

Разработка системы для анализа технических и организационных документов на соответствие нормативным требованиям атомной отрасли на основе больших лингвистических моделей

Д.Д. Шемонаев, П.В. Анни, М.Г. Жабицкий, Ю.Н. Конев

искусственный интеллект, естественный язык, интерфейс, федеральные нормы и правила, атомная промышленность

Аннотация— В данной работе рассматривается проектирование, разработка и испытания прототипа прикладной системы с применением технологий искусственного интеллекта для автоматизированного анализа документации атомной отрасли, включая проверку соответствия регламентирующим документам и нормативным требованиям. Система использует технологии больших языковых моделей (LLM) и поисково-дополненной генерации (RAG). Рассмотрены варианты применения данного стека технологий для использования специализированных доменов знаний (массива нормативной документации и технических инструкций атомной отрасли) для качественной оперативной экспертизы рабочих документов с применением искусственного интеллекта. Повышение качества и скорости квалифицированных экспертиз для сложных высокотехнологичных отраслей может зачастую иметь критическую важность для реализации серьезных проектов.

В связи с наличием значимой чувствительной информации технического и экономического характера в отрасли, важным является организация работ без внешнего доступа к анализируемой проектной документации. В работе рассмотрены способы организации работ с применением как облачных, так и внутриконтурных LLM. Разработан прототип инструмента оркестрации разнородных моделей с доступом в закрытый контур компании через VPN-каналы. Выполнено сравнение решения задачи интеллектуальным ассистентом эксперта в облачном и внутриконтурном режиме. Описан разработанный прототип системы, функционирующей без подключения к внешним сетям, обеспечивающей высокий уровень конфиденциальности и импортозамещения. Прототип обеспечивает распределение ролей предметного пользователя (эксперта) и администратора, предоставляя возможность защищённого взаимодействия с системой. Разработанный интеллектуальный ассистент реализует продвинутый контекстный поиск и аналитику/экспертизу проектной документации на основании информации из специализированных доменов знаний ограниченного доступа на базе применения технологии RAG. Реализация решения выполнена на базе цифрового стенда-полигона Высшей инженерной школы МИФИ и рассчитана на применение для образовательных задач, учебно-практических кейсах индустриальных партнеров и для выполнения работ в интересах заказчиков из производственной сферы, в первую очередь – предприятий Росатома (включая, но не ограничиваясь).

Ключевые слова— RAG, Text Mining, LLM, векторизация, база данных, большие языковые модели,

I. ВВЕДЕНИЕ

На современном этапе технологического развития информационные технологии обеспечивают прорывное расширение возможностей и значительное повышение производительности труда даже для специалистов, ранее слабо вовлеченных в цифровую сферу. Более того, масштаб и эффективность преобразований особенно велики именно для таких видов деятельности. Общий тезис о создании умных систем, обеспечивающих автоматизацию принятия решений и обработку больших объемов данных, в принципе универсальный, для таких направлений деятельности имеет особую значимость.

Ярким примером является экспертиза человекочитаемых документов. Работа с документацией для развитого промышленного производства является одной из основ. Вся деятельность в бизнесе, производстве, общественных отношениях и в значительной степени даже для конечных пользователей связана с документированием и использованием его результатов. Документирование сопровождает практически любое действие на жизненном цикле искусственных объектов, процессы в естественной природной среде и во взаимодействиях людей и организаций (субъектов с определенными ролями). Применительно к производственной деятельности документированию подлежит полный жизненный цикл сложного инженерного объекта, в том числе столь сложных и ответственных, как объекты использования атомной энергии. Это концептуальный проект, проектная документация, документирование процесса строительства и ввода в эксплуатацию, эксплуатационная документация, документация по выводу из эксплуатации.

Важно различать при этом виды и формы представления информации. Выделим два класса информационных сущностей – знания и данные. При некоторой неопределенности формального разграничения, в абсолютном большинстве случаев кластеризация и классификация единицы информации между ними вполне однозначна. Другая, ортогональная классификация – метод обработки информации. До недавнего времени однозначно разделялись человекочитаемая информация, знания (документы) и машинообрабатываемая информация (базы данных).

При этом формы хранения и передачи информации, а также ее первичная обработка интегрированы в процесс цифровизации достаточно давно, с 60-х годов. Но основные операции предыдущей технологической парадигмы – сканирование, распознавание, хранение в цифровом виде никак не были связаны с уровнем когнитивной деятельности – пониманием смысла документа и работой со знаниями. Даже развивавшиеся в последнее десятилетие технологии нейронных сетей и машинного обучения были ограниченным, паллиативным и недостаточно эффективным решением для полной задачи содержательной цифровой обработки человекочитаемой информации (документов).

