

Оценка и обратная связь больших языковых моделей по транскрибированным симулированным вербальным взаимодействиям врач-пациент

Источник: Frontiers in AI — Medicine

Дата публикации: 2025-05-01

Оригинал: <https://www.frontiersin.org/articles/10.3389/frai.2026.1936749>

LLM

диагностика

клиническая коммуникация

медицинское образование

оценка навыков

стандартизированные пациенты

Введение

Устные взаимодействия между врачом и пациентом являются фундаментальной частью клинической практики и медицинского образования, однако их оценка требует значительных временных затрат и подвержена межэкспертной вариабельности. Мы оценили возможность точной оценки транскрибированных стандартизированных устных взаимодействий между пациентом и врачом со стороны больших языковых моделей (LLM) и генерации полезной текстовой обратной связи.

Методы

В рамках этого одноцентрового исследования 12 стандартизированных встреч между пациентом и студентом-медиком были независимо оценены моделями ChatGPT-4o, Gemini 3 Pro и врачом-рецензентом с использованием золотого стандарта рубрики, состоящей из 55 элементов сбора анамнеза и 19 элементов презентации случая, оценивающих диагностическое мышление, планирование лечения и консультирование пациента. Согласованность

оценок оценивалась с использованием коэффициента каппа Коэна и внутриклассовых коэффициентов корреляции. Текстовая обратная связь, сгенерированная LLM, была независимо оценена моделями ChatGPT, Gemini, Claude и врачом-рецензентом.

Результаты

В среднем баллы по всем экспертам были схожими: было выявлено 35–36 из 55 элементов сбора анамнеза и 13–14 из 19 элементов презентации случая. Согласованность была высокой для элементов сбора анамнеза ($\kappa = 0,84–0,89$; внутриклассовые коэффициенты корреляции = $0,93–0,97$) и более низкой для элементов презентации случая ($\kappa = 0,62–0,72$; внутриклассовые коэффициенты корреляции = $0,61–0,83$). Обе LLM сгенерировали подробную, точную и клинически полезную текстовую обратную связь.

Обсуждение

Эти результаты демонстрируют, что LLM могут надежно оценивать смоделированные взаимодействия между врачом и пациентом и могут поддерживать стандартизированную оценку клинического общения для целей исследований, улучшения качества и медицинского образования.