https://doi.org/10.29188/2222-8543-2026-19-2-10-21

Галлюцинации больших языковых моделей в клинической урологии: систематическая оценка 18 моделей и система генерации с дополненной выборкой (RAG)

ПРОСПЕКТИВНОЕ СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ

И.А. Шадёркин[1], В.А. Шадёркина[2]
[1] ООО «Робоскоп Патолоджи», Москва, Россия
[2] UroWeb.ru, Москва, Россия

Контакт: *Шадёркин Игорь Аркадьевич, info@uroweb.ru*

Аннотация:

Введение. *Большие языковые модели (БЯМ) – программы на основе искусственного интеллекта, обученные на массивах текстовых данных и способные генерировать связные текстовые ответы на произвольные запросы – за последние два года совершили революцию в информационном взаимодействии во многих областях, включая медицину. Однако урология представляет особый вызов для БЯМ по нескольким причинам: узкоспециализированная терминология, многообразие клинических сценариев, необходимость учета актуальных клинических рекомендаций и высокая значимость точных числовых данных.*

Цель. *Оценить клиническую точность и безопасность 18 БЯМ при ответах на вопросы по урологии и сравнить их с системой генерации с дополненной выборкой (RAG), основанной на верифицированной доказательной базе знаний.*

Материалы и методы. *Проспективное сравнительное исследование включало два направления: оценку фактической точности 17 БЯМ на 28 специфических вопросах по урологии с верифицированными эталонными ответами (115 фактов, допустимое отклонение 2%); клиническую оценку 18 БЯМ на 120 клинических случаях из реальных консультаций в сравнении с эталонными ответами врача-уролога. Модели стратифицированы по 4 уровням: флагманские (Уровень 1, n=7), сильные (Уровень 2, n=4), открытые (Уровень 3, n=4) и компактные (Уровень 4, n=3). RAG-система: Qwen3-80B + векторная база данных (2,23 млн фрагментов, 17 947 документов, 19 доменов). Оценка качества выполнена с помощью метода «языковая модель как эксперт» (Gemini 2.5 Flash), 2260 клинических оценок. Статистический анализ: дисперсионный анализ (ANOVA), критерий Cohen's d, 95% доверительные интервалы.*

Результаты. *RAG превзошла все «чистые» БЯМ по точности: 4,83 vs 4,09 балла (лучшая модель, Claude Opus 4.6), Cohen's d = 1,05 (p<0,001). Различия между Уровнями 1, 2 и 3 статистически незначимы (Cohen's d=0,07–0,15), тогда как Уровень 4 показал клинически неприемлемую точность (2,52/5). Наибольший эффект RAG достигнут при онкоурологии (Δ=+1,47–1,89) и хирургических методиках (Δ=+1,77). Все «чистые» БЯМ продемонстрировали значительное число проблем безопасности (0,85–3,09 на кейс), при этом эмпатия оставалась стабильно высокой (4,18–4,85), создавая у пациента ложное ощущение компетентности.*

Заключение. *«Чистые» БЯМ всех уровней демонстрируют высокий уровень галлюцинаций в клинической урологии. RAG-система устраняет этот недостаток с большим эффектом. Флагманский статус модели не гарантирует клинической точности. RAG является перспективной архитектурой для систем поддержки принятия врачебных решений в урологии.*

Ключевые слова: *большие языковые модели; галлюцинации; искусственный интеллект; генерация с дополненной выборкой; RAG; урология; доказательная медицина; клиническая точность; безопасность; клинические рекомендации.*

Для цитирования: *Шадёркин И.А., Шадёркина В.А. Галлюцинации больших языковых моделей в клинической урологии: систематическая оценка 18 моделей и система генерации с дополненной выборкой (RAG) как решение. Экспериментальная и клиническая урология. 2026;19(2):10-21; https://doi.org/10.29188/2222-8543-2026-19-2-10-21*

https://doi.org/10.29188/2222-8543-2026-19-2-10-21

Hallucinations of large language models in clinical urology: systematic evaluation of 18 models and a RAG system as a solution

PROSPECTIVE COMPARATIVE STUDY

I.A. Shaderkin[1], V.A. Shaderkina[2]
[1] Roboskop Pathology LLC, Moscow, Russia
[2] UroWeb.ru, Moscow, Russia

Contacts: *Igor A. Shaderkin, info@uroweb.ru*

Summary:

Introduction. *Large Language Models (LLMs), artificial intelligence–based programs trained on arrays of text data and capable of generating coherent text responses to arbitrary queries, have revolutionized information interaction in many fields over the past two years, including medicine. However, urology poses a special challenge for LLMs for several reasons: highly specialized terminology, a variety of clinical scenarios, the need to take into account current clinical recommendations, and the high importance of accurate numerical data.*

Objective. *To evaluate the clinical accuracy and safety of 18 LLMs in answering urological questions and compare them with a retrieval-augmented generation (RAG) system based on a verified evidence base.*
Materials and methods. *A prospective comparative study included two directions: factual accuracy assessment of 17 LLMs on 28 specific urological questions with verified ground truth (115 facts, 2% tolerance); clinical evaluation of 18 LLMs on 120 clinical cases from real consultations compared with reference urologist answers. Models were stratified into 4 tiers: flagship (Tier 1, n=7), strong (Tier 2, n=4), open-source (Tier 3, n=4), and small (Tier 4, n=3). RAG system: Qwen3-80B + ChromaDB (2.23M chunks, 17.947 documents, 19 domains). Quality assessment: LLM-as-judge (Gemini 2.5 Flash), 2.260 clinical evaluations. Statistics: ANOVA, Cohen's d, 95% confidence intervals.*
Results. *RAG outperformed all bare LLMs in accuracy: 4.83 vs 4.09 (best model, Claude Opus 4.6), Cohen's d = 1.05 (p<0.001). Differences between Tier 1, Tier 2, and Tier 3 were negligible (Cohen's d = 0.07–0.15), while Tier 4 demonstrated clinically unacceptable accuracy (2.52/5). The greatest RAG advantage was observed in oncourology (Δ=+1.47–1.89) and surgical techniques (Δ=+1.77). All bare LLMs demonstrated substantial safety issues (0.85–3.09 per case), while empathy remained consistently high (4.18–4.85), creating a false impression of competence.*
Conclusion. *Bare LLMs of all tiers demonstrate a high hallucination rate in clinical urology. The RAG system eliminates this limitation with a large effect size. Flagship model status does not guarantee clinical accuracy.*

Key words: *large language models; hallucinations; artificial intelligence; RAG; retrieval-augmented generation; urology; evidence-based medicine; clinical accuracy; safety.*

For citation: *Shaderkin I.A., Shaderkina V.A. Hallucinations of large language models in clinical urology: systematic evaluation of 18 models and a RAG system as a solution. Experimental and Clinical Urology. 2026;19(2):10-21; https://doi.org/10.29188/2222-8543-2026-19-2-10-21*

Глоссарий терминов

Поскольку тема исследования находится на стыке медицины и информационных технологий, ниже приведены ключевые термины, которые встречаются в тексте:

