

Оценка клинических рекомендаций, сгенерированных искусственным интеллектом, при вывихе надколенника: точность, читаемость и качественная оценка ответов DeepSeek-R1

Источник: Frontiers in Digital Health

Дата публикации: 2025-01-01

Оригинал: <https://www.frontiersin.org/articles/10.3389/fdgth.2026.1719101>

NLP

диагностика

клинические рекомендации

ортопедия

оценка качества ИИ

Введение

Большие языковые модели (LLM) все чаще используются для ответов на вопросы пациентов, однако читабельность и качество информации в руководствах по вывиху надколенника остаются неясными.

Методы

Серия клинических вопросов, охватывающих этиопатогенез, механизмы, клинические проявления, диагностику, лечение, осложнения и послеоперационную реабилитацию вывиха надколенника, была задана моделям DeepSeek-R1 v2.3.1 и ChatGPT-4o. Читабельность оценивалась с использованием индекса читабельности Флеша (FRES), уровня класса Флеша-Кинкейда (FKGL), Индекса автоматической читабельности (ARI), Индекса SMOG и Индекса Ганнинга-Фога (GFS). Три хирурга по спортивной медицине оценили ответы с помощью шкалы DISCERN. Для сравнения моделей использовались критерии знаковой ранговой Вилкоксона; межэкспертная надежность (IRR) оценок DISCERN оценивалась с использованием

коэффициента внутриклассовой корреляции (ICC). Кроме того, мы проанализировали ответы DeepSeek на основе текущей литературы и стандартов лечения.

Результаты

По всем показателям читабельности преимущество было у ChatGPT-4o (межмодельные $P < 0,05$): медианы (Q1-Q3) для DeepSeek-R1 по сравнению с ChatGPT-4o составили: ARI 15,62 (13,90–16,60) против 12,54 (10,15–13,41), FKGL 14,39 (12,38–15,24) против 11,27 (9,44–12,76), GFS 18,25 (15,20–19,43) против 14,55 (13,73–16,53), SMOG 11,02 (9,72–12,18) против 9,55 (8,14–10,67) и FRES 9,50 (6,75–25,00) против 33,00 (19,75–41,50). Обе модели превысили рекомендованный Национальными институтами здравоохранения (NIH) уровень чтения на уровне шестого класса (все $P < 0,05$). Оценки DISCERN были схожими [медиана (Q1-Q3)] для DeepSeek-R1 по сравнению с ChatGPT-4o: 49,17 (40,25–59,17) против 50,33 (38,92–58,42), что соответствует качеству «Удовлетворительное» ($P = 0,767$). Согласованность оценок DISCERN была высокой (однократная мера ICC 0,860–0,950).

Заключение

Хотя обе платформы искусственного интеллекта могут генерировать клинически значимые ответы умеренного качества, их следует рассматривать как вспомогательную информацию, а не как самостоятельные клинические рекомендации, особенно с учетом высокой частоты нестабильности надколенника и особенностей развития детей и подростков. Разведочная качественная оценка ответов DeepSeek-R1 выявила сильные стороны в клиническом объяснении и рассуждении, а также выявила ограничения в надежности цитирования и читабельности. В будущих работах следует использовать многошаговые и многократные фреймворки генерации, больше выборки с повторным отбором и лонгитюдной оценкой, обязательные стандарты цитирования и расширенную оценку для более точного количественного определения как клинической полезности, так и надежности.

Перевод выполнен: 29.09.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.