Однако, в последние годы произошли революционные изменения. Появление в доступе у массового пользователя технологии искусственного интеллекта в форме больших лингвистических моделей (LLM) принципиально сняло монополию работы со знаниями, смыслами и документами путем чтения и понимания. Это позволяет качественно изменить парадигму когнитивной деятельности, прежде всего в специальных областях.

Одним из исключительных видов деятельности человека-специалиста была экспертиза документации, выполняемая в огромных масштабах в различных сферах и отраслях. При этом, чем сложнее и потенциально опаснее является предметная область, тем больший объем экспертиз она требует. Экспертиза документации – достаточно специфическая деятельность, которая с одной стороны весьма рутинна, а с другой – требует от эксперта высшего уровня квалификации (следовательно, весьма недешева). На практике склонностью и интересом к данному виду деятельности обладает весьма ограниченная часть специалистов, а активное вовлечение высокорейтинговых экспертов за счет высокой оплаты труда резко удорожает процесс. Но формальный подход, возложение обязанностей глубокой содержательной экспертизы на начинающих сотрудников неизбежно ведет к снижению качества. Поэтому мы имеем классический конфликт «цена-качество» в острой форме. И неожиданное для пользователей появление (буквально «как чертик из коробки») прорывной технологии больших лингвистических моделей (Large Language Models, LLM) дало впервые надежду на разрешение этого противоречия путем разработки на их основе интеллектуальных ассистентов предметных экспертов.

II. ПОСТАНОВКА ЗАДАЧИ

Отрасль атомной энергетики представляет собой одну из наиболее комплексных и ответственных сфер человеческой деятельности, предъявляющих высокие требования к качеству работы с информацией [1 - 3]. Здесь необходима глубокая экспертиза и анализ значительного количества технической документации — нормативных актов, отчетов по эксплуатации оборудования, регламентирующих материалов и технических характеристик. Полноценная

интерпретация этих сведений, безусловно, способствует повышению производительности и минимизации рисков.

Однако традиционная практика ручной обработки подобных документов связана с высокими рисками возникновения ошибок, длительными временными издержками, необходимостью привлечения большого числа специалистов высокой квалификации. Именно поэтому внедрение методов искусственного интеллекта представляется оптимальным решением проблемы автоматизации рабочих процессов в данной сфере. Современные технологии машинного обучения способны распознавать текстовые материалы, выделять важные аспекты и формировать точные ответы на запросы пользователей [4 - 7]. Сегодня особое внимание уделяется разработке автономных платформ для внутренней корпоративной среды, позволяющих эффективно управлять конфиденциальной документацией без обращения к внешним ресурсам интернета. Подобные платформы приобретают дополнительную ценность благодаря требованиям импортозамещения отечественного программного обеспечения, необходимым для защиты стратегически значимых направлений промышленности.

Отдельной проблемой является информационная безопасность. Подлежащая анализу документация включает для высокотехнологичных отраслей, в частности для Росатома информацию, являющуюся не подлежащей распространению интеллектуальной собственностью, коммерческой тайной, а иногда – и государственной тайной. Обработка таких материалов в облачных сетях недопустима. Разработка собственной интеллектуальной системы обработки корпоративных данных служит важным вкладом в повышение уровня информационной безопасности и обеспечение быстрого доступа к необходимой информации внутри организации.

Основной задачей настоящего исследования выступает создание изолированной локальной системы на базе технологий искусственного интеллекта, предназначенной для автоматизированной обработки конфиденциальных текстовых материалов и функционирующей автономно вне сети Интернет. Разработка подобного решения приобретает особую значимость ввиду актуальной потребности в обеспечении независимости отечественного ПО, используемого в стратегически значимых промышленных сферах [8-10]. Значимость предлагаемой системы заключается главным образом в отсутствии зависимости от сторонних онлайн-сервисов, возможности гарантированного уровня безопасности и оперативного доступа к необходимой информации благодаря отсутствию обмена данными через глобальную сеть.

III. ОПИСАНИЕ АРХИТЕКТУРЫ СИСТЕМЫ

Разработанная архитектура была спроектирована на основании требований, сформулированным в дискуссиях с заинтересованными представителями заказчиков, преимущественно из атомной отрасли (АО

АСЭ, АО ВНИИАЭС, Концерн Росэнергоатом). Состав верхне-уровневых нефункциональных требований:

- Обеспечение конфиденциальности и безопасность данных.
- Обеспечение независимости от внешних сервисов.
- Обеспечение быстрой доступности и достаточной скорости обработки данных.
- Требование обеспечения экономия ресурсов при регулярно высоких уровнях загрузки системы.
- Обеспечение управление данными без привлечения внешних услуг и сервисов.
- Обеспечение бесперебойности работы, в том числе при отключениях, сбоях и запретах на использование внешней инфраструктуры или сервисов для обеспечения непрерывности бизнес-процессов.
- Возможность развития инструмента внутренними ресурсами организации.
- Масштабируемость - увеличения объемов обработки системы без привлечения внешней инфраструктуры или сервисов.

