ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2027. Т. 9, № 1(30)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.8 DOI: 10.54708/SIIT-2027-no1-p3 EDN: ZEKJCK

# Интеллектуальная система обработки нормативных документов в сфере образования на базе RAG-архитектур

**М. М. Гаянова • А. М. Вульфин • С. М. Валиев**

*Уфимский университет науки и технологий*

Современная сфера образования сталкивается с необходимостью обработки постоянно растущего объема нормативных документов. Традиционные методы ручного анализа и существующие системы электронного документооборота ориентированы преимущественно на работу с мета-данными и не обеспечивают семантического анализа содержания документов. В работе представлена интеллектуальная система автоматизированного анализа образовательных стандартов, построенная на архитектуре RAG (Retrieval-Augmented Generation – генерация, дополненная извлечением данных). Описана предметно-ориентированная методика ее построения для русскоязычных нормативных документов: правила разбиения текста с сохранением привязки фрагмента к документу, редакции и структурной единице, правила формирования контекста с атрибуцией к первоисточнику и протокол оценки качества, разделяющий дословность цитаты, релевантность фрагмента и корректность ответа. Система обеспечивает загрузку документов в форматах PDF, DOCX и TXT, их предварительную обработку, векторизацию моделью multilingual-e5-large и формирование ответов с использованием языковой модели Qwen2-72B и векторного поиска на основе библиотеки FAISS; программный интерфейс построен на фреймворке FastAPI. Апробация выполнена на корпусе из 1200 документов десяти образовательных организаций и наборе из 40 запросов трех уровней сложности. Доля дословно подтвержденных цитат составила 97,3 %, среднее время векторного поиска – 148 мс, среднее время порождения ответа – 2,7 с, субъективная оценка сокращения времени выполнения типовых задач – 43–47 %. F1-мера качества поиска составила 0,92 на простых запросах, 0,82 на запросах средней сложности и 0,69 на сложных запросах, требующих агрегации требований из нескольких документов; ограничение качества на междокументных запросах связано с неучетом транзитивных связей между стандартами. Реализация системы в виде веб-сервиса представляет практическую ценность для образовательных организаций и органов управления образованием.

*Нейросетевой анализ; образовательные стандарты; нормативные документы; RAG-архитектура; семантический поиск; обработка естественного языка; веб-сервис; векторные представления.*

## Введение

Современные образовательные стандарты, включая федеральные государственные образовательные стандарты (далее – ФГОС), представляют собой объемные нормативные документы, требующие программных средств автоматизированного анализа и обработки [Кож25, Кит23]. Ежегодно в Российской Федерации публикуется порядка пяти тысяч новых нормативно-методических документов, относящихся к сфере образования, что существенно затрудняет их ручную обработку и своевременный учет [Каш25]. Актуальность настоящей работы обусловлена быстрым ростом объема образовательной документации и необходимостью оперативного доступа к положениям нормативных актов, учебных программ и методических материалов [Кож25].

Гаянова М. М., Вульфин А. М., Валиев С. М. Интеллектуальная система обработки нормативных документов в сфере образования на базе RAG-архитектуры // СИИТ. 2027. Т. 9, № 1(30). С. 3-17. DOI: 10.54708/SIIT-2027-no1-p3. EDN: ZEKJCK.

Gayanova M. M., Vulfin A. M., Valiev S. M. “An intelligent system for processing regulatory documents in the field of education based on the RAG architecture” // SIIT. 2027. Vol. 9, no. 1(30), pp. 3-17. DOI: 10.54708/SIIT-2027-no1-p3. EDN: ZEKJCK. (In Russian).

Целью работы является разработка интеллектуальной системы автоматизированного анализа образовательных стандартов, обеспечивающей загрузку и индексацию документов, семантический поиск с использованием векторных представлений текста и интерактивный пользовательский интерфейс, позволяющий получать ответы на произвольные запросы пользователя на основе контекста, извлеченного из нормативной базы [Шем25, Lew20].

В ходе работы решаются следующие задачи: разработка модуля предварительной обработки документов, включающего преобразование форматов и очистку текста; реализация векторного поиска на основе библиотеки FAISS (Facebook AI Similarity Search – программная библиотека для поиска по сходству в многомерных пространствах) [Абр25, Joh21]; построение программного интерфейса на базе фреймворка FastAPI [Кош24]; интеграция большой языковой модели семейства Qwen2 для порождения ответов на естественном языке [Yan24].

Научная новизна работы состоит не в применении готовых моделей, а в предметно-ориентированной методике построения RAG-системы для русскоязычных нормативных документов сферы образования и в ее экспериментальном обосновании. Методика задает: правила предобработки и разбиения нормативных документов, при которых каждый фрагмент сохраняет привязку к документу-источнику, его редакции и структурной единице (пункт, компетенция, дисциплина); правила формирования контекста, при которых в ответ включаются только фрагменты с атрибуцией к первоисточнику; протокол оценки качества, разделяющий три уровня проверки (дословность цитаты, релевантность извлеченного фрагмента, корректность ответа) и три уровня сложности запроса (поиск внутри одного документа, сопоставление двух документов, агрегация трех и более документов). Совокупность этих правил позволяет получать проверяемые ответы по русскоязычной нормативной базе и количественно оценивать границы применимости подхода: эксперимент показывает, что качество сохраняется на внутридокументных запросах и снижается при агрегации требований из нескольких стандартов. Использованные модели и библиотеки (multilingual-e5-large, FAISS, Qwen2-72B) рассматриваются как средство реализации и самостоятельного вклада не образуют.

## Анализ предметной области

Существующие системы электронного документооборота (СЭД), такие как «1С:Документооборот» и Directum, представляют собой надежные программные средства автоматизации типовых процессов. Они обеспечивают регистрацию входящей и исходящей корреспонденции, контроль исполнительской дисциплины по резолюциям, версионирование договоров с привязкой к контрагентам, а также настройку маршрутов согласования документов. Вместе с тем в типовой конфигурации их архитектура ориентирована на обработку метаданных – реквизитов, статусов, классификаторов и маршрутов движения документов – и не предусматривает семантического анализа текстового содержания; отраслевые расширения и модули поиска по содержанию в настоящей работе не рассматривались.

В качестве иллюстрации ограничений традиционных СЭД рассмотрим типовую задачу образовательного учреждения: требуется выявить все рабочие программы дисциплин (РПД), содержание которых обеспечивает формирование заданной компетенции, установленной ФГОС высшего образования. В системе «1С:Документооборот» такой поиск возможен исключительно при условии, что ответственным сотрудником вручную заполнено соответствующее поле карточки документа, содержащее код компетенции. Программное обеспечение указанного класса в типовой конфигурации не выполняет автоматизированного анализа содержания РПД и не способно самостоятельно определить соответствующую компетенцию: результат зависит от полноты ручного заполнения карточки. Кроме того, семантически различные, но лексически близкие формулировки – например, «обучающийся должен знать язык Python» и «обучающийся знаком с языком Python» – не различаются стандартными механизмами поиска, тогда как с точки зрения оценки сформированности компетенции разница между полным владением материалом и поверхностным ознакомлением является существенной.

