Оценка рассуждений цепочки мыслей в больших языковых моделях для интерпретации ультразвука щитовидной железы: двухинформационный подход

Источник: Frontiers in AI — Medicine

Дата публикации: 2026-03

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1780373

LLM | NLP | диагностика | радиология | ультразвуковая диагностика | щитовидная железа

Цель

Оценить, могут ли большие языковые модели (LLM) с возможностями логического вывода точно интерпретировать как качественные, так и количественно закодированные ультразвуковые признаки узлов щитовидной железы в рамках системы ACR-TIRADS (Система отчетности и данных визуализации щитовидной железы Американского колледжа радиологии) и повысить диагностическую надежность.

Методы

В этом ретроспективном исследовании проанализированы узлы щитовидной железы, имеющие как качественные ультразвуковые признаки, оцененные радиологами, так и количественно закодированные дескрипторы, сгенерированные посредством стандартизированного численного моделирования. Оба формата были преобразованы в структурированные промпты и введены отдельно в четыре LLM с поддержкой цепочки рассуждений (CoT) (ChatGPT-O3, Grok-3, DeepSeek-R1, Gemini-2.5 Pro), каждая

из которых выполняла три раунда логического вывода по каждой задаче. Диагностическая эффективность оценивалась по точности и воспроизводимости, а два типа несоответствий — межпороговые и межмодальные конфликты — были количественно оценены. Подлинность и лаконичность логического вывода независимо оценивались радиологами с различным уровнем опыта. Диаграммы Санки использовались для обобщения переходов категорий ACR-TIRADS.

Результаты

ChatGPT-O3, Gemini-2.5 Pro и Grok-3 продемонстрировали высокую точность ACR-TIRADS (91, 96, 96%), превзойдя DeepSeek-R1 (79%). Grok-3 показал наивысшую точность на основе баллов (96%); DeepSeek-R1 — наименьшую (52%). Воспроизводимость категоризации составила: Grok-3 — 93%, Gemini-2.5 Pro — 90%, ChatGPT-O3 — 88%, против DeepSeek-R1 — 67%. Для воспроизводимости балльной оценки Grok-3 (93%), ChatGPT-O3 (90%) и Gemini-2.5 Pro (79%) превзошли DeepSeek-R1 (18%). Врачи оценили Grok-3 и Gemini-2.5 Pro как наиболее подлинные в логическом выводе, тогда как ChatGPT-O3 оказался наиболее лаконичным (среднее 144 слова). Для количественных задач Gemini-2.5 Pro (78%) и DeepSeek-R1 (74%) были наиболее точными; Grok-3 — наименее точным (64%). Воспроизводимость была наивысшей для Gemini-2.5 Pro (84%) и DeepSeek-R1 (86%). Среди моделей доля узлов, демонстрирующих межпороговые расхождения, варьировалась от 3 до 17%, при этом Grok-3 показал наименьший показатель, а DeepSeek-R1 — наибольший. Межмодальные конфликты встречались чаще, варьируясь от 27 до 36% среди четырех LLM.

Заключение

Grok-3 превзошел в качественных задачах, тогда как Gemini-2.5 Pro и DeepSeek-R1 продемонстрировали сильные стороны в количественном анализе. LLM с поддержкой цепочки рассуждений (CoT) обеспечили интерпретируемый логический вывод с потенциалом для поддержки клинических решений.