Термин	Пояснение
Большая языковая модель (БЯМ) (англ. large language model, LLM)	Программа на основе искусственного интеллекта, обученная на огромных массивах текста, способная генерировать связные ответы на произвольные вопросы. Примеры: ChatGPT, Claude, Gemini.
Галлюцинация	Ситуация, когда модель генерирует фактически неверный, но внешне правдоподобный ответ. Модель не «лжет» намеренно – она не имеет доступа к достоверным источникам в момент генерации.
RAG (генерация с дополненной выборкой)	Метод, при котором модель отвечает не «по памяти», а на основе документов, найденных в верифицированной базе знаний. Каждый тезис подтверждается ссылкой на источник.
«Чистая» БЯМ (англ. bare LLM)	Языковая модель, используемая напрямую, без подключения к базе знаний. Отвечает исключительно на основе информации, полученной при обучении.
Векторная база данных	Хранилище, в котором тексты представлены в виде числовых векторов (эмбеддингов). Это позволяет искать документы по смыслу, а не только по точному совпадению слов.
Эмбеддинг (векторное представление текста)	Числовое представление фрагмента текста, отражающее его смысл. Тексты с близким смыслом имеют близкие эмбеддинги.
Фрагмент (чанк)	Небольшой отрезок текста (~1000 символов), на который разбиваются документы перед загрузкой в базу. Поиск происходит по фрагментам, а не по целым документам.
Семантический поиск	Поиск по смыслу запроса, а не по точному совпадению слов. Позволяет находить «лечение аденомы» при запросе «терапия ДГПЖ».
BM25	Классический алгоритм поиска по ключевым словам, аналогичный поиску в Яндекс или Google. Используется в комбинации с семантическим поиском.
«Языковая модель как эксперт» (LLM-as-judge)	Метод оценки, при котором одна языковая модель оценивает качество ответов другой модели по заданным критериям.
Reasoning-модель (рассуждающая модель)	Модель, которая перед ответом проводит пошаговое рассуждение (цепочку мыслей), что повышает качество ответа на сложные вопросы.
Токен	Единица текста, примерно соответствующая слогу или короткому слову. Используется для измерения объема текста и стоимости запросов к модели.
Промпт (запрос к модели)	Текстовая инструкция, которую получает модель перед генерацией ответа. Определяет роль модели, формат и содержание ответа.
Fine-tuning (дообучение)	Метод адаптации модели к задаче путем дополнительного обучения на специализированных данных. В отличие от RAG, не позволяет обновлять знания без переобучения.

ВВЕДЕНИЕ

Большие языковые модели (БЯМ) – программы на основе искусственного интеллекта, обученные на массивах текстовых данных и способные генерировать связные текстовые ответы на произвольные запросы – за последние два года совершили революцию в информационном взаимодействии во многих областях, включая медицину [1, 2]. Выпуск ChatGPT (OpenAI) в ноябре 2022 г. стал поворотным моментом: впервые система искусственного интеллекта продемонстрировала способность вести осмысленные медицинские диалоги, интерпретировать результаты обследований и формулировать дифференциальные диагнозы [3, 4]. К 2026 г. экосистема моделей включает десятки разработчиков – OpenAI (GPT-4o, GPT-4.1), Anthropic (Claude Opus 4.6, Claude Sonnet 4.5),

Google (Gemini 2.5 Pro, Gemini 2.5 Flash), Meta (Llama 4), Alibaba (Qwen3), DeepSeek, Mistral и другие [5].

Пациенты все чаще обращаются к БЯМ за медицинскими консультациями, минуя традиционный врачебный контакт. По данным J.W. Ayers и соавт., ответы ChatGPT на медицинские вопросы пациентов оцениваются как более эмпатичные, чем ответы врачей на медицинских форумах [3]. Параллельно растет интерес клиницистов к использованию моделей как инструмента поддержки принятия решений [4].

Однако фундаментальный принцип работы БЯМ – вероятностная генерация текста на основе статистических закономерностей обучающего корпуса – делает их подверженными «галлюцинациям»: генерации фактически неверных, но внешне правдоподобных утверждений [6]. В клинической практике это может приводить к назначению неадекватной терапии, пропуску тревожных симптомов («красных флагов»), неверной интерпретации результатов обследований и в конечном счете к вреду для пациента [7].

Исследования галлюцинаций в медицине пока немногочисленны. K. Singhal и соавт. показали, что Med-PaLM 2 достигает 85% точности на вопросах формата USMLE, однако этот показатель далек от клинической приемлемости [8]. H. Nori и соавт. продемонстрировали, что GPT-4 решает 84% медицинских задач, но в реальных клинических сценариях точность значительно снижается [9]. В урологии публикации ограничиваются оценкой ChatGPT на сертификационных экзаменах или узких вопросах [10–14], при этом систематическая оценка галлюцинаций не проводилась.

Урология представляет особый вызов для БЯМ по нескольким причинам. Во-первых, узкоспециализированная терминология: корректное использование терминов «ГМП», «НМИРМП», «HoLEP» требует понимания клинического контекста. Во-вторых, многообразие клинических сценариев: от консервативного лечения инфекций мочевых путей до робот-ассистированной простатэктомии. В-третьих, необходимость учета актуальных клинических рекомендаций Европейской ассоциации урологов (EAU) и Американской урологической ассоциации (AUA), которые регулярно обновляются и могут отсутствовать в обучающем корпусе модели. В-четвертых, высокая значимость точных числовых данных: уровни простатспецифического антигена (ПСА, нг/мл), стадии опухолей по TNM, дозировки препаратов [15].

Альтернативным подходом является генерация с дополненной выборкой (RAG) – архитектура, при которой модель генерирует ответ не «по памяти», а на основе релевантных документов, извлеченных из верифицированной базы знаний [16]. RAG-система работает в два этапа: 1) поисковый модуль находит наиболее релевантные фрагменты документов по запросу, используя комбинацию поиска по смыслу и поиска по ключевым словам; 2) языковая модель синтезирует ответ на основе найденного контекста с обязательным цитированием источников. RAG потенциально устраняет галлюцинации и обеспечивает объяснимость – врач может проверить, из какого источника взят каждый тезис [16–21].

Цель исследования – провести систематическую оценку клинической точности и безопасности 18 БЯМ разных поколений и масштабов при ответах на вопросы по урологии, определить зависимость качества ответов от класса модели (4 уровня) и сравнить результаты с RAG-системой, основанной на верифицированной доказательной базе знаний из 17 947 документов.

МАТЕРИАЛЫ И МЕТОДЫ

Дизайн исследования

Проспективное сравнительное исследование, проведенное в апреле 2026 г. Исследование включало два взаимодополняющих направления.

Направление 1 (специфические вопросы): оценка фактической точности ответов на узкие специфические вопросы по урологии с верифицированными эталонными ответами. Врачом-урологом составлено 35 вопросов, охватывающих: эпидемиологические данные (распространенность, заболеваемость), пороговые значения (ПСА, объем предстательной железы, размеры камней), дозировки препаратов, стадии опухолевого процесса по TNM, показания к хирургическому лечению. Для 28 вопросов сформулирован эталонный набор фактов – 115 верифицируемых утверждений в формате {содержание утверждения, числовое значение, единица измерения, уровень риска}. Эталонные факты извлечены автоматически из базы знаний RAG-системы и верифицированы врачом-урологом. Семь вопросов исключены из-за недостаточности эталонных данных. Каждая из 17 моделей отвечала на все 28 вопросов (476 вызовов).

Ответы проверялись автоматически: каждый факт классифицировался как точное совпадение, частичное совпадение, пропуск или ошибка. Допустимое отклонение для числовых значений – 2% (не 10%, так как клинически значимые различия – например, 3,0 *vs* 3,5 нг/мл ПСА – маскируются при 10% допуске).

Направление 2 (клинические случаи): оценка клинического качества ответов на реальные вопросы пациентов. Из 2733 онлайн-консультаций врача-уролога стратифицированно отобрано 120 клинических случаев по 15 клиническим доменам (6–9 кейсов на домен): нижние мочевые пути (симптомы нарушения функции нижних мочевыводящих путей, доброкачественная гиперплазия предстательной железы, гиперактивный мочевой пузырь), мочекаменная болезнь, инфекции мочевых путей, рак предстательной железы, рак мочевого пузыря, рак почки, андрология, сексуальная медицина, фармакология, диагностика, хирургические методики, детская урология, урогинекология,

травма мочеполовых органов, прочее. Каждый кейс содержал вопрос пациента (формулировка без изменений) и эталонный ответ врача-уролога.

Модели

Протестировано 18 БЯМ, стратифицированных по 4 уровням (табл. 1). Стратификация основана на количестве параметров, стоимости вызовов и рыночном позиционировании:

• уровень 1 (флагманские, 7 моделей) – наиболее мощные коммерческие модели;

• уровень 2 (сильные, 4 модели) – модели среднего сегмента;

• уровень 3 (открытые, 4 модели) – модели с открытым исходным кодом;

• уровень 4 (компактные, 3 модели) – модели с 8 млрд параметров.

