Повышение эффективности конфигурации RAG для работы больших языковых моделей с клиническими рекомендациями на примере аллергического ринита

К.Ю. Мокшин, Е.В. Боброва, М.Г. Жабицкий

***Аннотация* — В статье рассматриваются подходы к повышению эффективности больших языковых моделей (Large Language Models, далее – LLM) при решении специализированных медицинских задач на основе клинических рекомендаций Министерства здравоохранения Российской Федерации. Основное внимание уделено сравнению различных вариаций архитектуры Retrieval Augmented Generation (далее – RAG), применяемой для снижения фактологических ошибок и повышения релевантности ответов моделей. В качестве примера предметной области использованы клинические рекомендации по аллергическому риниту, а в качестве языковых моделей — российские модели семейства GigaChat. В рамках экспериментального исследования было реализовано шестнадцать конфигураций взаимодействия с моделями, отличающихся мощностью LLM, форматом исходных документов: PDF и предварительно адаптированный Markdown, а также стратегиями поиска по базе знаний: по ключевым словам, векторный и гибридный поиск. Оценка качества ответов проводилась путём сравнения с эталонными ответами на основе клинических рекомендаций с использованием метрик BLEU и METEOR. Результаты показали, что наилучшие показатели достигаются при использовании более мощной модели в сочетании с RAG на основе векторного или гибридного поиска и базы знаний, сформированной из структурированного машиночитаемого текста. При этом адаптация клинических рекомендаций из оригинального формата в адаптированный вид Markdown оказывает положительное влияние на результаты работы RAG. Полученные выводы подтверждают целесообразность применения RAG в медицинских информационных системах и подчёркивают значимость предварительной подготовки нормативных документов и их перевод в машиночитаемый формат для повышения качества работы больших языковых моделей.**

***Ключевые слова* — большие языковые модели, LLM, клинические рекомендации, RAG, промпт-инжиниринг, медицинский искусственный интеллект, нейронные сети, медицинская информатика**

Статья получена 25 февраля 2026.

Мокшин Кирилл Юрьевич, НИЯУ МИФИ, ассистент ВИШ НИЯУ МИФИ, KYMokshin@mephi.ru

Боброва Елизавета Витальевна, НИЯУ МИФИ, преподаватель ВИШ НИЯУ МИФИ, EVBobrova@mephi.ru

Жабицкий Михаил Георгиевич, НИЯУ МИФИ, заместитель директора ВИШ НИЯУ МИФИ, jabitsky@mail.ru

I. Введение

Большие языковые модели (далее – LLM) – это системы искусственного интеллекта, которые позволяют обрабатывать текст, сформулированный человеком, и генерировать его с помощью процедуры определения вероятности возникновения следующей лексической единицы в последовательности [1]. Таким образом, LLM может генерировать правдоподобный текст на языке, который будет понятен человеку. Языковая модель является "большой", если она обучалась на обширных массивах данных: статьях, документах, книгах, веб-ресурсах и так далее. Обучение делится на несколько этапов [2], первым из которых является предварительным или pre-train – когда в систему загружают множество данных. LLM за счёт изменения и оптимизации своих параметров при обучении на большом количестве материалов способны генерировать «осмысленный» текст не благодаря осознанию обучающей выборки в человеческом понимании, а благодаря тому, что алгоритм получает знания о том, как комбинируются лексические единицы, с какой частотой они могут использоваться в тех или иных контекстах. Вторым этапом является дообучение или fine-tuning – метод адаптации модели отвечать на конкретные вопросы, что реализуется с помощью обучения с учителем на специально подготовленных экспертами инструкциях. Третий этап – alignment или выравнивание. Ответы основной модели ранжируются человеком и используются для другой модели вознаграждения, которая, в свою очередь, используется в качестве источника сигнала о корректном или ошибочном ответе при обучении с подкреплением. Таким образом, результаты работы основной LLM проверяются и модель с каждой итерацией учится их улучшать. Обученные LLM могут генерировать текст, отвечать на вопросы и выполнять различные задачи обработки естественного языка, имитируя человеческий стиль общения.

Одной из ключевых особенностей текущего этапа развития медицинской сферы как в Российской Федерации, так и во всём мире, является широкое применение информационных технологий. В качестве примера могут выступать клинические системы,

основанные на LLM для помощи в диагностике [1], применение искусственного интеллекта для анализа и интерпретации медицинских изображений [3], роботизированные хирургические комплексы [4], применение IoT носимых устройств для сбора показателей состояния здоровья пациента для их дальнейшего внесения в медицинскую карту и принятия решений по лечению на основе анамнеза [5]. LLM позволяют поддерживать решение множества задач в клинической практике: доступное изложение назначений и другой медицинской документации для упрощения коммуникации между врачом и пациентом; автоматизация составления врачебной отчётности через консолидацию неструктурированных письменных и устных записей; быстрое справочное обеспечение по клиническим рекомендациям и другим документам; поддержка учёных в области медицины в написании кода для анализа и визуализации данных в научных исследованиях; помощь в обучении медицинских специалистов в плане разъяснения учебных материалов и симуляции врачебной практики [6]. Существуют LLM-решения со специализацией для работы с вопросами из медицинской сферы, одним из которых является чат-бот Med-PaLM 2 от Google [7] или раздел Health в ChatGPT от OpenAI для общения с пользователем на темы, связанные с медициной, с учётом анализа персональных данных о состоянии здоровья [8].

Однако несмотря на все преимущества и примеры использования LLM в медицинской практике, существует ряд проблем у подобных решений. Одними из них являются вопросы правового регулирования использования LLM в клинической практике [9]. Результаты работы LLM-систем влияют на лечение и здоровье пациентов, от чего зависит состояние общественного здравоохранения. Исследования юридических аспектов внедрения таких решений в медицинскую практику затрагивают регулирование безопасности ответов LLM, соблюдение этических стандартов, обеспечение конфиденциальности данных пациента [10]. Кроме этого, использование решений на основе LLM для формулировки диагностических гипотез может быть неэффективным, поскольку в научной литературе отмечают ряд технических сложностей в применении LLM в медицинской сфере. Это неактуальность наборов с обучающими данными; фактические ошибки в ответах из-за использования моделями вероятностных связей между словами, а не подлинного смысла информации; эффект «чёрного ящика», поскольку тяжело проследить чёткие шаги алгоритма работы модели для получения того или иного ответа [11]. Также, например, в статье журнала Nature группа исследователей обозначает, что при применении LLM в сфере здравоохранения тяжело добиться воспроизводимости ответа, его стандартизации, обеспечить работу только на актуальных и релевантных данных [12]. Предложенные LLM гипотезы могут не учитывать ключевые особенности пациента и контекста, такие как возраст или географическое расположение, что может привести к назначению неподходящего или недоступного лечения, а разрозненные или устаревшие данные повышают риск неадекватных или неприменимых рекомендаций.

