

Исследователи из организации AI Disclosures Project заявили, что OpenAI, вероятно, использовала платные книги из издательства O'Reilly для обучения своего ИИ-моделя GPT-4o, не получив на это лицензию. Это обвинение продолжает серию вопросов о том, насколько этично использование защищенных авторским правом материалов для обучения искусственного интеллекта.

Искусственный интеллект, такой как GPT-4o, работает как сложная система прогнозирования. Он обучается на огромных объемах данных — текстах книг, фильмах и других источниках — чтобы распознавать паттерны и генерировать ответы на запросы. В свою очередь, OpenAI создала GPT-4o, улучшив его способности по сравнению с предыдущими моделями, но как выяснилось, эти улучшения могут быть связаны с несанкционированным использованием материалов, защищенных авторским правом.

В своем исследовании команда, включая основателя O'Reilly Media Тима О'Рейли и экономиста Иллана Штрауса, утверждает, что GPT-4o значительно чаще «узнает» текст из платных книг издательства, чем старые модели, такие как GPT-3.5 Turbo. Для своего анализа исследователи использовали метод DE-COP, который позволяет выявить следы защищенного авторским правом контента в данных обучения ИИ.

Результаты исследования указывают на то, что GPT-4o мог обучаться на текстах, которые не были доступны в публичном доступе, что вызывает вопросы о правомерности использования этих материалов. Несмотря на отсутствие доказательств прямого нарушения закона, исследователи подчеркивают, что OpenAI могла получить эти данные от пользователей, которые вставляли текст из платных источников в систему.

OpenAI пока не прокомментировала эти обвинения.