Все модели (кроме локальной) вызывались через единый программный интерфейс (OpenRouter API). Параметры генерации унифицированы: степень случайности 0,3, максимальная длина ответа 2048 токенов (~1500 слов).

RAG-система

Использована двухслойная RAG-система, разработанная специально для урологии (рис. 1). Архитектура системы:

Рис. 1. Упрощенная схема архитектуры RAG-системы для урологии. Вопрос врача поступает в систему, где проходит через два слоя поиска: по карточкам знаний (навигация по темам) и по фрагментам документов (детальный поиск). Найденный контекст подается в языковую модель, которая синтезирует ответ с цитированием источников. Ответ может быть выдан в формате для врача или конвертирован в формат для пациента
Fig. 1. Simplified architecture diagram of a RAG system for urology. The doctor's question enters the system, where it passes through two layers of search: by knowledge cards (topic navigation) and by document fragments (detailed search). The found context is fed into a language model that synthesizes the response with a citation of the sources. The response can be given in the format for the doctor or converted to the format for the patient

Таблица 1. Стратификация протестированных БЯМ
Table 1. Stratification of tested LLMs

Уровень Tier	Категория Category	Модели Models	N	Стоимость за 1 млн токенов Cost per 1 million tokens
Уровень 1 (флагманские) Tier 1 (Flagship)	Коммерческие флагманские Commercial flagship	GPT-4o, GPT-4.1, Claude Sonnet 4.5, Claude Opus 4.6, Gemini 2.5 Pro, Gemini 2.5 Flash, Qwen3-80B (локальная)	7	$0,15–75
Уровень 2 (сильные) Tier 2 (strong)	Сильные коммерческие Strong commercial	Claude Sonnet 4, DeepSeek V3, Qwen3 Max, Mistral Large	4	$0,5–5
Уровень 3 (открытые) Tier 3 (open)	С открытым исходным кодом Open source	Llama 4 Maverick, Llama 3.3 70B, Qwen3.5-35B, DeepSeek R1	4	$0,1–1
Уровень 4 (компактные) Tier 4 (compact)	Компактные (8 млрд параметров) Compact (8 billion parameters)	GPT-4o-mini, Llama 3.1 8B, Qwen3 8B	3	$0,05–0,15

Слой 1 – карточки знаний (навигационный). 2281 структурированная карточка, покрывающая всю урологию в 19 доменах. Каждая карточка содержит: определение, ключевые факты, эпидемиологию, лечение (консервативное, медикаментозное, хирургическое), резюме рекомендаций EAU/AUA, уровень доказательности.

Слой 2 – поиск по документам (детальный). 2 227 761 текстовый фрагмент из корпуса, проиндексированных с помощью комбинированного поиска: поиск по смыслу (семантический, через векторные представления текста) и поиск по ключевым словам (алгоритм BM25). Двуязычный поиск – русский и английский.

База знаний включает 17 947 обработанных документов: фундаментальные учебники (Campbell-Walsh Urology 12th ed., Hinman's Atlas, Oxford Handbook of Urology) и монографии, клинические рекомендации EAU и AUA (актуальные версии), систематические обзоры Cochrane, более 14 000 полнотекстовых статей PubMed, 92 330 рефератов из медицинских баз данных (PubMed, OpenAlex, Europe PMC, Semantic Scholar), 1245 статей журнала «Экспериментальная и клиническая урология», углубленные исследовательские обзоры. Общий объем: 19 доменов, примерно 977 фрагментов текста на каждую тему.

Техническая платформа. Система полностью работает на локальном компьютере врача (MacBook Pro, процессор Apple M4 Max, 128 ГБ оперативной памяти) с использованием программы LM Studio для запуска языковой модели Qwen3-80B. Генерация ответов и поиск по базе данных происходят без подключения к интернету. Данные пациента не покидают компьютер врача, что соответствует требованиям Федерального закона № 152-ФЗ «О персональных данных» и сохраняет врачебную тайну. Это принципиальное преимущество перед облачными сервисами (ChatGPT, Claude, Gemini), при использовании которых медицинские данные пациентов передаются на серверы сторонних организаций.

Оценка качества

Применена методология «языковая модель как эксперт», включавшая четыре этапа:

1. Извлечение структурированного контрольного списка (чек-листа) из эталонного ответа врача (диагноз, препараты и дозировки, обследования, «красные флаги», направления к специалистам, рекомендации, предупреждения, план наблюдения).

2. Автоматическая сверка ответа модели с чек-листом.

3. Обнаружение проблем безопасности – опасных рекомендаций в ответе модели, отсутствующих в ответе врача.

4. Общая оценка по 5-балльной шкале.

«Чистые» БЯМ оценивали по критериям «ответ пациенту»: точность, полнота, безопасность, эмпатия, практическая применимость. **RAG-система** оценивалась по критериям «инструмент для врача»: точность, полнота, качество доказательной базы и цитирования, полезность для принятия клинического решения, безопасность, релевантность.

Ключевое методологическое решение: RAG и «чистые» БЯМ анализировали по разным шкалам. RAG-ответ – это клинический документ для принятия решения врачом (с цитатами, уровнями доказательности LE/GR, библиографией), а не прямой ответ пациенту. Пациент не может оценить качество клинической информации – это может только врач. Поэтому прямое сравнение возможно только по метрикам точности и безопасности, присутствующим в обеих шкалах.

Модель-оценщик: Gemini 2.5 Flash. Всего выполнено 2260 клинических оценок (2140 «чистые» БЯМ + 120 RAG).

Слой эмпатии

Для демонстрации возможности конвертации RAG-ответа (документ для врача) в формат для пациента применен «слой эмпатии»: RAG-ответ → языковая модель → ответ для пациента с сохранением медицинской точности, но понятным и эмпатичным языком.

Статистический анализ

Дисперсионный анализ (ANOVA) для сравнения уровней по точности. 95% доверительные интервалы. Cohen's d для оценки величины эффекта при попарных сравнениях. Классификация эффектов: пренебрежимый (<0,2), малый (0,2–0,5), средний (0,5–0,8), большой (≥0,8). Уровень значимости α=0,05.

РЕЗУЛЬТАТЫ

Направление 1. Фактическая точность (специфические вопросы)

Средняя полнота воспроизведения фактов по всем «чистым» БЯМ составила 35,5%. RAG-система достигла 71,3% – двукратное превосходство. Анализ типов ошибок показал, что наиболее частый тип – пропуск фактов (65%), тогда как генерация фактически неверных числовых значений составила лишь 1,8%. Таким образом, основная проблема «чистых» БЯМ – не «придумывание» неверных данных, а неспособность воспроизвести специфическую информацию, отсутствующую в обучающем корпусе.

RAG-система, напротив, извлекает конкретные факты из верифицированных источников, что кардинально снижает долю пропусков.

Направление 2. Клиническая оценка (120 кейсов)

Всего получено 2147 ответов «чистых» БЯМ и 120 ответов RAG. Все RAG-ответы сгенерированы успешно (0 ошибок), среднее время генерации – 98 секунд.

Результаты по всем 18 моделям и RAG-системе представлены в табл. 2.

Сравнение по уровням

Средняя точность по уровням составила: Уровень 1 – 3,55, Уровень 2 – 3,48, Уровень 3 – 3,38, Уровень 4 – 2,52 балла из 5. RAG-система достигла 4,83 балла (95% ДИ 4,73–4,94), значимо превзойдя все модели.

Дисперсионный анализ подтвердил статистически значимые различия: F(3, 2136) = 81,07, p<0,001. При этом попарные сравнения показали, что различия между Уровнями 1, 2 и 3 статистически незначимы (Cohen's d: 0,07; 0,15; 0,08 – все пренебрежимые). Флагманские коммерческие модели (стоимостью $15–75 за 1 млн токенов) не обеспечивают клинически значимого преимущества перед сильными ($0,5–5) или открытыми ($0,1–1). Уровень 4 значимо уступает Уровню 1 (d=0,96, средний эффект) – компактные модели клинически неприемлемы (табл. 3, рис. 2).