Существуют способы решения некоторых обозначенных задач, в частности, связанных со снижением фактологических ошибок в ответах и с обеспечением работы модели только на конкретных релевантных данных. В качестве одного из методов для достижения эффективности исследователи выделяют fine-tuning на подобранных экспертами инструкциях [13]. Модель учится на примерах, свойственных для специально выбранной отрасли, что позволяет получить настройку её параметров так, чтобы она решала задачи в определённой сфере с высокой точностью. Например, группа американских врачей в статье 2025 года, посвящённой сравнению методов дообучения модели в клинической медицине использовала разные наборы инструкций для конкретных задач, которые должна была решать LLM в эксперименте [14]. Для решения задачи классификации описаний пациентов с инфекцией мочевыводящих путей и без неё проводили fine-tuning LLM на 800 сгенерированных текстах, половина из которых подразумевала наличие болезни, а другая её отсутствие. Для задачи имитации клинического мышления LLM дообучалась на выборке из 4095 вопросов экзаменационных медицинских тестов и ответов на них, часть из которых была корректной, а другая содержала ошибки. Для получения краткого изложения медицинского текста использовались материалы из сгенерированных 4500 эталонных изложений врачебных записей. Таким образом, для решения определённой задачи LLM проходила дообучение на данных и примерах, характерных именно для неё. В эксперименте использовались две open-source модели: Llama-3 8B и Mistral 7B. После дообучения они давали заметный прирост качества при решении простых задач, описанных выше, по сравнению с базовыми версиями. При всех преимуществах fine tuning имеет свои недостатки, один из которых – это потребность в значительных ресурсах, так как для дообучения требуются дорогостоящие GPU [15]. Эту проблему можно решить с помощью LoRA – экономичному методу дообучения, который изменяет не все параметры модели, а лишь добавляет небольшие дообучаемые матрицы [16]. Однако здесь нужно добиться необходимого уровня решения задач моделью без её полного переобучения, что может вызывать технические сложности. Также при fine-tuning существует риск адаптации модели к обучающей выборке. LLM будет выдавать ответы низкого качества при работе с данными, которых не было в обучающих материалах, и модель вообще может утратить способности к решению задач, с которыми до этого справлялась успешно [17].

Существует и другая технология повышения качества ответов моделью, которая называется Retrieval Augmented Generation (далее – RAG). Здесь модель не подвергается дообучению, а получает доступ к контексту из внешних источников данных. Для работы RAG реализуется процесс, состоящий из этапа поиска

релевантных материалов в имеющихся хранилищах данных и этапа генерации текста моделью, на вход которой передаются найденные чанки [18]. В качестве данных может выступать внутренняя документация, корпоративные данные, нормативные акты и так далее. С помощью описанной технологии решается проблема устаревших или недостаточных материалов, которыми располагает модель, поскольку на них обучалась. RAG подходит для решения задач, где необходимы узкоспециализированные знания, которые являются уникальными для области и часто меняются. Цепочка шагов применения RAG начинается с индексации – разделения исходного документа на фрагменты – чанки, преобразование их в векторное представление в виде эмбеддингов и сохранение в базе данных для быстрого поиска [19]. Для хранения эмбедингов используются векторные базы данных, поиск информации в которых осуществляется не по совпадению значений, как в реляционных, а по сходству векторов [20]. После индексации следует этап поиска нужных чанков в базе по пользовательскому запросу. Задаваемое топ-k найденных чанков используется для подстановки в системный промпт, который соединяет исходный вопрос пользователя и найденный по нему контекст. Далее LLM генерирует ответ, который впоследствии проверяется на качество и возвращается пользователю. Из достоинств RAG стоит выделить отсутствие необходимости в вычислительных затратах на дообучение модели, обеспечение доступа к актуальным документам, сниженный риск галлюцинаций за счёт использования проверенных источников. Недостатки подхода включают в себя затраты на предварительный отбор и подготовку используемых для поиска материалов. В литературе упоминаются успешные случаи использования RAG для решения медицинских задач. Например, в обзорной статье 2025 года по применению RAG в медицине группа учёных из США отмечает, что подход позволяет осуществлять поддержку врачей для установки диагностических гипотез, используя похожие клинические случаи и релевантные научные публикации на примере диагностики глаукомы, хронической болезни почек и инфекционных заболеваниях [21].

В научной литературе встречаются случаи объединения подходов fine-tuning и RAG для улучшения показателей модели. Например, исследование «Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation» 2025 года в журнале Bioengineering посвящено сравнению эффективности различных методов адаптации LLM, в том числе их комбинации [22]. Целью являлся выбор оптимального способа оптимизации LLM для решения медицинских задач. В статье приведены тезисы, доказывающие, что применение fine-tuning может привести к повышению точности ответов системы на медицинские вопросы, в то же время RAG обеспечивает актуальность ответов за счёт доступа к внешним знаниям. Гибридный подход тоже показал приемлемые результаты, но требовал излишних вычислительных мощностей.

Подводя итог рассмотрения разных способов повышения эффективности LLM для решения задач, можно прийти к выводу, что на данный момент универсального решения не существует. Ответы высокого качества особенно важны для вопросов из медицинской сферы. В рамках исследования было решено детально сравнить разные способы повышения эффективности работы LLM на примере RAG как одного из распространённых методов улучшения ответов модели. Так как большая часть научной литературы посвящена экспериментам с зарубежными LLM и англоязычными источниками знаний, было предложено использовать российские модели семейства GigaChat и клинические рекомендации Министерства здравоохранения РФ, чтобы локализовать результаты для возможности применения в отечественной практике. Модели GigaChat предназначены для решения задач широкого плана и не адаптированы исключительно для клинической диагностики в отличие от других решений, например, ClinicalGPT [23]. Предполагается, что использование именно таких моделей может позволить провести более чистый эксперимент: качество ответов модели можно будет оценивать именно как результат применения соответствующих методов без значительного влияния предварительного обучения модели на специализированных медицинских знаниях.

