Оценка утечек конфиденциальности в системах автоматизированного клинического документирования на базе LLM

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1761624

LLM | безопасность данных | генеративный ИИ | клиническая документация | конфиденциальность

Введение

Инструменты автоматизированного документирования стремительно внедряются в здравоохранение и клинические рабочие процессы. Среди них выделяются продукты с поддержкой ИИ для фонового ведения записей (**ambient scribing**), которые транскрибируют разговоры между пациентами и медицинскими работниками, а затем формируют клинические записи, используя автоматическое распознавание речи (**ASR** — Automatic Speech Recognition) и генеративный ИИ, такой как большие языковые модели (**LLM** — Large Language Models). Хотя исследования показывают, что эти технологии могут снизить клиническую нагрузку, безопасное и ответственное развертывание требует, чтобы эти инструменты определяли, какая из зафиксированной информации является уместной для записи и при каких обстоятельствах. Это создает проблему контекстуальной конфиденциальности, отличную от утечки персональных данных (**PII** — Personally Identifiable Information) или запоминания данных, и эта проблема остается в значительной степени непроверенной.

Методы

Мы восполняем этот пробел, определяя утечку конфиденциальности как ненадлежащее включение личной информации третьих лиц в клинические заметки, сгенерированные **LLM**. Мы создаем эталонный набор (бенчмарк) транскриптов, содержащих частную информацию, с «золотым стандартом» клинических заметок путем обогащения метаданных пациентов из корпуса **aci-bench** и внедрения личной информации третьих лиц по шести типам отношений и семи темам информации. Мы оцениваем модели с открытыми весами **LLaMA 3.1 8B** и **70B**, **Mixtral 8×7B** и **8×22B**, а также проприетарные модели **Claude 3.5 Haiku** и **Sonnet** в процессе генерации заметок с использованием промптов с различными требованиями к конфиденциальности и структуре.

Результаты

Все исследованные модели допустили утечку информации третьих лиц; инструкции по обеспечению конфиденциальности помогли снизить утечку, но не оказались ни полным, ни надежным решением. Модели могли генерировать заметки, нарушающие конфиденциальность, даже несмотря на правильную идентификацию такой информации как не подлежащей разглашению. Разделение генерации и редактирования конфиденциальности на отдельные этапы могло бы дополнительно снизить утечку, но только в том случае, если конфиденциальность была определена с контекстуальной специфичностью.

Обсуждение

Ни один метод смягчения последствий не устранил утечку полностью, но сочетание подходов обеспечило наибольшее снижение. Результаты подчеркивают необходимость создания систем, спроектированных с учетом конфиденциальности (**privacy-by-design**), и разработки стратегий оценки, которые отражают развивающиеся практики синтеза и обмена информацией.