УДК 004.8.621

ПОДГОТОВКА ИНЖЕНЕРОВ ПО ТЕЛЕКОММУНИКАЦИЯМ В ВОЕННОЙ СФЕРЕ НА ОСНОВЕ МЕТОДА RETRIEVAL-AUGMENTED GENERATION

А. Ю. САВИЦКИЙ,
к. воен. наук, старший преподаватель кафедры связи

А. В. БАРДАШЕВИЧ,
курсант 4-го курса военного факультета

Белорусский государственный университет информатики и радиоэлектроники

В статье обоснована необходимость разработки локального и защищенного интеллектуального обучающего средства подготовки специалистов связи в военной сфере. Предложена архитектура системы на основе метода Retrieval-Augmented Generation (RAG), функционирующая в закрытой вычислительной среде и опирающаяся исключительно на доверенные источники: учебные пособия, руководства по эксплуатации стационарных (полевых) узлов связи и техническому обслуживанию. Представлены элементы архитектуры, а также критерии отбора и обработки информации. Сделан акцент на обеспечении достоверности ответов, информационной безопасности и возможности оперативного обновления знаний без переобучения модели.

Ключевые слова: эмбеддинг, чанк, семантический поиск, релевантная информация.

ВВЕДЕНИЕ

Современный этап развития Вооруженных Сил Республики Беларусь характеризуется увеличением объема и сложности информации, требующей освоения специалистами связи. Цифровизация, новые поколения радиотехнического оборудования и ухудшение условий применения средств связи в условиях радиоэлектронного противодействия требуют от инженеров по телекоммуникациям не только глубоких знаний, но и способности оперативно принимать точные решения. Традиционные формы обучения при этом не обеспечивают необходимой гибкости, что обуславливает поиск адаптивных методов подготовки.

Перспективным направлением является использование технологий искусственного интеллекта [1]. Однако прямое применение крупных языковых моделей в военном образовании ограничено рядом факторов: обучением на гражданских данных, невозможностью верификации ответов [2] и сложностью оперативного обновления знаний без переобучения. Данные обстоятельства однозначно указывают на **необходимость разработки локального, автономного и защищенного интеллектуального обучающего средства, способного функционировать в закрытой вычислительной среде и опираться исключительно на доверенные источники**. Наиболее перспективным подходом к реализации такого решения является метод RAG, адаптированный под требования военного образования и информационной безопасности. В связи с этим **целью исследования** является повышение качества подготовки инженеров по телекоммуникациям посредством разработки архитектуры локальной защищенной системы на основе метода RAG. **Объектом исследования** выступает процесс профессионального обучения в военных учебных заведениях, характеризующийся высокими требованиями к достоверности информации, а **предметом** – реализация архитектуры RAG-системы в закрытой вычислительной среде, позволяющая обновлять знания без переобучения модели.

АРХИТЕКТУРА RAG-СИСТЕМЫ

Архитектура RAG-системы представляет собой гибридную систему, объединяющую модуль семантического поиска и модуль генерации текста. В отличие от традиционных языковых моделей RAG не полагается на параметрические знания, заложенные при обучении, а динамически извлекает наиболее подходящую информацию из собственного корпуса знаний перед каждым ответом. Архитектура состоит из трех функционально взаимосвязанных элементов (рис. 1): базы знаний, сервера и пользовательского интерфейса [3].

Пользовательский интерфейс реализован как веб-приложение на основе фреймворка Streamlit, обеспечивающего простоту развертывания, кросс-платформенность и минимальные требования к клиентскому устройству. Приложение доступно через любой современный браузер и не требует установки дополнительного программного обеспечения.

Интерфейс представляет собой поле ввода текстового запроса, элементы управления (очистка, история, настройки) и область отображения сгенерированного ответа (рис. 2). Доступ к интерфейсу возможен только после входа в учетную запись. После завершения сессии диалог удаляется.

Все взаимодействия между клиентом и сервером осуществляются по протоколу HTTPS, данные передаются в формате JSON. Это обеспечивает совместимость с любыми современными браузерами и возможность последующей интеграции в LMS-платформы с соблюдением принципов информационной безопасности (локальность, контролируемость воздействия и изоляция базы знаний) [4].