II. Метод исследования

В качестве цели исследования выступает сравнительный анализ вариаций подхода RAG для повышения эффективности применения LLM в практике получения ответов на специализированные медицинские вопросы в условиях российской системы здравоохранения. В качестве объекта исследования выступает процесс прохождения теста с медицинскими вопросами с применением информационных технологий. Предмет исследования – выявление эффективной методики применения больших языковых моделей и RAG для формирования ответов на медицинские вопросы с интеграцией клинических рекомендаций Министерства здравоохранения Российской Федерации. Эксперимент является продолжением работы по модульной переработке медицинских документов как метода повышения релевантности ответов LLM [24]. В обсуждении результатов указанной работы предлагалось провести следующую итерацию исследования для улучшения качества ответов модели за счёт применения архитектур RAG при работе с медицинскими нормативными документами.

Эксперимент был направлен на сравнительный анализ различных вариаций применения подхода RAG с целью повышения релевантности ответов LLM на вопросы, сформулированные на основе клинических рекомендаций Министерства здравоохранения Российской Федерации. Экспериментальная часть исследования была реализована с использованием платформы, представляющей собой решение с открытым исходным кодом для разработки приложений искусственного интеллекта за счёт интеграции Backend-

as-a-Service и инструментов класса LLMOps. Платформа обеспечивала единое окружение для управления большими языковыми моделями, базами знаний, механизмами поиска и архитектурой с включением RAG.

На первом этапе эксперимента был вручную сформирован опросник, состоящий из одиннадцати вопросов по содержанию клинических рекомендаций Министерства здравоохранения Российской Федерации, посвящённых аллергическому риниту [25]. Вопросы подбирались таким образом, чтобы определить эффективность модели при разных вариациях поставленных задач. Были выделены следующие группы:

- фактологические вопросы для проверки, как себя показывает модель при базовом поиске по тексту;
- уточняющие и сравнительные для проверки понимания контекста вопроса и клинических рекомендаций моделью;
- сценарные с описанием ситуации в медицинской практике, решение которой отсутствует в источнике, но может быть предложено исходя из контекста описания подобных случаев в документе; данные вопросы могут дать возможность оценить способности модели к синтезу и анализу.

Для каждого вопроса авторами исследования были подготовлены эталонные ответы. При их формировании использовался исключительно текст соответствующих клинических рекомендаций, что позволило рассматривать данные ответы как релевантные при последующей оценке качества.

Таблица 1 – Содержание медицинского опросника, на котором проводился эксперимент

Группа вопросов	**Цель проверки**	**Вопрос**
Фактологические (базовый поиск по тексту)	Противоречие источнику	Что такое эрозивный рефлюкс-эзофагит согласно рекомендациям?
	Определение	Что такое аллергический ринит (АР) согласно рекомендациям?
	Кодировка МКБ	Какие коды МКБ-10 соответствуют аллергическому риниту?
	Основные аллергены	Перечислите основные этиологические факторы (аллергены), вызывающие АР.
	Классификация	На какие виды делится АР по характеру течения и по степени тяжести?
	Симптомы	Каковы пять основных симптомов АР?
Уточняющие и сравнительные (понимание контекста)	Дифференциация	Чем симптомы сезонного АР (САР) отличаются от симптомов круглогодичного АР (КАР)?
	Пересечение болезней	Какова связь между аллергическим ринитом и бронхиальной астмой (БА)? Почему АР считается фактором риска?
	Диагностика	Каковы критерии установления диагноза АР? Что важнее: жалобы или результаты анализов?
	Методы обследования	В чем разница между диагностической ценностью кожных проб и анализа на специфические IgE?
Сценарные (синтез и анализ)	Сценарий для врача	На приеме мужчина 27 лет с подозрением на АР. Пациент против медикаментозного лечения и просит назначить гомеопатию или фитотерапию. Что делать с пациентом, основываясь на рекомендациях?

Документы были представлены в двух форматах: исходный формат PDF и вручную обработанный и структурированный формат Markdown. Во втором файле были произведены предварительные манипуляции, включающие в себя удаление списка ссылок, титульной страницы, оглавления, приложений. Оба варианта документов были разбиты на чанки. Для их векторизации применялась embedding-модель семейства GigaChat от Сбера с контекстным окном в 512 токенов и размером вектора, равным 1024 координатам. Полученные векторные представления документов были загружены в векторную базу данных, которая в дальнейшем использовалась для реализации различных стратегий поиска в рамках RAG.

В рамках эксперимента было собрано шестнадцать вариаций взаимодействия с большими языковыми моделями, отличающихся:

- типом используемой модели: «лёгкая» модель GigaChat 2 Lite и более мощная GigaChat 2 Max, сравнение показателей которых по различным метрикам представлено в таблице 2;
- способом предоставления контекста: прямая подстановка длинного контекста или использование RAG; блок-схема, описывающая архитектуру обеих экспериментальных сборок показана на рисунках 1 и 2;
- форматом исходных данных для наполнения базы знаний: PDF или Markdown;
- типом поиска в базе знаний: по ключевым словам, векторный или гибридный.

Таблица 2 – Сравнение показателей моделей семейства GigaChat по различным метрикам [26, 27]

Метрика	**GigaChat 2 Lite**	**GigaChat 2 Max**

Общие знания		
MMLU (5-shot)	0,72	0,86
ruMMLU (5-shot)	0,66	0,80
Следование инструкциям		
IFEval(en) (0-shot)	0,80	0,89
IFEval(ru) (0-shot)	0,73	0,83

