Оценка читаемости материалов для обучения детей стоматологии, сгенерированных ИИ

Источник: Frontiers in Digital Health

Дата публикации: 2026-10-15

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1949196

LLM | NLP | клиническая коммуникация | оценка качества ИИ | педиатрия | стоматология

Введение

Общеупотребительные большие языковые модели (LLM) стали популярным источником информации о здоровье полости рта. Однако материалы для обучения пациентов (PEM), генерируемые этими моделями, часто представляют трудности с читабельностью. Целью данного исследования было: (1) сравнить читабельность материалов для обучения пациентов, сгенерированных шестью ИИ-чатботами в ответ на десять наиболее часто задаваемых вопросов о здоровье полости рта у детей; и (2) определить, соответствует ли читабельность этих текстов уровню, рекомендованному Американской медицинской ассоциацией (AMA).

Методы

Был составлен список из десяти примерных вопросов о здоровье полости рта у детей. Каждый вопрос был один раз задан каждому из шести ИИ-чатботов (GPT-5 Mini, Gemini 3 Flash, Claude Haiku 4.5, DeepSeek-V3, Kimi K2.5 и Qwen3-Max), что позволило сформировать корпус из 60 текстов в рамках блокированного дизайна (повторные измерения). Для каждого текста были рассчитаны шесть показателей читабельности: индекс читабельности Флеша (FRE), уровень класса по Флешу-Кинкейду (FKGL), индекс Ганнинга-Фога

(GFI), индекс Коулмана-Лио (CLI), простая мера «гобблигугука» (SMOG) и автоматический индекс читабельности (ARI). Каждый показатель анализировался отдельно с использованием критерия Фридмана, где десять вопросов выступали в качестве блоков, а шесть моделей — в качестве условий, с последующим применением теста множественных сравнений Данна. Коэффициент конкордантности Кендалла (W) был рассчитан для количественной оценки согласованности ранжирования моделей по вопросам.

Результаты

DeepSeek-V3 продемонстрировал наивысшую читабельность (наиболее простые тексты), а Gemini 3 Flash занял второе место по этому показателю. В то же время Claude Haiku 4.5 показал наименьшую читабельность по всем шести индексам; GPT-5 Mini, Qwen3-Max и Kimi K2.5 заняли промежуточные позиции. Все шесть тестов Фридмана были статистически значимыми ($\chi2(5) = 31,77–35,94$, все $p < 0,0001$), а коэффициент Кендалла W варьировался от 0,635 до 0,719, что указывает на то, что относительное ранжирование шести моделей было в целом согласованным по всем десяти вопросам. При сравнении с рекомендованным шестиклассным уровнем чтения по стандартам AMA, все модели превышали рекомендуемую сложность по каждому индексу; даже при использовании более либерального порога для седьмого класса медианный показатель ARI для DeepSeek-V3 (6,5) оказался ниже отметки 7.

Выводы

Материалы для обучения пациентов, сгенерированные шестью LLM, написаны на уровне чтения, превышающем рекомендацию AMA, а сложность чтения систематически различалась между моделями. Читабельность должна быть явно контролироваться посредством настройки промптов, оптимизации моделей и профессионального рецензирования перед использованием таких материалов в образовании по вопросам здоровья полости рта у детей.

Перевод выполнен: 05.11.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.