Функциональные требования (определяют, какие функции система должна выполнять для достижения поставленных целей)

- Система должна выполнять анализ проектной документации атомной отрасли на предмет соответствия Федеральным нормам и правилам в сфере атомной энергетики
- Система должна обеспечивать возможность обработки различных форматов документов (PDF, DOCX, TXT и др.).
- Система должна обеспечивать поддержку работы с большими объемами данных (десятки тысяч страниц).
- Объем хранимых данных ограничен только техническими характеристиками сервера, для прототипа объём доступного дискового пространства составляет 7 терабайт.
- Система должна обеспечивать использование технологии RAG (Retrieval Augmented Generation) путем интеграции модели RAG для извлечения релевантных фрагментов текста из нормативных актов и сопоставления их с проектной документацией.
- Система должна обеспечивать на уровне прототипа окно ввода не менее 8 тысяч токенов
- Система должна обеспечивать локальное хранение данных.
- Система должна иметь два режима работы: локальный, без выхода в интернет с развернутыми в контуре моделями и облачный.
- Система должна обеспечивать вариативность инструментов -возможность интеграции с разными LLM и векторными БД.
- Система должна иметь интерфейс администратора для загрузки файлов в векторную базу данных и управления правами пользователей.
- Система должна обеспечивать возможность индивидуализированных настроек под специфические потребности предметной области путем наполнения векторных баз данных.
- Система должна обеспечивать обработку уникальных

настраиваемых форматов документов, использование специализированных языков и терминологий.

Разработанная архитектура представлена на рисунке 1.

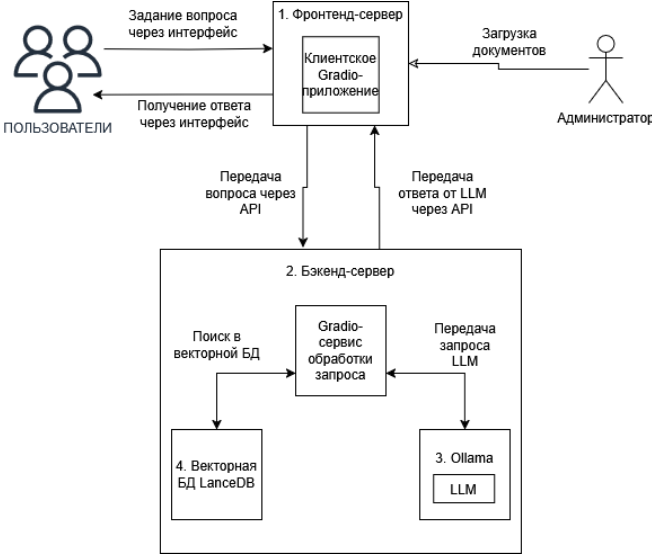


Рис. 1. Архитектура системы

В качестве локальной большой лингвистической модели использована Llama3, которая обладает достаточным количеством параметров и производительностью, достаточной для данного проекта. При этом система имеет удобный функционал подключения (смены и выбора) других языковых моделей.

Для хранения векторных представлений выбрана векторная база Lance DB. Её можно запустить локально на сервере проекта и достаточно просто использовать. База данных допускает размещение в Docker-контейнер для упрощения развертывания, замены, переноса и масштабирования системы. Для выбора векторной БД критериями были простота настройки и скорость работы. Скорость оценивалась по среднему времени отклика. Рассматривалось 4 БД: Milvus, LanceDB, FAISS и ChromaDB. Результаты - в Таблице I.

Таблица I – Сравнение баз данных

	Простота использования	Производительность	Особенности
FAISS	–	+	Сложна в эксплуатации и администрировании, подходит для масштабных проектов.
Milvus	–	+	Множество интеграций (например, с TensorFlow и PyTorch)
LanceDB	+	+	Меньше документации/сообществ, более молодое решение
ChromaDB	+ –	+	Могут возникать сложности с масштабированием при работе с Docker

Milvus медленнее LanceDB; FAISS быстрее, но сложнее установка и настройка; ChromaDB по скорости сопоставима с LanceDB. ChromaDB имеет ограничения

в масштабируемости, требует ручной настройки при работе с Docker. LanceDB оказалась оптимальным выбором по скорости и простоте настройки. Оказалось удобным сохранять документы в формате Lance, который обеспечивает быстрое извлечение данных.

