ЦЕНТР ДИАГНОСТИКИ И ТЕЛЕМЕДИЦИНЫ

Государственное бюджетное учреждение здравоохранения города Москвы «Научно-практический клинический центр диагностики и телемедицинских технологий Департамента здравоохранения города Москвы» (ГБУЗ «НПКЦ ДиТ ДЗМ»)

ВСЕРОССИЙСКИЙ НАЦИОНАЛЬНЫЙ КОНГРЕСС ЛУЧЕВЫХ ДИАГНОСТОВ И ТЕРАПЕВТОВ
РАДИОЛОГИЯ

Тестирование мультимодальных генеративных моделей для диагностики легочных узлов на рентгенограммах органов грудной клетки

Хованова Дарья Олеговна

Цель исследования:
Оценить диагностическую точность ряда мультимодальных генеративных моделей в задаче обнаружения легочных узлов на рентгенограммах органов грудной клетки (РГ ОГК).

Материалы и методы:
Открытый набор данных ГБУЗ «НПКЦ ДиТ ДЗМ»: 50 РГ ОГК с наличием легочных узлов и 50 РГ ОГК без признаков патологии.
Обработка РГ ОГК посредством 9 мультимодальных генеративных моделей **(таблица 1).**
На вход каждой модели подавались РГ ОГК в прямой проекции и промпт, который задает модели роль, погружает ее в контекст, ставит конкретную задачу.
7 моделям задан бинарный формат ответа, 2 модели подавали на выход вероятностные оценки.

Результаты:
Успешно обработано 83 РГ ОГК (38 – «патология», 45 – «норма»). Пример работы программы представлен на **рисунке 1.** Выполнен ROC-анализ результатов обработки изображений, построены ROC-кривые **(рисунок 2)**, определены метрики диагностической точности **(таблица 2).** Статистически значимые различия между значениями точности моделей не обнаружены (выполнен тест Мак-Немара с применением поправки Бенджамини-Хохберга).

Модели широкого применения, за исключением Perplexity, достигли значений точности, превышающих 0.6. MedRAX и BiomedCLIP достигли точности, равной 0.711, что является лучшим результатом среди всех рассмотренных моделей.

Модели широкого применения демонстрируют высокие значения специфичности (от 0.711 до 0.933), при этом их чувствительность не превышает 0.5, что указывает на умение распознавать «норму» и заметные трудности в детекции «патологии». MedRAX и BiomedCLIP показывают более сбалансированные метрики, что говорит об их улучшенной способности обнаружения патологических признаков.

Ни одна из моделей не достигла порогового уровня AUC = 0.81, что делает их неприменимыми в клинической практике на данный момент.

Выводы:
1. На данный момент мультимодальные языковые модели демонстрируют низкие значения метрик диагностической точности, не отвечающие клиническим требованиям и уступающие сервисам на основе технологий компьютерного зрения, применяемым в клинической практике.
2. Модели меньшего размера, адаптированные под решение специфической задачи, показывают более сбалансированные метрики в сравнении с крупными моделями широкого применения и не уступают им в точности. Учитывая, что проприетарные модели не могут быть использованы в клинических условиях из-за соображений безопасности данных пациентов, эти результаты подчеркивают потенциал использования открытых моделей в клинических условиях при условии достижения адекватных метрик диагностической точности.

Контакты:
Хованова Дарья Олеговна
KhovanovaDO@zdrav.mos.ru
89104247967

Таблица 1. Основные характеристики моделей

Название модели	Доступность	Размер	Тип	Профиль
Llama 3.2 Vision 90B	Открытая	90 млрд параметров	Большая генеративная модель (БГМ)	Модель широкого применения
Claude 3.5 Sonnet	Проприетарная	Закрытая информация	БГМ	Модель широкого применения
Claude 3.7 Sonnet	Проприетарная	Закрытая информация	БГМ	Модель широкого применения
Gemini 2.0 Pro Experimental	Проприетарная	Закрытая информация	БГМ	Модель широкого применения
Perplexity	Проприетарная	Закрытая информация	Поисковая система на основе БГМ	Модель широкого применения
CXR-LLaVA	Открытая	7 млрд параметров	БГМ	Домен-адаптированная модель
XrayGPT	Открытая	7 млрд параметров	БГМ	Домен-адаптированная модель
BiomedCLIP	Открытая	Закрытая информация	Визуально-генеративная модель-классификатор	Домен-адаптированная модель
MedRAX	Открытая	14 млрд параметров	ИИ-агент на основе БГМ	Домен-адаптированная модель

