Ученые из Microsoft разработали новую языковую модель BitNet b1.58 2B4T, которая может работать даже на обычных процессорах. Это облегченная версия искусственного интеллекта (ИИ) с двумя миллиардами параметров, обученная на четырех триллионах токенов. Модель уже доступна в открытом доступе на платформе Hugging Face и, по оценкам, способна запускаться, например, на чипе Apple M2.

BitNet использует 1-битные веса, представленные всего тремя значениями: -1, 0 и +1. За счет этого модель требует гораздо меньше памяти и вычислительных ресурсов по сравнению с аналогами, работающими с 16- или 32-битными данными. Несмотря на простоту, в точности BitNet уступает более сложным моделям, но масштаб тренировочных данных — эквивалент более 33 миллионов книг — частично компенсирует это.

В тестах BitNet показала достойные результаты, сравнившись с другими компактными ИИ, такими как Gemma 3 от Google и Qwen 2.5 от Alibaba. При этом она использует всего 400 МБ оперативной памяти — менее трети от объема, необходимого ближайшему аналогу.

Чтобы добиться таких результатов, модель должна работать через специальную библиотеку bitnet.cpp, доступную на GitHub. Без нее, например, при использовании стандартной библиотеки transformers, эффективность будет значительно ниже.

Такие легкие модели могут стать решением для запуска ИИ на устройствах с ограниченными ресурсами, сократив нагрузку на дата-центры и сделав технологии более доступными для обычных пользователей.