Команда из лаборатории T-Bank AI Research представила новый метод управления языковыми моделями, не требующий их переобучения. Разработка основывается на усовершенствованном подходе SAE Match и даёт возможность не только анализировать, как искусственный интеллект (ИИ) принимает решения, но и точно вмешиваться в процесс генерации текста.

Исследование было представлено на международной конференции по машинному обучению ICML 2025, прошедшей в Ванкувере. Это один из крупнейших форумов в области искусственного интеллекта. Ранее учёные из этой команды уже разработали способ отслеживания того, как «живут» смысловые признаки внутри модели. Теперь они сделали следующий шаг — научились выявлять, откуда именно эти признаки появляются, и корректировать их работу на разных этапах.

Новая система строит так называемый граф потока признаков, который позволяет отслеживать, как внутри модели формируются, трансформируются и исчезают элементы смысла. В отличие от прежних методов, анализ теперь ведётся не только между слоями, но и внутри самих компонентов модели — между модулями внимания и логики. Это помогает понять, использует ли модель информацию из контекста или из своих внутренних знаний.

Самое важное, что новый подход позволяет воздействовать на эти признаки — усиливать одни и подавлять другие. В результате можно изменять стиль, тему или тональность текста без изменения параметров модели. Такие вмешательства возможны на нескольких уровнях одновременно, что делает управление более точным и стабильным.

Метод не требует дополнительных данных и работает с уже обученными моделями. Это делает его особенно ценным для исследовательских и коммерческих проектов, где ресурсы ограничены.

Разработка может помочь создать более предсказуемые и безопасные ИИ-системы — например, для фильтрации нежелательного контента без необходимости полной перенастройки модели.

Все права защищены

save pdf date >>> 05.12.2025