Иллюзия готовности: почему медицинский искусственный интеллект всё ещё не готов стать врачом

Дата публикации: 08.10.2025

Искусственный интеллект всё чаще рассматривается как инструмент, способный изменить медицину, сделав диагностику точнее, а помощь — доступнее. Однако новое исследование Microsoft Research в области Health & Life Sciences показывает, что, несмотря на впечатляющие результаты в тестах, «врач» на базе ИИ пока не готов к приёму пациентов. За кажущейся компетентностью скрываются уязвимости, которые могут привести к неверным диагнозам в реальной практике.

Современные мультимодальные медицинские системы искусственного интеллекта — такие как GPT-5, Gemini-2.5 Pro, OpenAI-o3 и o4-mini — демонстрируют нестабильное поведение при стрессовых испытаниях. При незначительных изменениях в подсказках модели меняют ответы, иногда дают правильный результат по чистой догадке, а порой уверенно рассуждают на основе ложных визуальных деталей. Исследователи отмечают, что такие модели часто создают «иллюзию понимания»: они формулируют убедительные, но не

всегда достоверные объяснения.

Проблема кроется в самой методике оценки. Бенчмарки, широко используемые для измерения прогресса, фокусируются на сопоставлении с эталонными ответами, а не на способности к клиническому мышлению. Это приводит к тому, что модели обучаются «угадывать правильный ответ», а не действительно понимать медицинский контекст. По мере совершенствования архитектур ИИ растут показатели точности, но одновременно увеличивается риск скрытых уязвимостей.

В исследовании под названием «Иллюзия готовности: стресс-тестирование крупных мультимодальных медицинских моделей» (arXiv, 2025) был предложен новый подход к оценке. Учёные создали серию стресс-тестов, которые проверяли устойчивость моделей к шуму, зависимости от визуального ввода и способность рассуждать при изменении структуры данных.

В тестах использовались шесть медицинских систем ИИ, проверенных по шести наборам данных: фильтрованным пунктам JAMA (1141 задач), NEJM (743 задачи), клинически подобранным случаям NEJM, требующим визуальных данных (175 задач), и отдельному набору из 40 визуальных замен. Модели оценивались как по «текстовым» задачам, так и по мультимодальным случаям, где требовалось анализировать изображение и описание одновременно.

Результаты оказались неоднозначными. Например, удаление изображений из заданий NEJM резко снижало точность GPT-5 с 80,9% до 67,5%, а Gemini-2.5 Pro — с 79,9% до 65,0%. OpenAI-o3 показала схожее снижение — с 80,9% до 67,0%. При этом GPT-4o продемонстрировал аномальное поведение — его точность даже выросла при подмене изображений. Это свидетельствует о том, что модель могла угадывать ответы по тексту, а не использовать визуальные данные.

Тесты, требующие визуального ввода, показали, что даже лучшие модели нередко выдают ответы выше случайного уровня (около 20%) без анализа изображений. Для клинического применения такой эффект недопустим: врач не может «угадывать» диагноз, опираясь лишь на догадки. При нарушении структуры подсказок и изменении порядка ответов большинство моделей также снижали точность, что указывает на поверхностное распознавание паттернов, а не на понимание сути.

Особенно показательные результаты были получены в тестах с контрфактуальными визуальными заменами, когда изображение подменялось похожим, но нерелевантным. Точность всех моделей упала почти вдвое: GPT-5 — с 83% до 51%, Gemini-2.5 Pro — с 80% до 47%, OpenAI-o3 — с 76% до 52%. Это демонстрирует, что ИИ не способен надёжно отличать медицински значимую

информацию от отвлекающих факторов.

Анализ рассуждений также выявил интересный феномен. Во многих случаях модели давали правильный ответ, но логика их объяснений была ошибочной или содержала вымышленные детали. Например, описывалась несуществующая опухоль на изображении, которая «подтверждала» диагноз. Такие случаи создают опасное впечатление уверенности, особенно если система работает в роли помощника врача.

Авторы исследования подчёркивают: даже высокий процент правильных ответов не гарантирует клинической пригодности. Результаты бенчмарков могут маскировать нестабильное поведение, чрезмерное упрощение рассуждений и зависимость от шаблонов. Реальная медицина требует систем, способных не просто распознавать паттерны, но и адекватно работать в условиях неопределённости, неполноты данных и изменчивости контекста.

Исследователи предлагают новую систему оценки медицинского ИИ, основанную на трёх принципах: систематическое стресс-тестирование (в том числе с контрфактуальными сценариями), прозрачная документация логики и визуальных зависимостей, а также отчётность по надёжности и устойчивости решений. Только при соблюдении этих условий, по мнению авторов, искусственный интеллект сможет претендовать на доверие в медицинской практике.

Таким образом, «врач» на базе искусственного интеллекта действительно становится умнее — но не надёжнее. Пока системы демонстрируют эффектный рост показателей, на деле они остаются уязвимыми перед неожиданностями реального мира. Чтобы заменить или даже дополнить живого врача, ИИ предстоит не просто научиться отвечать правильно, но и понимать, почему его ответ верен.

> **Ссылка:** «Иллюзия готовности: стресс-тестирование моделей крупных пограничных состояний на мультимодальных медицинских эталонах» DOI: 10.48550/arxiv.2509.18234.