ВНЕДРЕНИЕ RAG-СИСТЕМЫ В УЧЕБНЫЙ ПРОЦЕСС ПОДГОТОВКИ ИНЖЕНЕРОВ ПО ТЕЛЕКОММУНИКАЦИЯМ

Бардашевич А.В.

Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь

Федоренко В.А.

Аннотация. Рассмотрен вопрос внедрение локальной защищённой системы на основе архитектуры Retrieval-Augmented Generation в учебный процесс подготовки инженеров по телекоммуникациям. Кратко описана архитектура решения, функционирующего в изолированной среде и опирающегося на верифицированные источники информации. Обоснованы преимущества подхода для обеспечения достоверности ответов, информационной безопасности и адаптивности образовательного контента. Определены ключевые направления интеграции системы в различные этапы профессиональной подготовки специалистов связи.

Цифровизация Вооружённых Сил Республики Беларусь требует от инженеров связи способности оперативно работать с большими объёмами технической информации, что актуализирует задачу совершенствования методов их профессиональной подготовки. Традиционные методы обучения не обеспечивают необходимой гибкости, в связи с чем перспективным направлением представляется использование технологий искусственного интеллекта, в частности больших языковых моделей. Однако их прямое применение сопряжено с рисками утечки данных и генерации недостоверной информации. Решением является внедрение локальной RAG-системы, функционирующей в изолированном контуре и опирающейся исключительно на верифицированные источники. Архитектура системы построена по модульному принципу и включает три функциональных уровня [1].

Первый уровень – база знаний – реализована на основе реляционной СУБД PostgreSQL с расширением pgvector, обеспечивающим хранение и эффективный поиск векторных представлений текстовых фрагментов в закрытой вычислительной среде [2]. Выбор данной платформы обусловлен ее надежностью, масштабируемостью, поддержкой стандарта SQL и возможностью развертывания в изолированных сетях без доступа к глобальному интернету. Процесс формирования базы состоит из четырех последовательных этапов. На этапе отбора источников используются только документы официального статуса (учебные пособия, руководства по эксплуатации, нормативная документация) в машинном формате (PDF, DOCX, TXT). Второй этап предполагает предварительную обработку: конвертацию в UTF-8, удаление колонтитулов и номеров страниц, нормализацию пробельных символов и декодирование специальных знаков. Третий этап – разбиение текста на семантические целостные чанки (текстовый фрагмент) объемом 300 – 500 слов. Разбиение производится с учетом лингвистической структуры документа (границы абзацев и разделов), допускается частичное перекрытие соседних фрагментов (5-10%) для сохранения контекста, а формулы и таблицы обрабатываются как отдельные блоки. Завершающий этап включает генерацию эмбеддингов (числовой вектор, фиксированной размерности) с использованием предобученной мультиязычной модели, преобразующей каждый чанк в эмбеддинги. Такой подход позволяет обновлять корпус знаний путём добавления новых документов без необходимости переобучения языковой модели.

Второй уровень – серверная часть – реализован на языке Python с использованием асинхронного фреймворка FastAPI. Выбор данной технологии обусловлен высокой производительностью, автоматической генерацией документации и поддержки неблокирующих операций, что критично для обработки множественных запросов в реальном времени.

Сервер выступает центральным элементом системы и выполняет три ключевые функции. Первая функция – преобразование пользовательских запросов в эмбеддинг. Входящий текстовый запрос нормализируется и передается в локально развернутую модель эмбеддигов, которая преобразует текст в числовой вектор фиксированной размерности, идентичной векторам в базе знаний. Вторая функция – семантический поиск релевантных фрагментов. Сформированный вектор используется в SQL-запросе к СУБД PostgreSQL с расширением pgvector. Поиск осуществляется путём вычисления метрики косинусного сходства между вектором запроса и векторами чанков в базе. Система извлекает топ-5 наиболее релевантных фрагментов, обеспечивая баланс между полнотой контекста и нагрузкой на модель [3]. Третья функция – генерация ответа языковой моделью. Найденные фрагменты объединяются в единый контекст и встраиваются в системный промт, содержащий строгую инструкцию использовать только предоставленную информацию. Промт передается в языковую модель и ответ возвращается клиенту в структурированном формате JSON. Важной особенностью является механизм контроля релевантности, предотвращающий недостоверные ответы. При семантическом поиске устанавливается минимальный порог косинусного сходства (по умолчанию 0,75). Если один фрагмент не достигает установленного порога, система явно сообщает пользователю об отсутствии информации в базе знаний, вместо генерации ответов.

