysis of links and indicators of national projects, which allows for automatic detection and verification of semantic links "event-indicator" in national project documents, significantly increasing the efficiency of analytical work. This approach is based on the use of the Retrieval-Augmented Generation (RAG) system, which combines a locally adapted language model with vector search technologies. The work demonstrates that the integration of the RAG approach with vector search and taking into account the project ontology allows achieving the required accuracy and relevance of the analysis. The system is particularly valuable not only for its ability to generate interpretable justifications for the identified links, but also for its ability to identify key events that affect the achievement of indicators for several national projects at once, including those whose impact on the implementation of these indicators is not obvious. The proposed solution opens up new opportunities for the digitalization of public administration and can be adapted for other tasks, such as identifying risks in the implementation of events and generating new events.

Keywords: RAG systems, large language models, national projects, semantic search, automation, national goals, artificial intelligence in public administration.

Acknowledgments: The work was completed within the framework of the State assignment for 2025, passport No. 6381-25 on scientific, methodological and resource support of the education system on the topic: "Scientific and methodological support for work on the analysis of the activities of public finance management of the Russian Federation using artificial intelligence".

For citation: Kashirina I.L., Kirillov V.V., Albychev A.S., Starichkova Yu.V., Magomedov Sh.G., Chervyakov A.A. Integration of the RAG system for automation of search links of indicators and activities of national projects. *Modeling, Optimization and Information Technology*. 2025;13(3). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=2001 DOI: 10.26102/2310-6018/2025.50.3.027

Введение

Современная система государственного управления сталкивается с растущей сложностью координации мероприятий в рамках национальных проектов. С каждым годом увеличивается объем информации, связанной с описанием целей, задач, мероприятий и ожидаемых результатов, растет число контролей.

Управление национальными проектами осуществляется в государственной автоматизированной информационной системе «Управление» (далее – ГАС «Управление»). Развитие ГАС «Управление» проводится Федеральным казначейством в рамках реализации постановления Правительства Российской Федерации от 25 декабря 2009 года № 1088 «О государственной автоматизированной информационной системе «Управление». ГАС «Управление» представляет собой единую государственную информационную систему, обеспечивающую сбор, учет, обработку и анализ данных, содержащихся в государственных и муниципальных информационных ресурсах,

аналитических данных, данных официальной государственной статистики, а также иных сведений, необходимых для обеспечения поддержки принятия управленческих решений.

Одной из ключевых задач стратегического планирования достижения национальных целей является привязка мероприятий национальных проектов к показателям национальных проектов. Данная задача возложена в первую очередь на руководителей национальных проектов и Аналитический центр при Правительстве Российской Федерации. Привязка мероприятий к показателям обеспечивает основу для расчета уровня достижения показателей, определения ключевых мероприятий, выполнение которых влияет сразу на несколько национальных целей и, соответственно, требующих отдельного контроля, а также определения мероприятий, влияние которых на достижения показателей неочевидно.

Современные большие языковые модели (LLM) демонстрируют уникальную способность к пониманию и обработке естественного языка, что открывает принципиально новые возможности для автоматизации аналитических процессов в сфере государственного управления. Применение LLM позволяет анализировать масштабные массивы текстовых данных, выявлять скрытые закономерности и устанавливать контекстуальные связи между описаниями мероприятий и ключевыми показателями национальных проектов.

Таким образом, в современных условиях цифровой трансформации государственного управления актуальным становится применение передовых технологий искусственного интеллекта для автоматизации аналитических процессов.

Целью данного исследования является разработка подхода для обнаружения и формализации ранее не выявленных взаимосвязей между проводимыми мероприятиями и целевыми показателями с использованием больших языковых моделей. Достижение указанной цели является важным этапом решения задачи формирования единого плана выполнения национальных проектов Российской Федерации.

Одним из перспективных направлений для достижения поставленной цели является использование систем Retrieval-Augmented Generation (RAG) [1], которые принципиально меняют подходы к обработке больших массивов текстовой информации. Основное преимущество RAG-систем заключается в их способности динамически обогащать контекст запроса к языковой модели самыми актуальными данными из внешних источников, что существенно повышает качество анализа.