Сервер. Серверная часть представляет собой центральный компонент RAG-системы, обеспечивающий обработку пользовательских запросов, взаимодействие с базой знаний и генерацию ответов (рис. 3). Реализация выполняется на языке Python с использованием фреймворка FastAPI, отличающегося высокой производительностью, автоматической генерацией документации и поддержкой асинхронных операций [5].

Сервер выполняет три ключевые функции:

1. *Преобразование запроса в эмбеддинг* – числовое векторное представление текста. Входящий текстовый запрос пользователя передается в модель эмбеддингов. Результат – числовой вектор той же размерности, что и векторы чанков в базе.

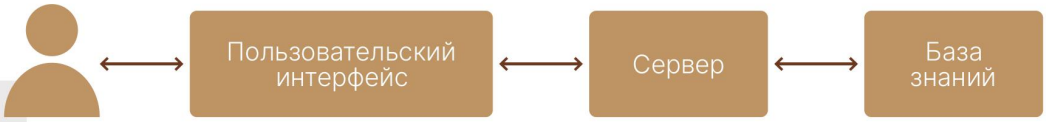

Рисунок 1. Архитектура RAG-системы

Рисунок 2. Вариант пользовательского интерфейса RAG-системы

Рисунок 3. Основные функции сервера RAG-системы

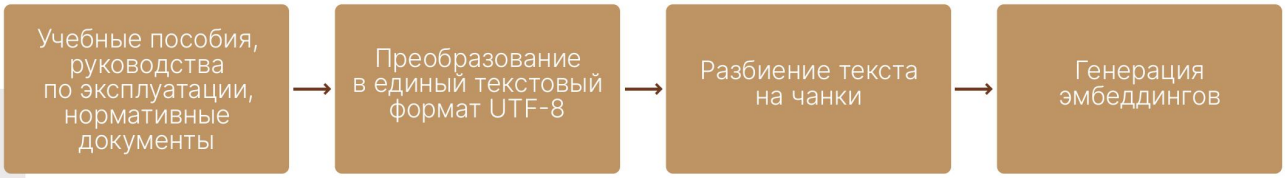

Рисунок 4. Процесс формирования базы знаний

2. *Семантический поиск.* Сформированный на первом этапе вектор используется в SQL-запросе к PosgreSQL с расширением pgvector для поиска наиболее релевантных фрагментов (чанков) из базы знаний.
3. *Генерация ответа.* Найденные фрагменты объединяются в единый контекст и встраиваются в шаблон промпта, содержащий строгую инструкцию использовать только предоставленную информацию. Промпт передается в языковую модель, и ответ возвращается клиенту в структурированном формате (JSON).

Важно подчеркнуть, что **система не способна генерировать решения, выходящие за рамки предоставленных документов. Однако подобное ограничение в условиях военного обучения является преимуществом, так как гарантирует предоставление только точной информации**.

Особое внимание уделено обеспечению информационной безопасности. Доступ к системе предоставляется только после успешной аутентификации пользователя по логину и паролю, привязанным к его учетной записи. Аутентификация реализована на основе стандарта JSON Web Token (JWT) – механизма, позволяющего безопасно передавать данные между сторонами с использованием цифровой подписи. После проверки учетных данных сервер выдает токен, подлинность, срок действия и принадлежность которого проверяются на каждом этапе взаимодействия с системой [6].

Управление доступом к функционалу осуществляется в соответствии с ролевой моделью: курсанты и прочие пользователи имеют право только на отправку запросов; преподаватели, офицеры и инженеры могут управлять содержимым базы знаний – добавлять редактировать и верифицировать источники; администраторы отвечают за техническую настройку и сопровождение системы. Такой подход исключает несанкционированный доступ и обеспечивает гибкость управления в соответствии с должностными полномочиями.

Тексты пользовательских запросов не сохраняются на сервере после завершения сессии. Для целей мониторинга и анализа фиксируются анонимизированные метрики – время формирования ответа и категория вопроса. Это полностью исключает возможность утечки конфиденциальной или значимой информации.