Дополнительную сложность создает регулярное обновление нормативной базы. При выходе нового приказа Министерства науки и высшего образования Российской Федерации, вносящего изменения в ФГОС, традиционные СЭД не выполняют автоматического переклассифицирования связанных документов. Требуется либо ручное переназначение классификаторов сотрудниками, либо доработка системы силами разработчиков. Смысловая связь между новым документом и ранее утвержденными учебными планами не устанавливается автоматически, что повышает риск использования устаревших положений стандарта. Попытки расширить функциональность СЭД путем введения дополнительных реквизитов и справочников приводят к значительному росту трудозатрат на ручное заполнение карточек и фактически нивелируют преимущества автоматизации.

В области автоматизированной обработки образовательных стандартов сформировались два основных подхода: тонкая настройка предобученных языковых моделей (fine-tuning – дополнительное обучение готовой модели на специализированном корпусе) и архитектура генерации, дополненной извлечением (Retrieval-Augmented Generation, далее – RAG), предложенная в работе [Lew20].

Первый подход предполагает дообучение предварительно обученной модели, например, BERT (Bidirectional Encoder Representations from Transformers – двунаправленные кодирующие представления на основе архитектуры «трансформер»), на предметных данных – текстах ФГОС, локальных нормативных актах образовательной организации, рабочих программах дисциплин. Тонкая настройка обеспечивает высокое качество на ограниченном наборе тем, однако имеет существенные практические ограничения. Для ее применения необходим размеченный учебный корпус, подготовка которого требует участия экспертов-предметников и значительных временных затрат. При каждом существенном изменении нормативной базы модель требуется переобучать, что сопряжено с расходом вычислительных ресурсов – графических ускорителей (GPU – Graphics Processing Unit – специализированный процессор для выполнения параллельных матричных операций) и оперативной памяти [Hu21].

Архитектура RAG реализует иной принцип: перед порождением ответа языковая модель выполняет запрос к внешней базе знаний, в которой хранятся векторизованные фрагменты документов; на основе извлеченных наиболее релевантных фрагментов формируется ответ [Lew20, Gao24]. Преимущества такой архитектуры применительно к задачам работы с нормативной базой состоят в следующем. Обновление базы знаний не требует переобучения модели — достаточно добавить новый документ в векторное хранилище, после чего система учитывает его при последующих ответах. Достоверность порождаемых ответов повышается за счет того, что модель опирается на извлеченные фрагменты исходных документов и в меньшей степени подвержена явлению галлюцинаций (hallucinations – порождение фактически недостоверного, но правдоподобного по форме текста). При отсутствии запрашиваемой информации в базе знаний корректно построенная RAG-система сообщает о невозможности ответа, а не порождает недостоверное содержание [Lew20].

Примеры применения RAG в смежных областях описаны в работах [Каш25, Шем25, Бой26], где показано, что использование векторного поиска совместно с языковой моделью позволяет существенно повысить точность поиска релевантных нормативных положений и сократить время подготовки справочных материалов. Обзор современных подходов к построению RAG-систем, включая методы разбиения текста на фрагменты, оценки релевантности и управления контекстом, представлен в работе [Gao24]. Вопросы структурной сегментации и гибридного лексико-векторного поиска по нормативно-технической документации рассматриваются в работе [Хру26], где показано, что наивное разбиение текста разрушает структуру таблиц и приводит к потере контекста родительских заголовков.

Сопоставление с близкими работами целесообразно проводить по признакам, которые непосредственно определяют качество ответа на запрос по нормативной базе: способ разбиения текста на фрагменты, механизм поиска, обработка междокументных (транзитивных)

связей, учет редакций документов и проверка ответа по первоисточнику. Сводное сопоставление классов решений, к которым относятся рассмотренные работы, приведено в табл. 1.

Таблица 1

**Сопоставление классов RAG-решений по функциональным признакам**

| Признак | | | Графовые расширения RAG [Gao24] | Предложенная система |
|---|---|---|---|---|
| Разбиение текста | по фиксированной длине с перекрытием | не регламентировано | по фиксированной длине | по границам абзацев и пунктов; 256–512 токенов, перекрытие 10–20 %; привязка к документу, редакции и структурной единице |
| Механизм поиска | плотные векторные представления, top-k | BM25 и плотные векторы, линейная комбинация оценок | плотные векторы и явный обход графа связей | плотные векторы multilingual-e5-large, индекс IndexFlatIP, косинусная мера, top-k |
| Междокументные связи | не обрабатываются | не обрабатываются | обрабатываются явно | описаны в логической модели; в поиске не используются (перспектива) |
| Учет редакций | не рассматривается | не рассматривается | не рассматривается | номер и дата редакции фиксируются в метаданных фрагмента; ответ содержит ссылку на источник |
| Проверка ответа | не выполняется | не выполняется | частично (ссылки на вершины графа) | ответ формируется по извлеченным фрагментам; цитаты проверяются на дословное совпадение |

Из табл. 1 видно, что рассмотренные классы решений закрывают отдельные признаки, но не их совокупность: универсальные RAG-схемы не обрабатывают междокументные связи и редакции документов, гибридный лексико-векторный поиск повышает устойчивость к неточным формулировкам, но не решает задачу агрегации требований, а графовые расширения требуют заранее размеченных связей и не описывают процедуру их получения. Остается ограничение: для русскоязычных образовательных стандартов не описан подход, который одновременно фиксирует редакцию документа, обеспечивает проверяемость цитат и позволяет количественно оценить деградацию качества при переходе от внутридокументных запросов к междокументным. На устранение этого ограничения направлены предложенная методика и ее экспериментальная апробация, описанные в последующих разделах.

Прямое применение универсальных моделей векторного представления текста (эмбеддеров – encoders, нейросетевых кодировщиков, отображающих текст в числовой вектор фиксированной размерности) и больших языковых моделей, обученных преимущественно на общеязыковых корпусах, к задаче анализа образовательных стандартов связано с рядом трудностей. Во-первых, предметные аббревиатуры, обозначающие универсальные, общепрофессиональные и профессиональные компетенции (УК, ОПК, ПК с порядковыми номерами), зачастую распознаются универсальными моделями как несемантические токены либо смешиваются с номерами документов, что приводит к неполноте результатов поиска и к ложным срабатываниям. Во-вторых, канцелярский стиль формулировок ФГОС, характеризующийся использованием глаголов с близким значением («способен анализировать», «готов применять», «владеет методами»), затрудняет различение уровней сформированности компетенций без специализированного дообучения модели на педагогическом корпусе. В-третьих, загрузка всей совокупности действующих стандартов, паспортов компетенций и рабочих программ в контекстное окно языковой модели невозможна вследствие ограничения на длину входной последовательности, измеряемой в токенах. Архитектура RAG решает это ограничение посредством векторного поиска, однако при необходимости агрегации информации из значительного числа различных стандартов – например, при сопоставлении требований к одной и той же компетенции в разных поколениях ФГОС – точность и полнота ответа могут снижаться.

Таким образом, актуальной остается задача построения гибридных решений, сочетающих RAG-архитектуру с дообученными на образовательных текстах моделями векторного представления, а также с правилами предметной области.

## Информационное обеспечение

Модель данных системы основана на концепции семантически обогащенных документов и описывается как граф сущностей и связей. Вершины графа соответствуют нормативному документу, фрагменту документа, понятию, связи понятий, учебному плану, дисциплине, оценочному средству и фонду оценочных средств; ребра задают отношения «содержит», «задает структуру», «оперирует», «участвует», «формируется в», «проверяется через», «опирается на», «входит в». Ключевые сущности и их взаимосвязи представлены на рис. 1.

