Сравнение способности трех больших языковых моделей оценивать риск систематической ошибки с помощью инструмента ROBINS-I

Источник: BMJ Digital Health & AI

Оригинал: https://bmjdigitalhealth.bmj.com/cgi/content/short/2/1/e000034?rss=1

LLM | анализ данных | валидация ИИ | клиническая методология

Цели

Данное исследование направлено на сравнение надежности и точности трех больших языковых моделей (LLM) (Claude, Gemini и GPT) при оценке риска систематической ошибки в нерандомизированных исследованиях с использованием инструмента **ROBINS-I** (Risk Of Bias In Non-randomized Studies of Interventions — оценка риска систематической ошибки в нерандомизированных исследованиях вмешательств).

Методы и анализ

Мы провели вторичный анализ 171 нерандомизированного исследования, которые ранее были оценены двумя независимыми группами экспертов-людей с помощью инструмента **ROBINS-I**. В анализ были включены только те исследования, в которых оценки экспертов на уровне отдельных областей (доменов) совпали. Каждое исследование дважды независимо оценивалось моделями Claude, Gemini и **GPT** (Generative Pre-trained Transformer — генеративный предварительно обученный трансформер) с использованием агентских структурированных реализаций инструмента **ROBINS-I**. Надежность (согласованность между двумя запусками одной и той же LLM) оценивалась с помощью процента согласия и коэффициента **Gwet’s AC1**.

Точность (согласованность с экспертами-людьми) оценивалась только для исследований с последовательными оценками LLM с использованием тех же метрик.

Результаты

Claude продемонстрировал высокую надежность во всех областях (согласие 79,5–98,0%, AC1 = 0,729–0,975). **Gemini** показал умеренно высокую надежность (согласие 76,7–100%, AC1 = 0,680–1,0). **GPT** в целом продемонстрировал более низкую надежность, хотя согласие на уровне отдельных областей варьировалось от 70,9% до 95,6% (AC1 = 0,596–0,944). Что касается точности, **Claude** показал в целом низкое соответствие экспертам-людям (согласие 14,4–68,5%; низкие значения AC1). **Gemini** продемонстрировал умеренную или высокую точность в нескольких областях, включая отклонения от намеченных вмешательств (79,6%, AC1 = 0,848) и измерение исходов (73,9%, AC1 = 0,702), показав самый высокий общий уровень согласия (40,0%, AC1 = 0,672). **GPT** показал переменную точность, наиболее высокую в измерении исходов (62,8%, AC1 = 0,571) и классификации вмешательств (57,8%, AC1 = 0,498), но низкую эффективность в отборе (14,3%, AC1 = -0,041) и в общем согласии (23,0%, AC1 = 0,267).

Выводы

Модель **Claude** была внутренне последовательной, но плохо согласованной с экспертами-людьми. **Gemini** достигла как высокой надежности, так и умеренно высокой точности, в то время как **GPT** имел более низкую надежность и смешанную точность. Современные готовые к использованию LLM не могут надежно проводить оценку риска систематической ошибки по методу **ROBINS-I**.
