Эффективность больших языковых моделей в предоставлении точной и понятной информации пациентам о сердечной недостаточности и кардиомиопатии

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1847603

LLM | диагностика | информирование пациентов | кардиология | оценка качества

Введение

Большие языковые модели (LLMs) все чаще используются пациентами, ищущими информацию о здоровье сердечно-сосудистой системы через цифровые платформы. Однако их точность и пригодность для предоставления рекомендаций по таким гетерогенным заболеваниям, как **кардиомиопатии** и **сердечная недостаточность**, остаются недостаточно оцененными. В данном исследовании проводилось систематическое сравнительное тестирование современных LLM на предмет клинической уместности и понятности ответов на вопросы пациентов, касающиеся сердечной недостаточности и кардиомиопатии.

Методы

Шесть известных чат-ботов на базе LLM были протестированы на 50 отобранных экспертами вопросах, охватывающих понимание заболевания и советы по образу жизни. Веб-платформа для оценки рандомизировала и скрывала ответы для проведения экспертизы двенадцатью рецензентами (кардиологами, студентами-медиками и автоматизированными системами

оценки на базе ИИ). Ответы оценивались по 5-балльной шкале Лайкерта в девяти областях, включая уместность, читабельность и эмпатию. Рецензенты также выбирали предпочтительную модель для каждого вопроса.

Результаты

Лингвистическая сложность и длина ответов существенно различались. **Gemini** предоставила наиболее читабельные ответы (индекс удобочитаемости Флеша — Кинкейда 11,3 ± 1,9), но при этом была одной из самых многословных (668,7 ± 116,1 слова). По результатам 2700 оценок, Gemini получила самый высокий средний совокупный балл (4,41 ± 0,77), продемонстрировав отличные показатели полноты и фактической достоверности, за ней следовал **Grok** (4,23 ± 0,76). Показатель избегания конфабуляций (галлюцинаций) был стабильно высоким у всех моделей (4,49 ± 0,02), в то время самый низкий балл был у показателя лаконичности (3,81 ± 0,05). Как правило, эксперты выбирали Gemini в качестве предпочтительного источника информации в 43,7% случаев, за ней следовал Grok (30,3%). Тенденции оценки различались в зависимости от группы экспертов: автоматизированные системы оценки выставили самые высокие средние баллы (среднее 4,58 ± 0,60), за ними следовали студенты (4,10 ± 0,88), в то время как эксперты были более консервативны (3,79 ± 0,93).

Обсуждение

Все LLM продемонстрировали хорошую точность, избегая распространения медицинской дезинформации, хотя наблюдается вариативность в читабельности и полноте ответов. Несмотря на то, что в ходе нашего слепого тестирования серьезные фактические ошибки или галлюцинации встречались редко, они не отсутствовали полностью.