Для взаимодействия с API использовалась библиотека ollama-python. API встроен в ollama, через API к ней обращается бэкенд сервер. Библиотека ollama-python обеспечивает удобный интерфейс для взаимодействия с API.

При создании пользовательского интерфейса использовался сервис Gradio. Это обеспечило понятное и доступное взаимодействие с моделью.

Инструментом достижения достаточной точности и релевантности ответов был метод RAG (Retrieval Augmented Generation). Этот стек технологий обеспечивает извлечение информации и генерацию текста с использованием больших языковых моделей.

Схема работы с векторной базой данных приведена на рисунке 2. Загрузка файлов в векторную базу данных доступна только администратору строго по его логину и паролю.

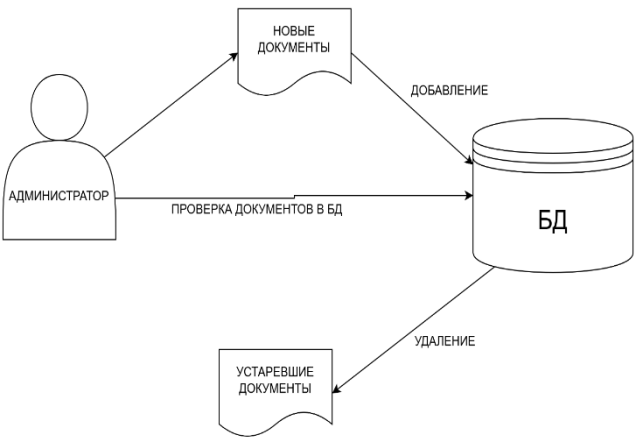


Рис.2. Схема работы с векторной базой данных

Интерфейсы системы для администратора и пользователя различны (в связи с различием функционала и прав).

Интерфейс администратора показан на рисунке 3. Администратор актуализирует базу знаний. Новые документы (например, полученные с сайта Ростехнадзора [11]), оперативно добавляются в базу, а устаревшие или утратившие силу – удаляются. Корректность экспертиз обеспечивается актуальностью базы знаний (набором файлов НТД).

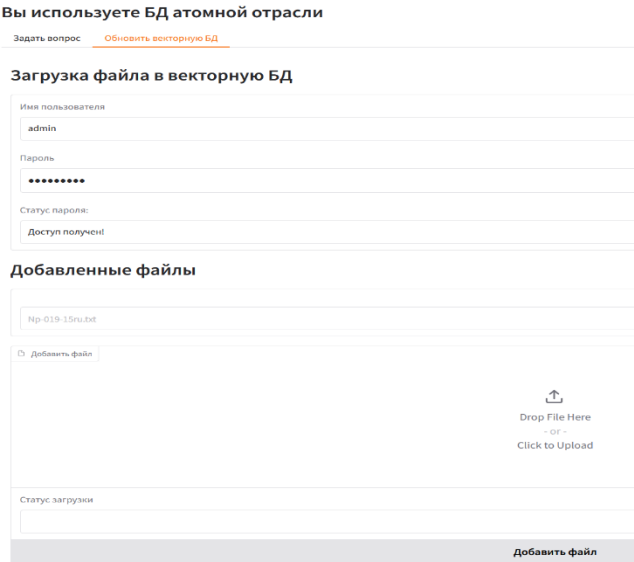


Рис.3. Интерфейс администратора

Интерфейс пользователя показан на рисунке 4. Пользователь может задать вопрос и загрузить файл для проверки на соответствие федеральным нормам права атомной отрасли. Также ему доступен выбор языковой модели.