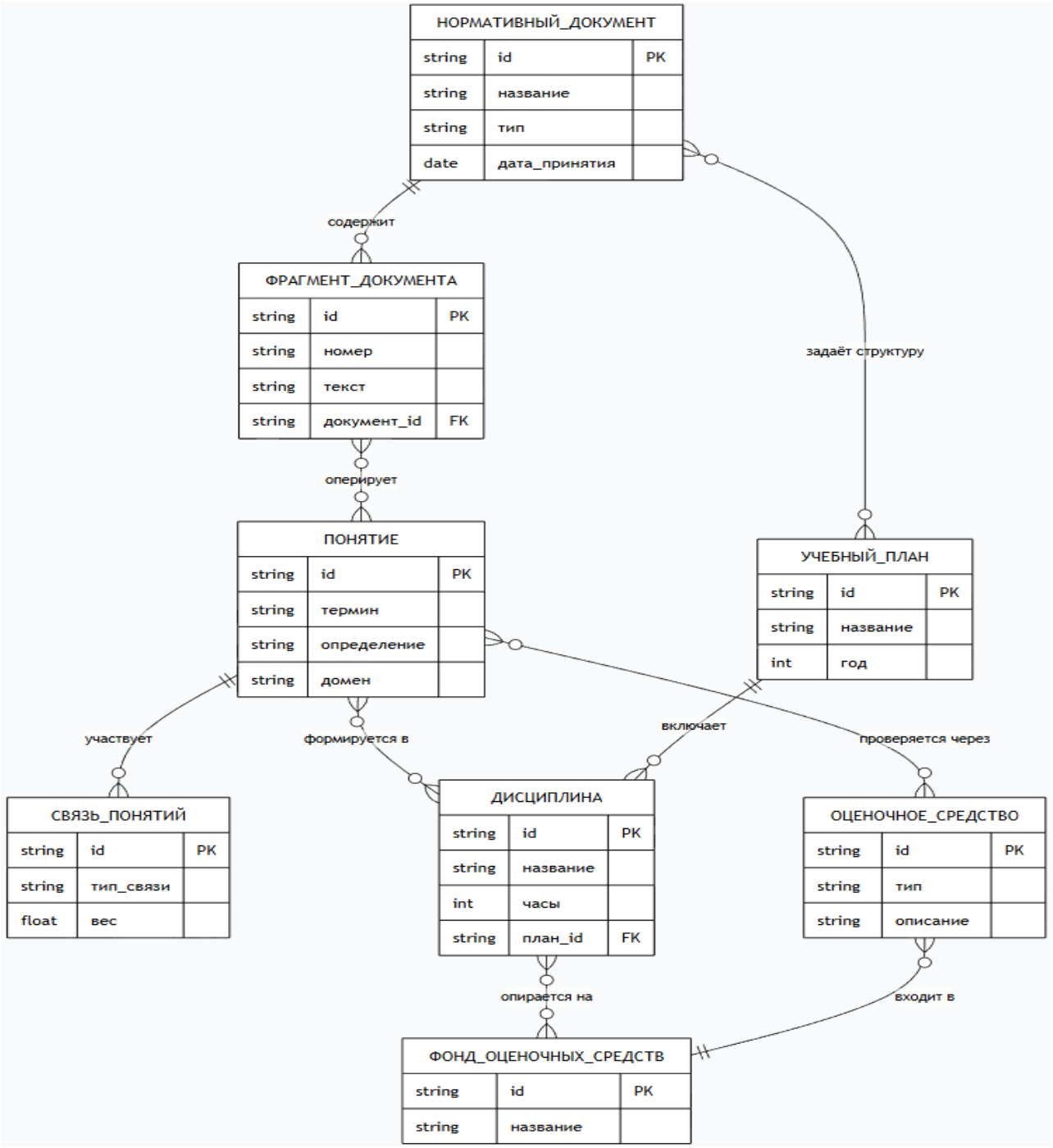



**Рис. 1.** Логическая модель данных системы

Модель является логической. На текущем этапе сущности, связи и метаданные документов (наименование, дата, автор, источник, номер и дата редакции) хранятся в реляционных таблицах PostgreSQL, векторные представления фрагментов – в индексе FAISS. Отдельная графовая система управления базами данных в предложенной системе не используется, и обход связей при поиске не выполняется: связи применяются для интерпретации результата (указание документа и структурной единицы, к которым относится фрагмент) и рассматриваются как резерв повышения качества на междокументных запросах. Графовая СУБД Neo4j использовалась только в базовой конфигурации, с которой проводилось сопоставление; в предложенной системе она не применяется.

Для хранения векторных представлений текста используется библиотека FAISS, обеспечивающая эффективный поиск ближайших соседей в многомерном пространстве [Абр25, Joh21, Зык25]. Для понимания требований к вычислительным ресурсам рассмотрим устройство базового индекса – так называемой плоской таблицы (Flat index).

Плоский индекс в FAISS, реализуемый классами IndexFlatL2 (использует евклидово расстояние) или IndexFlatIP (использует скалярное произведение), представляет собой простейшую структуру данных для поиска с полным перебором. По существу, это двумерный массив, в котором каждая строка соответствует одному вектору, а количество столбцов равно размерности вектора. Все векторы хранятся в несжатом виде, без предварительной кластеризации. Такая структура гарантирует точный результат при поиске, поскольку запрос последовательно сравнивается с каждым вектором базы. Платой за точность является временная сложность, линейно возрастающая с размером базы.

Объем оперативной памяти, занимаемый плоским индексом, определяется количеством векторов, их размерностью и типом данных. В стандартной конфигурации FAISS хранит элементы векторов в формате 32-разрядных чисел с плавающей точкой (float32), то есть каждый элемент занимает 4 байта. Для примера: индекс, содержащий 10 000 векторов размерностью 1024, потребует около 40 Мбайт памяти (10 000 × 1024 × 4 байта = 39 Мбайт); индекс из 10 миллионов векторов той же размерности потребует порядка 40 Гбайт, что может превысить объем оперативной памяти одного сервера.

Для крупных коллекций, в которых требуется экономия памяти или ускорение поиска, библиотека предоставляет альтернативные индексы: IVF (Inverted File – обратный файловый индекс с предварительной кластеризацией пространства) и HNSW (Hierarchical Navigable Small World – иерархический граф малых миров, обеспечивающий приближенный поиск логарифмической сложности). Оба алгоритма реализуют приближенный поиск ближайших соседей, сокращая время отклика и требования к памяти ценой умеренного снижения точности. Метаданные документов – наименование, дата создания, автор, источник – хранятся в реляционной системе управления базами данных PostgreSQL, тогда как FAISS обеспечивает вычислительную составляющую семантического поиска [Зык25].

## Программная реализация

В качестве основного языка разработки выбран Python, что обусловлено его развитой экосистемой программных библиотек для обработки естественного языка и машинного обучения. Серверная часть построена на фреймворке FastAPI, обеспечивающем высокую производительность за счет асинхронной архитектуры и автоматическое формирование спецификации программного интерфейса согласно стандарту OpenAPI [Кош24]. FastAPI используется для решения типовых задач веб-сервера: приема пользовательских запросов, валидации входных данных, управления конкурентными соединениями и обработки загружаемых архивов документов. Для извлечения текста из файлов различных форматов (DOCX, PDF, TXT) применяются специализированные библиотеки-парсеры (python-docx, PyPDF2 и встроенные средства Python соответственно).

