**УДК 004.8:61**

# ПРОЕКТИРОВАНИЕ ИНТЕГРАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В АРХИТЕКТУРУ ПРОТОТИПА МЕДИЦИНСКОГО ИИ-АССИСТЕНТА

**Самойленко Е. И.**[1] (студент)
**Научный руководитель – инженер ФПИН Лаврова А. К.**[1]
[1]Университет ИТМО
evasamoilenko@yandex.ru

## Введение

В условиях цифровой трансформации здравоохранения объём медицинских данных стремительно растёт, однако их практическая ценность снижается из-за высокой разрозненности. История здоровья пациента формируется из бумажных документов, PDF-файлов и данных различных информационных систем, что не позволяет врачу оперативно получить целостную картину состояния пациента. Фрагментация информации приводит к дублированию анализов, увеличению времени диагностики и снижению качества медицинских решений. В масштабах системы здравоохранения это ограничивает использование накопленных данных.

Существующие медицинские ИИ-ассистенты, как правило, включают модули извлечения данных, аналитики и пользовательского взаимодействия, однако данные компоненты часто функционируют несогласованно. Особую сложность представляет интеграция больших языковых моделей (LLM): при наивном подключении они используют избыточный контекст, дублируют вычисления и демонстрируют высокую ресурсоёмкость, а также подвержены риску генерации недостоверной информации [1].

Отечественные исследования подчёркивают потенциал LLM для персонализации коммуникации в системе «врач–пациент» и повышения доступности медицинской информации [2]. Вместе с тем авторы указывают на необходимость ограничения контекста и использования структурированных данных для повышения достоверности ответов [3]. Зарубежные работы также рассматривают гибридные архитектуры, в которых языковая модель отделена от прикладной логики и взаимодействует с системой через контролируемые механизмы извлечения данных [4]. Анализ существующих решений показывает, что, несмотря на востребованность LLM в медицине, проблема согласованной архитектурной интеграции остаётся нерешённой.

Таким образом, формируется двойная научно-практическая проблема: необходимость интеллектуального структурирования медицинских данных и создание архитектурного механизма безопасной и ресурсно-эффективной интеграции LLM.

## Основная часть

Предлагаемое решение основано на разработке многоуровневой модульной архитектуры медицинского ИИ-ассистента, в которой большая языковая модель выполняет строго ограниченную роль интеллектуального интерфейса. Ключевой принцип архитектуры — разделение ответственности между компонентами системы и исключение прямого доступа LLM к медицинским данным.

В центре системы находится серверное ядро, выполняющее функции оркестратора: маршрутизацию запросов, управление пользовательскими сессиями, контроль доступа и предоставление унифицированных API. Все взаимодействия между модулями осуществляются исключительно через данный слой, что обеспечивает слабую связанность и масштабируемость.

Обработка медицинских документов реализуется посредством модуля OCR/NER, который распознаёт текст и извлекает медицинские сущности: показатели, единицы измерения, даты и референсные значения. Полученные данные преобразуются в

унифицированный формат и сохраняются в базе данных. Аналитико-прогностический модуль работает только со структурированной информацией, анализирует динамику показателей и формирует прогнозы изменения состояния здоровья.

Интеграционный слой LLM функционирует как интеллектуальный интерфейс между пользователем и системой. Перед генерацией ответа специальный retriever-компонент извлекает из базы данных исключительно релевантные сведения, формируя ограниченный контекст. Таким образом реализуется модифицированный RAG-подход, при котором источником информации выступают только внутренние валидированные данные пациента, а не внешние документы или обобщённые знания модели. Языковая модель преобразует структурированные факты в связный текст, адаптированный к уровню пользователя, не выполняя самостоятельного поиска или логического вывода.

Предложенная архитектура является экономичной и технологически гибкой: языковая модель может быть заменена без изменения логики системы, а сервисная организация модулей позволяет масштабировать отдельные компоненты. Использование унифицированных форматов данных и централизованного API соответствует современным подходам к проектированию интеллектуальных систем и обеспечивает контролируемость генерации.

## Выводы

Основным результатом является разработка структурной схемы, в которой LLM выполняет функцию интерпретации, тогда как поиск, анализ и прогнозирование осуществляются специализированными модулями. Такое разделение ответственности снижает риск генерации недостоверной информации и повышает интерпретируемость системы.

Практическое применение результатов возможно при разработке прототипа медицинского ассистента для обработки пользовательских документов и формирования персонализированных объяснений медицинских показателей. Внедрение системы предполагает поэтапное тестирование: модульную проверку отдельных компонентов, интеграционные испытания полного пайплайна обработки данных и сценарное тестирование пользовательских запросов.

Предложенный подход может быть использован при создании цифровых медицинских сервисов, ориентированных на повышение качества диагностики, непрерывность наблюдения пациента и рациональное использование медицинских данных.

## Литература

1. Chow, J.C.L.; Li, K. Large Language Models in Medical Chatbots: Opportunities, Challenges, and the Need to Address AI Risks. Information 2025, 16, 549. https://doi.org/10.3390/info16070549.
2. Костров С.А., Потапов М.П., Аккуратов Е.Г. Персонализация коммуникации с пациентом: большие языковые модели. Пациентоориентированная медицина и фармация. 2025;3(2):68-79. https://doi.org/10.37489/2949-1924-0093.
3. Назаров Д. М., Бадаев Ф. И. Применение больших языковых моделей в сфере здравоохранения. Менеджер здравоохранения. 2025; 5:142–154. DOI: 10.21045/1811-0185-2025-5-142-154
4. Brati´c, D.; Šapina, M.; Jureˇci´c, D.; Žiljak Grši´c, J. Centralized Database Access: Transformer Framework and LLM/Chatbot Integration-Based Hybrid Model. Appl. Syst. Innov. 2024,7,17. https://doi.org/10.3390/asi7010017.