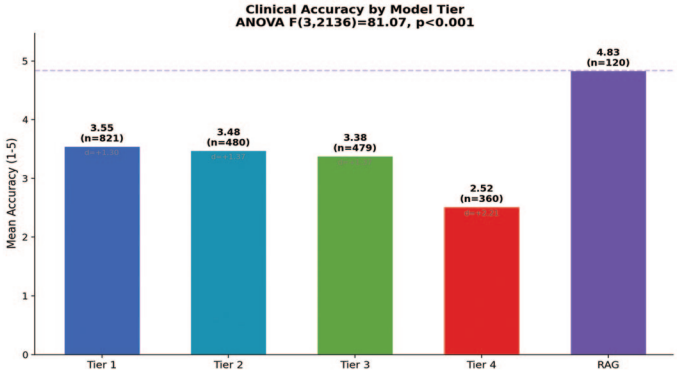

Рис. 2. Средняя точность БЯМ по уровням и RAG-система. Уровни 1, 2 и 3 статистически неразличимы. RAG значимо превосходит все уровни (большой эффект, d=1,30–2,21). Планки погрешностей – 95% доверительные интервалы
Fig. 2. The average accuracy of the LLMs by tiers and the RAG system. Tiers 1, 2, and 3 are statistically indistinguishable. RAG significantly exceeds all tiers (high effect, d=1.30–2.21). Error bars – 95% confidence intervals

Таблица 2. Результаты клинической оценки 18 БЯМ и RAG-системы (120 кейсов, 5-балльная шкала)
Table 2. Results of the clinical evaluation of 18 LLMs and RAG systems (120 cases, 5-point scale)

Модель Model	Уровень Tier	Точность Accuracy	Полнота Exhaustiveness	Безопасность Safety	Эмпатия Empathy	Практическая применимость Practical applicability
RAG (Qwen3+ChromaDB)	**RAG**	**4,83**	**4,88**	**4,92**	—	—
Claude Opus 4.6	1	4,09	3,28	4,01	4,70	4,12
Qwen3 Max	2	3,90	3,47	3,61	4,85	4,08
DeepSeek R1	3	3,89	3,39	3,78	4,75	4,12
Claude Sonnet 4.5	1	3,78	3,19	3,87	4,60	4,06
GPT-4.1	1	3,71	2,94	3,96	4,69	3,68
Qwen3.5-35B	3	3,58	2,83	3,37	4,55	3,38
Mistral Large	2	3,48	3,39	2,82	4,63	3,66
Gemini 2.5 Pro	1	3,46	2,38	3,36	4,63	2,98
Gemini 2.5 Flash	1	3,42	2,31	3,61	4,59	2,84
Llama 4 Maverick	3	3,28	2,50	3,22	4,24	3,12
DeepSeek V3	2	3,27	2,92	2,76	4,39	3,38
Qwen3-80B (локальная)	1	3,26	2,42	2,87	4,50	3,15
Claude Sonnet 4	2	3,24	2,67	3,00	4,29	3,52
GPT-4o	1	3,05	2,20	3,44	4,48	3,02
GPT-4o-mini	4	2,94	2,24	2,90	4,36	2,98
Qwen3 8B	4	2,88	2,60	2,66	4,18	3,04
Llama 3.3 70B	3	2,79	2,07	2,67	4,18	2,61
Llama 3.1 8B	4	1,73	1,35	2,40	2,90	1,64

Примечание. Модели упорядочены по убыванию точности. Прочерк (—) для RAG: эмпатия и практическая применимость оценивались по шкале для врача, где эти метрики не применяются
Note. The models are ordered in descending order of accuracy. Dash (—) for RAG: empathy and practical applicability were assessed on a scale for a doctor where these metrics do not apply

Таблица 3. Средние показатели по уровням (ANOVA F(3,2136)=81,07, p<0,001)
Table 3. Average indicators by tiers (ANOVA F(3,2136)=81,07, p<0,001)

Уровень Tier	Модель Model	Оценок, n Ratings, n	Точность, M±SD Accuracy, M±SD	95% ДИ 95% CI	Безопасность Safety	Проблем/кейс Problems/case study	Cohen's d vs Уровень 1 Cohen's d vs Tier 1
1	7	821	3,55±1,00	3,48–3,62	3,59	1,24	—
2	4	480	3,48±1,04	3,39–3,57	3,05	2,27	0,07 (пренебрежимый/ negligible)
3	4	479	3,38±1,07	3,29–3,48	3,26	2,09	0,15 (пренебрежимый/ negligible)
4	3	360	2,52±1,09	2,41–2,63	2,65	2,87	0,96 (средний/ average)
RAG	1	120	**4,83±0,59**	**4,73–4,94**	**4,92**	—	d vs RAG: 1,30 (большой/big)

Сравнение RAG с «чистыми» БЯМ

RAG превзошла каждую из 18 моделей с большим эффектом (Cohen's d от 1,05 до 3,76) (табл. 4, рис. 3). Даже лучшая «чистая» модель – Claude Opus 4.6 (точность – 4,09) – уступила RAG с d=1,05.

Анализ по доменам

RAG показала стабильное превосходство во всех 15 доменах без исключения (рис. 4). Наибольшая разница зарегистрирована в онкологии почек (+1,89), нижних мочевых путях (+1,91) и хирургических методиках (+1,77). Наименьшая – в детской урологии (+0,97) и травме (+1,01).

Домены с наибольшим превосходством RAG характеризуются высокой степенью специализации и необходимостью точных числовых данных (стадии TNM при раке почки, параметры ДУВЛ/УРС/ПНЛ при мочекаменной болезни). При раке простаты «чистые» БЯМ показали наихудший результат (точность – 2,90), тогда как RAG достигла 4,38 – значительное улучшение (рис. 5).