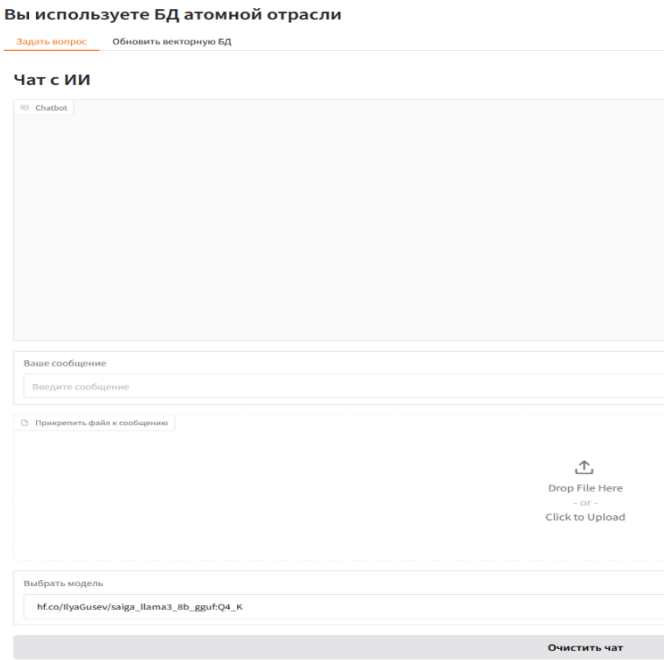


Рис.4. Интерфейс администратора

IV. МЕТОДИКА ТЕСТИРОВАНИЯ

В качестве анализируемых документов использованы документы по контролю проектирования водо-водяного ядерного реактора Westinghouse AP1000, которые есть в открытом доступе [12]. Рабочий цикл экспертизы документации включает в себя следующие шаги:

1. Загрузка актуального нормативного документа, на соответствие которому проводится экспертиза.
2. Автоматизированное создание чек-листа по конкретному нормативному документу. Полученный чек-лист может проверяться экспертом. В ходе испытаний выполнялась 100%-я проверка, в рабочем

режиме предполагается выборочная проверка сформированных чек-листов.

2. Загрузка файлов подлежащей проверки документации (проектная документация, технологические регламенты и др).

3. Активизируется функция заполнения чек-листа для оценки подлежащего экспертизе документа на соответствие требованиям нормативного документа. В ходе испытаний выполнялась 100%-я проверка всех заполненных чек-листов экспертами.

4. Рассчитывается доля корректно заполненных пунктов чек-листа (от общего числа пунктов). Этот параметр принимался в качестве показателя оценки соответствия подлежащего экспертизе документа требованиям НТД.

Подход через декомпозицию нормативно-технических документов к набору требований (пунктов чек листа) был принципиальным методическим прорывом, более «размытые» подходы давали кардинально худшие результаты экспертизы. С помощью приема формирования чек листа задача сводится к бинарной задаче проверки наличия в документах фрагментов, соответствующих критериям чек-листа, что оказалось внятной и реализуемой задачей для LLM.

В ходе испытаний системы были выполнены все шаги методики. В качестве нормативного документа (см. пункт 1) использовался документ НП-019-15. Созданный системой в результате автоматизированного анализа чек-лист представлен на рис.5.

№	Требование	Критерии соответствия
1	Обеспечена минимизация облучения персонала и населения	Уровень радиационного воздействия ниже норм
2	Установлены методы сбора ЖРО с учетом их физических и химических свойств	Есть обоснованные проектные решения
3	Осуществляется раздельный сбор ЖРО разных типов	Наличие специальных емкостей для каждого типа ЖРО
4	Обеспечены барьеры для предотвращения утечек	Установлена система сигнализации и мониторинга протечек
5	Выполняется контроль характеристик ЖРО перед переработкой	Есть протоколы входного контроля
6	Установлены методы переработки ЖРО для снижения радиационной опасности	Обоснованы методики переработки в проектной документации
7	Установлены лимиты активности ЖРО в хранилищах	Прописаны в проектной документации
8	Помещения оборудованы системами пожаротушения	Наличие средств пожаротушения и сигнализации
9	Организован радиационный контроль	Наличие приборов контроля и отчетов
10	Кондиционированные ЖРО соответствуют критериям захоронения	Паспорта соответствия для упаковок

Рис. 5. Чек-лист по нормативному документу НП-019-15

Была выполнена автоматизированная экспертиза 10 файлов проектной документации (ML11171A347.txt – ML11171A356.txt). Система заполнила полученный чек-лист для оценки этой проектной документации. Пример заполненного чек-листа представлен на рисунке 6.

№	Требование	Соответствие:	
		да/нет	Обоснование
1	Обеспечена минимизация облучения персонала и населения	Нет	Документ ML11171A350.txt не охватывает все нормы НП-019-15 по минимизации облучения.
2	Установлены методы сбора ЖРО с учетом их физических и химических свойств	Да	Методы сбора описаны в ML11171A347.txt, раздел 11.2.
3	Осуществляется раздельный сбор ЖРО разных типов	Да	Документ ML11171A348.txt включает данные о раздельном сборе, раздел 11.3.2.
4	Обеспечены барьеры для предотвращения утечек	Нет	Документы не содержат данные о барьерах.
5	Выполняется контроль характеристик ЖРО перед переработкой	Да	Контроль описан в ML11171A348.txt, раздел 11.3.4.
6	Установлены методы переработки ЖРО для снижения радиационной опасности	Да	Методы переработки детально описаны в ML11171A348.txt.
7	Установлены лимиты активности ЖРО в хранилищах	Нет	Документы не содержат данных о лимитах активности.
8	Помещения оборудованы системами пожаротушения	Нет	Документ ML11171A349.txt описывает пожарную защиту не полностью.
9	Организован радиационный контроль	Да	Радиационный контроль описан в ML11171A352.txt, раздел 11.5.
10	Кондиционированные ЖРО соответствуют критериям захоронения	Нет	Документ ML11171A356.txt не описывает соответствие критериям захоронения.

