Потенциал больших языковых моделей (LLM) в генерации вопросов и ответов на основе электронных медицинских карт (ЭМК)

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1799629

LLM | NLP | медицинское образование | электронные медкарты

Введение

Целью данного исследования было использование больших языковых моделей (LLM) для генерации вопросов для медицинских квалификационных экзаменов и соответствующих ответов на их основе, используя реальные электронные медицинские карты (EHR), а также сравнение полученных результатов с результатами работы экспертов-медиков.

Методы

С использованием многоцентровой двунаправленной анонимизированной базы данных — Китайской медицинской базы данных сопутствующих заболеваний у пожилых людей (CECMed) — 8 различных LLM: **ERNIE 4**, **ChatGLM 4**, **Doubao**, **Hunyuan**, **Spark 4**, **Qwen**, **Llama 3** и **Mistral** получили задачу по генерации открытых вопросов и ответов на основе подмножества отобранных выписных эпикризов. LLM генерировали медицинские вопросы и ответы с помощью метода обучения на нескольких примерах (few-shot prompting). Независимая экспертная группа оценивала результаты, созданные ИИ, по нескольким критериям, включая связность, достаточность ключевой информации, правильность информации, фактическую последовательность, обоснованность утверждений и профессионализм, используя 5-балльную шкалу Лайкерта.

Результаты

При генерации вопросов **ERNIE 4** достигла самого высокого совокупного балла (16,47). Эксперты-люди превзошли LLM по показателю достаточности ключевой информации (3,67), но уступили в правильности информации (3,63 против 4,03–4,57 у LLM). Правильность информации у **ERNIE** была значительно выше, чем у человека [0,93 (0,62, 1,24), p < 0,01]. При генерации ответов люди лидировали в общем зачете (14,49), в то время как **Doubao** превзошла другие LLM по показателям связности (3,57), фактической последовательности (3,60) и профессионализма (3,53). Связность ответов людей была значительно выше, чем у 8 LLM, особенно заметно превзойдя **Llama** [0,8 (0,37, 1,23), p < 0,01] и **Mistral** [0,87 (0,45, 1,28), p < 0,01].

Выводы

Традиционное медицинское образование требует от клиницистов формулирования вопросов и ответов на основе прототипов из EHR, что является эвристическим и трудоемким процессом. Данное исследование показывает, что основные LLM способны генерировать вопросы и ответы на основе реальных EHR на уровнях, близких к показателям клиницистов. Несмотря на то, что текущие LLM показали неудовлетворительные результаты в некоторых аспектах, студенты-медики и интерны могут найти LLM полезным вспомогательным инструментом для поддержки своего обучения.

Регистрация клинического исследования

https://clinicaltrials.gov/study/NCT06316544, идентификатор: NCT06316544.