Для порождения векторных представлений текста выбрана модель multilingual-e5-large – многоязычный кодировщик семейства E5 (Embeddings from bidirectional Encoder representations – векторные представления, получаемые на основе двунаправленного энкодера) [Wan24, Rei19]. Выбор обусловлен необходимостью обработки русскоязычной образовательной терминологии: модель обучена более чем на ста языках и демонстрирует высокое качество кросслингвального переноса. Архитектурной основой модели служит трансформер BERT; общее число обучаемых параметров составляет приблизительно 560 миллионов. В конфигурации с 32-разрядной точностью с плавающей точкой (FP32) модель занимает около 2,24 Гбайт оперативной или видеопамяти, в 16-разрядной конфигурации (FP16) – около 1,12 Гбайт, что допускает ее запуск на серверах с ограниченными вычислительными ресурсами. Важной особенностью практического использования multilingual-e5-large является обязательное добавление префиксов к входным текстам: для индексируемых документов — префикс passage:, для поисковых запросов пользователя – префикс query:. Указанный прием существенно повышает релевантность результатов семантического поиска [Wan24].

Для порождения ответов на естественном языке на основе извлеченного контекста используется большая языковая модель Qwen2-72B – модель с 72 миллиардами параметров, демонстрирующая результаты, сопоставимые с GPT-4 на ряде публичных тестовых наборов, и обеспечивающая приемлемое качество обработки русского языка [Yan24]. Модель поддерживает длинное контекстное окно, что важно для обработки пространных фрагментов нормативных документов. Для справки: в конфигурации FP16 веса Qwen2-72B занимают приблизительно 145 Гбайт видеопамяти, а для устойчивого инференса (inference – этап применения обученной модели для порождения выходных данных) требуется от четырех до восьми графических ускорителей класса NVIDIA A100. Применение методов квантизации (quantization – понижение разрядности представления весов модели), в частности INT4 или AWQ (Activation-aware Weight Quantization – квантизация весов с учетом активаций), позволяет снизить требования к видеопамяти до 35-40 Гбайт, однако даже в этом случае развертывание модели на локальном оборудовании требует высокопроизводительного графического ускорителя. В настоящей работе обращение к модели Qwen2-72B выполняется через удаленный программный интерфейс; при таком подходе вычислительная нагрузка переносится на инфраструктуру поставщика услуг, а потребитель оплачивает фактический объем использования.

Аппаратная конфигурация сервера, на котором развернута разрабатываемая система, включает центральный процессор с восемью вычислительными ядрами и 64 Гбайт оперативной памяти, необходимых для функционирования веб-сервера на базе FastAPI, выполнения задач извлечения и предварительной обработки текста, а также хранения индекса FAISS. Для выполнения вычислений модели multilingual-e5-large в режиме инференса используется графический ускоритель NVIDIA с объемом видеопамяти не менее 4 Гбайт. Поскольку порождение ответов языковой моделью Qwen2-72B выполняется удаленно, собственные требования сервера к видеопамяти определяются исключительно моделью кодирования. Векторный индекс загружается в оперативную память сервера, а метаданные документов сохраняются в реляционной базе данных PostgreSQL.

Основной алгоритм работы системы включает следующие этапы. На этапе загрузки пользователь передает серверу архив формата ZIP, содержащий документы образовательной тематики. FastAPI-приложение принимает архив, распаковывает его и извлекает текстовое содержимое файлов поддерживаемых форматов при помощи соответствующих парсеров. Извлеченный текст проходит очистку: удаляются служебные символы, лишние пробелы, колонтитулы, номера страниц и иные артефакты, несущественные с точки зрения семантического содержания. Далее очищенный текст подвергается разбиению на фрагменты (chunking – разбиение протяженного текста на смысловые отрезки ограниченного размера для последующей векторизации). Разбиение выполняется по границам абзацев и пунктов документа, а не по произвольным позициям символов; каждый фрагмент наследует идентификаторы документа, редакции и структурной единицы. Токенизация выполняется токенизатором модели multilingual-e5-

large (XLM-R), а фрагмент, превышающий предельную длину входной последовательности модели (512 токенов), усекается по границе предложения. Перекрытие смежных фрагментов обеспечивает сохранение смысловой целостности на границах разбиения.

После разбиения для каждого фрагмента средствами модели multilingual-e5-large формируется векторное представление – плотный числовой вектор фиксированной размерности (1024 элемента для указанной модели). Векторы сохраняются в индексе FAISS. Поисковый запрос пользователя векторизуется той же моделью с префиксом query:; полученный вектор подается на вход индекса, который возвращает набор наиболее близких в смысле косинусной меры фрагментов. Извлеченные фрагменты вместе с исходным запросом формируют контекст, передаваемый языковой модели Qwen2-72B, которая формирует итоговый ответ по предоставленному контексту с указанием документов-источников; это снижает вероятность возникновения галлюцинаций и повышает прослеживаемость ответа к первоисточнику [Lew20, Gao24].

Параметры разбиения и поиска задавались эмпирически на выборке для настройки (см. раздел «Тестирование и результаты»): использовались фрагменты размером от 256 до 512 токенов с перекрытием 10–20 %, а количество извлекаемых фрагментов top-k подбиралось так, чтобы контекст оставался представительным при ограничении длины запроса к языковой модели. Фактические значения параметров, использованные в контрольных прогонах, должны быть указаны явно: top-k = 5, температура генерации = 0,2, максимальная длина генерации = 1024 токена. Применен индекс FAISS IndexFlatIP с предварительной L2-нормализацией векторов, что эквивалентно поиску по косинусной мере. Шаблон инструкции языковой модели требует отвечать только на основе переданных фрагментов, указывать для каждого утверждения документ-источник и сообщать о невозможности ответа, если требуемое положение в контексте отсутствует. Для порождения ответов использовалась модель Qwen2-72B-Instruct, доступная через удаленный интерфейс; параметры генерации фиксировались до контрольных прогонов. Учет редакций обеспечивается метаданными: при загрузке документа фиксируются его номер и дата редакции, а в ответе указывается документ и редакция, к которой относится цитата. Соблюдение инструкции не является гарантией достоверности: фактическая достоверность ответов оценивалась по метрикам, описанным в разделе «Тестирование и результаты».

## Тестирование и результаты

Оценка качества разработанной системы проводилась по трем взаимодополняющим направлениям. Во-первых, фиксировались интегральные показатели функционирования системы в штатном режиме – доля дословно подтвержденных цитат, среднее время векторного поиска и среднее время порождения ответа; сводные значения представлены в табл. 2. Во-вторых, проводилась оценка качества поиска по метрикам точности (precision), полноты (recall) и F1-меры на едином наборе запросов и единой эталонной разметке (табл. 3); для базовой (baseline) гибридной конфигурации измерялось время отклика. В-третьих, оценивалась пользовательская удовлетворенность (CSAT – customer satisfaction score, балльная оценка субъективной удовлетворенности качеством сервиса).