Рис. 6. Чек-лист, заполненный для проектной документации

Контроль корректности заполнения чек-листов экспертами показал, что неверно заполнен пункт 4: документ ML11171A349.txt содержит данные о барьерах и обоснование их эффективности. Процент правильно заполненных пунктов чек-листа составил 90%, (9 из 10 пунктов). Сходные результаты (80 – 100% правильно выполненных элементарных проверок) были зафиксированы и в других случаях. На этапе прототипа системы это приемлемый результат. Система будет дорабатываться и в части цифровой среды, и в части методики применения..

V. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНОГО ИССЛЕДОВАНИЯ

В процессе валидации прототипа была проведена серия тестов с использованием реальной документации из атомного сектора. Цель заключалась в количественной и качественной оценке эффективности работы системы по автоматическому анализу проектной документации и проверке её соответствия действующим нормативным требованиям.

Основным показателем точности был выбран процент правильно интерпретированных пунктов чек-листа, сформированного на основе документа НП-019-15. В качестве эталонного набора использовались 10 текстовых файлов, содержащих проектные материалы, относящиеся к конструкции реакторной установки AP1000.

Алгоритм проверки включал:

- сопоставление пунктов нормативного чек-листа с фрагментами проектной документации;
- определение наличия или отсутствия соответствующих положений;
- интерпретацию модели в виде бинарной классификации (соответствует / не соответствует);
- верификацию результатов вручную специалистом.

По итогам эксперимента в среднем 9 из 10 пунктов были интерпретированы корректно, что дает оценку

показателя точности 90%. Ошибкой оказался один пункт, в котором модель не выявила в документации информацию о барьерах безопасности, хотя она там присутствовала. Анализ этой ошибки показал, что причина кроется в недостаточном контекстном охвате фрагментов документа при векторизации.

Дополнительно были зафиксированы следующие технические метрики:

- Среднее время отклика системы на запрос: ~1.3 сек;
- Среднее время генерации чек-листа: ~10.4 сек;
- Объём векторной базы после загрузки: 78 МБ;
- Утилизация оперативной памяти при нагрузке: ~7.6 ГБ;
- Нагрузка на CPU (Llama3 8b в режиме inferencing): до 85% на 4 ядрах.

Полученные результаты продемонстрировали применимость предложенного подхода и подтвердили его потенциальную эффективность при использовании в замкнутых высокозащищённых средах. Высокая степень локализации модели, отсутствие обращения к внешним API и простота администрирования платформы делают систему пригодной для дальнейшего промышленного внедрения и адаптации под иные нормативные базы.

VI. ПЕРСПЕКТИВЫ РАЗВИТИЯ СИСТЕМЫ

В рамках дальнейшего развития системы определён ряд приоритетных направлений, направленных на повышение эффективности, расширение функциональности и обеспечение надёжности работы в условиях промышленных и изолированных контуров информационных систем.

- Оптимизация производительности и времени отклика

Будет реализован комплекс мер по снижению времени генерации ответов и отклика системы. В план включены: переход к более лёгким квантованным моделям без потери качества (например, Llama3 8b Q4_K_M), адаптация процессов векторизации с уменьшением избыточной нагрузки, пересмотр параметров chunk'инга и слияния результатов, а также внедрение асинхронной обработки пользовательских запросов.

- Автоматизация обновления базы нормативных документов

Разработка и интеграция механизма автоматического отслеживания изменений на порталах Ростехнадзора и других регуляторов с последующей актуализацией базы знаний. Это позволит системе самостоятельно поддерживать релевантность нормативной базы без участия администратора.

- Поддержка альтернативных языковых моделей

Расширение модели inference-ядра с возможностью горячей замены моделей (Llama, Mistral, Gemma, отечественные LLM). Будет реализован механизм динамического выбора модели в зависимости от профиля задачи и доступных вычислительных ресурсов.

- Формализация оценки качества ответов

Добавление встроенной метрики оценки качества

(precision, recall, f1) на основе внутренних эталонных задач, включая контрольные чек-листы и сценарии проверки документации. Это позволит регулярно проводить аудит качества модели без участия человека.

- Повышение устойчивости и масштабируемости

Контейнеризация всех компонентов с возможностью масштабирования в рамках частного кластера. Добавление поддержки балансировки нагрузки и работы в режиме нескольких обработчиков запросов с общей базой векторов.