Рисунок 1 – Блок-схема экспериментальной сборки для поиска с длинным контекстом

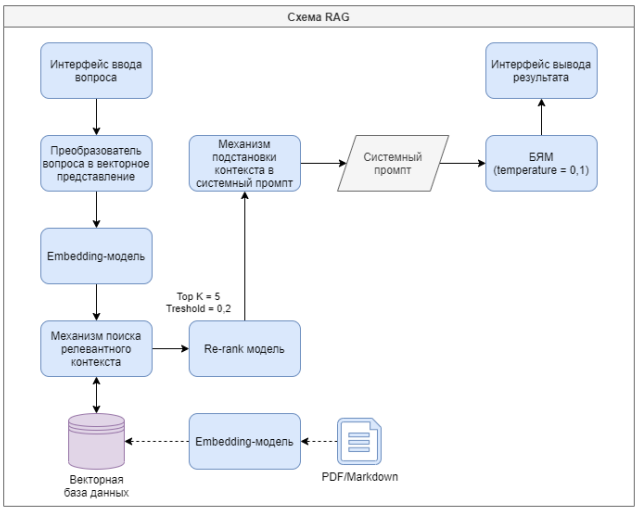

Рисунок 2 – Блок-схема экспериментальной сборки с RAG

Первые четыре вариации предполагали использование системного промпта с подстановкой длинного контекста клинических рекомендаций без применения RAG отдельно для PDF и Markdown форматов для обеих моделей. Оставшиеся двенадцать вариаций реализовывали подход RAG. В них использовались различные стратегии поиска по базе знаний: поиск по ключевым словам; векторный поиск; гибридный поиск, объединяющий ключевые слова и векторное сходство. Для всех RAG-конфигураций в системный промпт модели подставлялись 5 наиболее релевантных чанков, найденных в базе знаний, которые были ранжированы с помощью re-rank модели Mixedbread. Эксперименты проводились отдельно для баз знаний, сформированных из PDF-документов и из Markdown-документов, а также для обеих версий языковой модели GigaChat. Полный перечень вариаций представлен в таблице 3.

Таблица 3 – Перечень вариаций экспериментальных сборок

Подход	**Формат исходной документации**	**Использование базы знаний**	**Тип поиска по базе знаний**	**Модель**
Long context	PDF	Нет	Нет	GigaChat Lite
				GigaChat Max
	Markdown			GigaChat Lite
				GigaChat Max
RAG	PDF	Да	Key words Top K = 5	GigaChat Lite
				GigaChat Max
	Markdown			GigaChat Lite
				GigaChat Max
	PDF		Vector search Top K = 5	GigaChat Lite
				GigaChat Max
	Markdown			GigaChat Lite
				GigaChat Max
	PDF		Hybrid search Top K = 5	GigaChat Lite
				GigaChat Max
	Markdown			GigaChat Lite
				GigaChat Max

Для каждой из шестнадцати экспериментальных вариаций были получены ответы на все одиннадцать вопросов опросника. Таким образом, общее количество сгенерированных ответов позволило провести сопоставимый и воспроизводимый анализ качества генерации в различных конфигурациях. Полученные ответы сравнивались с заранее подготовленными эталонными ответами. Сравнение проводилось с использованием совокупности статистических метрик, позволяющих оценить текстовое сходство ответов модели с эталонными.

Системный промпт, который использовался в алгоритме: *«Необходимо предоставлять ответ по следующим требованиям: 1) На основании клинических рекомендаций из загруженного контекста ={{#context#}} нужно предоставить ответ на вопрос, указанный в {{#sys.query#}}, не добавляя в ответ лишней информации. 2) Ответ должен быть адаптирован для использования врачом в качестве дополнительного источника знаний. 3) В конце каждой рекомендации должна быть краткая ссылка на источник из загруженного контекста. 4) Запрещено включать в ответ информацию, которая отсутствует*

в загруженном контексте. 5) Если объект вопроса не упоминается в загруженном контексте, нужно отвечать "Отсутствуют знания для предоставления достоверного ответа"».

III. Результаты исследования

Для оценки результатов работы LLM используют различные метрики, универсальная отсутствует. В статье, посвящённой описанию использования подхода RAG в программах для медицинского применения, авторы предлагают среди прочего для оценки сгенерированного текста с помощью LLM метрики BLEU, группу метрик ROUGE и METEOR [19]. В пользу двух первых метрик также высказываются исследователи из МИФИ в своей статье о коллапсе LLM в медицинских приложениях [28]. Несмотря на то, что данные метрики были изначально созданы для оценки качества машинных переводов текстов моделями [29], судя по работам в научном сообществе, их также применяют и для оценки алгоритмов RAG.

Статистические метрики, которые были выбраны для оценки результатов эксперимента:

- метрика BLEU (Bilingual Evaluation Understudy) измеряет точность n-грамм между сгенерированным текстом и эталонными (референсными) текстами, со штрафом за краткость (brevity penalty) для слишком коротких выводов [19]:

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

где BP — это штраф за краткость, w_n — веса (обычно 1/N), а p_n — модифицированная точность n-грамм;

- группа метрик ROUGE (Recall Oriented Understudy for Gisting Evaluation) позволяет оценивать покрытие n-грамм эталонного текста в сгенерированном [30];
- метрика METEOR (Metric for Evaluation of Translation with Explicit Ordering) сопоставляет сгенерированный текст с эталонным текстом, используя точные совпадения, совпадения по основе слова (stem), синонимы и перефразирования [19]:

$$METEOR = (1 - Penalty) \times \frac{F_{mean}}{\alpha \times P + (1 - \alpha) \times R}$$

где F_{mean} — гармоническое среднее точности (precision) и полноты (recall), Penalty учитывает фрагментацию совпадений, а α — параметр, регулирующий баланс между точностью и полнотой.

Выбранные метрики были применены к ответам LLM, полученным в результате эксперимента, через библиотеку Evaluate от Hugging Face для языка программирования Python. Каждый из полученных моделями ответов сравнивался с эталонным с помощью обозначенных метрик. Были сформированы шесть матриц по BLEU, ROUGE-1, ROUGE-2, ROUGE-L, METEOR по 11 вопросам для 16 экспериментальных сборок. По каждой из них были вычислены суммарное значение весов, максимальное, среднее и медианное. Из полученных результатов было решено убрать результаты применения группы метрик ROUGE, поскольку распределение весов в результирующей матрице получилось недостаточно информативным. Визуализация общих результатов была проведена с помощью библиотеки matplotlib для метрики BLEU (рисунки 3-6) и метрики METEOR (рисунки 7-10).

