Российские учёные обнаружили новое свойство больших языковых моделей, которое может ускорить их работу на 10-15% без потери качества. Об этом сообщили в пресс-службе Института искусственного интеллекта, AIRI. Открытие позволит тратить меньше вычислительных ресурсов на развёртку и инференс нейросетей.

Генеральный директор AIRI Иван Оселедец отметил, что эффект, который выявили исследователи, выглядит очень контринтуитивно и противоречит многим представлениям о глубоком обучении. Однако именно это свойство позволяет повысить эффективность работы языковых моделей. Исследователи из AIRI, SberAI и «Сколтеха» изучили устройство двух десятков языковых моделей с открытым исходным кодом и выявили высокую линейную зависимость в числовых представлениях данных, что упрощает архитектуру моделей.

Андрей Белёвцев, старший вице-президент Сбербанка, заявил, что банк планирует тестировать и, в случае успеха, внедрять эту идею в свои флагманские модели. По его словам, такие находки в ИИ-архитектурах помогают частично компенсировать вычислительный голод. Учёные разработали алгоритм-регуляризатор и выложили его в открытый доступ, чтобы поделиться этим важным открытием с научным сообществом.