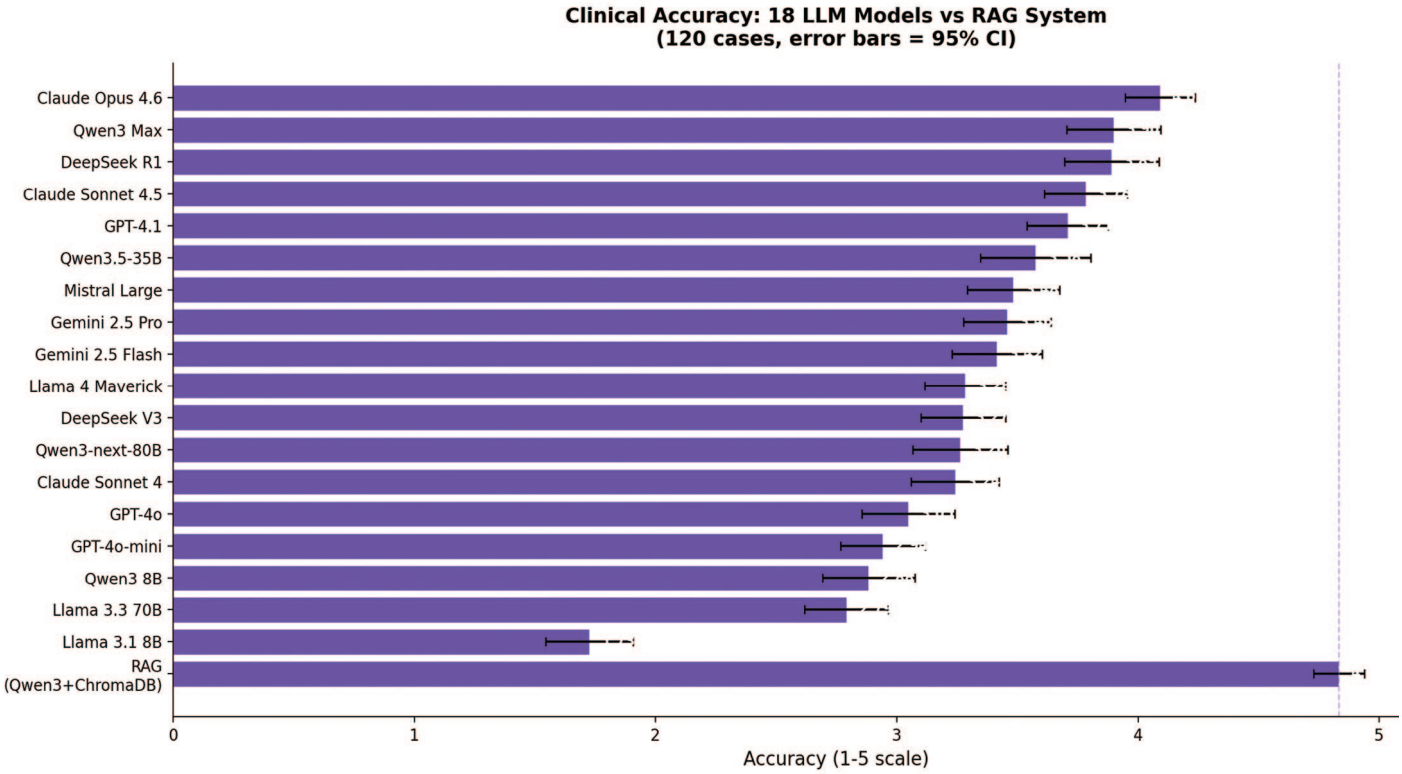

Рис. 3. Точность всех 18 БЯМ и RAG-системы (горизонтальная диаграмма, 95% доверительные интервалы). Фиолетовый цвет – RAG. Пунктирная линия – средняя точность RAG
Fig. 3. Accuracy of all 18 LLMs and RAG systems (horizontal diagram, 95% confidence intervals). The color purple – RAG. The dotted line is the average accuracy of the RAG

Таблица 4. Величина эффекта RAG vs «чистые» БЯМ (Cohen's d)
Table 4. The magnitude of the effect of RAG vs «pure» LLMs (Cohen's d)

Модель Model	Уровень Tier	Точность Accuracy	Точность RAG RAG Accuracy	Cohen's d	Эффект Effect
Claude Opus 4.6	1	4,09	4,83	1,05	большой / big
Qwen3 Max	2	3,90	4,83	1,07	большой / big
DeepSeek R1	3	3,89	4,83	1,08	большой / big
Claude Sonnet 4.5	1	3,78	4,83	1,32	большой / big
GPT-4.1	1	3,71	4,83	1,44	большой / big
Gemini 2.5 Pro	1	3,46	4,83	1,66	большой / big
GPT-4o	1	3,05	4,83	2,21	большой / big
GPT-4o-mini	4	2,94	4,83	2,34	большой / big
Llama 3.1 8B	4	1,73	4,83	3,76	большой / big

Примечание. Показаны все модели Уровня 1 и по одной модели каждого уровня с наибольшим/наименьшим d. Все 18 сравнений показали большой эффект
Note. All models of Level 1 and one model of each level with the highest/lowest d are shown. All 18 comparisons showed a great effect

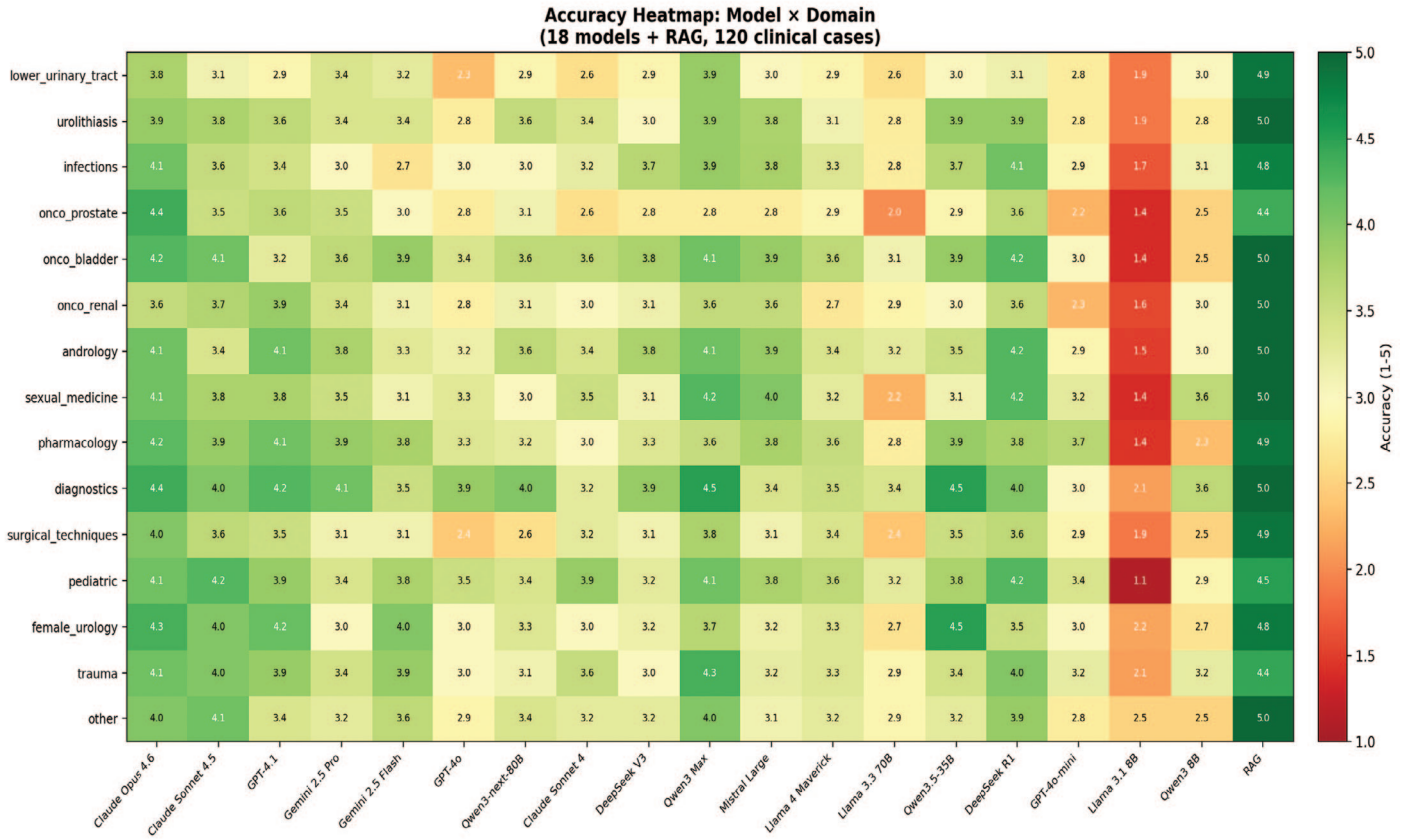

Рис. 4. Точность 18 БЯМ и RAG-системы по 15 клиническим доменам (тепловая карта). RAG (крайний правый столбец) демонстрирует стабильное превосходство во всех доменах
Fig. 4. Accuracy of 18 LLMs and RAG systems for 15 clinical domains (heat map). RAG (rightmost column) demonstrates stable superiority in all domains