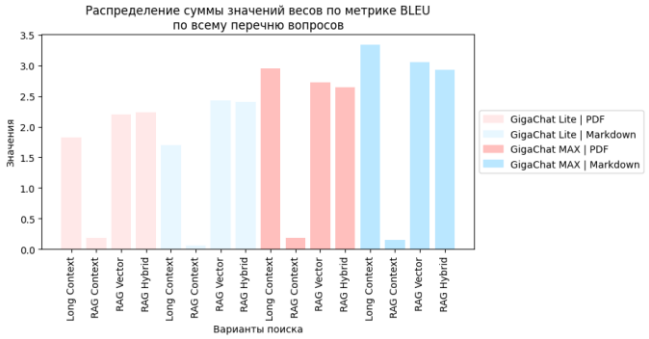

Рисунок 3 – Распределение суммы значений весов по метрике BLEU по всему перечню вопросов

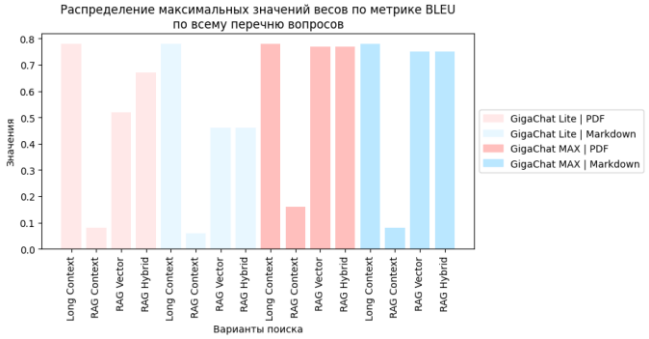

Рисунок 4 – Распределение максимальных значений весов по метрике BLEU по всему перечню вопросов

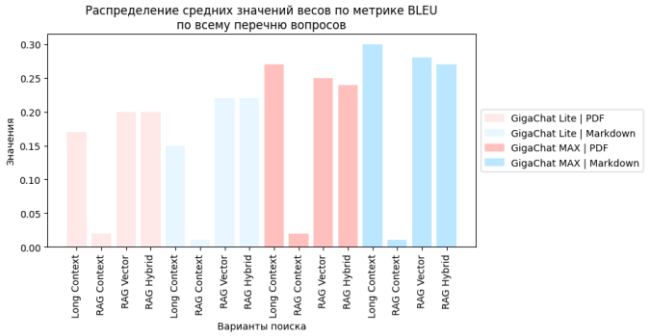

Рисунок 5 – Распределение средних значений весов по метрике BLEU по всему перечню вопросов

Рисунок 6 – Распределение медианных значений весов по метрике BLEU по всему перечню вопросов

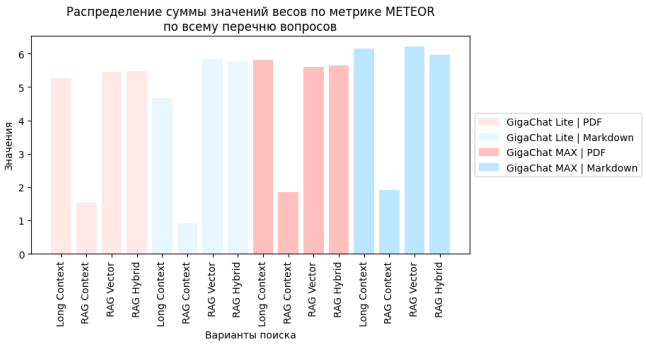

Рисунок 7 – Распределение суммы значений весов по метрике METEOR по всему перечню вопросов

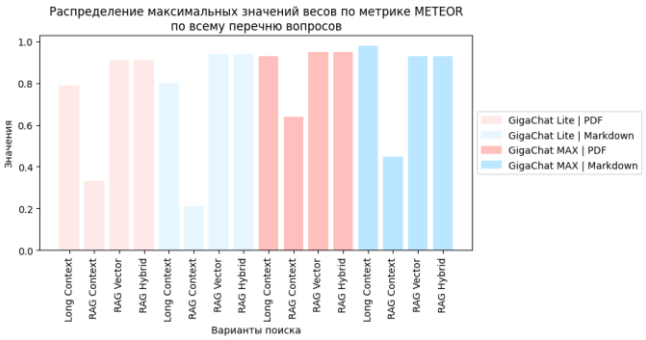

Рисунок 8 – Распределение максимальных значений весов по метрике METEOR по всему перечню вопросов

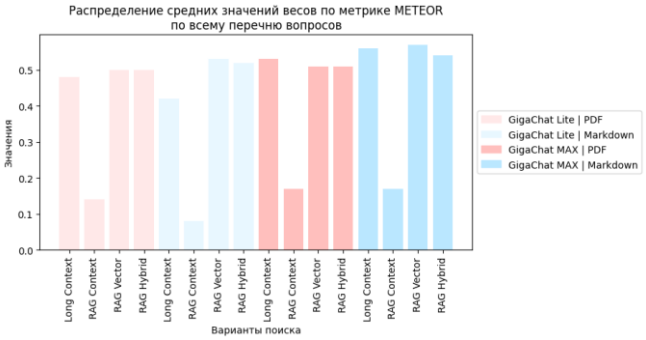

Рисунок 9 – Распределение средних значений весов по метрике METEOR по всему перечню вопросов

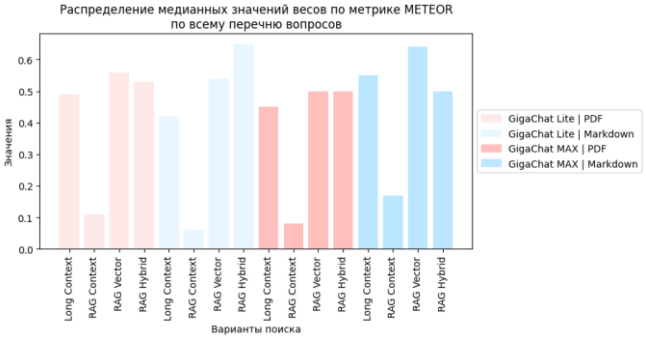

Рисунок 10 – Распределение медианных значений весов по метрике METEOR по всему перечню вопросов

Также были получены распределения средних значений весов по обеим метрикам для ответа на сложный вопрос со сценарием, что показано на рисунках 11 и 12.