Материалы и методы

Определение семантических связей, отражающих различные типы отношений между разными текстами, представляет собой одну из ключевых задач в области обработки естественного языка[1]. Существующие подходы к определению семантической близости и взаимосвязей можно условно разделить на несколько категорий.

1. Статистические методы охватывают анализ частотности слов, TF-IDF (Term Frequency-Inverse Document Frequency) и другие метрики, позволяющие оценить значимость терминов в текстах [2]. Данный подход отличается низкой вычислительной сложностью и хорошей интерпретируемостью результатов. Однако существенным ограничением является необходимость использования одинаковой терминологии для обозначения связанных концепций, поскольку метод является неэффективным при отсутствии общих ключевых слов в анализируемых текстах.

[1] Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. *Автоматическая обработка текстов на естественном языке и анализ данных.* Москва: Изд-во НИУ ВШЭ; 2017. 269 с.

2. Методы, основанные на векторном представлении [3], обеспечивают возможность отображения текстов в виде векторов в многомерном пространстве, где семантически близкие тексты располагаются ближе друг к другу и для определения семантической близости вычисляется косинусное расстояние или применяется другая метрика сходства. Преимущество этого подхода заключается в способности улавливать семантические нюансы и учитывать контекст. Тем не менее, качество векторных представлений может оказаться недостаточным для выявления сложных и неочевидных связей между текстами.

3. Использование больших языковых моделей (LLM) [4] представляет собой наиболее мощный инструмент для глубокого семантического анализа, способный выявлять неявные связи и предоставлять их обоснования. Однако применение таких моделей сопряжено с высокими вычислительными затратами, риском генерации правдоподобных, но фактически неверных утверждений, а также проблемами с устойчивостью получаемых ответов.

Автоматическое выявление перекрестных связей между показателями и мероприятиями национальных проектов представляет собой комплексную задачу, требующую глубокого семантического анализа и точного понимания специфического контекста. Национальные проекты содержат узкоспециализированную терминологию, нормативные формулировки и специфические показатели эффективности. Стандартные языковые модели могут не обладать достаточными знаниями в этой предметной области. Кроме того, информация о проектах способна обновляться, могут корректироваться их цели, показатели и мероприятия. А базовые языковые модели ограничены знаниями, полученными во время обучения.

В качестве инструмента для решения перечисленных проблем может выступать RAG-система, позволяющая дополнить возможности языковой модели релевантными документами из базы знаний, содержащими актуальную информацию о национальных проектах, их структуре, терминологии и методиках расчета показателей.

Концептуальная основа RAG-систем заключается в интеграции механизмов извлечения релевантной информации из структурированных баз данных с генеративными возможностями современных языковых моделей, что позволяет создавать аналитические решения, обладающие как фактологической точностью, так и контекстуальной релевантностью [5].

RAG-системы представляют собой подход к генерации контента, который объединяет два ключевых механизма: механизм поиска информации (Retrieval) и механизм генерации (Generation) [6]. В отличие от традиционных языковых моделей, RAG дает возможность динамически извлекать релевантную информацию из внешних источников данных непосредственно в процессе генерации ответа. Это принципиально расширяет контекстуальные возможности искусственного интеллекта, позволяя системе не полагаться исключительно на изначально заложенные при обучении знания, а оперативно получать и обрабатывать актуальную дополнительную информацию. При этом, в отличие от интеграции с внешними поисковыми системами, где модель может получить доступ к непроверенной, противоречивой или недостоверной информации из открытых источников, RAG-системы оперируют с предварительно отобранными и валидированными данными.

Архитектура RAG-системы предполагает формирование векторной базы данных для ее реализации. В рамках решения задачи анализа взаимосвязей между показателями национальных проектов и соответствующими мероприятиями для построения такой базы использовалась библиотека с открытым исходным кодом FAISS, предназначенная для быстрого и эффективного поиска схожих эмбеддингов в условиях ограниченных вычислительных ресурсов.

