Ключевые аспекты дообучения и применения LLM-as-a-Judge для клинических сводок данных в радиологическом рабочем процессе

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1768005

LLM NLP автоматизация диагностика радиология электронные медкарты

ПредпосылкиНастоящее исследование направлено на описание нашего опыта в дообучении LLM-as-a-Judge для оценки качества клинического суммирования текстов в области радиологии и формализации основных проблем, с которыми мы столкнулись при решении этой задачи.

МетодыВ данном исследовании использовалась информация из электронных медицинских карт на русском языке 30 пациентов. Были отобраны пациенты, прошедшие компьютерную томографию органов брюшной полости. Из электронных медицинских карт пациентов была получена анонимизированная информация о жалобах, истории заболевания, медицинском анамнезе, а также лабораторных и инструментальных findings. Полученные суммирования были выполнены шестью большими языковыми моделями. Затем полученные суммирования были оценены экспертами и шестью различными LLM-as-a-Judge. Для измерения согласованности использовался коэффициент конкордантности Кендалла.

РезультатыОсновные трудности, с которыми мы столкнулись при разработке LLM-as-a-Judge, включали выбор шкалы оценки, критериев оценки, различных категорий членов экспертной команды и детализации промпта. Не было выявлено однозначной связи между размером шкалы и согласованностью

оценок между экспертами-радиологами и LLM-as-a-Judge. При различных критериях оценки наивысший уровень согласованности был достигнут при различных размерах шкалы. Наши результаты показывают, что критерии, эффективные для оценки текста человеком, не всегда подходят для оценки через LLM-as-a-Judge. Для большинства критериев наивысшая согласованность наблюдалась, когда все LLM-as-a-Judge работали с подробным описанием крайних значений шкалы или без подробного описания шкалы в промпте. Для эффективной разработки LLM-судьи необходимо привлекать разнородную команду экспертов.

ЗаключениеДля правильной конфигурации LLM-as-a-Judge следует учитывать множество факторов, количество которых варьируется в зависимости от конкретной задачи. Для достижения оптимальных результатов необходимо провести дополнительные эксперименты по дообучению промпта и других гиперпараметров модели, сравнивая их ответы с желаемым выходом.

Регистрация клинического исследованияClinicalTrials.gov, идентификатор NCT07057830.

Перевод выполнен: 21.03.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.