

Doi: 10.52341/20738080_2026_143_4_67

КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ ПРИМЕНЕНИЯ ЛОКАЛЬНОГО ИИ-АССИСТЕНТА НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ПОДГОТОВКИ ОТВЕТОВ НА ОБРАЩЕНИЯ ГРАЖДАН В МЕДИЦИНСКИЕ ОРГАНИЗАЦИИ



ИГЛОВИКОВ Н.Ю.,

к.м.н., доцент, доцент кафедры урологии ФГБВОУ ВО «Военно-медицинская академия имени С.М.Кирова» Минобороны России, подполковник мед. службы, iglovikov@yandex.ru



МАРКЕВИЧ В.Ю.,

д.м.н., доцент, заместитель начальника ФГБВОУ ВО «Военно-медицинская академия имени С.М.Кирова» Минобороны России – начальник Лечебно-диагностического центра (клинического, многопрофильного, высоких технологий) Минобороны России, полковник мед. службы, vitaly-med@mail.ru



ГРЕБНЕВ Д.Г.,

к.м.н., преподаватель кафедры военно-полевой хирургии ФГБВОУ ВО «Военно-медицинская академия имени С.М.Кирова» Минобороны России, подполковник мед. службы, grebnev83@yandex.ru



БАГНЕНКО А.С.,

к.м.н., доцент, заместитель начальника кафедры ЧЛХ и хирургической стоматологии ФГБВОУ ВО «Военно-медицинская академия имени С.М.Кирова» Минобороны России, ведущий научный сотрудник ФГБУ НМИЦ «Центральный научно-исследовательский институт стоматологии и челюстно-лицевой хирургии» Минздрава России, лауреат премии Правительства Российской Федерации в области медицинской науки, подполковник мед. службы, bagnenkoa.s.mfs@mail.ru



ШАКИРЗЯНОВ Р.С.,

начальник смены отдела технического обеспечения АСУ (Центра информационных технологий) ФГБВОУ ВО «Военно-медицинская академия имени С.М.Кирова» Минобороны России, xxstrim@mail.ru

Работа посвящена разработке концептуальной модели применения ассистента на основе больших языковых моделей для подготовки ответов на обращения граждан в медицинской организации. Ключевой особенностью предложенного подхода является развертывание интеллектуальной системы на локальном сервере медицинской организации без передачи данных во внешние сервисы. Это позволяет учитывать требования к защите персональных данных и контролю информационных потоков. В статье обсуждены также потенциальные организационные эффекты, ограничения и риски внедрения данной модели в административную практику медицинских организаций.

Ключевые слова: искусственный интеллект, большие языковые модели, локальный сервер, обращения граждан, медицинская организация, защита персональных данных, документооборот.

CONCEPTUAL MODEL OF THE USE OF A LOCAL AI ASSISTANT BASED ON LARGE LANGUAGE MODELS TO PREPARE RESPONSES TO THE PUBLIC APPEALS TO MEDICAL INSTITUTIONS

Iglovikov N., Markevich V., Grebnev D., Bagnenko A., Shakirzyanov R.

The paper is dedicated to the development of the conceptual model of the use of an assistant based on large language models to prepare responses to the public appeals to medical institutions. The key feature of the proposed approach is the intellectual system installing on a medical institution's local server without transferring data to external services. This ensures compliance with the requirements of personal data protection and control of information flows. The article also discusses potential organizational effects, restrictions and risks of implementing this model in administrative practice of the medical institutions.

Key words: artificial intelligence, large language models, local server, public appeals, medical institution, personal data protection, document management.

Введение

Рост числа обращений граждан в медицинские организации (МО), жёсткие нормативные сроки подготовки ответов и ограниченные кадровые ресурсы делают обработку этих обращений всё более сложной задачей. Расширение электронных каналов взаимодействия – порталов государственных услуг, электронной почты и региональных информационных систем – увеличило объём поступающих сообщений, но не решило проблему их оперативной обработки [1, 2].

В этих условиях большие языковые модели (large language model – LLM) рассматриваются как перспективный инструмент автоматизации подготовки документов. Современные открытые модели способны генерировать связные тексты, унифицировать стиль деловой переписки и структурировать входящую информацию [3, 4]. Однако их использование в МО ограничено требованиями российского законодательства, запрещающими передачу персональных данных на внешние серверы без специального правового основания. Поскольку обращения граждан содержат персональные и нередко медицинские данные, применение облачных ИИ-сервисов создаёт существенные правовые риски [5]. Следовательно, наиболее рациональным представляется использование языковой модели, функционирующей локально в защищённом информационном контуре МО без передачи данных во внешние сети.

Несмотря на практическую реализуемость такого подхода, в отечественной литературе по организации здравоохранения он практически не освещён: существующие публикации сосредоточены преимущественно на клиническом применении искусственного интеллекта (ИИ) или в чат-ботах для пациентов [5, 6].

