Бенчмаркинг больших языковых моделей и клиницистов с использованием локально сгенерированных сценариев первичной медико-санитарной помощи в Кении

Источник: BMJ Digital Health & AI

Оригинал: https://bmjdigitalhealth.bmj.com/cgi/content/short/2/1/e000081?rss=1

LLM | клиническая практика | медицинская безопасность | первичная медико-санитарная помощь | сравнительный анализ

Цель

Определить способность больших языковых моделей (LLMs — Large Language Models) точно отвечать на открытые вопросы в сфере здравоохранения (в отличие от вопросов с множественным выбором) в условиях ограниченных ресурсов.

Методы и анализ

Мы провели сравнительный анализ пяти LLM (GPT-4.1, Gemini-2.5-Flash, DeepSeek-R1, MedGemma и o3) с показателями кенийских клиницистов, используя набора данных из 507 случайно выбранных клинических случаев (из более крупного пула в 5107 сценариев), охватывающих 12 категорий сестринских компетенций. Группы врачей в условиях «слепого» тестирования оценивали ответы по 5-балльной шкале Лайкерта, используя рубрикатор из 11 областей, охватывающий точность, безопасность, контекстуальную уместность и коммуникацию. Мы суммировали средние

баллы и использовали байесовскую порядковую логистическую регрессию для оценки вероятности получения высоких оценок (≥4) и для проведения попарных сравнений между LLM и клиницистами.

Результаты

Средние оценки клиницистов были ниже, чем у LLM, в 9 из 11 областей: 2,86 против 4,25–4,72 (соответствие клиническим рекомендациям), 2,76 против 4,25–4,73 (экспертные знания), 2,96 против 4,30–4,73 (логическая связность) и 2,58 против 4,16–4,68 (низкая частота пропуска критически важной информации). В областях, связанных с безопасностью, LLM получили более высокие оценки: минимальная степень возможного вреда — 3,16 против 4,29–4,68; низкая вероятность причинения вреда — 3,68 против 4,54–4,81. Показатели были схожими в отношении низкого содержания нерелевантного контента (4,28 против 4,25–4,35) и избегания демографической предвзятости (4,86 против 4,91–4,94). В байесовских моделях LLM имели >90% вероятность получения оценок ≥4 в большинстве областей, тогда как клиницисты превышали этот порог только в категориях контекстуальной релевантности и демографической/социально-экономической предвзятости. Попарные сравнения показали широкое перекрытие доверительных интервалов среди LLM, при этом модель o3 имела численное лидерство в большинстве областей, за исключением контекстуальной релевантности, демографической/социально-экономической предвзятости и соответствия вопросу. Генерация всех ответов LLM стоила от 3,86 до 8,68 долларов США на одну модель (0,008–0,017 доллара США за один клинический случай) по сравнению с 3,35 долларами США за один случай, подготовленный клиницистом.

Выводы

В контролируемых задачах на основе клинических сценариев LLM генерировали ответы, которые были более точными, безопасными и структурированными, чем ответы клиницистов, что позволяет предположить потенциал использования LLM в качестве вспомогательных инструментов поддержки знаний и безопасности. Полученные результаты обосновывают необходимость дальнейшей оценки в условиях реального взаимодействия с пациентами для определения эффективности, безопасности и возможности интеграции в клинические рабочие процессы, особенно в системах здравоохранения с ограниченными ресурсами.