Система на основе FAISS использует иерархическую кластеризацию векторов, при которой все элементы векторной БД распределяются по семантическим кластерам с выделенными центроидами. Процесс поиска оптимизируется за счет двухэтапной процедуры: сначала определяется набор наиболее релевантных кластеров, затем осуществляется точный поиск внутри отобранных групп. Такой подход обеспечивает ускорение идентификации семантических связей между объектами за счет сокращения пространства поиска на первом этапе. FAISS использует продвинутые методы квантования векторов и эффективных алгоритмов приближенного поиска ближайших соседей, что делает эту архитектуру наиболее подходящей для задач с высокими требованиями к скорости обработки запросов.

Общий подход к поиску связей между мероприятиями и показателями национальных проектов с использованием RAG-системы включает следующие основные этапы.

1. Генерация векторных представлений (эмбеддингов) текстовых описаний мероприятий и показателей.
2. Поиск релевантной информации в векторной базе знаний для обогащения промта.
3. Формирование промта с учетом найденного контекста.
4. Инференс модели. Промпт передается языковой модели для генерации итоговой оценки сематической связи между показателем и мероприятием.

Особенностью предлагаемого в данном исследовании подхода является использование RAG-системы для реализации современной техники few-shot обучения [7]. Данная техника предполагает включение в промпт языковой модели одного или нескольких релевантных образцов из предварительно размеченного датасета. Каждый образец содержит пару «показатель-мероприятие» с экспертной аннотацией о наличии или отсутствии семантической связи между элементами. Перед анализом каждой новой пары RAG-система автоматически отбирает из векторной базы наиболее схожие примеры, которые служат контекстуальными ориентирами для модели при принятии решения о взаимосвязи анализируемых сущностей.

На Рисунке 1 представлена архитектура предлагаемой RAG-системы.

Рисунок 1 – Архитектура RAG-системы
Figure 1 – RAG system architecture

Таким образом, база данных FAISS будет содержать векторные представления показателей и мероприятий, а также метаинформацию о характере их взаимосвязи. Это позволит модели учитывать исторический контекст экспертных оценок связей между схожими показателями и мероприятиями, что особенно важно в условиях специализированной терминологии национальных проектов.

Важной особенностью FAISS является возможность возвращать не только ближайшего соседа, но и заданное количество ближайших соседей (второго, третьего и так далее по порядку близости). Предлагаемый алгоритм использования RAG для поиска в векторной базе k наиболее релевантных исходному запросу пар «показатель-мероприятие», имеет следующий вид.

Этап 1. Преобразование текста в эмбеддинги.

На начальном этапе исходные тексты показателей и мероприятий конвертируются в векторы эмбеддингов с использованием предобученной модели LaBSE («Language Agnostic BERT Sentence Embeddings»), специально оптимизированной для работы с русскоязычными текстами.

Этап 2. Построение векторных индексов.

Для оптимизации поиска ближайших соседей создаются два отдельных индекса FAIIS: один для векторов показателей и другой для векторов мероприятий. Использование отдельных индексов позволяет независимо искать ближайшие аналоги для каждого типа сущности.

Этап 3. Поиск похожих показателей и мероприятий.

Для каждого входного показателя выполняется поиск топ-10k наиболее схожих показателей. Аналогично, для каждого мероприятия выполняется поиск топ-10k наиболее схожих мероприятий. Результатом поиска являются списки топ-10k ближайших соседей для каждого показателя и мероприятия, а также соответствующие оценки косинусного сходства.

Этап 4. Нормализация оценок схожести.

Для обеспечения сопоставимости оценок косинусного сходства, полученных для показателей и мероприятий, применяется метод квантования [8]. Оценки делятся на квантили (десятые доли), что позволяет выровнять их распределение и преобразовать в шкалу от 0 до 1, где 1 соответствует максимальной схожести.

Этап 5. Генерация комбинаций пар показателей и мероприятий.

Результаты поиска (топ-10k показатели и топ-10k мероприятия) комбинируются с использованием декартова произведения, формируя все возможные пары «показатель-мероприятие» для дальнейшей обработки.

Этап 6. Фильтрация и расчет комбинированной оценки.

