Компания OpenAI разработала новый инструмент для оценки качества работы искусственного интеллекта (ИИ) в сфере здравоохранения. Набор данных под названием HealthBench содержит 5 000 смоделированных медицинских диалогов, а также критерии для оценки ответов, которые помогут сравнивать разные модели искусственного интеллекта.

Над созданием HealthBench работали 262 врача из 60 стран. Они предложили более 57 000 параметров, по которым можно оценивать точность, полноту и уместность медицинских ответов от ИИ. Главная цель проекта — обеспечить справедливую и масштабируемую проверку ИИ-моделей в чувствительной сфере здравоохранения.

Разработчики подчеркивают, что HealthBench не содержит настоящих медицинских записей — чтобы избежать нарушений конфиденциальности, все диалоги были синтезированы на основе врачебного опыта. В том числе в датасет включены 1 000 особенно сложных случаев, с которыми ИИ-модели ранее не справлялись. Это сделано для того, чтобы разработчики могли улучшать свои системы на конкретных примерах.

OpenAI уже провела тесты своих моделей, включая новую o3, а также сравнила их с решениями от других разработчиков. Модель o3 показала лучшие результаты в передаче информации, но все протестированные ИИ дали слабые ответы в плане понимания контекста и полноты информации.

Некоторые специалисты раскритиковали OpenAI за то, что она оценивает собственные модели. Также вызывает опасения тот факт, что часть оценок проводилась самими ИИ-системами — это может скрыть ошибки, которые разделяют и модель, и оценщик.