Помимо точности: оценка надежности больших языковых моделей для медицинского тестирования

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1832829

LLM | автоматизация оценки | медицинское образование | надежность ИИ | эндокринология

Введение

Большие языковые модели (**LLM** — Large Language Models) демонстрируют хорошие результаты при прохождении медицинских экзаменов, однако их почти всегда оценивают как участников тестирования, судя по точности их ответов. Нам гораздо меньше известно об их надежности в качестве инструментов автоматизации оценки — например, при извлечении метаданных экзаменационных заданий — а также о том, могут ли бенчмарки (эталонные тесты), основанные на точности в рамках одного прогона, адекватно характеризовать такие инструменты.

Методы

Двенадцать основных **LLM** (8 отечественных, 4 международных) оценивались на предмет извлечения восьми полей метаданных из эндокринологического экзамена, состоящего из 60 вопросов; каждая модель выполняла идентичную задачу в трех независимых прогонах в реалистичных веб-условиях. Показатели эффективности разделялись на коэффициент завершения (**completion rate**), условную точность (**conditional accuracy**) и показатель успеха задачи (**TSR** — Task Success Rate), определяемый как

правильное заполнение всех полей; при этом основным результатом считалась дисперсия между прогонами. Условная точность и **TSR** сравнивались внутри моделей, а отечественные и международные модели сравнивались по показателю **TSR**.

Результаты

Условная точность была единообразно высокой, однако показатель **TSR** имел резко бимодальное распределение: пять моделей набрали 90% и выше, а семь — ниже 70%, при этом промежуточных значений не было. Решающее значение имела надежность, а не средняя точность. Несколько моделей с идеальной условной точностью потерпели неудачу, когда один или два из трех их прогонов полностью провалились — этот паттерн был бы упущен при однократной оценке. Сбои охватывали как поведение моделей, так и ограничения на уровне платформы, последние из которых невозможно устранить с помощью промптинга (составления запросов). Происхождение модели не позволяло предсказать производительность. В одном из пунктов четыре модели независимо изменили назначенный экзаменатором когнитивный уровень на один из своих собственных — это «скрытая перемаркировка», имеющая прямые последствия для целостности банка заданий.

Заключение

Для автоматизированного извлечения метаданных оценки именно надежность, а не точность, определяет, может ли **LLM** быть внедрена, при этом национальное происхождение не является значимым предиктором. Такие инструменты следует оценивать путем многократных прогонов с отчетом о дисперсии между ними и внедрять посредством целевого пилотного тестирования в рамках рабочего процесса, который оставляет семантические и нормативные поля на усмотрение человека.