Каждая пара «показатель-мероприятие» проходит проверку на наличие в размеченных данных. Если такая пара уже существует, вычисляется среднее арифметическое оценок косинусного сходства показателя и мероприятия, которое становится «комбинированной оценкой» релевантности пары. Пары, для которых хотя бы одна из оценок косинусного сходства (показателя или мероприятия) ниже заданного порога, отбрасываются.

Этап 7. Сортировка и возврат топ-k результатов.

Релевантные пары «показатель-мероприятие» сортируются по убыванию комбинированной оценки. Алгоритм возвращает топ-k пар с наивысшими оценками, представляющих наиболее релевантные комбинации показателя и мероприятия.

Рассмотрим пример работы описанного алгоритма. Пусть на вход алгоритма поступает показатель «Количество рационализаторских предложений, поданных работниками предприятий, нарастающим итогом» и мероприятие «Подтверждены

компетенции рационализатора и пройдена сертификация работников предприятий-участников национального проекта, в том числе студентов образовательных организаций». Результатом алгоритма будет топ-*k* наиболее схожих пар из размеченных данных, каждая из которых будет представлена в следующем формате:

[{'показатель': 'Количество работников предприятий (включая подрядные и образовательные организации), вовлеченных в движение рационализаторов, нарастающим итогом',

'мероприятие': 'Сформирован перечень субъектов Российской Федерации по количеству вовлеченных предприятий-участников национального проекта в движение рационализаторов.',

'схожесть_показателя': 0,9,

'схожесть_мероприятия': 0,8,

'комбинированная_оценка': 0,85

'связь': нет}]

Предлагаемый подход демонстрирует отличие от классических методов анализа взаимосвязей, используя эффективный механизм на основе few-shot обучения для выявления скрытых семантических связей между разнотипными текстами.

Результаты

В рамках исследования, посвященного интеграции RAG-систем для автоматизации поиска и анализа показателей и мероприятий и национальных проектов (НП), был проведен комплекс экспериментов, направленных на оценку эффективности различных подходов к выявлению связей между показателями и результатами. Для объективного анализа эффективности предлагаемого подхода эксперименты были разделены на две основные категории: без использования RAG-системы и с ее применением. Это позволило не только проверить гипотезу о повышении точности выявления скрытых связей при интеграции RAG-систем, но и оценить их практическую ценность в контексте обработки данных национальных проектов.

Эксперименты проводились на двух наборах данных.

Первый набор содержал описания 224 показателей и 780 мероприятий национальных проектов (НП). Таким образом, исходный датасет включал 174720 возможных пар «Показатель – Мероприятие», из которых 1232 пары были связаны в рамках паспортов национальных проектов, то есть имели разметку «да».

Второй датасет содержал 10 показателей и 64 мероприятия, то есть, всего 640 вариантов возможных пар. Из них 212 были размечены экспертами. 110 пар имели разметку «да» и 102 пары – разметку «нет».

В качестве метрик качества для оценки моделей использовалась полнота (recall), характеризующая долю найденных моделью связей из всех размеченных, и точность (precision), характеризующая, какую долю занимают размеченные связи среди всех найденных моделью. Постановка задачи требовала, чтобы результат модели по каждой из этих метрик превысил 70 %. Еще одной важной характеристикой, оцениваемой в экспериментах, являлась производительность модели, то есть количество связей, которое модель обрабатывает в секунду, так как полная онтология всех национальных проектов содержит около 2 миллионов потенциальных связей «Показатель – Мероприятие», перебор которых может оказаться существенно затратным по времени. Эксперименты проводились с использованием видеокарты H100. Для ускорения перебора использовался метод квантизации – то есть понижения точности представления весов модели с чисел с плавающей точкой до более компактных форматов [9]. Кроме того, применялась параллельная обработка группы связей (батчинг), представляющая

собой технику объединения нескольких запросов в единый пакет для одновременной обработки. Это позволило эффективно использовать параллельные вычислительные возможности GPU за счет векторизации операций.