Цель исследования

Разработать концептуальную модель применения локального ИИ-ассистента на основе LLM для поддержки подготовки ответов на обращения граждан в МО.

Материалы и методы

Работа носит концептуально-аналитический характер. Эмпирические данные в узком смысле – измеренные показатели конкретного внедрения – в ней не представлены: задача состояла не в оценке уже

реализованной системы, а в формализации принципов, на которых такая система должна строиться.

Процессный анализ (этап 1).

На 1-м этапе рассмотрена структура административного процесса обработки обращений граждан как последовательности управленческих операций. Цель – выделить те стадии, где основная трудоёмкость приходится не на принятие решения, а на текстовую работу с документом.

Анализ нормативно-правовой базы (этап 2).

На 2-м этапе проанализированы требования федеральных законов № 59-ФЗ («О порядке рассмотрения обращений граждан Российской Федерации» от 02.05.2006 [последняя редакция], далее – 59-ФЗ) и № 152-ФЗ («О персональных данных» от 27.07.2006 [последняя редакция], далее – 152-ФЗ) применительно к вопросу о том, какая архитектура информационных систем допустима при обработке данных из обращений граждан в МО.

Концептуальное моделирование (этап 3).

На 3-м этапе, на основе синтеза результатов предыдущих этапов и анализа научной литературы, сформулированы принципы модели, её функциональные роли и сценарии взаимодействия специалиста с системой.

Аналитическую базу составили нормативные правовые акты, научные публикации по цифровизации здравоохранения и государственного управления, а также смежные отечественные работы в области применения LLM в медицине.

Результаты и обсуждение

Анализ процесса показывает, что является достаточным проследить путь одного обращения. Регистрация, определение ответственного исполнителя, анализ содержания, формирование текста ответа, согласование, оформление и отправка – на каждом из этих шагов задействован живой сотрудник, и каждый шаг нормативно ограничен по времени (59-ФЗ устанавливает 30-дневный срок, а ряд категорий обращений – срок более короткий).

Теперь добавим к этому обстоятельство, которое на практике наблюдается часто, но редко попадает в научные публикации: один гражданин вправе направить одно и то же обращение одновременно в несколько инстанций – непосредственно в учреждение, в орган управления здравоохранением, через депутатский запрос, через надзорный орган. Каждый из этих параллельных каналов порождает самостоятельный за-

прос, требующий самостоятельного ответа в установленные сроки. Для крупной МО в периоды повышенной конфликтности это означает не двукратное, а многократное увеличение нагрузки на тех же исполнителей.

При этом содержательная сложность большинства обращений невелика: значительная их доля представляет собой запросы разъяснительного характера, жалобы на организацию приёма или просьбы о предоставлении документов. Ответы на них подчиняются устоявшимся шаблонам, и именно здесь возникает пространство для технологической поддержки.

Когда речь заходит о применении ИИ-инструментов в текстовой работе, наиболее очевидным решением являются доступные облачные сервисы. Они удобны, не требуют инфраструктурных затрат и обеспечивают высокое качество генерации. Однако в контексте обращений граждан их применение создаёт неустранимую правовую проблему.

Обращение гражданина в МО – это документ, который почти всегда содержит персональные данные субъекта (фамилия, дата рождения, контактные сведения), а нередко и сведения о состоянии его здоровья. Их передача на серверы стороннего провайдера – в том числе иностранного облачного сервиса – квалифицируется как передача персональных данных третьим лицам и требует либо согласия субъекта, либо иного законного основания (152-ФЗ, ст. 6). Обеспечить это в условиях рутинного делопроизводства практически невозможно, а правовые последствия нарушения несопоставимы с административным удобством.

Дополнительным существенным ограничением при использовании облачных решений является требование локализации персональных данных, закреплённое в ст. 18.1 152-ФЗ. В соответствии с указанной нормой при сборе персональных данных граждан Российской Федерации оператор обязан обеспечивать их запись, систематизацию, накопление, хранение и извлечение с использованием баз данных, находящихся на территории России. Использование облачных сервисов, инфраструктура которых может быть размещена за пределами Российской Федерации или не может обеспечивать прозрачного контроля за географией хранения данных, затрудняет соблюдение данного требования и повышает правовые риски для МО. В этой связи локальное развёртывание ИИ-систем в пределах информационного контура учреждения представляется более соответствующим действующему нормативному регулированию.

Наиболее правомерным и управленчески безопасным вариантом в рассматриваемом контексте является локальное развёртывание: языковая модель и все компоненты системы размещаются на серверном оборудовании самой МО и функционируют без выхода в сеть Интернет. Тексты обращений не покидают информационного контура учреждения.

Локальное развёртывание само по себе не является гарантией информационной безопасности. Она обеспечивается комплексом организационных и технических мер – разграничением доступа, журналированием операций, шифрованием хранилища. Локальная архитектура создаёт необходимую основу для применения этих мер, а не заменяет их [7].