Серверная часть может быть развернута локально – в изолированной вычислительной сети без подключения к интернету. В этом случае модель эмбеддингов и языковая модель запускаются в Docker-контейнерах – изолированной и переносимой среде выполнения, содержащей приложение со всеми необходимыми зависимостями. Все данные при этом остаются в пределах закрытого контура. А это соответствует требованиям информационной безопасности, предъявляемым к военным информационным системам.

База знаний. Формирование базы знаний – это процесс преобразования неструктурированных или слабоструктурированных текстовых документов в единое цифровое хранилище, пригодное для семантического поиска (рис. 4). В рамках RAG-архитектуры эта база знаний служит внешним источником достоверной информации, на которую опирается языковая модель при генерации ответов, и реализуется на основе реляционной системы управления базами данных PostgreSQL с расширением pgvector, обеспечивающим поддержку векторных операций [7]. Выбор PostgreSQL обусловлен его надежностью, масштабируемостью, соответствием стандарту SQL и возможностью развертывания в закрытых сетях, что критически важно для военного применения. Процесс состоит из четырех последовательных этапов: 1) сбора источников, 2) предварительной обработки, 3) разбивки на чанки, 4) генерации эмбеддингов.

На **первом этапе** формируется набор документов, отвечающих требованиям достоверности и соответствия предметной области. Отбор источников осуществляется по следующим критериям:

1. Документ должен обладать официальным статусом, то есть быть утвержденным в установленном порядке. К таким источникам относятся учебные пособия, руководства по эксплуатации стационарных (полевых) узлов связи и другая техническая документация.
2. Содержание документа должно напрямую относиться к деятельности войск связи. В частности, рассматриваются материалы, посвященные радиорелейным и тропосферным линиям связи, абонентским сетям, подвижным и стационарным средствам связи, методам обеспечения электромагнитной совместимости, вопросам противодействия средствам радиоэлектронной борьбы, а также расчетам параметров линий связи.
3. Предпочтение отдается документам, изданным в течение последних десяти лет, что позволяет обеспечить актуальность используемой информации.
4. Документ должен быть не только описательным, но и содержать расчетные, нормативные и инструктивные материалы, необходимые для принятия инженерных решений в реальных условиях эксплуатации.
5. Источник должен быть доступен в цифровом виде – в форматах PDF, DOCX, или TXT, пригодных для автоматизированной обработки. Сканированные копии, не содержащие машиночитаемого текстового слоя (без OCR), к использованию не допускаются.

Следует отметить, что достоверность ответов системы напрямую зависит от качества исходных документов: если в руководстве по эксплуатации содержится ошибка, она может быть воспроизведена в ответе [1]. Поэтому особое значение имеет строгий отбор источников на основе указанных критериев.

На **втором этапе** собранные документы, независимо от исходного формата (PDF, DOCX, TXT), преобразуются в единый текстовый формат UTF-8. Для этого применяются специализированные библиотеки: Pdfplumber или PyPDF2 – для извлечения текста из PDF с сохранением структуры абзацев, Python-docx – для работы с документами Microsoft Word, для текстовых файлов (TXT) используется стандартное чтение через встроенные средства Python без дополнительных библиотек. Преобразование документов осуществляется программным (автоматизированным) способом, без ручного вмешательства [6].

В ходе обработки происходит удаление колонтитулов и номеров страниц, нормализация пробельных символов, декодирование специальных знаков и технических обозначений. Результатом является чистый, машиночитаемый текст, готовый к дальнейшей обработке.

На **третьем этапе** исходный текст разбивается на фрагменты – чанки, что обусловлено ограниченной способностью языковых моделей и алгоритмов поиска эффективно обрабатывать длинный текст. Оптимальный размер чанка выбирается эмпирически и обычно составляет 300–500 слов. Такой объем обеспечивает достаточную семантическую целостность (фрагмент содержит законченную мысль), совместимость с ограничениями контекстного окна моделей эмбеддингов, высокую точность поиска.

Разбиение выполняется не механически, а с учетом лингвистических и структурных признаков: границы совмещаются с концами абзацев или разделов, допускается частичное перекрытие соседних чанков (5–10%) для сохранения контекста, фрагменты, содержащие формулы, таблицы или списки, обрабатываются как единые блоки.