1. *Эксперименты без использования RAG*

На первом этапе исследования с использованием первого датасета было проведено тестирование нескольких популярных языковых моделей открытого доступа [10] с целью выбора базовой модели для проведения серии экспериментов в рамках данного исследования. Поскольку первый датасет не содержал разметки «нет», для его оценки использовалась только метрика Recall, которая показывает долю корректно найденных моделью связей из общего числа связей в паспорте. В ходе экспериментов тестировались различные запросы, созданные с применением техники промпт-инжиниринга. При этом отбирались только те запросы, которые генерировали не более 150 % связей от исходного количества размеченных связей в датасете. Кроме того, в данной серии экспериментов исследовалось добавление подробного текстового описания характеристики проведенного мероприятия в промпт языковой модели.

Как видно из Таблицы 1, наибольшую полноту (70 %) продемонстрировала модель Qwen2.5-72B, использующая характеристику мероприятия, однако скорость обработки составила лишь 1 связь в секунду (86 400 связей в сутки), что ограничивает ее практическое применение из-за высоких вычислительных затрат. Модель Llama-3.1-70B показала полноту 65 % при скорости 3 связи в секунду. Оптимизированные для русского языка модели, такие как Vikhr-Llama3.1-8B, достигли скорости 16 связей в секунду, однако их полнота (53 %) оказалась недостаточной для решения задачи. Результаты эксперимента продемонстрировали две закономерности: включение дополнительных характеристик в запрос повышает точность извлечения связей, но снижает общую производительность системы; крупные модели (70B-72B параметров) обеспечивают более высокую точность результатов, однако существенно уступают в скорости обработки данных.

Таблица 1 – Сравнение языковых моделей на первом датасете
Table 1 – Comparison of language models on the first dataset

Модель	Recall	Скорость (связей/сек)
Llama-3.1-8B без характеристики	0,58	15
Llama-3.1-8B с характеристикой	0,60	12
Vikhr-Llama3.1-8B без характеристики	0,53	16
Vikhr-Llama3.1-8B с характеристикой	0,55	13
Llama-3.1-70B без характеристики	0,63	3
Llama-3.1-70B с характеристикой	0,65	2
Qwen2.5-72B без характеристики	0,68	2
Qwen2.5-72B с характеристикой	0,7	1

Для дальнейших экспериментов в рамках данного исследования в качестве базовой модели была выбрана Llama-3.1-8b, продемонстрировавшая наиболее оптимальное соотношение полноты и производительности, а также генерирующая логичные обоснования наличия связи. В Таблице 2 представлены результаты валидации этой модели на втором датасете, содержащем 212 пар, размеченных экспертами. Поскольку данный датасет содержал разметку как для случаев наличия, так и отсутствия связи между показателями и мероприятиями, для размеченной части датасета были рассчитаны метрики точности (Precision), полноты (Recall) и общей доли правильных ответов модели (Accuracy).

Таблица 2 – Результаты модели llama-3.1-8b на втором датасете
Table 2 – Results of the llama-3.1-8b model on the second dataset

Модель	Accuracy	Precision	Recall
Llama-3.1-8b-Instruct без характеристики	0,716	0,656	0,775
Llama-3.1-8b-Instruct с характеристикой	0,717	0,631	0,783

Анализ результатов моделей на втором датасете выявил два важных аспекта. Включение характеристики мероприятия в запрос способствует повышению полноты, однако при этом снижает точность модели. Кроме того, в ходе экспериментов со вторым датасетом не удалось достичь требуемого одновременного превышения порога 0,7 для метрик точности и полноты.

2. *Эксперименты с использованием RAG*

Как уже было отмечено, первый датасет содержал только положительные примеры связей. В этом случае RAG-система использовалась для поиска релевантных контекстов, близких к входному запросу. Сравнение с базовой версией модели, представленное в Таблице 3, показывает, что использование RAG повысило полноту (recall) на 17 % относительно исходного значения.

Таблица 3 – Результаты модели Llama-3.1-8B на первом датасете
Table 3 – Results of the Llama-3.1-8B model on the first dataset

Модель	RAG	Recall
Llama-3.1-8B (с характеристикой)	–	0,60
Llama-3.1-8B (с характеристикой)	+	**0,70**

Базовая модель без RAG демонстрировала recall 0,6, что ниже требуемого порога. Добавление RAG-системы, которая обогащала промпты релевантными положительными примерами, позволило повысить полноту до 0,7. Это подтверждает, что даже при отсутствии отрицательной разметки RAG улучшает способность модели находить корректные связи за счет семантического поиска.

