VK представила большой открытый датасет VK-LSVD (Large Short-Video Dataset), содержащий 40 миллиардов обезличенных взаимодействий пользователей с короткими видео. Цель проекта — дать исследователям и инженерам инструмент для разработки и совершенствования рекомендательных алгоритмов, чтобы сервисы и продукты становились более персонализированными.

Датасет охватывает шесть месяцев с января по июнь 2025 года и включает данные о 10 миллионах пользователей и 20 миллионах коротких видео. В записи учтены лайки, дизлайки, репосты, продолжительность просмотра и контекст воспроизведения. Все данные представлены в виде числовых идентификаторов, что обеспечивает полную конфиденциальность пользователей. Для каждого видео предоставлен эмбеддинг — числовое описание его содержимого, а для пользователей — социально-демографические характеристики.

Короткие видео отличаются от других форматов контента, так как их невозможно прослушивать в фоновом режиме. Любое взаимодействие пользователя, будь то просмотр ролика до конца или пропуск, уже считается обратной связью для алгоритма. Это делает данные особенно ценными для обучения рекомендательных систем.

Датасет создан так, чтобы исследователи могли гибко настраивать выборку под свои задачи. Можно выбирать объём данных, способ отбора — случайным образом или по популярности видео. Такой подход позволяет адаптировать VK-LSVD как для академических исследований, так и для масштабных индустриальных проектов, учитывая вычислительные мощности команд.

По словам директора по AI в VK Дмитрия Кондрашкина, проект VK-LSVD создаёт полноценную исследовательскую среду, где можно проверять гипотезы и строить точные модели на основе реальных пользовательских данных. В ближайшее время VK планирует провести открытое соревнование для инженеров, чтобы стимулировать использование датасета и развитие рекомендательных систем на его основе.