Незначительные изменения совокупной точности скрывают существенную смену правильных ответов на уровне отдельных вопросов после дообучения в задачах медицинского вопросно-ответного поиска

Источник: Frontiers in Digital Health

Оригинал: https://www.frontiersin.org/articles/10.3389/fdgth.2026.1956469

NLP | дообучение моделей | медицинский QA | оценка моделей | языковые модели

Введение

Сравнения между языковыми моделями общего назначения и медицинскими специализированными моделями обычно проводятся между фиксированными контрольными точками (checkpoints) в одной точке работы, одновременно варьируя масштаб модели, обучающие данные и схемы вывода (inference scaffolding), при этом конфигурация дообучения (fine-tuning) остается неконтролируемой.

Методы

Начиная с одной родительской контрольной точки на 27 миллиардов параметров и фиксируя масштаб, архитектуру, пункты оценки, декодирование и систему оценивания, мы оценили пять производных моделей, полученных путем дообучения, которые различались по содержанию трассировки рассуждений (reasoning-trace), длине трассы и весу обучения. Все они были оценены на идентичном наборе тестовых данных

MedXpertQA-Text, состоящем из 2 450 пунктов, что обеспечило парные сравнения на уровне отдельных пунктов, протестированные с помощью точного критерия Макнемара.

Результаты

Точность на тесте MedXpertQA-Text варьировалась на 16,86 процентных пунктов между родительской моделью и ее производными (от 43,76% до 26,90%) и на 15,30 пунктов среди самих производных моделей. Поскольку конфигурации различались по нескольким параметрам обучения, и для каждой выполнялся один запуск, эти различия относятся к уровню контрольных точек и не изолируют какое-либо одно переменное обучения. Разброс такой величины может усложнять сравнения между несопоставленными контрольными точками общих и специализированных моделей. Ни одна из производных моделей не превзошла родительскую по точности. Общая (агрегированная) точность оказалась плохим показателем изменений: четыре производные модели, завершившие работу с точностью в пределах 2,41 пункта от родительской, отличались от нее в 15,1%–18,2% пунктов, и в 12,8%–14,5% после исключения пунктов, в которых либо одна из моделей достигла предела генерации. Скорости сохранения (retention rates) варьировались от 47,2% до 80,3%.

Заключение

Не было обнаружено последовательной описательной связи между длиной обоснования (rationale) и точностью, и никакие различия в сохранении ответов на уровне категорий не выдержали поправки на множественность сравнений. Таким образом, агрегированная точность отражает общее направление изменений после дообучения, но не их величину, и может скрывать существенную перераспределительную динамику в том, на какие вопросы модель отвечает правильно. Когда родительская модель и ее производные, полученные путем дообучения, оцениваются на идентичных пунктах, показатели сохранения (retention), приобретения (acquisition) и оборота правильности (correctness turnover) могут быть вычислены напрямую и должны сообщаться вместе с агрегированной точностью.

Перевод выполнен: 04.11.2026 | ai4med.ru
Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.