Рисунок 11 – Распределение средних значений весов по метрике BLEU по вопросу со сценарием

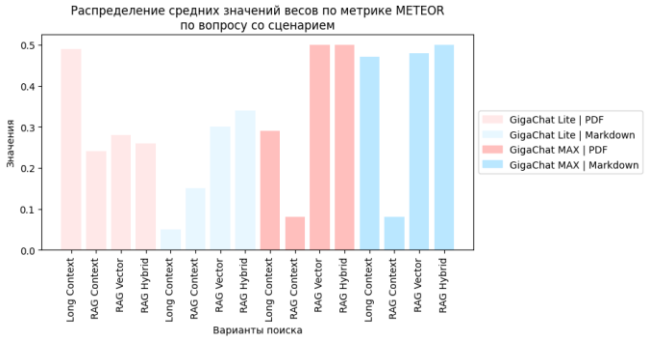

Рисунок 12 – Распределение средних значений весов по метрике METEOR по вопросу со сценарием

IV. Обсуждение результатов

С помощью результатов проведённого исследования можно сформулировать несколько тезисов, подтверждающих гипотезу о том, что вариации подходов одного выбранного метода для повышения эффективности LLM по-разному влияют на качество её ответов.

1. Распределение максимальных значений весов по обеим метрикам всего множества вопросов позволяет определить, что единичные ответы с наибольшей степенью схожести с эталонными удалось получить за счёт использования промпта с длинным контекстом, а также более мощной LLM при применении RAG с векторным и гибридным поиском: от 0,8 до 1. Учитывая тот факт, что загрузка всего документа с клиническими рекомендациями в контекст промпта является ресурсозатратной задачей для обработки моделью, лучшие показатели единичной схожести можно отнести к RAG и GigaChat Max – в контекст их системного промпта добавляются всего 5 чанков, что значительно снижает количество потребляемых моделью токенов.

2. Распределение средних значений весов по обеим метрикам всего множества ответов позволяет выявить, что худшие результаты даёт использование RAG с поиском по ключевым словам: от 0 до 0,18. Можно сделать вывод, что данный вид поиска, предназначенный для документов с большим количеством специализированных сокращений и терминологии, показывает низкую эффективность для текстов клинических рекомендаций по аллергическому риниту.

3. Распределение средних значений весов также показывает, что лучшие результаты выдаёт применение

более мощной модели GigaChat Max с RAG с векторным и гибридным поиском в сочетании с использованием базы знаний, наполнение которой осуществлялось из адаптированного текста клинических рекомендаций в формате Markdown. По сравнению с поиском по базе знаний, обогащённой данными из PDF: приблизительный прирост метрики BLEU = 0,04, приблизительный прирост METEOR = 0,05. Можно сделать вывод о том, что формат и адаптация исходного файла с клиническими рекомендациями влияет на итоговый результат работы модели с алгоритмом RAG.

4. Аналогичный вывод можно сделать и для использования разного формата клинических рекомендаций при использовании менее мощной модели GigaChat Lite: приблизительный прирост распределения средних значений весов при применении адаптированного Markdown для BLEU = 0,02, для METEOR = 0,02, приблизительный прирост распределения суммы значений весов при применении адаптированного Markdown для BLEU = 0,2, для METEOR = 0,1.

5. По приросту распределения средних значений весов: по BLEU приблизительно = 0,05, по METEOR приблизительно = 0,02, а также по приросту их суммы по обеим метрикам: по BLEU приблизительно = 0,8, по METEOR приблизительно = 0,2, нельзя сделать очевидного вывода о том, насколько сильно влияние применения более мощной модели в алгоритме RAG для работы с клиническими рекомендациями. Предлагается в будущих исследованиях попробовать применить другие метрики для анализа результирующей выборки данных, а также сформировать новую, в которой будет сделан акцент именно на сравнение влияния мощности модели для решаемой задачи. Однако в текущем эксперименте, лучшие показатели были у более мощной модели.

6. По распределению значений весов для гибридного и векторного поиска в алгоритме RAG для обеих моделей нельзя дать однозначного ответа о том, какой из них эффективнее. Полученные результаты показывают приблизительно одинаковые результаты.

7. Оценивая результаты работы алгоритмов для ответа на сложный вопрос со сценарием распределение средних значений по обеим метрикам, можно сделать вывод, что значительно лучшие результаты показала более мощная модель с применением RAG. При этом по метрике METEOR даже использование исходного документа для загрузки в базу знаний в формате PDF дало высокие результаты, не отличающиеся от Markdown, по метрике BLEU всё же виден прирост при использовании машиночитаемого формата документа.

В итоге, лучшие показатели остались за применением более мощной LLM GigaChat Max с RAG с применением векторного или гибридного поиска при предварительной адаптации текста клинических рекомендаций по аллергическому риниту в формат Markdown с очисткой лишних данных.

Для будущих исследований актуально проведение аналогичного исследования для клинических рекомендаций других нозологий, например, связанных с орфанными заболеваниями. Состав базы знаний будет более разнообразным, так как документы будут относиться к разным группам заболеваний. Также представляет интерес сравнение российских LLM для решения поставленной задачи с зарубежными: смогут ли другие решения показывать аналогичные результаты при работе с вопросами и медицинскими документами на русском языке.

V. Заключение

В ходе исследования был проведён сравнительный анализ различных вариаций применения подхода RAG для повышения эффективности LLM для получения ответов на специализированные медицинские вопросы. Полученные результаты показали, что качество ответов LLM в значительной степени определяется конкретной конфигурацией RAG. Наименее эффективным оказался поиск по ключевым словам, тогда как векторный и гибридный поиск демонстрировали более высокие и стабильные показатели сходства с эталонными ответами, особенно при использовании более мощной модели GigaChat Max.

Разработан и экспериментально обоснован метод повышения качества ответов на медицинские вопросы, основанный на структурной адаптации клинических рекомендаций и их интеграции в гибридную архитектуру RAG. Он позволил повысить релевантность и нормативную точность медицинских ответов по сравнению с использованием исходных текстов или внутренних знаний модели. В результате исследования установлено, что предварительная адаптация исходных PDF-файлов клинических рекомендаций в обработанный машиночитаемый формат Markdown положительно влияет на итоговую релевантность ответов независимо от мощности модели.

Полученные выводы подтверждают практическую применимость RAG для медицинских информационных систем и определяют направления дальнейших исследований, связанные с расширением набора нозологий, сравнением различных LLM и использованием дополнительных метрик, ориентированных на клиническую значимость ответов.

Библиография

[1] Shool S., Adimi S., Amleshi R.S. et al. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making. 2025. №1(25). P.117.

[2] Пичугов А. А., Намиот Д. Е., Зубарева Е. В. Современные методы обучения больших языковых моделей с минимумом данных: От одного примера к абсолютному нулю–академический обзор. International Journal of Open Information Technologies. 2025. №6(13). С.114-124.

