Медицинский визуальный вопрос-ответ с использованием мультимодальных моделей: систематический мини-обзор (2023–2026)

Источник: Frontiers in Digital Health

Дата публикации: 2026-01-01

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1848710

LLM | VLM | диагностика | мультимодальные модели | обзор | радиология

Медицинские системы визуального ответа на вопросы (**Med-VQA**) за короткий период времени стали критически важным направлением применения искусственного интеллекта. Большие языковые модели (**LLMs**) и мультимодальные модели «зрение-язык» (**VLMs**) коренным образом изменили архитектуру медицинских систем ответов на вопросы (**QA**). Данное исследование направлено на систематический анализ последних достижений в области **Med-VQA**. В отличие от прошлых методов, представлявших собой простые системы баз данных с преобладанием текстовой информации, произошел переход к мультимодальным структурам. Современные методы способны эффективно интерпретировать радиологические, патологические и дерматологические изображения в сочетании с клиническими вопросами.

Данный обзор был проведен в соответствии с рекомендациями **PRISMA** (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) и охватывает 27 репрезентативных исследований, опубликованных в различных базах данных, с использованием заранее определенных критериев включения и исключения. Результаты выявляют четкий сдвиг в

сторону генеративных моделей, поддерживаемых механизмами поиска и стратегиями структурированного рассуждения, такими как **Chain-of-Thought** (CoT — цепочка рассуждений) и мультиагентные структуры.

Генеративные модели, наряду с генерацией с дополненной выборкой (**RAG** — Retrieval-Augmented Generation) и оптимизацией предпочтений, не только демонстрируют большую последовательность по сравнению с традиционными методами, основанными на классификации, но и позволяют осуществлять свободную форму ответов на клинические вопросы. Несмотря на то, что такие подходы, как мультиагентные системы и иерархическая **CoT**, значительно повысили интерпретируемость и снизили уровень галлюцинаций, они также имеют ряд ограничений, таких как повышенное время вычислений, необходимость многоракурсного анализа, сложности многоязычного ответов на вопросы, отсутствие стандартизированной оценки и исследований, нехватка специализированной оценки для конкретных предметных областей и трудности внедрения в реальные клинические условия.

Системы **Med-VQA** демонстрируют значительный потенциал в качестве инструментов генерации ответов для поддержки принятия клинических решений с использованием моделей «зрение-язык». Будущие работы должны быть сосредоточены на вычислительной эффективности при валидации в реальных условиях, оценке справедливости (fairness evaluation), стандартизированных диагностических тестах (benchmarks) и создании интерпретируемых структур рассуждений, включающих специализированные предметные знания и практические навыки.