- Улучшение интерфейсов и взаимодействия с пользователем

Реализация пакетной обработки документов, визуализация прогресса генерации чек-листов, расширение интерфейса администрирования для ведения логов, аудита действий и контроля версий загруженных нормативов.

Эти направления обеспечат дальнейшее повышение стабильности, скорости работы и адаптивности системы, а также позволят подготовить её к промышленной эксплуатации в условиях повышенных требований к безопасности и автономности.

VII. ЗАКЛЮЧЕНИЕ

Эта статья освещает создание специализированной платформы для анализа документации атомного сектора, включающей верификацию соответствия законодательным нормам и стандартам. Был разработан опытный образец-прототип локализованной системы, использующей искусственный интеллект (большие лингвистические модели) совместно с технологией RAG.

Реализация внутриконтурной системы для обработки текстовых материалов (файлов) с ограниченным доступом показала потенциал внедрения технологий искусственного интеллекта в специализированных сферах и обозначила вектор дальнейшего совершенствования. Среди перспектив — интеграция продвинутых лингвистических моделей и автоматизированная синхронизация базы знаний с официальными ресурсами Ростехнадзора.

БЛАГОДАРНОСТИ

Авторы выражают благодарность Высшей инженеринговой школе НИЯУ МИФИ за помощь в подготовке публикации.

БИБЛИОГРАФИЯ

- [1] Rothman D. Transformers for Natural Language Processing. – Birmingham: Packt Publishing, 2024.
- [2] Wei J. et al. "Emergent abilities of large language models", arXiv preprint arXiv:2206.07682 (2022).
- [3] Ivanova S. D., Shemonaev D. D. Electronic educational resources, Web of Scholar. – 2017. – No. 5(14). – P. 59-60. – EDN ZEQMFMH.
- [4] От национальных проектов к системе стратегического планирования. Координация действий министерств и ведомств. М. Ю. Мокшин, В. Г. Марача, М. Г. Жабицкий, О. В. Бойко, International Journal of Open Information Technologies. – 2022. – Т. 10, № 8. – С. 113-120. – EDN UIHCNL

- [5] Jain S. M. Hugging face, Introduction to transformers for NLP: With the hugging face library and models to solve problems. – Berkeley, CA: Apress, 2022. – С. 51-67.
- [6] Ivanova S. D., Shemonaev D. D. New reality 2020: distance education and formation of the electronic learning environment in the university, Modern European Researches. – 2020. – No. 3. – P. 10-14. – EDN QUSXWY.
- [7] Предобработка текста в NLP. Available: <https://python-school.ru/blog/nlp/nlp-text-preprocessing> Retrieved: Jun, 2025
- [8] Pedregosa F. et al. Scikit-learn: Machine learning in Python, The Journal of machine Learning research. – 2011. – Т. 12. – С. 2825-2830.
- [9] Шемонаев, Д. Д., Иванова С. Д. Влияние цифровых технологий на социальное взаимодействие: вербальные и невербальные коммуникации, Политехнический молодежный журнал. – 2023. – № 9(86). – DOI 10.18698/2541-8009-2023-9-933. – EDN OZNCGW.
- [10] Technical Sharing – AI with RAG. Available: <https://www.manifera.com/technical-sharing-at-manifera-introduce-rag/> Retrieved: Jun, 2025.
- [11] Федеральная служба по экологическому, технологическому и атомному надзору (Ростехнадзор). Available: <http://cntr.gosnadzor.ru/> Retrieved: Jun, 2025.
- [12] Westinghouse AP1000 Design Control Document. Available: <https://www.nrc.gov/docs/ml1117/ml11171a500.html> Retrieved: Jun, 2025.

Статья получена 29 июля 2025

Шемонаев Дмитрий Дмитриевич, выпускник магистратуры ВИШ НИЯУ МИФИ, dmitryshemonaev@gmail.com

Анни Павел Витальевич, инженер учебного центра MinIO Inc., pavel.anni@gmail.com

Жабицкий Михаил Георгиевич, заместитель директора ВИШ НИЯУ МИФИ, jabitsky@mail.ru

Конев Юрий Николаевич, заместитель начальника отдела ВИШ НИЯУ МИФИ, YNKonev@mephi.ru

Analyzing system development for technical and organizational documents compliance with regulatory requirements of the nuclear industry based LLM technology

D.D. Shemonaev, P.V. Anni, M.G. Zhabitsky, Yu.N. Konev

Abstract - This paper discusses the design, development and testing of a prototype of an application system using artificial intelligence technologies for automated analysis of nuclear industry documentation, including verification of compliance with regulatory documents and regulatory requirements. The system uses large language model (LLM) and search-augmented generation (RAG) technologies. Options for using this technology stack for the use of specialized knowledge domains (an array of regulatory documents and technical instructions of the nuclear industry) for high-quality operational examination of working documents using artificial intelligence are considered. Improving the quality and speed of qualified expertise for complex high-tech industries can often be critical for the implementation of serious projects.