Среди практикующих специалистов распространено убеждение, что локальный запуск языковых моделей требует специализированного суперкомпьютерного оборудования. Это не так. Современные открытые модели серии Llama, Mistral и их производные распространяются свободно и в квантизованном формате запускаются на стандартном серверном оборудовании с достаточным объёмом оперативной памяти без дорогостоящих GPU-ускорителей. Для задач административной переписки, не требующих мультимодальных возможностей, этого вполне достаточно.

Организационные и инфраструктурные аспекты локального развёртывания (выбор программного стека, требования к оборудованию, этапы внедрения и опыт реализации проекта в Военно-медицинской академии имени С.М.Кирова Минобороны России) будут подробно рассмотрены в отдельной публикации.

Концептуальная модель

Предлагаемая модель строится на нескольких принципах, ни один из которых не является факультативным.

Первый – безусловное сохранение ответственности специалиста: ИИ-ассистент формирует проект документа, но ни в коем случае не направляет ответ самостоятельно.

Второй – локальное развёртывание без передачи данных во внешние сети, что, как показано выше, является не опцией, а нормативным требованием.

Третий – встраивание системы в существующий регламент документооборота без его пересмотра: организационная ломка при внедрении – самостоятельный риск, заслуживающий отдельного управления.

В рамках модели ИИ-ассистент выполняет четыре функциональные роли:

- извлечение и структурирование информации: система анализирует текст обращения и выделяет тему, участников, характер запроса, временные параметры; результат – структурированное представление, упрощающее маршрутизацию;

- рекомендация исполнителя: ассистент предлагает подразделение или специалиста; окончательное решение остаётся за ответственным сотрудником;

- формирование проекта ответа с использованием типовых формулировок деловой переписки;

- документальное оформление: унификация стиля, реквизиты, структура.

Как это работает на практике?

Первый сценарий (поддержка подготовки ответа) ориентирован на типовые информационные запросы. Система анализирует обращение, формирует проект, специалист проверяет и при необходимости редактирует. *Второй сценарий* (редакторская поддержка) – для случаев, когда содержательную часть ответа должен сформулировать именно специалист: юридически чувствительные ситуации, жалобы с медицинской составляющей, конфликтные обращения. Здесь ассистент подключается после того, как специалист написал суть ответа, и берёт на себя стилистическую унификацию и документальное оформление.

Принципиально, что оба сценария предполагают участие специалиста на финальной стадии. Это не техническое ограничение системы, это – сознательный архитектурный принцип. Должностное лицо, подписывающее документ, несёт за него ответственность, и никакой ИИ эту ответственность принять не может.

Ожидаемые эффекты и ограничения

Реалистичная оценка ожидаемых эффектов требует честности в отношении того, что модель делает, а что – нет. Она снижает трудоёмкость рутинных текстовых операций при работе с типовыми запросами и стандартизирует документное оформление. Она не решает проблему кадрового дефицита, не ускоряет принятие сложных управленческих решений и не устраняет содержательных ошибок, если специалист не понимает сути обращения [3, 5].

Основной технологический риск (генерация неточных или юридически некорректных формулировок) делает обязательный экспертный контроль не рекомендацией, а условием применения системы [6]. Правовые риски, связанные с обработкой персональных данных, требуют не только локальной архитектуры, но и соответствующих

организационных мер – политики доступа, журналирования, регламента работы с системой.

Наконец, организационный риск, который на практике недооценивается: сотрудники нередко воспринимают внедрение ИИ-инструментов как угрозу занятости или попытку снять с них ответственность. Преодоление этого сопротивления – самостоятельная управленческая задача [8].

Следует подчеркнуть, что настоящая работа носит концептуальный характер. Количественная оценка эффектов возможна только в рамках пилотного внедрения с заранее зафиксированными базовыми показателями. Это – следующий необходимый шаг.

Выводы

1. Рост числа обращений граждан через электронные каналы в сочетании с жёсткими нормативными сроками ответа формирует устойчивую административную нагрузку на МО. Значительная её часть приходится на рутинные текстовые операции, которые принципиально поддаются технологической поддержке.

2. Применение облачных LLM-сервисов при работе с обращениями граждан в МО противоречит требованиям 152-ФЗ, поскольку обращения содержат персональные данные, передача которых третьим лицам требует специального правового основания. Локальное развёртывание языковой модели – это не технический выбор, а нормативно обусловленный архитектурный принцип.

3. Предложенная концептуальная модель включает четыре функциональные роли ИИ-ассистента и два сценария его взаимодействия со специалистом. В обоих сценариях ответственность за итоговый документ сохраняется за подписывающим его должностным лицом. Это – принципиальное условие, а не рекомендация.

4. Обязательный экспертный контроль результатов работы ассистента является единственным надёжным механизмом управления риском генерации неточных или юридически некорректных формулировок. Автономное направление ответов на обращения без участия специалиста данной моделью не предусмотрено и не рекомендуется.

Литература