Приведенные в табл. 2 значения интерпретируются следующим образом. Доля дословно подтвержденных цитат – отношение числа цитат, текст которых совпадает с текстом исходного документа, к общему числу цитат, приведенных в ответах на 40 запросов; всего проверено 75 цитат. Совпадение устанавливалось автоматическим сопоставлением с текстом документа, расхождения верифицировались вручную. Показатель характеризует прослеживаемость ответа к первоисточнику, но не свидетельствует о том, что цитата отвечает на заданный вопрос и относится к действующей редакции документа. Метрики precision, recall и F1 в табл. 3 характеризуют релевантность извлеченных фрагментов, а не корректность сгенерированного ответа; корректность ответа оценивалась экспертами отдельно.

Таблица 2

**Результаты тестирования**

| Показатель | Целевое значение | Достигнутое значение |
|---|---|---|
| Доля дословно подтвержденных цитат | > 95 % | 97,3 % |
| Среднее время векторного поиска | < 200 мс | 148 мс |
| Среднее время порождения ответа (с обращением к внешнему интерфейсу модели) | < 3 с | 2,7 с |
| Субъективная оценка сокращения времени выполнения типовых задач (по опросу пользователей) | > 40 % | 43–47 % |

Среднее время векторного поиска измерялось на индексе FAISS, содержащем векторные представления всех загруженных документов, по результатам повторных измерений каждого запроса; среднее время порождения ответа включает обращение к внешнему интерфейсу модели Qwen2-72B и сетевую задержку. Сокращение трудозатрат определялось по результатам экспертного опроса пользователей при выполнении типовых задач и представляет собой субъективную оценку экономии времени: хронометраж фактического выполнения заданий не проводился. Диапазон 43–47 % отражает разброс индивидуальных оценок респондентов, а не доверительный интервал.

Для оценки качества поиска был сформирован тестовый набор из сорока эталонных запросов трех уровней сложности: десять простых, двадцать средней сложности и десять сложных. Простые запросы касались нахождения отдельных положений одного нормативного документа. Запросы средней сложности предполагали сопоставление положений двух стандартов — например, соотнесение универсальных компетенций ФГОС высшего образования по направлению «Педагогическое образование» с трудовыми функциями профессионального стандарта «Педагог профессионального обучения». Сложные запросы требовали построения матрицы соответствия между положениями трех и более документов – например, между требованиями к информационно-коммуникационной компетентности учителя в ФГОС высшего образования, в профессиональном стандарте педагога и в международной рамке UNESCO ICT CFT (Information and Communication Technology Competency Framework for Teachers – международная рамка компетенций учителей в области информационно-коммуникационных технологий).

Для каждого запроса экспертами предварительно был подготовлен эталонный набор релевантных фрагментов, с которым сопоставлялись результаты работы поискового компонента системы. При вычислении метрик использовались стандартные определения: precision – доля действительно релевантных фрагментов среди возвращенных системой, recall – доля действительно релевантных фрагментов, найденных системой, в общем числе релевантных в эталоне, F1 – гармоническое среднее precision и recall [Es24]. Эти метрики относятся к качеству извлечения и не заменяют экспертную оценку корректности итогового ответа. Так, если система возвратила тринадцать фрагментов, из которых одиннадцать являются релевантными по эталону, и в эталоне определено пятнадцать релевантных фрагментов, то precision = 11 / 13 ≈ 0,846, recall = 11 / 15 ≈ 0,733, F1 = 2 · 0,846 · 0,733 / (0,846 + 0,733) ≈ 0,785.

В качестве базовой конфигурации, относительно которой проводилось сопоставление, использована гибридная поисковая схема, сочетающая алгоритм ранжирования BM25 (Best Matching 25 – метод ранжирования документов на основе взвешенной частоты терминов с учетом длины документа) и нейросетевой кодировщик предложений all-MiniLM-L6-v2. Оценки релевантности обеих подсистем нормировались методом min–max по каждому запросу и объединялись линейно с весами 0,4 (BM25) и 0,6 (векторный поиск); веса подбирались на выборке для настройки. Кодировщик all-MiniLM-L6-v2 выбран как распространенная компактная модель, используемая в качестве базовой в прикладных RAG-системах; поскольку он обучен преимущественно на англоязычных данных, его применение к русскоязычному корпусу заведомо не оптимально, что должно учитываться при интерпретации сравнения. Ответы

в базовой конфигурации формировались той же языковой моделью Qwen2-72B и по тому же шаблону инструкции, поэтому различие конфигураций ограничено поисковым компонентом. Для сложных запросов, требующих учета связей между документами, в базовой конфигурации применялась графовая СУБД Neo4j: нормативные акты представлялись вершинами графа, а ребра классифицировались типами «пересекается», «уточняет», «противоречит»; источником связей служили те же 500 перекрестных связей, размеченных экспертами. Базовая конфигурация развертывалась на сервере с центральным процессором, 64 Гбайт оперативной памяти и графическим ускорителем NVIDIA A10 (24 Гбайт видеопамяти). Время ответа в базовой конфигурации составило 0,4 с для простого запроса и 2,3 с для сложного; эти значения включают поиск и порождение ответа, то есть соответствуют полному времени ответа и сопоставимы с суммой времени поиска и генерации предложенной системы. Показатели precision, recall и F1 для базовой конфигурации в настоящей работе не измерялись: сопоставление ограничено временем отклика и качественным анализом ошибок, поэтому выводы о преимуществе предложенной схемы по качеству поиска не формулируются, а полное сопоставительное исследование двух конфигураций составляет ближайшую задачу.

Экспериментальный корпус был сформирован на материале десяти образовательных организаций: трех вузов, четырех профессиональных образовательных организаций (колледжей) и трех центров дополнительного профессионального образования. Всего в систему было загружено 1200 документов, включая ФГОС, профессиональные стандарты, локальные нормативные акты и методические письма.

Обучение моделей в работе не выполнялось: веса multilingual-e5-large и Qwen2-72B не изменялись, поэтому вместо термина «обучающая выборка» далее используется термин «выборка для настройки параметров». Настраивались параметры разбиения текста (размер фрагмента и величина перекрытия), количество извлекаемых фрагментов top-k и шаблон инструкции языковой модели. Настройка выполнялась на обучающей части каждого цикла, после чего параметры фиксировались до завершения контрольного прогона; изменение параметров по результатам контрольных запросов не допускалось. Настройка и контроль выполнялись по схеме стратифицированной пятикратной кросс-валидации (five-fold cross-validation – процедура оценки, при которой данные циклически разбиваются на пять частей, одна из которых используется для контроля, а остальные – для настройки): 40 запросов делились на пять частей по восемь запросов с сохранением пропорций уровней сложности (два простых, четыре запроса средней сложности и два сложных); в каждом цикле восемь запросов составляли контрольную часть, а остальные 32 использовались для настройки параметров. Значения, приведенные в табл. 3, агрегированы по всем 40 запросам: каждый запрос оценивался ровно один раз – в том цикле, в котором он входил в контрольную часть; настройка на контрольных запросах не производилась, что исключает использование правильных ответов при подборе параметров.