[3] Трошина Е.А., Захарова С.М., Цыгулева К.В. и др. Применение искусственного интеллекта в ультразвуковой диагностике узловых образований щитовидной железы. Клиническая и экспериментальная тиреоидология. 2024. №1(20). С.15-29.

[4] Marcus H.J., Ramirez P.T., Khan D.Z. et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nature medicine. 2024. №1(30). P.61-75.

[5] Tettey F., Parupelli S.K., Desai S. A Review of Biomedical Devices: Classification, Regulatory Guidelines, Human Factors, Software as a Medical Device, and Cybersecurity. Biomedical Materials & Devices. 2024. №2. P.316–341.

[6] Clusmann J., Kolbinger F.R., Muti H.S. et al. The future landscape of large language models in medicine. Nature Communications medicine. 2023. №3(1). P.141.
[7] Singhal K., Tu T., Gottweis J. et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025. №31(3). P.943-950.
[8] OpenAI. Introducing ChatGPT Health. URL: https://openai.com/ru-RU/index/introducing-chatgpt-health (дата обращения: 01.02.2026)
[9] Андрейченко А.Е., Гусев А.В. Перспективы применения больших языковых моделей в здравоохранении. Национальное здравоохранение. 2023. №4(4). С.48-55.
[10] Костров С.А., Потапов М.П. Большие языковые модели в медицине: актуальные этические вызовы. Медицинская этика. 2025. №2(13). С.23-34.
[11] Боброва Е.В., Маканов А.Ж., Основин С.С. Генерация врачебных заключений и классификация по Bethesda с использованием глубокого обучения. International Journal of Open Information Technologies. 2023. №10(11). С.119-129.
[12] Busch F., Hoffmann L., Rueger C. et al. Current applications and challenges in large language models for patient care: a systematic review. Nature Communications Medicine. 2025. №5(26). P.1-13.
[13] Church K. W., Chen Z., Ma Y. Emerging trends: A gentle introduction to fine-tuning. Natural Language Engineering. 2021. №6(27). С.763-778.
[14] Savage T., Ma S.P., Boukil A. et al. Fine-Tuning Methods for Large Language Models in Clinical Medicine by Supervised Fine-Tuning and Direct Preference Optimization: Comparative Evaluation. Journal of Medical Internet Research. 2025. №27. P.1-9.
[15] Rangan K., Yin Y. A fine-tuning enhanced RAG system with quantized influence measure as AI judge. Nature Scientific Reports. 2024. №1(14). P.1-17.
[16] Xu L., Xie H., Qin S.J. et al. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2026. P.1-20.
[17] Montesinos López O. A., Montesinos López A., Crossa J. Overfitting, model tuning, and evaluation of prediction performance. Multivariate statistical machine learning methods for genomic prediction. Springer International Publishing. 2022. P.109-139.
[18] Zhao P., Zhang H., Yu Q. et al. Retrieval-augmented generation for AI-generated content: A survey. Data Science and Engineering. 2026. P.1-29.
[19] Abo El-Enen M., Saad S., Nazmy T. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. Neural Computing and Applications. 2025. №33(37). P.28191-28267.
[20] Taipalus T. Vector database management systems: Fundamental concepts, use-cases, and current challenges. Cognitive Systems Research. 2024. №85. P.1-13.
[21] Neha F., Bhati D., Shukla D.K. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. AI. 2025. №9(6). P.226.
[22] Pingua B., Sahoo A., Kandpal M. et al. Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation. Bioengineering. 2025. №12. P.687.
[23] Назаров Д.М., Бадаев Ф.И. Применение больших языковых моделей в сфере здравоохранения. Менеджер здравоохранения. 2025. №5. P.142-154.
[24] Фролов Е.М., Жабицкий М.Г., Мокшин К.Ю. Модульная переработка нормативных текстов как метод повышения релевантности ответов большой языковой модели: пилотное исследование на примере клинических рекомендаций по артериальной гипертензии. International Journal of Open Information Technologies. 2025. №8(13). P.94-104.
[25] Министерство здравоохранения Российской Федерации. Рубрикатор клинических рекомендаций. Аллергический ринит. URL: https://cr.minzdrav.gov.ru/preview-cr/261_2 (дата обращения: 05.02.2026)
[26] Mamedov V., Kosarev E., Leleytner G. et al. GigaChat Family: Efficient Russian Language Modeling Through Mixture of Experts Architecture Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025. №3.
[27] Sber Developer. Описание моделей GigaChat. URL: https://developers.sber.ru/docs/ru/gigachat/models/gigachat-2-lite (дата обращения: 05.02.2026)
[28] Боброва Е.В., Зайцев К.С., Свириденко Д.К. Исследование раннего и позднего коллапса языковых моделей в медицинских приложениях. International Journal of Open Information Technologies. 2025. №8(13). С.51-59.
[29] Ray P.P. Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023. №3(3). P.1-19.
[30] Lin C. Y. ROUGE: A package for automatic evaluation of summaries. Text summarization branches out. Conference: In Proceedings of the Workshop on Text Summarization Branches Out. 2004. P.74-81.

Enhancing RAG Configuration Efficiency for Large Language Models Working with Clinical Guidelines: The Case of Allergic Rhinitis

K.Y. Mokshin, E.V. Bobrova, M.G. Zhabitsky

***Abstract* — The article examines approaches to improving the effectiveness of Large Language Models (LLMs) in solving specialized medical tasks based on clinical guidelines issued by the Ministry of Health of the Russian Federation. Particular attention is paid to comparing different variations of the Retrieval-Augmented Generation (RAG) architecture, which is used to reduce factual errors and improve the relevance of model responses. Clinical guidelines on allergic rhinitis were selected as the domain-specific case study, and Russian GigaChat family models were used as the language models. Within the experimental study, sixteen configurations of model interaction were implemented, differing in LLM capacity, source document format (PDF and pre-adapted Markdown), and knowledge base retrieval strategies (keyword-based, vector-based, and hybrid search). The quality of responses was evaluated by comparing them with reference answers derived from the clinical guidelines using BLEU and METEOR metrics. The results demonstrated that the best performance is achieved when using a more powerful model in combination with a RAG approach based on vector or hybrid retrieval and a knowledge base constructed from structured, machine-readable text. The findings confirm the feasibility of applying RAG in medical information systems and highlight the importance of prior preparation and structuring of regulatory documents to improve the performance of large language models.**

***Keywords* — large language models, LLM, clinical guidelines, RAG, prompt engineering, medical artificial intelligence, neural networks, medical informatics.**

REFERENCES

[1] Shool S., Adimi S., Amleshi R.S. et al. A systematic review of large language model (LLM) evaluations in clinical medicine. BMC Medical Informatics and Decision Making. 2025. #1(25). P.117.