Третий уровень – пользовательский интерфейс – реализован как веб–приложение на фреймворке Srteamlit. Выбор данной технологии обусловлен простотой развертывания, кроссплатформенностью и минимальными требованиями к клиентскому устройству. Приложение доступно через любой современный браузер и не требует установки дополнительного программного обеспечения на рабочих местах курсантов и преподавателей. Структура интерфейса включает поле ввода текстового запроса, панель элементов управления и область отображения сгенерированного ответа. Важной функциональной особенностью является обязательное цитирование источников: каждый ответ сопровождается метаданными об используемых документах и конкретных разделах, что позволяет оперативно перейти к исходному тексту для самостоятельной верификации информации. Доступ к системе строго регламентирован: вход возможен только после успешной аутентификации по логину и паролю, привязанных к учётной записи. В соответствии с ролевой моделью, функционал интерфейса адаптируется под права пользователя. После окончания сессии история диалога автоматически удаляется, что исключает накопление конфиденциальной информации на клиентской стороне и обеспечивает конфиденциальность работы.

Безопасность системы обеспечивается комплексом взаимосвязанных мер. Локальное развертывание в Docker-контейнерах гарантирует изоляцию всех компонентов системы и исключает доступ к глобальной сети, что соответствует требованиям к военным информационным системам. Аутентификация на основе JWT-токенов обеспечивает безопасную передачу учетных данных и контроль сессий без хранения паролей на клиентской стороне. Ролевая модель доступа строго регламентирует функциональные возможности (курсанты –только запросы, преподаватели – управления базой знаний, администраторы – техническое сопровождение). Отсутствие сохранения текстов запросов после завершения сессии, что исключает накопление конфиденциальной информации. Фиксируются только анонимизированные метрики для мониторинга производительности.

В учебном процессе система интегрируется на всех этапах подготовки: от теоретического изучения дисциплин до практических занятий и самостоятельной работы. Система берет на себя функции оперативного информационного поиска и первичной структуризации данных, освобождая время для углубленного разбора практических задач и индивидуальной работы. В рамках самоподготовки курсанты используют систему для оперативного поиска тактико-технических характеристик и нормативных документов, что сокращает время работы с бумажными носителями. При выполнении лабораторных работ система выступает как справочно-консультативный модуль: обучающийся может оперативно запросить методику расчёта или допустимые значения параметров, сверяя их с полученными результатами. Обучающийся формулирует запросы, анализирует ответы с указанием источников и несет ответственность за принятое решение.

Эффективность применения обеспечивается чётким разграничением ролей. Преподаватель контролирует актуальность базы знаний и анализирует статистику запросов для корректировки учебного плана [4]. Техническую поддержку и обновление программного окружения осуществляет администратор учебного заведения. Такой подход формирует компетенции по работе с технической документацией, анализу информации и обеспечению информационной безопасности.

Использование RAG-системы направлено на формирование профессиональных компетенций по специализации «Системы и сети инфокоммуникаций» и профилизации «Системы телекоммуникаций специального назначения», включающих владение основами исследовательской деятельности, осуществлять поиск анализ и синтез информации; решение стандартных задач профессиональной деятельности на основе применения информационно-коммуникационных технологий; определение параметров поиска и хранения мультимедийных данных, осуществление логического и физического проектирования баз данных; организацию информационной безопасности и защиту государственной тайны; применение положения основных нормативных правовых актов Республики Беларусь в повседневной деятельности подразделений.

Внедрение RAG-архитектуры повышает качество подготовки инженеров за счёт гарантированной достоверности информации, сокращения времени поиска данных и возможности адаптации под конкретные типы аппаратуры связи. Система не заменяет преподавателя, а усиливает образовательный процесс, беря на себя функции оперативного информационного поиска и первичной структуризации данных. Дальнейшее развитие предполагает модульную адаптацию под профили воинских частей и интеграцию с существующими LMS-платформами.

Список использованных источников:

1. 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.

2. Pgvector: Vector similarity search for PostgreSQL. – URL: https://github.com/pgvector/pgvector (дата обращения: 25.01.2026).4. Кузнецов А.А., Левкович М.В. Кибербезопасность в условиях цифровизации: проблемы и решения. – Минск: НЦПИ РБ, 2022. – 154 с.

3. FastAPI: Modern, fast web framework for building APIs with Python. – URL: https://fastapi.tiangolo.com/ (дата обращения: 25.01.2026).

4. Корчагин, П. А. Использование больших языковых моделей в образовании / П. А. Корчагин // Вестник Казанского университета. – 2023. – № 4. – С. 45–52