

Расхождение между медицинскими знаниями и клиническим мышлением: оценка больших языковых моделей в пренатальной диагностике

Источник: Frontiers in Digital Health

Дата публикации: 2026-09-15

Оригинал: <https://www.frontiersin.org/articles/10.3389/fdgth.2026.1917573>

LLM

клиническое мышление

медицинское образование

оценка ИИ

пренатальная диагностика

ультразвуковая диагностика

Цель: Большие языковые модели (LLM) продемонстрировали впечатляющие результаты при прохождении различных медицинских экзаменов; однако их компетентность в области субспециализированного клинического мышления, особенно в пренатальной диагностике, остается плохо изученной.

Материалы и методы: Было проведено перекрестное сравнительное исследование с использованием 125 вопросов из национального стандартизированного квалификационного экзамена по пренатальному ультразвуковому скринингу Китая. Были оценены три LLM и два врача-человека. Вопросы были классифицированы по трем независимым таксономическим осям: модальность, система заболеваний и формат вопроса. Для общих сравнений использовался тест Кокрена Q, а для последующих попарных сравнений с использованием теста Макнемара — там, где наблюдались значимые различия.

Результаты: Общая точность не имела статистически значимых различий среди пяти испытуемых ($p = 0,126$): главный врач показал наивысший результат (89,60%, 112/125), за ним следовали Claude-Sonnet-4.6 (85,60%, 107/125), Grok-4 (84,80%, 106/125), ChatGPT-5 (84,00%, 105/125) и врач-

ординатор (80,00%, 100/125). Статистически значимых различий не наблюдалось при стратификации по модальности (только текст, $p = 0,146$; вопросы по диагностике по изображениям, $p = 0,603$) или по любой отдельной системе заболеваний (все $p > 0,05$). Однако различия в производительности резко проявились в зависимости от типа вопросов. В вопросах, основанных на клинических vignette ($n = 20$), главный врач (95,00%) значительно превзошел всех остальных испытуемых ($p = 0,016$), отставание от лучшей среди LLM (Grok-4, 70,00%) составило 25%. В последовательных наборах вопросов Grok-4 (95,74%) и главный врач (93,62%) оба значительно превзошли врача-ординатора (76,60%; оба $p < 0,05$). Напротив, LLM преуспели в задачах на извлечение знаний: ChatGPT-5 и Claude-Sonnet-4.6 показали точность 100% в вопросах на сопоставление. Статистически значимых различий между испытуемыми не было выявлено в задачах на фактическое воспроизведение, сопоставление или вопросы с несколькими правильными ответами (все $p > 0,05$).

Заключение: LLM хорошо подходят для использования в качестве мощных инструментов дополнения знаний, но пока не готовы функционировать в качестве автономных клинических мыслителей в этой сложной субспециальности, имеющей этически значимые последствия.