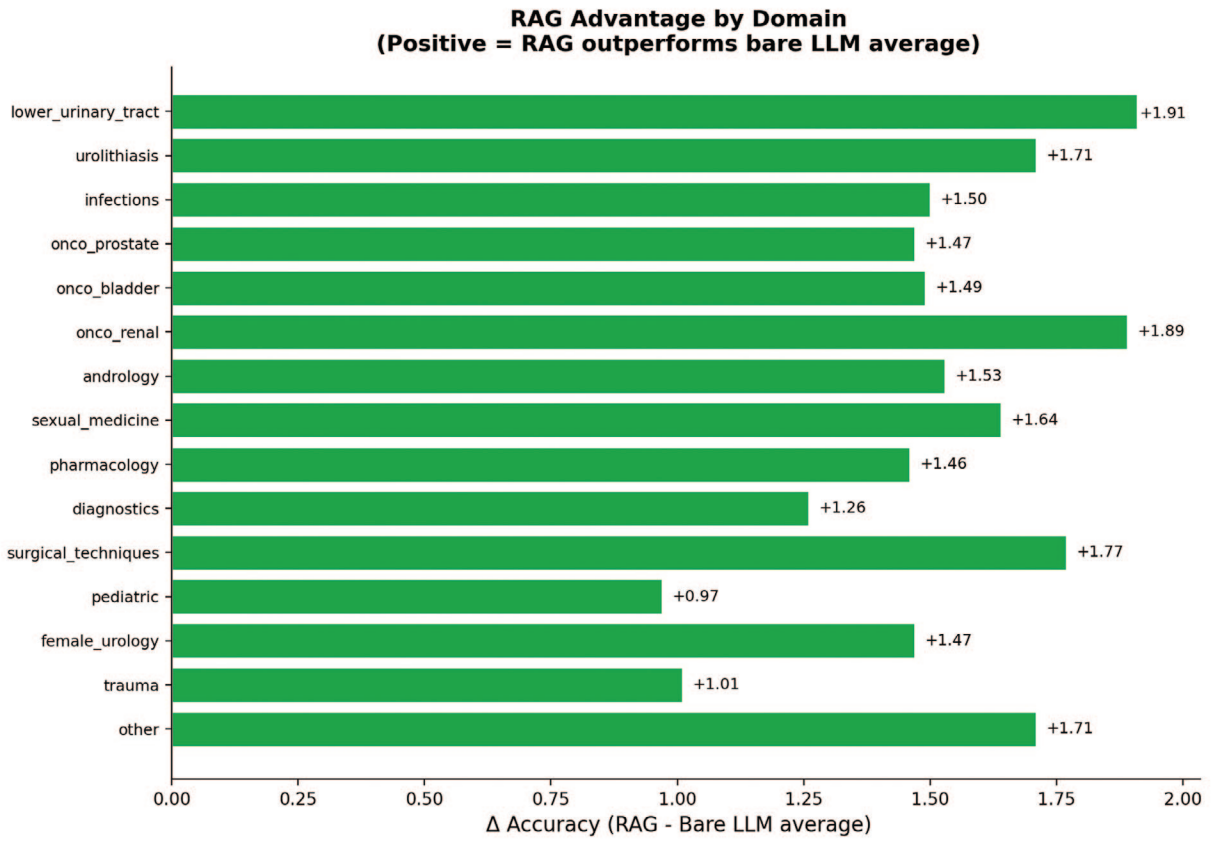

Рис. 5. Преимущество RAG над «чистыми» БЯМ по клиническим доменам (Δ точности). Все значения положительны – RAG превосходит «чистые» БЯМ во всех доменах. Наибольший эффект – нижние мочевые пути (+1,91) и рак почки (+1,89)
Fig. 5. The advantage of RAG over "pure" LLMs in terms of clinical domains (Δ accuracy). All values are positive – RAG surpasses "pure" LLMs in all domains. The greatest effect is lower urinary tract (+1.91) and kidney cancer (+1.89)

Безопасность

«Чистые» БЯМ продемонстрировали значительное число проблем безопасности: от 0,85 (GPT-4.1) до 3,09 (Qwen3 8B) проблем на кейс (табл. 5, рис. 6). Уровень 4 показал наихудший результат – в среднем 2,87 проблем на кейс, в том числе критические ошибки: Llama 3.1 8B – 70 критических проблем на 120 кейсов (58,3% кейсов).

Критические проблемы включали: рекомендацию противопоказанных препаратов (например, ингибиторы ФДЭ-5 при приапизме), пропуск «красных флагов» (гематурия без рекомендации обследования, острая задержка мочи без рекомендации немедленного обращения), неверные дозировки (превышение рекомендованных доз α-блокаторов), рекомендацию хирургического вмешательства без обсуждения консервативных альтернатив.

RAG-система достигла безопасности 4,92/5, поскольку каждый ответ основан на верифицированных источниках (рекомендации EAU/AUA, учебники).

Феномен «эмпатического прикрытия». Примечательно, что эмпатия «чистых» БЯМ была стабильно высокой: от 4,18 до 4,85 из 5. Это создает парадоксальную ситуацию: ответы звучат заботливо, профессионально и убедительно, что может создавать у пациента ложное впечатление клинической компетентности и задержать обращение за реальной медицинской помощью. Данный феномен – «эмпатическое прикрытие» некомпе-

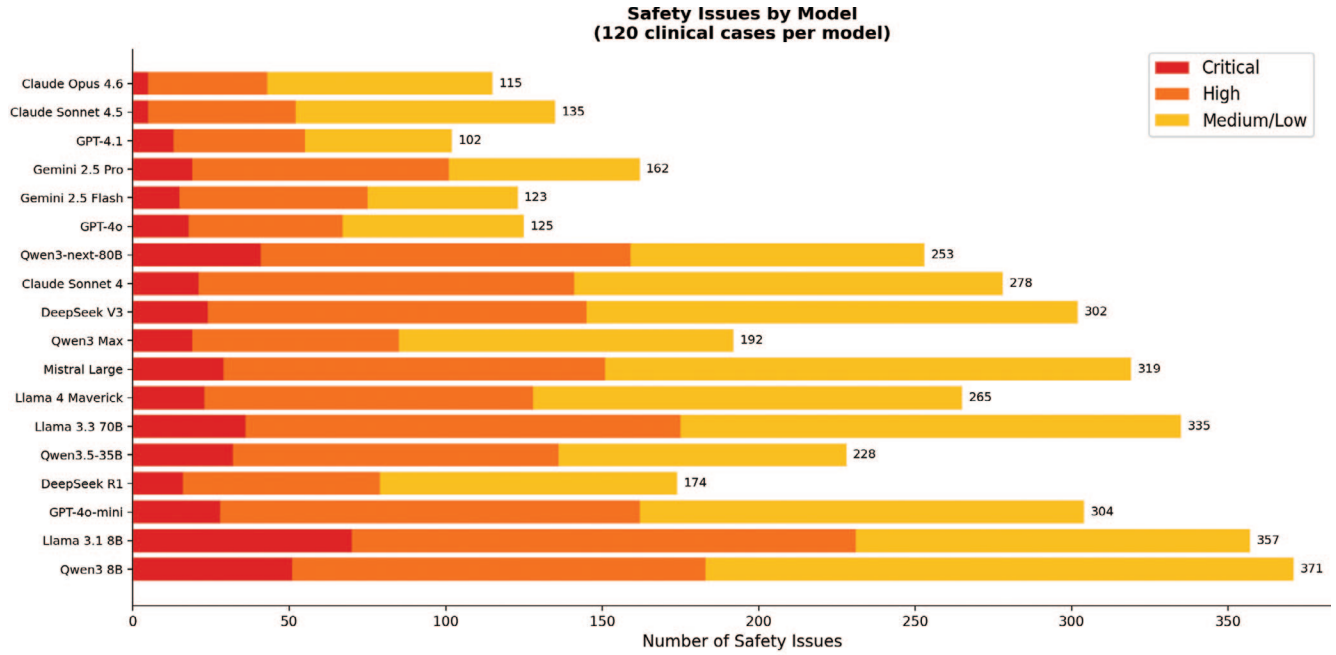

Рис. 6. Проблемы безопасности по моделям. Красный – критические (потенциально опасные для жизни), оранжевый – высокие (клинически значимые), желтый – средние/низкие
Fig. 6. Security issues by model. Red – critical (potentially life–threatening), orange – high (clinically significant), yellow – medium/low

Таблица 5. Проблемы безопасности по моделям
Table 5. Security issues by model

Модель Model	Уровень Tier	Всего проблем Total problems	Критические Critical	Высокие High	Средние/низкие Medium/Low	Проблем/кейс Problems/case study
GPT-4.1	1	102	13	42	47	0,85
Claude Opus 4.6	1	115	5	38	72	0,96
Gemini 2.5 Flash	1	123	15	60	48	1,04
Claude Sonnet 4.5	1	135	5	47	83	1,12
GPT-4o	1	125	18	49	58	1,19
Gemini 2.5 Pro	1	162	19	82	61	1,35
DeepSeek R1	3	174	16	63	95	1,46
Qwen3 Max	2	192	19	66	107	1,60
Qwen3-80B	1	253	41	118	94	2,14
GPT-4o-mini	4	304	28	134	142	2,53
Qwen3 8B	4	371	51	132	188	3,09
Llama 3.1 8B	4	357	70	161	126	2,98

тентности – является новым наблюдением, ранее не описывавшимся в литературе.

