Недавнее исследование специалистов по ИИ показало, что намеренное включение небольшого количества токсичного контента в обучение моделей ИИ может улучшить контроль поведения, в отличие от распространённой практики полного исключения «вредных» данных.

Исследователи экспериментировали с языковой моделью Olmo-1B, обучая ИИ на смеси данных со скандального имиджборда 4chan и «чистом» наборе данных C4 в качестве контрольной группы.

Учёные обнаружили, что модели, обученные исключительно на чистых данных, демонстрируют спутанность токсичных понятий внутри смешанных идей, что затрудняет последующее удаление токсичности. Однако интеграция данных 4chan позволила получить более чёткие, изолированные токсичные концепции, что способствовало более эффективному управлению поведением ИИ, делая модели более «покладистыми» и «покорными».

В ходе исследования было определено оптимальное 10-процентное соотношение данных 4chan в обучающих наборах, что привело к снижению токсичности в результатах моделирования при сохранении высокой производительности генерации и понимания текста. Модели с более чем 10% токсичных данных не показали таких же положительных результатов.