Ученые из Швейцарской политехнической школы Лозанны (EPFL) обнаружили в крупных языковых моделях (LLM) специализированные элементы, которые напоминают языковую сеть человеческого мозга. Эти «языковые» нейроны играют важную роль в обработке текста, и их отключение резко снижает способности моделей.

Команда из лабораторий NeuroAI и обработки естественного языка изучила 18 популярных языковых моделей. Ученые сравнивали активность нейронов при обработке осмысленных предложений и случайных списков слов. Нейроны, которые активнее реагировали на предложения, были названы «языковыми». Их оказалось менее 1% от общего числа — около 100 нейронов. Когда эти элементы удаляли, модели теряли способность генерировать связный текст и плохо справлялись с языковыми тестами. Удаление случайных нейронов таких последствий не вызывало.

Метод исследования заимствован из нейронауки, где подобные подходы помогают изучать функции мозга. Ученые удивились, что простая техника, используемая для анализа мозга, так эффективно выявила ключевые элементы в искусственном интеллекте (ИИ). Это открытие упрощает понимание того, как модели обрабатывают язык, без сложных методов машинного обучения.

Исследователи также проверили, есть ли в моделях нейроны, отвечающие за логическое мышление или социальное взаимодействие, как в сетях мозга, связанных с теорией разума. В некоторых моделях такие элементы нашлись, в других — нет. Ученые планируют выяснить, почему так происходит и как это связано с обучением моделей или данными.

В будущем команда хочет изучить мультимодальные модели, которые работают не только с текстом, но и с изображениями, видео и звуком.