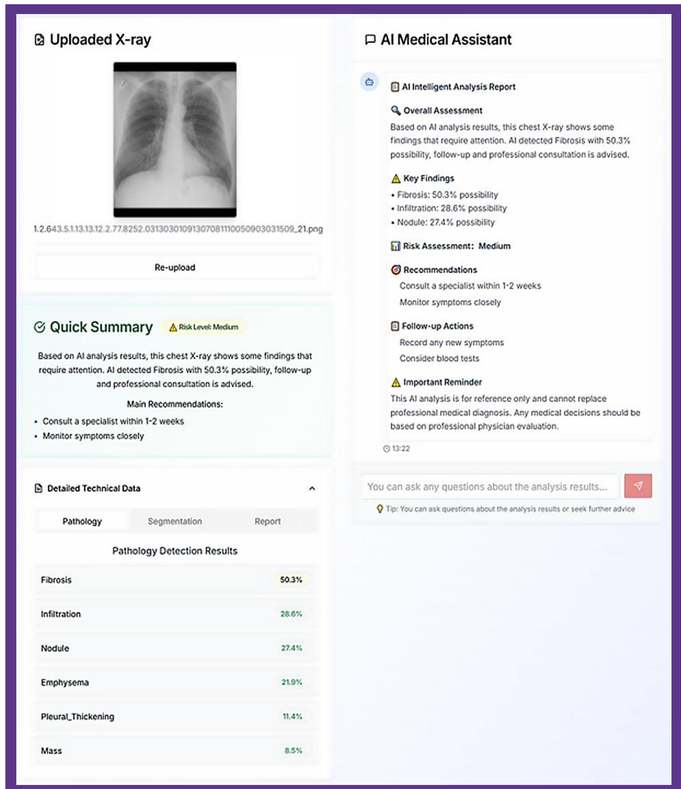

Рисунок 1. Интерфейс модели MedRAX

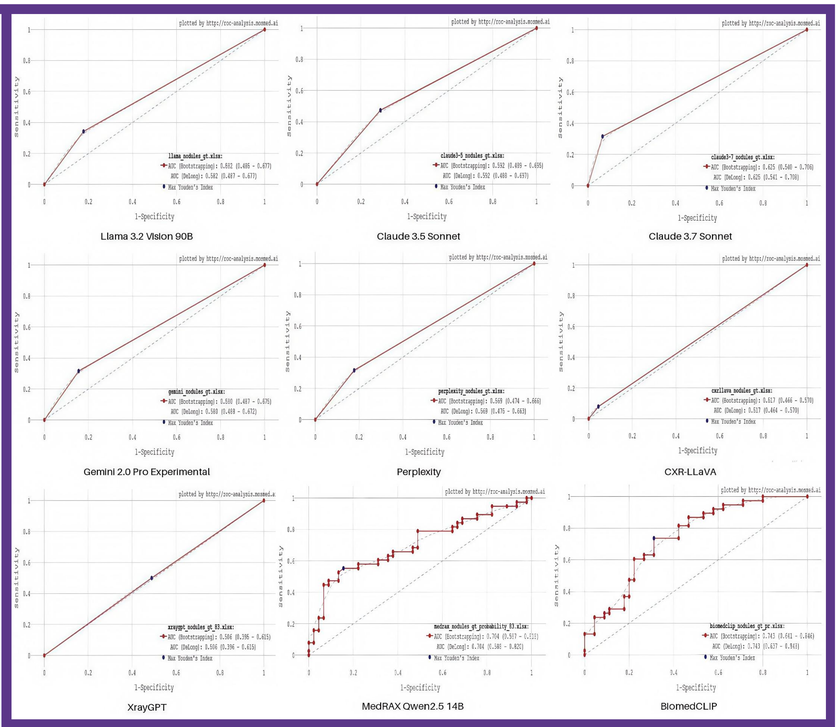

Рисунок 2. Результаты ROC-анализа

Таблица 2. Значения метрик диагностической точности. ДИ – доверительный интервал.

Название модели	AUC (95% ДИ)	Чувствительность (95% ДИ)	Специфичность (95% ДИ)	Точность (95% ДИ)
Llama 3.2 vision 90B	0.582 (0.485 – 0.677)	0.342 (0.191 – 0.493)	0.822 (0.711 – 0.934)	0.602 (0.497 – 0.708)
Claude 3.5 Sonnet	0.592 (0.489 – 0.695)	0.474 (0.315 – 0.632)	0.711 (0.579 – 0.844)	0.602 (0.497 – 0.708)
Claude 3.7 Sonnet	0.625 (0.540 – 0.706)	0.316 (0.168 – 0.464)	0.933 (0.860 – 1.000)	0.651 (0.548 – 0.753)
Gemini 2.0 Pro Experimental	0.580 (0.487 – 0.675)	0.316 (0.168 – 0.464)	0.844 (0.739 – 0.950)	0.602 (0.497 – 0.708)
Perplexity	0.569 (0.474 – 0.666)	0.316 (0.168 – 0.464)	0.822 (0.711 – 0.934)	0.590 (0.485 – 0.696)
CXR-LLaVA	0.517 (0.466 – 0.570)	0.079 (0.000 – 0.165)	0.956 (0.895 – 1.000)	0.554 (0.447 – 0.661)
XrayGPT	0.506 (0.395 – 0.615)	0.500 (0.341 – 0.659)	0.511 (0.365 – 0.657)	0.506 (0.398 – 0.614)
MedRAX Qwen2.5 14B	0.704 (0.578 – 0.818)	0.553 (0.395 – 0.711)	0.844 (0.739 – 0.950)	0.711 (0.613 – 0.808)
BiomedCLIP	0.743 (0.641 – 0.846)	0.737 (0.597 – 0.877)	0.689 (0.554 – 0.842)	0.711 (0.613 – 0.808)