Эталонную разметку выполняли эксперты-предметники, имеющие опыт разработки основных профессиональных образовательных программ и сопоставления компетенций ФГОС с трудовыми функциями профессиональных стандартов. Каждый объект независимо оценивали два эксперта; фрагмент считался релевантным, если он содержал положение, необходимое для обоснованного ответа на запрос, а не только совпадал по терминологии. Разногласия разрешал третий эксперт-арбитр, согласованность разметки оценивалась коэффициентом κ Коэна: $\kappa = 0{,}86$. Пятьсот размеченных перекрестных связей между положениями стандартов использовались в двух задачах: как эталон для оценки полноты на запросах средней и высокой сложности и как исходные данные графового компонента базовой конфигурации на Neo4j; в предложенной системе эти связи не использовались. Оценку удобства по шкале CSAT проводили те же эксперты, каждая оценка выставлялась по завершении цикла работы с системой. Если система возвращала фрагмент вне эталонного набора и оба эксперта признавали его

релевантным, эталон дополнялся, а метрики предложенной системы пересчитывались по единому правилу; для базовой конфигурации такой пересчет не выполнялся, поскольку показатели precision, recall и F1 для нее не измерялись.

Оценки CSAT собирались после каждого цикла: по три представителя от каждой из десяти организаций, то есть 30 участников за цикл и 150 оценок за пять циклов. Распределение 150 индивидуальных оценок по категориям шкалы CSAT необходимо указать перед публикацией: 1 балл — 0 оценок; 2 балла — 1; 3 балла — 7; 4 балла — 45; 5 баллов — 97 (всего 150 оценок; среднее 4,59 балла). Приведенные в табл. 3 групповые значения – средние по запросам соответствующего уровня сложности; итоговое значение 4,6 балла представляет собой среднее арифметическое по всем 150 индивидуальным оценкам, тогда как взвешивание групповых средних по числу запросов дает 4,5 балла. Расхождение объясняется разным числом оценок, приходящихся на запросы разных уровней сложности.

Набор из 40 запросов не позволяет оценить статистическую значимость различий между конфигурациями и переносить выводы на всю совокупность образовательных организаций. Полученные результаты рассматриваются как ограниченная апробация, а выводы – как предварительные, требующие расширения выборки запросов.

Таблица 3

**Результаты оценки качества поиска по уровням сложности запросов**

| Тип запросов | Precision | Recall | F1 | CSAT, балл |
|---|---|---|---|---|
| Простые ($n$ = 10) | 0,94 | 0,91 | 0,92 | 4,9 |
| Средней сложности ($n$ = 20) | 0,85 | 0,80 | 0,82 | 4,7 |
| Сложные ($n$ = 10) | 0,72 | 0,68 | 0,69 | 3,8 |
| Итого ($n$ = 40) | 0,84 | 0,80 | 0,81 | 4,6 |

Как видно из табл. 3, качество работы предложенной системы монотонно снижается с ростом сложности запросов. На простых запросах достигнут высокий уровень качества (F1 = 0,92, средняя оценка удовлетворенности CSAT – 4,9 балла), что соответствует практическим ожиданиям от системы семантического поиска. На запросах средней сложности F1 составляет 0,82, что является удовлетворительным результатом для практического применения. Существенное снижение наблюдается на сложных запросах, требующих агрегации информации из трех и более источников: F1 падает до 0,69, а пользовательская оценка – до 3,8 балла. Значения precision, recall и F1 рассчитаны как средние по запросам, поэтому итоговая F1-мера не совпадает с гармоническим средним итоговых precision и recall: для сложных запросов гармоническое среднее значений precision = 0,72 и recall = 0,68 составляет 0,70, тогда как среднее индивидуальных F1-мер равно 0,69.

Анализ ошибок, допущенных системой на сложных запросах, показал, что 63 % случаев снижения качества связано с игнорированием транзитивных связей между стандартами; к этой категории относились ответы, в которых пропущенный фрагмент был достижим только по цепочке ссылок между стандартами. Если стандарт A содержит ссылку на стандарт B, а стандарт B – ссылку на стандарт C, то непосредственная смысловая связь между A и C средствами чисто векторного поиска не устанавливается. Связи такого рода присутствуют в логической модели данных (рис. 1), однако в предложенной системе при поиске не обходятся. Для устранения указанного ограничения в последующих версиях системы предполагается интеграция обхода связей в графовой базе знаний и интерактивное уточнение запроса в диалоге с пользователем.

Характерные частные случаи работы системы иллюстрируют ее сильные и слабые стороны. На запросе о соответствии компетенций ФГОС высшего образования трудовым функциям профессионального стандарта система возвратила 12 фрагментов: 11 из 12 компетенций

исходного эталона и дополнительную компетенцию УК-1; компетенция ПК-6 была пропущена. До пересмотра эталона получаем TP = 11, FP = 1, FN = 1: precision = 11/12 = 0,917, recall = 11/12 = 0,917, F1 = 0,917. После экспертного признания УК-1 релевантной эталон расширен до 13 компетенций: TP = 12, FP = 0, FN = 1, поэтому precision = 12/12 = 1,000, recall = 12/13 = 0,923, F1 ≈ 0,960. Показатель recall не может быть равен 1,000, поскольку компетенция ПК-6 по-прежнему не найдена. Пересчет относится только к предложенной системе; показатели базовой конфигурации по этому примеру не приводятся, поскольку precision, recall и F1 для нее не измерялись. При формировании матрицы соответствия между тремя документами (ФГОС, профессиональный стандарт педагога и рамка UNESCO ICT CFT) система объединила базовый и продвинутый уровни международной рамки в одну категорию, тогда как эталон предполагал их раздельное представление. Эксперты отметили, что отсутствие детализации по уровням существенно снижает практическую ценность построенной матрицы. Этот случай отнесен к ошибкам агрегации, а не поиска: релевантные фрагменты были извлечены, но уровень детализации потерян при формировании ответа.

## Заключение

В работе разработана интеллектуальная система автоматизированного анализа образовательных стандартов, построенная на архитектуре RAG, и предложена предметно-ориентированная методика ее построения для русскоязычных нормативных документов сферы образования. В экспериментальной апробации на корпусе из 1200 документов десяти образовательных организаций доля дословно подтвержденных цитат составила 97,3 %, среднее время векторного поиска – 148 мс, среднее время порождения ответа – 2,7 с, субъективная оценка сокращения времени выполнения типовых задач – 43–47 %. F1-мера качества поиска составила 0,92 на простых запросах, 0,82 на запросах средней сложности и 0,69 на сложных запросах, требующих агрегации требований из трех и более документов. Показатели качества поиска приведены в табл. 3; сопоставление с базовой конфигурацией ограничено временем отклика, что отмечено в разделе «Тестирование и результаты».

Основное ограничение полученных результатов – небольшой объем проверочного набора запросов и снижение качества на междокументных запросах. Приведенные значения следует рассматривать как результат ограниченной апробации, а не как оценку применимости системы в масштабе отрасли; расширение набора запросов и проверка на корпусах других организаций являются предметом дальнейшей работы.

Применение кросс-лингвальных векторных представлений multilingual-e5-large и языковой модели Qwen2-72B позволило реализовать программное средство, функциональные возможности которого шире, чем у традиционных систем электронного документооборота: система выполняет поиск по содержанию нормативных документов и формирует ответ со ссылкой на первоисточник, тогда как СЭД работают с метаданными. Внедрение системы в деятельность образовательных организаций может сократить трудозатраты на поиск и анализ нормативных положений; количественная оценка экономической эффективности внедрения в работе не выполнялась и является предметом дальнейших исследований.