Рассуждающие модели (reasoning)

DeepSeek R1 (Уровень 3, открытая модель с механизмом пошагового рассуждения) заняла 3-е место среди всех 18 моделей: точность – 3,89, безопасность – 3,78, превзойдя большинство моделей Уровня 1 (GPT-4.1, GPT-4o, Gemini 2.5 Pro/Flash, Qwen3-80B) и все модели Уровня 2. Механизм пошагового рассуждения (chain-of-thought) позволяет модели анализировать клинический сценарий последовательными шагами, выявляя больше нюансов.

Роль локальной модели в RAG

Qwen3-80B в «чистом» режиме показала умеренные результаты: точность 3,26 (12-е место из 18), безопасность 2,87 (14-е место), 41 критическая проблема. Однако та же модель в составе RAG-системы достигла точности 4,83 и безопасности 4,92 – качественный скачок, обусловленный не заменой модели, а предоставлением верифицированного контекста. Ключевой вывод: проблема галлюцинаций – это проблема отсутствия доступа к достоверным данным, а не проблема «плохой» модели.

Слой эмпатии

RAG-ответы (документ для врача) были конвертированы в формат для пациента через «слой эмпатии» – отдельную языковую модель, переформулирующую врачебный документ в понятный и эмпатичный ответ. Конвертация 120/120 кейсов выполнена успешно (стоимость – $0,16, время – 20 минут). Медицинская точность при конвертации сохранена – все ключевые факты, дозировки и предупреждения переданы корректно.

ОБСУЖДЕНИЕ

Настоящее исследование представляет собой наиболее полную на сегодняшний день оценку БЯМ в клинической урологии: 18 моделей, 4 уровня, 120 клинических случаев по 15 доменам, 2260 оценок. Главный вывод: «чистые» БЯМ всех уровней производят клинически неприемлемое количество галлюцинаций, а RAG-система устраняет этот недостаток с большим эффектом.

Флагманский статус не гарантирует точности

Одно из наиболее неожиданных открытий – отсутствие клинически значимой разницы между Уровнем 1 (флагманские модели, стоимость $15–75 за 1 млн токенов) и Уровнями 2/3 ($0,1–5). Cohen's d между Уровнями 1 и 2 составил всего 0,07 (пренебрежимый). Практическое следствие: нет необходимости использовать наиболее дорогие модели для медицинских задач. Возможное объяснение: все модели обучены на сопоставимых корпусах, и урологический контент составляет в них относительно небольшую долю.

Безопасность – критическая проблема

Даже лучшие модели допускают опасные рекомендации: GPT-4.1 – 13 критических проблем на 120 кейсов (10,8%), Gemini 2.5 Pro – 19 (15,8%), Qwen3-80B – 41 (34,2%). При этом эмпатия моделей стабильно высока (4,18–4,85), создавая феномен «эмпатического прикрытия» – модель звучит убедительно, и пациент может не заметить фактически неверные рекомендации.

RAG как архитектурное решение

RAG-система достигла практически максимальных показателей (точность – 4,83/5, безопасность – 4,92/5). Преимущество обусловлено не качеством генерации, а качеством извлечения: система находит релевантные фрагменты из 17 947 документов и предоставляет их модели как контекст. Дополнительное преимущество – объяснимость: каждый тезис сопровождается ссылкой на конкретный источник (учебник, гайдлайн, обзор), что принципиально невозможно при использовании «чистых» БЯМ.

Критически важно, что система полностью работает локально, на компьютере врача, без передачи данных в интернет. Согласно Федеральному закону № 152-ФЗ «О персональных данных» и статье 13 Федерального закона № 323-ФЗ «Об основах охраны здоровья граждан» (врачебная тайна), передача персональных и медицинских данных пациентов сторонним организациям без информированного согласия запрещена. Использование облачных сервисов (ChatGPT, Claude, Gemini) для обработки данных пациентов нарушает эти требования. Локальная RAG-система обеспечивает полное соответствие законодательству – данные остаются на компьютере врача или в контуре клиники.

RAG как инструмент врача, а не пациента

Необходимо подчеркнуть: RAG-система, несмотря на высокие показатели точности и безопасности, является инструментом поддержки принятия врачебных решений, а не самостоятельным источником медицинских рекомендаций. Окончательное клиническое решение принимает врач, оценивая RAG-ответ в контексте конкретного пациента. Предоставление RAG-ответов напрямую пациентам без врачебной интерпретации сопряжено с рисками: пациент может неверно интерпретировать информацию, не учесть индивидуальные особенности, проигнорировать необходимость очного осмотра.

RAG vs дообучение модели (fine-tuning)

Альтернативным подходом является дообучение модели на урологическом корпусе. Однако RAG имеет преимущества:

1. Объяснимость – каждый факт привязан к источнику.

2. Актуальность – база знаний обновляется без переобучения.

3. Верифицируемость – врач проверяет любую рекомендацию.

4. Экономичность – добавление новых данных не требует ресурсоемкого дообучения.

Перспективы использования

RAG-архитектура открывает ряд перспектив для клинической практики урологии:

1. Формализованный медицинский текст. RAG-ответ может быть преобразован в формализованную медицинскую документацию: жалобы, анамнез заболевания, анамнез жизни, объективное обследование, план лечения. Первые эксперименты показывают, что локальные БЯМ вполне справляются с задачей генерации формализованного текста, что позволяет автоматически создавать записи для медицинской карты, назначения для медсестер, выписные эпикризы. Однако генерация текста в формате аттрактика (структурированного заключения) требует строгих промптов и жестких правил и остается предметом дальнейших исследований – не все локальные модели справляются с этой задачей достаточно надежно [22–25].

2. Расширение базы знаний: лекарственные препараты. Текущая база знаний содержит информацию о препаратах, упомянутых в урологических источниках. Однако пациенты с коморбидной патологией (сахарный диабет, артериальная гипертензия, ишемическая болезнь сердца) требуют информации о препаратах, выходящей за рамки урологии. В настоящее время ведется работа по созданию отдельного слоя базы данных по лекарственным препаратам (на основе Государственного реестра лекарственных средств Российской Федерации), который будет интегрирован в RAG-систему для учета межлекарственных взаимодействий и противопоказаний при коморбидности [22–25].

3. Расширение базы знаний: методы диагностики. Для полноценного использования системы в клинической практике потребуется дополнительный слой, посвященный методам диагностики: урофлоуметрия (параметры, интерпретация), ультразвуковое исследование (нормативные значения, патологические признаки), уродинамика, компьютерная томография, магнитно-резонансная томография. В существующей базе эти данные частично представлены в текстах учебников и гайдлайнов, однако отдельный структурированный слой обеспечит более точный поиск по диагностическим запросам [22–25].

4. Локальное развертывание в клинике. Полностью локальная архитектура (языковая модель + база знаний на компьютере врача или сервере клиники) позволяет развернуть систему во внутриклиническом контуре без подключения к интернету. Это особенно актуально для учреждений с ограничениями на передачу данных во внешние сети (военные госпитали, государственные клиники) [22–25].

Ограничения исследования

1. Метод «языковая модель как эксперт» не эквивалентен оценке врачом-экспертом. Возможна систематическая погрешность, связанная с предпочтениями модели-оценщика.

2. Различные шкалы оценки для «чистых» БЯМ и RAG не позволяют прямого сравнения по всем метрикам – сравнение проведено только по точности и безопасности.

3. Клинические кейсы основаны на онлайн-консультациях, что может не отражать полного спектра урологической патологии.

4. RAG-система тестировалась в фиксированной конфигурации; другие комбинации могут дать иные результаты.

5. Исследование проведено в апреле 2026 г.; быстрое развитие моделей может изменить абсолютные показатели, однако выявленные закономерности (Уровни 1–3 статистически неразличимы, RAG >> «чистые» БЯМ) обусловлены архитектурными ограничениями и вероятно сохранятся.

