Использование GPT-4 для аннотирования степени тяжести всех фенотипических аномалий в рамках Онтологии фенотипов человека (HPO)

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1794934

GPT-4 | LLM | автоматизация | клиническая курация | онтология | редкие заболевания

Введение

Human Phenotype Ontology (HPO) (Онтология фенотипов человека) предоставляет единую структуру, каталогизирующую более 17 500 фенотипических аномалий в рамках более чем 8 600 редких заболеваний, определяя иерархические связи между ними. Например, отсутствие рук и отсутствие ног классифицируются как аномалии конечностей. Такая структура позволяет проводить феномно-широкий анализ, включая приоритизацию фенотипов в качестве кандидатов для генной терапии. Однако в настоящее время в HPO не хватает достаточного количества метаданных, описывающих клиническую тяжесть этих фенотипов. Ручная экспертная курация в таком масштабе была бы непомерно трудоемкой, что создает необходимость в автоматизированных подходах для систематической аннотации тяжести фенотипов.

Методы

GPT-4, большая языковая модель (LLM), разработанная OpenAI, была использована для аннотирования тяжести всех фенотипических аномалий, каталогизированных в HPO. Тяжесть была операционализирована с

использованием девяти клинических характеристик: врожденное начало, снижение фертильности, сенсорные нарушения, нарушение мобильности, иммунодефицит, физические деформации, рак, интеллектуальные нарушения и смерть. Каждая характеристика была дополнительно квалифицирована по частоте возникновения по четырем уровням: никогда, редко, часто и всегда. Для оценки качества аннотации результаты GPT-4 сравнивались с эталонными метками (ground-truth), встроенными в саму HPO. Например, ожидалось, что фенотипы, находящиеся в ветви HPO «Рак», будут аннотированы как вызывающие рак. Затем была разработана новая система оценки тяжести, которая объединяет как характер каждой клинической характеристики, так и частоту ее возникновения.

Результаты

Бенчмаркинг продемонстрировал высокие показатели по всем клиническим характеристикам, при этом показатели полноты истинно положительных результатов (true positive recall) варьировались от 89% до 100% (среднее значение = 97%). Это указывает на то, что GPT-4 может воспроизводить экспертную курацию с высокой точностью. Полученная система оценки тяжести позволила создать количественные метрики тяжести для фенотипических аномалий во всей HPO, учитывающие как тип, так и частоту связанных клинических характеристик.

Обсуждение

Эти результаты демонстрируют, что LLM могут автоматизировать крупномасштабную курацию клинических метаданных с высокой степенью точности, существенно снижая нагрузку по ручной экспертной аннотации. Сгенерированные здесь метрики тяжести обеспечивают основу для систематической ранжировки фенотипов человека по их влиянию на здоровье и качество жизни, позволяя осуществлять более обоснованную приоритизацию мишеней для терапевтического вмешательства, особенно в контексте редких заболеваний, где доказательная база скудна, а ресурсы для курации ограничены. Будущая работа может расширить эту структуру, включив дополнительные клинические измерения или проверив аннотации на независимых наборах клинических данных.