Перспективные направления развития системы включают: интеграцию графовой базы знаний для обработки транзитивных связей между нормативными документами; дообучение модели векторных представлений на корпусе образовательных стандартов для повышения различимости предметных аббревиатур; расширение механизмов интерактивного уточнения запроса; разработку специализированных метрик оценки качества ответов в задачах сопоставления нормативных документов; применение агентных RAG-архитектур, в которых планирование запросов и самооценка промежуточных результатов выполняются самой системой [Рзя26].

## СПИСОК ЛИТЕРАТУРЫ | REFERENCES


[Абр25] Абрамович Р. К., Добрынин В. Ю., Платонов А. В. Объединение глубоких моделей и разреженных представлений в информационном поиске: обзор и анализ современных подходов // Информационные и математические технологии в науке и управлении. 2025. № 2 (38). С. 5–17. HRDWSX.

Abramovich R. K., Dobrynin V. Yu., Platonov A. V. Combining deep models and sparse representations in information retrieval: a review of modern approaches // Information and Mathematical Technologies in Science and Management. 2025. No. 2 (38). Pp. 5–17. HRDWSX. (In Russian).

[Бой26] Бойков Е. С., Колева Н. С., Кортенько Л. В. A legal support system for entrepreneurs based on a Retrieval-Augmented Generation architecture // Digital models and solutions. 2026. Т. 5, № 3. С. 28. HXTTUO.

Boykov E. S., Kolyeva N. S., Kortenko L. V. A legal support system for entrepreneurs based on a Retrieval-Augmented Generation architecture // Digital models and solutions. 2026. Vol. 5, no. 3. P. 28. HXTTUO.

[Зык25] Зыкина А. В., Мунько В. В., Шикин Е. О. Цифровой сервис для подготовки аттестационных листов обучающихся // Современные информационные технологии и ИТ-образование. 2025. Т. 21, № 3. С. 585–592. ZFSSVZ.

Zykina A. V., Munko V. V., Shikin E. O. Digital service for preparing students' certification sheets // Modern Information Technologies and IT-Education. 2025. Vol. 21, no. 3. Pp. 585–592. ZFSSVZ. (In Russian).

[Каш25] Каширина И. Л., Кириллов В. В., Албычев А. С. и др. Интеграция RAG-системы для автоматизации поиска связей показателей и мероприятий национальных проектов // Моделирование, оптимизация и информационные технологии. 2025. Т. 13, № 3 (50). EIMHGB.

Kashirina I. L., Kirillov V. V., Albychev A. S., et al. Integration of a RAG system for automating the search of links between indicators and measures of national projects // Modeling, Optimization and Information Technology. 2025. Vol. 13, no. 3 (50). EIMHGB. (In Russian).

[Кит23] Китаева О. И. Интеллектуальный анализ образовательных данных учебной дисциплины с использованием программы Orange // Информационные и математические технологии в науке и управлении. 2023. № 1 (29). С. 190–200. ONREBT.

Kitaeva O. I. Intellectual analysis of educational data of an academic discipline using the Orange software // Information and Mathematical Technologies in Science and Management. 2023. No. 1 (29). Pp. 190–200. ONREBT. (In Russian).

[Кож25] Кожевников И. С. Методика оценки соответствия содержания образовательных программ требованиям рынка труда // Моделирование, оптимизация и информационные технологии. 2025. Т. 13, № 4 (51). ALMYUB.

Kozhevnikov I. S. Methodology for assessing the compliance of educational programs content with labor market requirements // Modeling, Optimization and Information Technology. 2025. Vol. 13, no. 4 (51). ALMYUB. (In Russian).

[Кош24] Кошельков В. С., Грязев Т. А., Соколов М. А., Жуков Н. Н. Разработка обобщенной модели расчета времени запроса для веб-приложений на Express и FastAPI // Современные наукоемкие технологии. 2024. № 5. С. 57–63. HRYLAQ.

Koshelkov V. S., Gryazev T. A., Sokolov M. A., Zhukov N. N. Development of a generalized query time calculation model for Express and FastAPI web applications // Modern High Technologies. 2024. No. 5. Pp. 57–63. HRYLAQ. (In Russian).

[Рзя26] Рзянкин И. С., Барышев Р. А., Гучко А. А. Агентный подход к интеллектуальному поиску в библиотечных системах // Моделирование, оптимизация и информационные технологии. 2026. Т. 14, № 2 (53). KWPJOV.

Rzyankin I. S., Baryshev R. A., Guchko A. A. Agentic approach to intelligent search in library systems // Modeling, Optimization and Information Technology. 2026. Vol. 14, no. 2 (53). KWPJOV. (In Russian).

[Хру26] Хрущев С. А. Структурная сегментация и гибридный поиск по нормативно-технической документации // Информатика. Экономика. Управление. 2026. Т. 5, № 2. С. 2062–2069. UPIJYW.

Khrushchev S. A. Structural segmentation and hybrid search over normative and technical documentation // Informatics. Economics. Management. 2026. Vol. 5, no. 2. Pp. 2062–2069. UPIJYW. (In Russian).

[Шем25] Шемонаев Д. Д., Анни П. В., Жабицкий М. Г., Конев Ю. Н. Разработка системы для анализа технических и организационных документов на соответствие нормативным требованиям атомной отрасли на основе больших языковых моделей // Int. J. of Open Information Technologies. 2025. Т. 13, № 8. С. 20–27. NHKAFR.

Shemonaev D. D., Anni P. V., Zhabitsky M. G., Konev Yu. N. Development of a system for analyzing technical and organizational documents for compliance with nuclear industry regulatory requirements based on large language models // Int. J. of Open Information Technologies. 2025. Vol. 13, no. 8. Pp. 20–27. NHKAFR. (In Russian).

[Es24] Es S., James J., et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024. Pp. 150–158. 10.18653/v1/2024.eacl-demo.16.

Es S., James J., et al. RAGAS: Automated Evaluation of Retrieval Augmented Generation // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 2024. Pp. 150–158. 10.18653/v1/2024.eacl-demo.16.

[Gao24] Gao Y., Xiong Y., Gao X., et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2024. 10.48550/arXiv.2312.10997.

Gao Y., Xiong Y., Gao X., et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2024. 10.48550/arXiv.2312.10997.

[Hu21] Hu E. J., Shen Y., et al. LoRA: Low-Rank Adaptation of Large Language Models. 2021. 10.48550/arXiv.2106.09685.

[Joh21] Johnson J., Douze M., Jégou H. Billion-Scale Similarity Search with GPUs // IEEE Trans. Big Data. 2021. Vol. 7, i. 3. Pp. 535–547. 10.1109/TBDATA.2019.2921572. SDYLCV.

[Lew20] Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. Pp. 9459–9474. 10.48550/arXiv.2005.11401.

[Rei19] Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP). 2019. Pp. 3982–3992. 10.18653/v1/D19-1410.

[Wan24] Wang L., Yang N., et al. Multilingual E5 Text Embeddings: A Technical Report. 2024. 10.48550/arXiv.2402.05672.

[Yan24] Yang A., Yang B., Hui B. et al. Qwen2 Technical Report. 2024. 10.48550/arXiv.2407.10671.

Hu E. J., Shen Y., et al. LoRA: Low-Rank Adaptation of Large Language Models. 2021. 10.48550/arXiv.2106.09685.

Johnson J., Douze M., Jégou H. Billion-Scale Similarity Search with GPUs // IEEE Trans. Big Data. 2021. Vol. 7, i. 3. Pp. 535–547. 10.1109/TBDATA.2019.2921572. SDYLCV.

