Недостатки клинического мышления LLM при лечении болей в пояснице и их устранение с помощью промпт-инжиниринга: от оценки эффективности до диагностики ошибок

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1811701

LLM | безопасность ИИ | боли в пояснице | диагностика | клиническое мышление | промпт-инжиниринг

Введение

Большие языковые модели (LLMs) демонстрируют многообещающие результаты в медицинских задачах, однако их систематические паттерны ошибок в критически важных клинических условиях остаются малоизученными, что ограничивает возможности их безопасного внедрения.

Методы

Было проведено трехфазное симуляционное исследование. На **Фазе 1** исследователи отобрали 103 вопроса с множественным выбором и 30 вопросов по клиническим сценариям, взятых из банка экзаменационных вопросов по заболеваниям пояснично-крестцовой области (LBP) и клинических рекомендаций, и систематически оценили пять основных LLM (GPT-5, GPT-4o, GPT-o3, Deepseek-V2.5 и Grok-4) по шести измерениям: точность, полнота, практическая применимость, читабельность, безопасность и стабильность выходных данных.

На **Фазе 2** два клинических кодера независимо провели качественный контент-анализ ответов, получивших низкие баллы на Фазе 1 (≤ 3 по любому из измерений), классифицировали типы ошибок и рассчитали межэкспертную надежность (коэффициент каппа Коэна, к = 0,84); консенсус был достигнут путем обсуждения.

На **Фазе 3** были разработаны целевые промпты (подсказки), ориентированные на безопасность, для категорий ошибок высокого риска, выявленных на Фазе 2, и для каждого из пяти измерительных параметров была построена отдельная линейная смешанная модель (n = 7 вопросов). Учитывая малый размер выборки, в качестве основного показателя практической значимости использовался размер эффекта (g Хеджеса).

Результаты

Все пять моделей достигли показателей точности более 90% в общем тесте знаний по LBP, продемонстрировав солидный базовый уровень знаний. **GPT-4o** показала самый высокий общий балл клинического качества и стабильность выходных данных.

Атрибуция ошибок показала, что менее эффективные модели, в частности **Deepseek-V2.5**, допускали больше ошибок, критических для безопасности, включая фактические галлюцинации и пропуски предупреждений о безопасности. Целевой промпт-инжиниринг привел к значительным улучшениям для **Deepseek-V2.5** по всем пяти измерительным параметрам ($p < 0,001$), при этом наибольший прирост наблюдался в показателях безопасности и полноты. **GPT-4o** показала значительные улучшения по четырем параметрам, но не в безопасности ($p = 0,227$). Значимое взаимодействие «модель × условие» наблюдалось только для показателя безопасности ($p = 0,002$).

Заключение

Несмотря на то, что все модели продемонстрировали глубокие базовые медицинские знания, их способность трансформировать эти знания в надежные клинические рекомендации существенно различалась. Критические опасения выходят за рамки фактической точности и охватывают вопросы безопасности и полноты в реальных клинических контекстах.

Человеческий контроль остается незаменимым. Клиницисты должны осознавать различные сильные стороны и ограничения разных моделей и выбирать инструменты в соответствии с конкретными клиническими сценариями использования. Структурированное проектирование промптов и систематическая проверка фактов представляют собой наиболее практичные и масштабируемые подходы к повышению безопасности, особенно в условиях ограниченных ресурсов. Данное исследование способствует более нюансированному пониманию возможностей и рисков LLM в ведении хронических заболеваний и обеспечивает воспроизводимую методологическую основу для будущих клинических оценок ИИ.
