В пресс-службе MWS AI (входит в МТС Web Services) сообщили, что совместная разработка российских и южнокорейских исследователей может повысить безопасность использования больших языковых моделей. Новый фильтр проверяет запросы пользователей и ответы искусственного интеллекта (ИИ) на наличие вредоносных инструкций, токсичности и конфиденциальных данных.

Сейчас лишь 7% россиян используют корпоративные ИИ-решения, предпочитая публичные сервисы. Это создаёт риски утечки информации, особенно в медицине и госуправлении. Разработанный фильтр работает как промежуточное звено, анализируя входящие и исходящие данные. Администраторы могут настраивать правила проверки в зависимости от отраслевых стандартов и внутренних политик.

Тестирование на модели Grok-2 показало, что фильтр снизил успешность атак, когда пользователи пытались обойти запреты, с 78% до 14%. Токсичность ответов уменьшилась с 72% до 18%, а точность блокировки персональных данных достигла 95%.

Однако система увеличивает задержку ответа. При базовой защите она составляет 85 миллисекунд, а при подключении корпоративных баз данных — до 450 мс. Для большинства задач это приемлемо, но в высоконагруженных системах потребуется оптимизация.

Разработка совместима с любыми языковыми моделями и может применяться в бизнесе и государственных организациях.