Lewis P., Perez E., Piktus A. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. Pp. 9459–9474. 10.48550/arXiv.2005.11401.

Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proc. 2019 Conf. Empirical Methods in Natural Language Processing (EMNLP). 2019. Pp. 3982–3992. 10.18653/v1/D19-1410.

Wang L., Yang N., et al. Multilingual E5 Text Embeddings: A Technical Report. 2024. 10.48550/arXiv.2402.05672.

Yang A., Yang B., Hui B. et al. Qwen2 Technical Report. 2024. 10.48550/arXiv.2407.10671.

## ОБ АВТОРАХ | ABOUT THE AUTHORS

**ГАЯНОВА Майя Марсовна**
Уфимский университет науки и технологий, Россия.
maya.gayanova@yandex.ru
ORCID: 0009-0002-2513-6791.
Доцент каф. вычислительной математики и кибернетики. Дипл. математик (Башкирск. гос. ун-т). Канд. техн. наук (Уфимск. гос. авиац. техн. ун-т). Иссл. в обл. моделей и методов ИИ в сложных системах обработки информации и управления.



**GAYANOVA Maya Marsovna**
Ufa University of Science and Technology, Russia.
maya.gayanova@yandex.ru
ORCID: 0009-0002-2513-6791.
Assoc. Prof., Dept. of Computational Mathematics and Cybernetics. Degree in Mathematics (Bashkir State Univ.). Research interests: models and methods of artificial intelligence in complex information processing and control systems.

**ВУЛЬФИН Алексей Михайлович**
Уфимский университет науки и технологий, Россия.
vulfin.am@ugatu.su ORCID: 0000-0001-5857-2413.
Проф. каф. вычислительной техники и защиты информации. Дипл. инж.-программист (Уфимск. гос. нефт. техн ун-т). Д-р техн. наук по методам и системам защиты информации, информационной безопасности (Уфимск. гос. авиац. техн. ун-т). Иссл. в обл. технологий интелл. анализа данных.



**VULFIN Alexey Mikhailovich**
Ufa University of Science and Technology, Russia.
vulfin.am@ugatu.su ORCID: 0000-0001-5857-2413.
Prof., Dept. of Computer Engineering and Information Security. Degree in Software Engineering (Ufa State Petroleum Technol. Univ.). Dr. Techn. Sci. in Information Protection and Information Security (Ufa State Aviat. Techn. Univ.). Research interests: data mining technologies.

**ВАЛИЕВ Салават Марсович**
Уфимский университет науки и технологий, Россия.
salawat.valieff@yandex.ru
ORCID: 0009-0005-0490-8934.
Аспирант по науч. спец. «Управление в орг. системах». Магистр по прогр. инженерии (Уфимск. ун-т науки и технологий). Готовит дисс. в обл. интелл. анализа данных и машинного обучения.



**VALIEV Salavat Marsovich**
Ufa University of Science and Technology, Russia.
salawat.valieff@yandex.ru
ORCID: 0009-0005-0490-8934.
PhD student in Management in Organizational Systems". Master's degree in Software Engineering (Ufa Univ. of Science and Technology). Research interests: data mining and machine learning.

## МЕТАДАННЫЕ | METADATA

**Заглавие:** Интеллектуальная система обработки нормативных документов в сфере образования на базе RAG-архитектуры.

**Авторы:** Гаянова М. М., Вульфин А. М., Валиев С. М.

**Аннотация:** Современная сфера образования сталкивается с необходимостью обработки постоянно растущего объема нормативных документов. Традиционные методы ручного анализа и существующие системы электронного документооборота ориентированы преимущественно на работу с метаданными и не обеспечивают семантического анализа содержания документов. В работе представлена интеллектуальная система автоматизированного анализа образовательных

**Title:** An intelligent system for processing regulatory documents in the field of education based on the RAG architecture.

**Authors:** Gayanova M. M., Vulfin A. M., Valiev S. M.

**Abstract:** The modern field of education is faced with the need to process an ever-growing volume of regulatory documents. Traditional methods of manual analysis and existing electronic document management systems are focused primarily on working with metadata and do not provide semantic analysis of the content of documents. The paper presents an intelligent system for automated analysis of educational standards based on

стандартов, построенная на архитектуре RAG (Retrieval-Augmented Generation – генерация, дополненная извлечением данных). Описана предметно-ориентированная методика ее построения для русскоязычных нормативных документов: правила разбиения текста с сохранением привязки фрагмента к документу, редакции и структурной единице, правила формирования контекста с атрибуцией к первоисточнику и протокол оценки качества, разделяющий дословность цитаты, релевантность фрагмента и корректность ответа. Система обеспечивает загрузку документов в форматах PDF, DOCX и TXT, их предварительную обработку, векторизацию моделью multilingual-e5-large и формирование ответов с использованием языковой модели Qwen2-72B и векторного поиска на основе библиотеки FAISS; программный интерфейс построен на фреймворке FastAPI. Апробация выполнена на корпусе из 1200 документов десяти образовательных организаций и наборе из 40 запросов трех уровней сложности. Доля дословно подтвержденных цитат составила 97,3 %, среднее время векторного поиска – 148 мс, среднее время порождения ответа – 2,7 с, субъективная оценка сокращения времени выполнения типовых задач – 43–47 %. F1-мера качества поиска составила 0,92 на простых запросах, 0,82 на запросах средней сложности и 0,69 на сложных запросах, требующих агрегации требований из нескольких документов; ограничение качества на междокументных запросах связано с неучетом транзитивных связей между стандартами. Реализация системы в виде веб-сервиса представляет практическую ценность для образовательных организаций и органов управления образованием.

**Ключевые слова:** Нейросетевой анализ; образовательные стандарты; нормативные документы; RAG-архитектура; семантический поиск; обработка естественного языка; веб-сервис; векторные представления.

**Язык:** Русский.

Статья поступила в редакцию 22 июля 2026 г.

the RAG architecture (Retrieval-Augmented Generation). The paper describes a domain-oriented methodology for building such a system for Russian-language regulatory documents: rules for splitting text while keeping each fragment linked to the document, its revision and a structural unit, rules for composing the context with attribution to the primary source, and an evaluation protocol that separates the verbatim accuracy of a quotation, the relevance of a retrieved fragment and the correctness of the answer. The system uploads documents in PDF, DOCX and TXT formats, preprocesses them, vectorizes them with the multilingual-e5-large model and generates answers using the Qwen2-72B language model and FAISS-based vector search; the soft-ware interface is built on the FastAPI framework. The approbation was carried out on a corpus of 1200 documents from ten educational organizations and a set of 40 queries of three complexity levels. The share of verbatim confirmed quotations was 97.3 %, the average vector search time was 148 ms, the average answer generation time was 2.7 s, and the subjective estimate of the reduction in task completion time was 43-47 %. The F1 score was 0.92 for simple queries, 0.82 for medium-complexity queries and 0.69 for complex queries requiring aggregation of requirements from several documents; the quality limitation on cross-document queries is caused by the failure to account for transitive links between standards. The implementation of the system as a web service is of practical value for educational organizations and educational authorities.

**Key words:** Neural network analysis; educational standards; regulatory documents; RAG architecture; semantic search; natural language processing; web service; vector representations.

**Language:** Russian.

The article was received by the editors on 22 July 2026.