Четвертый этап – генерация векторных представлений при помощи предобученной модели эмбеддингов. Каждый чанк преобразуется в вектор фиксированной размерности (1536 или 484 соответственно), который сохраняется вместе с исходным текстом. Полученный набор «текст + вектор» составляет готовую базу знаний, которая в дальнейшем загружается в векторную базу данных для выполнения семантического поиска.

Важным преимуществом предложенной архитектуры является возможность оперативного обновления базы знаний: для внесения новой информации достаточно добавить соответствующие документы в исходный набор и повторно выполнить этапы предварительной обработки и генерации эмбеддингов. При этом языковая модель не требует дообучения, что значительно снижает трудозатраты на сопровождение системы.

Для повышения надежности функционирования системы и снижения вероятности предоставления неточной информации реализован комплекс механизмов контроля качества. Каждый сгенерированный ответ сопровождается указанием конкретных документов и разделов, на основе которых он был сформирован. Это позволяет пользователю самостоятельно верифицировать информацию и перейти к исходному документу для уточнения деталей.

При семантическом поиске устанавливается минимальный порог косинусного сходства между вектором запроса и векторами чанков (по умолчанию 0,75). Если ни один фрагмент не достигает установленного порога, система не включает его в контекст для генерации ответа, что

предотвращает использование слабо релевантной информации. Если в результате поиска не найдено фрагментов, превышающих порог релевантности, или общий объем извлеченного контекста недостаточен для формирования обоснованного ответа, система явно сообщает пользователю об отсутствии соответствующей информации в базе знаний.

Система допускает модульную адаптацию под конкретные воинские части, учебные заведения или под отдельные средства связи. Для каждой такой подсистемы формируется собственная специализированная база знаний, включающая только руководства и технические описания, соответствующие профилю деятельности подразделения. Благодаря различным уровням доступа обновление и верификация содержимого осуществляются офицерами, инженерами по телекоммуникациям или преподавателями, назначенными ответственными за актуальность информации в рамках своей воинской части или подразделения.

Таким образом, представленная архитектура обеспечивает полную функциональность RAG-системы: от приема запроса до выдачи обоснованного и достоверного ответа, основанного исключительно на достоверных источниках.

МЕТОДИКА ПРИМЕНЕНИЯ RAG-СИСТЕМЫ В ОБРАЗОВАТЕЛЬНОМ ПРОЦЕССЕ

Разработанная система не является автономным инструментом, заменяющим преподавателя, а **выступает в качестве интеллектуальной среды поддержки принятия решений**. Ее внедрение в учебный процесс военных учебных заведений предполагает применение на всех этапах подготовки инженеров по телекоммуникациям: от теоретического изучения дисциплин до практических занятий и самостоятельной работы. **Система занимает промежуточное положение между традиционными источниками знаний и преподавателем. Она берет на себя функцию оперативного информационного поиска и первичной структуризации данных, что высвобождает время для углубленного разбора практических задач и индивидуальной работы** [9].

Практическое применение системы реализуется через несколько взаимосвязанных форм. В рамках самоподготовки курсанты используют систему для быстрого поиска актуальных тактико-технических характеристик, нормативов развертывания узлов связи и инструкций по эксплуатации, что сокращает время работы с бумажными документами и позволяет сосредоточиться на осмыслении материала. При выполнении лабораторных и практических занятий система выступает в роли справочно-консультативного модуля: обучающийся может оперативно запросить методику расчета или допустимые значения параметров, сверяя их с полученными результатами.

Эффективность применения системы обеспечивается четким разграничением ролей участников образовательного процесса. Обучающийся формулирует запросы, анализирует полученные ответы и несет ответственность за принятое решение, развивая компетенцию работы с большими массивами технической информации. Преподаватель выступает координатором: контролирует актуальность наполнения базы знаний, анализирует статистику запросов для выявления проблемных зон в подготовке и корректирует учебный план. Техническую поддержку и обновление программного окружения осуществляет администратор учебного центра в рамках установленной ролевой модели доступа.