6. Отсутствие независимой врачебной оценки всех 2260 ответов – ограничение, диктуемое масштабом. Независимая валидация выборки экспертами – приоритет для дальнейших исследований.

7. Система не содержит отдельного слоя по лекарственным препаратам и методам диагностики, что ограничивает полноту ответов при коморбидных пациентах.

Практические рекомендации

1. «Чистые» БЯМ любого уровня не следует использовать как источник медицинской информации без верификации врачом.

2. RAG-системы обеспечивают клинически приемлемую точность и могут применяться как инструмент поддержки принятия врачебных решений.

3. Для RAG-систем достаточно моделей уровня 2–3 – нет необходимости в наиболее дорогих флагманских моделях.

4. Компактные модели (уровень 4, ≤8 млрд параметров) непригодны для медицинского применения.

5. При коммуникации с пациентом RAG-ответ может быть конвертирован в понятный формат через «слой эмпатии» без потери медицинской точности.

ЗАКЛЮЧЕНИЕ

1. «Чистые» БЯМ всех уровней демонстрируют высокий уровень галлюцинаций в клинической урологии: средняя точность 2,52–3,55 из 5 баллов, частота проблем безопасности – 1,24–2,87 на кейс. Ни одна из 18 моделей не достигла клинически приемлемого уровня (≥4,5/5).

2. Флагманский статус модели не определяет клиническую точность: различия между уровнями 1, 2 и 3 статистически незначимы (Cohen's d=0,07–0,15). При этом уровень 4 (≤8 млрд параметров) значимо уступает (d=0,96).

3. RAG-система на базе верифицированной базы знаний (17 947 документов, 2,23 млн фрагментов, 19 доменов) превосходит все «чистые» БЯМ с большим эффектом (Cohen's d = 1,05–3,76, p<0,001), достигая точности 4,83/5 и безопасности 4,92/5.

4. Проблема галлюцинаций – не проблема конкретной модели, а архитектурная проблема: та же модель Qwen3-80B, которая в «чистом» режиме показывает точность 3,26 и 41 критическую проблему, в составе RAG достигает точности 4,83 и безопасности 4,92.

5. Эмпатия «чистых» БЯМ (4,18–4,85) создает у пациента ложное ощущение компетентности («эмпатическое прикрытие»), что может задержать обращение за медицинской помощью.

6. Рассуждающие модели (DeepSeek R1) показывают результаты, сопоставимые с флагманскими, что открывает перспективу комбинации механизмов рассуждения с RAG-архитектурой.

7. RAG-система является инструментом поддержки принятия врачебных решений и не должна использоваться как самостоятельный источник рекомендаций для пациентов.

ЛИТЕРАТУРА/REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med.* 2019;25(1):44–56. https://doi.org/10.1038/s41591-018-0300-7.
2. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med.* 2023;29(8):1930–40. https://doi.org/10.1038/s41591-023-02448-8.
3. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med.* 2023;183(6):589–96. https://doi.org/10.1001/jamainternmed.2023.1838.
4. Lee P, Goldberg C, Kohane I. The AI Revolution in Medicine: GPT-4 and Beyond. Pearson; 2023.
5. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, et al. A survey of large language models. *arXiv.* 2023;2303.18223.
6. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *arXiv.* 2023;2311.05232.
7. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare.* 2023;11(6):887. https://doi.org/10.3390/healthcare11060887.
8. Singhal K, Azizi S, Tu T, Said S, Levi R, Mahdavi SS, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172–80. https://doi.org/10.1038/s41586-023-06291-2.
9. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv.* 2023;2303.13375.
10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLOS Digit Health.* 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198.
11. Huh S. Are ChatGPT's knowledge and ability accurate enough for use in medical education? *J Med Educ.* 2023;42(4):461–4. https://doi.org/10.3946/kjme.2023.276.
12. Bak D, Choi B, Paciorek M, Kim JS. Assessment of GPT-4 for urology board-style questions. *Urology.* 2024;189:61–6. https://doi.org/10.1016/j.urology.2024.02.030.
13. Plaza L von der, Borkowska EM, Reitz D, Nestler S, Goutzonis N, Kalogirou TE, et al. Accuracy of ChatGPT-4 regarding clinical questions on bladder cancer. *BJU Int.* 2024;133(4):496–503. https://doi.org/10.1111/bju.16350.
14. Coentro LMQ, Ziegelmann G, Warren G, James BC. Accuracy of ChatGPT in providing recommendations for urological care. *Urology.* 2024;192:1–6. https://doi.org/10.1016/j.urology.2024.06.025.
15. Campbell-Walsh Urology. 12th edition. Wein AJ, Kavoussi LR, Partin AW, Peters CA, editors. Philadelphia: Elsevier; 2021.
16. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-augmented generation for large language models: a survey. *arXiv.* 2024;2312.10997.
17. Xiong G, Jin Q, Lu Z, Zhang A. Benchmarking retrieval-augmented generation for medicine. *arXiv.* 2024;2402.13178.
18. Chen J, Lin H, Han X, Sun L. Benchmarking LLM evaluation: are LLMs good judges? *arXiv.* 2026;2406.12603.
19. Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
20. Zheng L, Chiang WL, Sheng L, Zhuang S, Wu Z, Zhuang Y, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Adv Neur Inf Proc Syst.* 2023;36.
21. Bornet P, Ducret B, Wulle P, Khatibi A, Pignaton N, Gudel M, et al. The role of AI and LLMs in transforming evidence-based medicine. *J Med Internet Res.* 2024;26:e54378. https://doi.org/10.2196/54378.
22. European Association of Urology. EAU Guidelines. Edition presented at the EAU Annual Congress 2025. Arnhem: EAU Guidelines Office; 2025.
23. Wang X, Wang Z, Gao X, Ding Y, Guo Y, Zeng X. Performance of ChatGPT on urology licensure examinations: evaluating accuracy, consistency, and the impact of prompts. *World J Urol.* 2024;42(1):136. https://doi.org/10.1007/s00345-023-04681-4.
24. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst.* 2023;47(1):33. https://doi.org/10.1007/s10916-023-01925-2.
25. Zaremba P, Dembowski J, Rola R. Large language models in urology: a systematic review of current applications and future directions. *Urol Int.* 2025;109(1):12–24. https://doi.org/10.1159/000539021.

Сведения об авторах:

Шадёркин И.А. – к.м.н., уролог, генеральный директор ООО «Робоскоп Патолоджи», Москва, Россия; РИНЦ Author ID: 695560, https://orcid.org/0000-0001-8669-2674, Scopus Author ID: 55022298600

Шадёркина В.А. – уролог, научный редактор урологического информационного портала UroWeb.ru, Москва, Россия; РИНЦ Author ID: 880571, https://orcid.org/0000-0002-8940-4129

Information about authors:

Shaderkin I.A. – PhD, urologist, Roboskop Pathology LLC, Moscow, Russia; RSCI Author ID: 695560, https://orcid.org/0000-0001-8669-2674, Scopus Author ID: 55022298600

Shaderkina V.A. – urologist, Scientific editor of the urological information portal UroWeb.ru, Moscow, Russia; RSCI Author ID: 880571, https://orcid.org/0000-0002-8940-4129

Вклад авторов:
Шадёркин И.А. – концепция исследования, дизайн, проведение исследования, анализ, написание текста, 50%
Шадёркина В.А. – концепция исследования, дизайн, сбор данных, анализ, написание текста, 50%

Authors' contributions:
Shaderkin I.A. – research concept, design, conducting research, analysis, writing, 50%
Shaderkina V.A. – research concept, design, data collection, analysis, writing, 50%

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest. The authors declare no conflict of interest.

Финансирование. Статья подготовлена без финансовой поддержки.

Financing. The article was made without financial support.

Статья поступила: 01.05.2026

Received: 01.05.2026

Результаты рецензирования: 17.05.2026

Peer review: 17.05.2026

Исправления получены: 20.05.2026

Corrections received: 20.05.2026

Принята к публикации: 25.05.2026

Accepted for publication: 25.05.2026