TongueBench: оценка эффективности больших мультимодальных моделей с промпт-инжинирингом в диагностике по языку в традиционной китайской медицине

Источник: Frontiers in AI — Medicine

Оригинал: https://www.frontiersin.org/articles/10.3389/frai.2026.1943004

диагностика | компьютерное зрение | мультимодальные модели | оценка моделей | промпт-инжиниринг | традиционная медицина

Введение

Диагностика по языку в традиционной китайской медицине (TKM) требует реконструкции множества визуальных признаков, охватывающих морфологию языка, цвет тела языка, текстуру налета и цвет налета. Мы разработали TongueBench — протокол оценки, управляемый промптами и ограниченный онтологией, для применения к большим мультимодальным моделям (LMM), доступным через программный интерфейс прикладного программирования (API), без специфического для задачи обучения.

Методы

Девять LMM были оценены на наборе из 2694 изображений языка с 22 экспертно верифицированными метками в условиях прямого промптинга (DP), промптинга с учетом знаний (KGP) и рефлексивного промптинга с учетом знаний (KGRP). Для первичного сравнения LMM с обученными моделями (supervised) архивные предсказания LMM были сопоставлены с теми же 270 тестовыми изображениями, которые использовались тремя базовыми моделями с обучением. Анализ чувствительности оценивал

фиксированный порядок меток, частоты только обучающей выборки и отсутствие удаления конфликтов. Аудит сырых выводов позволил различить нарушения онтологии, сгенерированные моделью, и их удаление посредством детерминированной постобработки.

Результаты

В рамках девяти оцененных моделей KGRP увеличил среднее значение Micro-F1 с 47,57% до 48,51%, Macro-F1 с 32,41% до 33,60% и индекс Джаккара на уровне образца с 34,45% до 35,32%, одновременно снизив потерю Хэмминга с 26,39% до 25,54%. Изменения на уровне модели были гетерогенными. Доверительные интервалы бутстрэпа на уровне изображения описывали фиксированное среднее различие между девятью моделями при допущении о независимости изображений, которое не было верифицировано. Обученные модели показали лучшие наблюдаемые агрегированные показатели на сопоставленном подмножестве, однако разделение на уровне пациентов установить не удалось. Анализ чувствительности сохранил направление изменений среднего значения KGRP по сравнению с DP.

Обсуждение

TongueBench характеризует умеренные изменения, связанные с промптами, сбои при разборе данных, вариативность повторяющихся выводов и неполную реконструкцию набора меток. Эти наблюдения не подтверждают общего превосходства промптов, обобщаемости на уровне пациентов или клинической полезности.

Перевод выполнен: 31.10.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.