[2] Pichugov A. A., Namiot D. E., Zubareva E. V. Sovremennye metody obuchenija bol'shih jazykovyh modelej s minimumom dannyh: Ot odnogo primera k absoljutnomu nulju–akademicheskij obzor. International Journal of Open Information Technologies. 2025. #6(13). S.114-124.

[3] Troshina E.A., Zaharova S.M., Cyguleva K.V. i dr. Primenenie iskusstvennogo intellekta v ul'trazvukovoj diagnostike uzlovyh obrazovanij shhitovidnoj zhelezy. Klinicheskaja i jeksperimental'naja tireoidologija. 2024. #1(20). S.15-29.

[4] Marcus H.J., Ramirez P.T., Khan D.Z. et al. The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring. Nature medicine. 2024. #1(30). P.61-75.

[5] Tettey F., Parupelli S.K., Desai S. A Review of Biomedical Devices: Classification, Regulatory Guidelines, Human Factors, Software as a Medical Device, and Cybersecurity. Biomedical Materials & Devices. 2024. #2. P.316–341.

[6] Clusmann J., Kolbinger F.R., Muti H.S. et al. The future landscape of large language models in medicine. Nature Communications medicine. 2023. #3(1). P.141.

[7] Singhal K., Tu T., Gottweis J. et al. Toward expert-level medical question answering with large language models. Nature Medicine. 2025. #31(3). P.943-950.

[8] OpenAI. Introducing ChatGPT Health. URL: https://openai.com/ru-RU/index/introducing-chatgpt-health (data obrashhenija: 01.02.2026)

[9] Andrejchenko A.E., Gusev A.V. Perspektivy primenenija bol'shih jazykovyh modelej v zdravoohranenii. Nacional'noe zdravoohranenie. 2023. #4(4). S.48-55.

[10] Kostrov S.A., Potapov M.P. Bol'shie jazykovye modeli v medicine: aktual'nye jeticheskie vyzovy. Medicinskaja jetika. 2025. #2(13). S.23-34.

[11] Bobrova E.V., Makanov A.Zh., Osnovin S.S. Generacija vrachebnyh zakljuchenij i klassifikacija po Bethesda s ispol'zovaniem glubokogo obuchenija. International Journal of Open Information Technologies. 2023. #10(11). S.119-129.

[12] Busch F., Hoffmann L., Rueger C. et al. Current applications and challenges in large language models for patient care: a systematic review. Nature Communications Medicine. 2025. #5(26). P.1-13.

[13] Church K. W., Chen Z., Ma Y. Emerging trends: A gentle introduction to fine-tuning. Natural Language Engineering. 2021. #6(27). S.763-778.

[14] Savage T., Ma S.P., Boukil A. et al. Fine-Tuning Methods for Large Language Models in Clinical Medicine by Supervised Fine-Tuning and Direct Preference Optimization: Comparative Evaluation. Journal of Medical Internet Research. 2025. #27. P.1-9.

[15] Rangan K., Yin Y. A fine-tuning enhanced RAG system with quantized influence measure as AI judge. Nature Scientific Reports. 2024. #1(14). P.1-17.

[16] Xu L., Xie H., Qin S.J. et al. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2026. P.1-20.

[17] Montesinos López O. A., Montesinos López A., Crossa J. Overfitting, model tuning, and evaluation of prediction performance. Multivariate statistical machine learning methods for genomic prediction. Springer International Publishing. 2022. P.109-139.

[18] Zhao P., Zhang H., Yu Q. et al. Retrieval-augmented generation for AI-generated content: A survey. Data Science and Engineering. 2026. P.1-29.

[19] Abo El-Enen M., Saad S., Nazmy T. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. Neural Computing and Applications. 2025. #33(37). P.28191-28267.

[20] Taipalus T. Vector database management systems: Fundamental concepts, use-cases, and current challenges. Cognitive Systems Research. 2024. #85. P.1-13.

[21] Neha F., Bhati D., Shukla D.K. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. AI. 2025. #9(6). P.226.

[22] Pingua B., Sahoo A., Kandpal M. et al. Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation. Bioengineering. 2025. #12. P.687.

[23] Nazarov D.M., Badaev F.I. Primenenie bol'shih jazykovyh modelej v sfere zdravoohranenija. Menedzher zdravoohranenija. 2025. #5. P.142-154.

[24] Frolov E.M., Zhabickij M.G., Mokshin K.Ju. Modul'naja pererabotka normativnyh tekstov kak metod povyshenija relevantnosti otvetov bol'shoj jazykovoj modeli: pilotnoe issledovanie na primere klinicheskih rekomendacij po arterial'noj gipertenzii. International Journal of Open Information Technologies. 2025. #8(13). P.94-104.

[25] Ministerstvo zdravoohranenija Rossijskoj Federacii. Rubrikator klinicheskih rekomendacij. Allergicheskij rinit. URL: https://cr.minzdrav.gov.ru/preview-cr/261_2 (data obrashhenija: 05.02.2026)

[26] Mamedov V., Kosarev E., Leleytner G. et al. GigaChat Family: Efficient Russian Language Modeling Through Mixture of

Experts Architecture Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. 2025. #3.
[27] Sber Developer. Opisanie modelej GigaChat. URL: https://developers.sber.ru/docs/ru/gigachat/models/gigachat-2-lite (data obrashhenija: 05.02.2026)
[28] Bobrova E.V., Zajcev K.S., Sviridenko D.K. Issledovanie rannego i pozdnego kollapsa jazykovyh modelej v medicinskih prilozhenijah. International Journal of Open Information Technologies. 2025. #8(13). S.51-59.
[29] Ray P.P. Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. BenchCouncil Transactions on Benchmarks, Standards and Evaluations. 2023. #3(3). P.1-19.
[30] Lin C. Y. ROUGE: A package for automatic evaluation of summaries. Text summarization branches out. Conference: In Proceedings of the Workshop on Text Summarization Branches Out. 2004. P.74-81.