Использование RAG-системы направлено на формирование профессиональных компетенций по специализации «Системы и сети инфокоммуникаций» и профилизации «Системы телекоммуникаций специального назначения». В число компетенций входят:

- владение основами исследовательской деятельности, поиска, анализа и синтеза информации;
- решение стандартных задач профессиональной деятельности на основе применения ИКТ;
- определение параметров поиска и хранения мультимедийных данных, осуществление логического и физического проектирования баз данных;
- организация информационной безопасности и защиты гостайны;
- применение положения основных нормативных правовых актов Республики Беларусь в повседневной деятельности подразделений.

Оценка образовательного эффекта внедрения системы предлагается на основе комплекса показателей: 1) сокращение времени поиска необходимой технической информации при сравнении работы с бумажными носителями и через RAG-интерфейс; 2) процент запросов, по которым система предоставила исчерпывающую информацию, подтвержденную источником; 3) уменьшение количества фактических ошибок при расчете параметров связи в ходе лабораторных работ. Таким образом, методика применения RAG-системы предполагает ее использование как инструмента повышения оперативности и достоверности информационной поддержки, что способствует повышению качества подготовки инженеров по телекоммуникациям.

ЭКСПЕРИМЕНТАЛЬНАЯ АПРОБАЦИЯ СИСТЕМЫ

Для оценки работоспособности и эффективности предложенного решения развернут функциональный прототип RAG-системы на военном факультете БГУИР. Эксперимент проводился в три этапа: формирование базы знаний, настройка параметров обработки (преподаватель кафедры связи) и тестирование качества ответов группами курсантов (233701, 233702).

База знаний сформирован из 47 документов общим объемом около 1 200 страниц, включающих учебные пособия по организации связи, руководства по эксплуатации радиорелейных станций и нормативные документы. Документы отобраны в соответствии с критериями, описанными в разделе «База знаний», и предварительно обработаны для удаления артефактов форматирования.

В качестве модели эмбеддингов использовалась предобученная модель Paraphrase-Multilingual-MiniLM-L12-v2, обеспечивающая баланс между точностью семантического сопоставления и скоростью обработки технических текстов. Для генерации ответов применялась локально развернутая модель Llama-3-8B-Instruct, выбранная по критерию соотношения качества генерации и требований к вычислительным ресурсам. Параметры разбиения на чанки определены эмпирически: размер фрагмента – 400 слов, перекрытие соседних чанков – 8%. Это позволило сохранить семантическую целостность технических описаний при высокой точности поиска. При формировании ответа система извлекала 5 наиболее релевантных фрагментов, что обеспечило достаточную полноту контекста без перегрузки промта.

Тестирование проводилось на наборе из 30 типовых запросов, сформулированных преподавателем кафедры связи. Они охватили три категории: фактографические («Тактико-технические характеристики Р-414МБРП»), процедурные («Порядок развертывания узла связи в полевых условиях») и расчетные («Методика расчета зоны прямой видимости»). Для каждого запроса фиксировались время формирования ответа, полнота информации и соответствие источнику.

Результаты показали, что среднее время отклика системы составило 3,2 секунды, что приемлемо для интерактивного использования в учебном процессе. 91% ответов содержал исчерпывающую информацию, подтвержденную цитированием источника. В 7% случаев система указывала на отсутствие данных в базе знаний, тогда как при использовании языковой модели без Retrieval-компонента этот показатель превышал 67% из-за генерации неточных или обобщенных формулировок. Сравнение с традиционным полнотекстовым поиском продемонстрировало преимущество семантического подхода: RAG-система корректно обрабатывала запросы, сформулированные в свободной форме, и находила релевантные фрагменты даже без точного совпадения терминов, тогда как полнотекстовый поиск требовал точного ввода ключевых слов и не учитывал контекст.

Качественный анализ ответов выявил, что система стабильно исключает формирование недостоверных ответов: в 100% случаев информация либо подтверждалась ссылкой на документ, либо система ясно указывала на отсутствие данных в базе знаний. Это подтверждает эффективность механизма ограничения генерации строго предоставленным контекстом. При этом зафиксировано, что качество ответов напрямую зависит от полноты исходной базы: запросы по темам, слабо представленным в корпусе документов, получали менее детализированные ответы, что подчеркивает важность методического сопровождения и регулярного обновления источников.