Due to the availability of significant sensitive information of a technical and economic nature in the industry, it is important to organize work without external access to the analyzed design documentation. The paper discusses methods of organizing work using both cloud and in-loop LLMs. A prototype of a tool for orchestrating heterogeneous models with access to the company's closed loop via VPN channels has been developed. A comparison of the solution of the problem by the intelligent assistant of the expert in the cloud and intra-contour modes is carried out. The developed prototype of a system operating without connection to external networks, providing a high level of confidentiality and import substitution, is described. The prototype provides for the separation of the roles of the subject user (expert) and the administrator, providing the possibility of secure interaction with the system. The developed intelligent assistant implements advanced contextual search and analytics/examination of project documentation based on information from specialized knowledge domains of limited access based on the use of RAG technology. The solution was implemented on the basis of a digital test site of the Higher School of Engineering MEPhI and is designed for use for educational tasks, educational and practical cases of industrial partners and for performing work in the interests of customers from the production sector, primarily Rosatom enterprises (including, but not limited to).

Keywords— RAG, Text Mining, LLM, vectorization, database, large language models, artificial intelligence, natural language, interface, federal rules and regulations, nuclear industry

BIBLIOGRAPHY

- [1] Rothman D. Transformers for Natural Language Processing. – Birmingham: Packt Publishing, 2024.
- [2] Wei J. et al. “Emergent abilities of large language models”, arXiv preprint arXiv:2206.07682 (2022).
- [3] Ivanova S. D., Shemonaev D. D. Electronic educational resources, Web of Scholar. – 2017. – No. 5(14). – P. 59-60. – EDN ZEQMFMH.
- [4] Ot nacional'nyh proektov k sisteme strategicheskogo planirovaniya. Koordinatsiya deystvii ministerstv i vedomstv / M. Ju. Mokshin, V. G. Maracha, M. G. Zhabitskiy, O. V. Bojko, International Journal of Open Information Technologies. – 2022. – T. 10, # 8. – S. 113-120. – EDN UIHCNL
- [5] Jain S. M. Hugging face, Introduction to transformers for NLP: With the hugging face library and models to solve problems. – Berkeley, CA: Apress, 2022. – S. 51-67.
- [6] Ivanova S. D., Shemonaev D. D. New reality 2020: distance education and formation of the electronic learning environment in the university, Modern European Researches. – 2020. – No. 3. – P. 10-14. – EDN QUSXWY.
- [7] Predobrabotka teksta v NLP. Available: <https://python-school.ru/blog/nlp/nlp-text-preprocessing> Retrieved: Jun, 2025
- [8] Pedregosa F. et al. Scikit-learn: Machine learning in Python, The Journal of machine Learning research. – 2011. – T. 12. – S. 2825-2830.
- [9] Shemonaev, D. D., Ivanova S. D. Vliyanie cifrovyyh tekhnologiy na social'noe vzaimodeystvie: verbal'nye i neverbal'nye kommunikatsii, Politehnicheskij molodezhnyj zhurnal. – 2023. – # 9(86). – DOI 10.18698/2541-8009-2023-9-933. – EDN OZNCGW.
- [10] Technical Sharing – AI with RAG. Available: <https://www.manifera.com/technical-sharing-at-manifera-introduce-rag/> Retrieved: Jun, 2025.
- [11] Federal'naja sluzhba po jekologicheskomu, tehnologicheskomu i atomnomu nadzoru (Rostehnadzor). Available: <http://cntr.gosnadzor.ru/DesignControlDocument> Retrieved: Jun, 2025. Westinghouse AP1000 Design Control Document. Available: <https://www.nrc.gov/docs/ml1117/ml11171a500.html> Retrieved: Jun, 2025.

Received: 29 July 2025

Shemonaev Dmitry Dmitrievich, graduate of the master's program of the Higher School of Economics of the National Research Nuclear University MEPhI, dmitryshemonaev@gmail.com

Pavel Annie, Engineer of the MinIO Inc. Training Center, pavel.anni@gmail.com

Zhabitskii Mikhail Georgievich, Deputy Director of the Higher School of Economics of the National Research Nuclear University MEPhI, jabitsky@mail.ru

Konev Yuri Nikolaevich, Deputy Head of the Department of the Higher School of Economics of the National Research Nuclear University MEPhI, YNKonev@mephi.ru