Второй используемый датасет включал положительные и отрицательные примеры связей (Таблица 4). RAG-система использовала негативное сэмплирование, добавляя в промпты как подтвержденные, так и отклоненные экспертами связи. Это привело к одновременному росту точности (precision) и полноты (recall).

Таблица 4 – Результаты модели Llama-3.1-8B на втором датасете
Table 4 – Results of the Llama-3.1-8B model on the second dataset

Модель	RAG	Precision	Recall
Llama-3.1-8b (с характеристикой)	–	0,631	0,783
Llama-3.1-8b (с характеристикой)	+	**0,76**	**0,82**

Добавление отрицательных примеров в промпты через использование RAG-системы позволило повысить precision до 0,76 (улучшение на 20 %) и recall до 0,82 (улучшение на 5,1 %). Это свидетельствует о том, что модель научилась точнее фильтровать нерелевантные связи, сохраняя высокую полноту.

Обе серии экспериментов продемонстрировали, что RAG-подход способен улучшить результаты генеративных моделей при отыскании связей между показателями и мероприятиями национальных проектов.

На Рисунке 2 приведен пример успешного применения разработанной модели, которая автоматически выявила неочевидную связь между показателем «Снижение совокупного объема выбросов» и мероприятием «Построены объекты ж/д инфраструктуры Центрального транспортного узла в целях организации диаметральных маршрутов» и привела логическое обоснование наличия такой связи.

Заключение

Анализ связей между показателями и мероприятиями федеральных проектов требует комплексного подхода, сочетающего различные методы автоматизации. Исследование показало, что применение RAG-систем улучшает баланс между точностью и полнотой выявляемых связей, особенно при использовании отрицательной разметки для few-shot обучения. В то время как модели без RAG демонстрируют либо высокую полноту с низкой точностью, либо умеренные метрики, ограниченные сложностью выявления косвенных взаимосвязей, интеграция RAG позволяет частично преодолеть эти проблемы.

Эксперименты с языковыми моделями выявили, что крупные архитектуры, такие как Qwen2.5-72B, обеспечивают хорошее качество анализа, однако их практическое применение ограничивается высокими требованиями к вычислительным ресурсам. Модели меньшего размера, такие как Llama-3.1-8b нуждаются в дальнейшей доработке с использованием RAG для повышения точности.

Разработанная RAG-система отвечает поставленной цели исследования и представляет собой эффективный инструмент для автоматизации анализа показателей и мероприятий национальных проектов, способствующий повышению прозрачности и эффективности государственного управления, сокращению временных затрат на обработку документов и минимизации субъективных ошибок.

Рисунок 2 – Примеры отыскания неочевидной связи между мероприятием и показателем из разных национальных проектов

Figure 2 – Examples of finding a non-obvious connection between an event and an indicator from different national projects