Таким образом, экспериментальная апробация подтвердила практическую реализуемость и эффективность предложенной архитектуры: **система обеспечивает приемлемую скорость работы, высокую достоверность ответов и устойчивость к генерации непроверенной информации, что соответствует требованиям военного образования**.

ЗАКЛЮЧЕНИЕ

В статье обоснована целесообразность применения метода Retrieval-Augmented Generation как средства повышения качества подготовки инженеров по телекоммуникациям в условиях военного образования. Показано, что в отличие от универсальных крупных языковых моделей метод RAG позволяет формировать ответы исключительно на основе официально утвержденных источников, таких как учебные пособия, руководства по эксплуатации стационарных (полевых) узлов связи и нормативная документация.

Ключевыми преимуществами метода являются: обеспечение информационной безопасности за счет локального развертывания и изоляции базы знаний; повышение достоверности ответов благодаря прямой привязке к авторитетным источникам; возможность оперативного обновления базы знаний без переобучения языковой модели; адаптивность под специфику конкретных подразделений

или типов аппаратуры за счет модульного формирования базы знаний. При этом метод сознательно ограничивает генеративные способности модели: он не создает новую информацию, а лишь ретранслирует и структурирует существующую.

Таким образом, метод RAG представляет собой практически реализуемый, безопасный и гибкий подход, способный повысить качество и эффективность подготовки инженеров по телекоммуникациям.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Корчагин, П. А. Использование больших языковых моделей в образовании / П. А. Корчагин // Вестник Казанского университета. – 2023. – № 4. – С. 45–52.
2. Причины галлюцинаций в современных языковых моделях и методы их снижения с использованием семантических сетей и графов знаний / Белорусский государственный технологический университет. – URL: https://elib.belstu.by/handle/123456789/73031 (дата обращения: 15.03.2026).
3. Lewis, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis [et al.] // Advances in Neural Information Processing Systems. – 2020. – Vol. 12. – P. 9459–9474.
4. Streamlit: The fastest way to build data apps in Python. – URL: https://streamlit.io/ (дата обращения: 16.02.2026).
5. FastAPI: Modern, fast web framework for building APIs with Python. – URL: https://fastapi.tiangolo.com/ (дата обращения: 25.01.2026).
6. Таненбаум, Э. Современные операционные системы / Э. Таненбаум, Х. Бос; пер. с англ. – 4-е изд. – Санкт-Петербург: Питер, 2015. – 1120 с.: ил.
7. Pgvector: Vector similarity search for PostgreSQL. – URL: https:// github.com/pgvector/pgvector (дата обращения: 25.01.2026).
8. 7. Ротман, Д. RAG и генеративный ИИ: создание собственных RAG-пайплайнов / Д. Ротман. – М.: ДМК Пресс, 2025. – 320 с.
9. Интеграция RAG-системы для автоматизации поиска связей в документах // Вестник МОИТВиВТ. – URL: https://moitvivt.ru/journal/pdf?id=2001 (дата обращения: 15.03.2026).
10. Оболенский, Д. М. Использование метода RAG и больших языковых моделей в интеллектуальных образовательных экосистемах / Д. М. Оболенский, В. И. Шевченко // Экономика. Информатика. – 2024. – Т. 51, № 3. – С. 699–709.

TRAINING TELECOMMUNICATIONS ENGINEERS IN THE MILITARY BASED ON THE RETRIEVAL-AUGMENTED GENERATION METHOD

A. SAVITSKY, A. BARDASHEVICH

Belarusian State University of Informatics and Radioelectronics

The article substantiates the need to develop a local and secure intelligent training tool for military communications specialists. A system architecture based on the Retrieval-Augmented Generation method is proposed, operating in a closed computing environment and relying exclusively on trusted sources: educational materials, operation manuals for stationary (field) communication nodes, and technical documentation. The architectural components are presented, along with criteria for selecting and processing information. Emphasis is placed on ensuring response reliability, information security, and the capability for rapid knowledge base updates without model retraining.

Keywords: embedding, chunk, semantic search, relevant information.

Поступила в редакцию 18.02.2026.
Принята к опубликованию 19.03.2026.