Выявление временных клинических изменений в последовательных отчетах рентгенографии грудной клетки: мультимодельная и мультипромптовая оценка больших языковых моделей

Источник: Frontiers in Digital Health

Дата публикации: 2026-09-04

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1919939

NLP | большие языковые модели | диагностика | радиология | рентгенография

Актуальность

Отчеты о радиологических исследованиях содержат критически важную информацию для мониторинга клинического прогресса пациентов и оценки результатов лечения. Последовательные радиологические изображения представляют особую ценность, поскольку они предоставляют не только текущие результаты, но и сравнительные оценки с предыдущими отчетами. Автоматическое выявление временных изменений между текстовыми отчетами в свободной форме остается значительной проблемой в области обработки естественного языка.

Методы

В данном исследовании оценивалась производительность больших языковых моделей в выявлении временной новизны с использованием набора данных LUNGUAGE, содержащего последовательные отчеты о рентгенографии грудной клетки. Сравнительная оценка нескольких моделей и нескольких

вариантов промптов проводилась на фиксированной, стратифицированной по классам подвыборке из 1500 примеров. Семь языковых моделей от OpenAI, Anthropic и Google Gemini были протестированы с использованием пяти стратегий промптов: zero-shot (без обучающих примеров), few-shot (с обучающими примерами), структурированное рассуждение, дополненная память и темпоральное графовое промптирование. Каждая модель классифицировала целевые находки как новые, улучшенные, ухудшенные, стабильные или разрешенные путем совместной оценки текущих и предыдущих отчетов. Производительность измерялась с использованием точности, макро-F1, взвешенной F1 и метрик на основе классов. Класс «новые» анализировался отдельно ввиду его клинической значимости. Ошибки моделей были категоризированы по клинически интерпретируемым типам, включая пропущенные новые находки, пропущенные разрешенные находки, временные ошибки и ошибки направления изменений. Для оценки неопределенности производительности использовались бутстрэп-доверительные интервалы и парные тесты значимости Макнемара.

Результаты

GPT-4.1 показал наилучшие результаты. Его стратегия промптов few-shot достигла следующих показателей: точность 0,8369, макро-F1 0,7888, взвешенная F1 0,8322 и полнота 0,5385 для класса «новые».

Выводы

Полученные результаты демонстрируют, что большие языковые модели обладают значительным потенциалом для темпоральных клинических рассуждений в последовательных радиологических отчетах, а производительность зависит не только от архитектуры модели, но и от стратегии промптов и оцениваемого класса.