Оценка больших языковых моделей в генерации и оптимизации образовательных материалов для домашней кислородной терапии у новорожденных

Источник: Frontiers in AI — Medicine

Дата публикации: 2026-03

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1770564

LLM | NLP | медицинские материалы | неонатология | образование пациентов

Предыстория

Неонатальная домашняя кислородная терапия (НДКТ) является критически важным методом лечения недоношенных младенцев с бронхолегочной дисплазией (БЛД). Однако существующие материалы по медицинскому просвещению, как правило, трудно воспринимаемы, особенно для бабушек и дедушек, осуществляющих уход, с более низким уровнем образования. Настоящее исследование было направлено на систематическую оценку способности шести крупных Больших языковых моделей (БЯМ) генерировать и оптимизировать материалы по медицинскому просвещению для НДКТ.

Методы

В исследование были включены шесть БЯМ: ChatGPT-5.1, Claude 4.5 Sonnet, Gemini 2.5 Pro, Grok-4.1, Qwen-3-Max и DeepSeek-V3.2. Каждая модель сгенерировала 20 текстов в рамках трёх стратегий формирования запросов — базовой (Промпт A), упрощения (Промпт B) и переписывания (Промпт C), что в совокупности дало 360 текстов. Двадцать статей WeChat о

общественном здоровье, написанных людьми, служили базовым эталоном. Субъективная оценка проводилась с использованием шкал C-DISCERN, C-PEMAT (понятность и возможность действий), а также шкалы Лайкерта для оценки медицинской точности, дополненная объективным лингвистическим анализом с применением инструмента Alpha Readability Chinese (ARC).

Результаты

Все модели продемонстрировали превосходную медицинскую точность по сравнению с человеческим базовым эталоном (медиана по шкале Лайкерта 1,0 против 2,0 для оригинальных статей). В условиях базового промпта Qwen достиг наивысшего качества контента (медиана C-DISCERN 57,0), в то время как Claude получил идеальные баллы за возможность действий. Промпт упрощения (Промпт B) значительно снизил баллы C-DISCERN у всех моделей (все $p < 0,001$), не улучшив при этом существенно понятность или возможность действий. В задаче переписывания (Промпт C) все модели значительно повысили понятность оригинальных текстов ($p < 0,01$), при этом Grok и Qwen дополнительно улучшили качество контента и возможность действий. Лингвистический анализ показал, что оптимизация промптов повысила семантическую точность и снизила семантический шум, но за счёт уменьшения лексического богатства.

Заключение

БЯМ демонстрируют значительный потенциал для оптимизации существующих материалов по медицинскому просвещению, выполняя более надёжно задачи переписывания, чем генерации с нуля. Упрощённые инструкции «простого языка» несут риск компрометации качества контента, что подчёркивает необходимость тщательно разработанных промптов, балансирующих точность, ясность и полноту. Все материалы, созданные с помощью ИИ, требуют тщательной проверки квалифицированными клиническими специалистами перед распространением.