Проведенное исследование имеет, в том числе и теоретическую значимость, расширяя представления о возможностях адаптации языковых моделей для задач государственного управления. Внедрение таких инструментов способствует переходу к управлению, основанному на данных, укрепляя связь между оперативной деятельностью и долгосрочными национальными целями. Результаты исследования могут быть применены и в других областях, например, в задаче определения рисков реализации мероприятий и генерации новых мероприятий.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 06–12 December 2020, Online.* 2020. https://doi.org/10.48550/arXiv.2005.11401
2. Mishra A., Vishwakarma S. Analysis of TF-IDF Model and Its Variant for Document Retrieval. In: *2015 International Conference on Computational Intelligence and Communication Networks (CICN), 02–14 December 2015, Jabalpur, India*. IEEE; 2015. P. 772–776. https://doi.org/10.1109/CICN.2015.157
3. Mikolov T., Sutskever I., Chen K., Corrado G.S., Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, 05–08 December 2013, Lake Tahoe, NV, USA*. 2013. https://doi.org/10.48550/arXiv.1310.4546
4. Ouyang L., Wu J., Jiang X., et al. Training Language Models to Follow Instructions with Human Feedback. In: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, 28 November – 9 December 2022, New Orleans, LA, USA.* 2022. https://doi.org/10.48550/arXiv.2203.02155
5. Guu K., Lee K., Tung Z., Pasupat P., Chang M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. arXiv. URL: https://doi.org/10.48550/arXiv.2002.08909 [Accessed 13th May 2025].
6. Gao Yu., Xiong Yu., Gao X., et al. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv. URL: https://doi.org/10.48550/arXiv.2312.10997 [Accessed 13th May 2025].
7. Brown T.B., Mann B., Ryder N., et al. Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, 06–12 December 2020, Online.* 2020. https://doi.org/10.48550/arXiv.2005.14165
8. Eremeev M., Vorontsov K.V. Lexical Quantile-Based Text Complexity Measure. In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing, RANLP 2019, 02–04 September 2019, Varna, Bulgaria.* INCOMA Ltd.; 2019. P. 270–275. https://doi.org/10.26615/978-954-452-056-4_031
9. Jin R., Du J., Huang W., et al. A Comprehensive Evaluation of Quantization Strategies for Large Language Models. In: *Findings of the Association for Computational Linguistics, ACL 2024, 11–16 August 2024, Bangkok, Thailand.* Association for Computational Linguistics; 2014. P. 12186–12215.
10. Izacard G., Grave E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021,*

19–23 April 2021, Online. Association for Computational Linguistics; 2021. P. 874–880. https://doi.org/10.48550/arXiv.2007.01282

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Каширина Ирина Леонидовна, доктор технических наук, профессор, кафедра технологий искусственного интеллекта, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: kash.irina@mail.ru
ORCID: 0000-0002-8664-9817

Irina L. Kashirina, Doctor of Engineering Sciences, Professor, Department of Artificial Intelligence Technologies, MIREA – Russian Technological University, Moscow, the Russian Federation.

Кириллов Вадим Витальевич, ассистент, кафедра технологий искусственного интеллекта, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: kirillov@mirea.ru
ORCID: 0009-0003-7550-8917

Vadim V. Kirillov, assistant, Department of Artificial Intelligence Technologies, MIREA – Russian Technological University, Moscow, the Russian Federation.

Албычев Александр Сергеевич, заместитель руководителя, Федеральное Казначейство, заведующий кафедрой «Государственные финансовые технологии», МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: albychevas@roskazna.ru
ORCID: 0000-0001-9632-7806

Alexander S. Albychev, Deputy Head, Federal Treasury, Head of the Department of State Financial Technologies, MIREA – Russian Technological University, Moscow, the Russian Federation.

Старичкова Юлия Викторовна, кандидат технических наук, доцент, заведующий кафедрой технологий искусственного интеллекта, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: starichkova@mirea.ru
ORCID: 0000-0003-1804-0761

Julia V. Starichkova, Candidate of Engineering Sciences, Docent, Head of the Department of Artificial Intelligence Technologies, MIREA – Russian Technological University, Moscow, the Russian Federation.

Магомедов Шамиль Гасангусейнович, доктор технических наук, доцент, директор Института искусственного интеллекта, МИРЭА – Российский технологический университет, Москва, Российская Федерация.
e-mail: magomedov_sh@mirea.ru
ORCID: 0000-0001-8560-1937

Shamil G. Magomedov, Doctor of Engineering Sciences, Docent, Director Institute of Artificial Intelligence, MIREA – Russian Technological University, Moscow, the Russian Federation.

Червяков Александр Александрович, кандидат технических наук, начальник Управления развития информационных систем, Федеральное Казначейство, Москва, Российская Федерация.
e-mail: achervyakov@roskazna.ru
ORCID: 0000-0002-5638-8361

Alexander A. Chervyakov, Candidate of Engineering Sciences, Head of the Information Systems Development Department, Federal Treasury, Moscow, the Russian Federation.

Статья поступила в редакцию 18.06.2025; одобрена после рецензирования 11.07.2025; принята к публикации 29.07.2025.

The article was submitted 18.06.2025; approved after reviewing 11.07.2025; accepted for publication 29.07.2025.