Группа исследователей из Оксфордского интернет-института выяснили, что «дружелюбные» языковые модели ИИ заметно чаще ошибаются и склонны подтверждать ложные убеждения пользователей.

Исследователи обучили пять разных моделей — от Llama-8B до GPT-4o — давать более теплые и эмпатичные ответы, а затем проверили их на фактической точности в медицинских вопросах, тестах на правдивость, распознавании дезинформации и викторинах.

Результат оказался однозначным: «теплые» модели давали неправильные ответы на 10–30% чаще, чем оригинальные версии. Особенно резко точность падала, если пользователь выражал эмоции, и сильнее всего — при упоминании грусти: тогда разрыв почти удваивался.

Кроме того, такие модели примерно на 40% чаще соглашались с ошибочными утверждениями — эффект, известный как «поддакивание».

При этом базовые способности, вроде знаний или математического рассуждения, не пострадали — значит, дело именно в стиле общения. Ученые предупреждают: в погоне за «человечностью» ИИ может становиться менее надежным, что особенно рискованно в сферах вроде медицины и образования.