Numenta Accelerates Large Language Models with Intel® Xeon® CPU Max Series

Numenta, a pioneer in applying brain-based principles to develop innovative AI solutions, has made breakthrough advances in AI and Deep Learning that enable customers to achieve 10 to more than 100X performance improvement across broad use cases, such as natural language processing and computer vision. Numenta demonstrated their custom-trained large language models can run 20X faster for large documents (long sequence lengths) when they run on Intel® Xeon® CPU Max Series processors with high bandwidth memory (HBM) on the processor vs current generation AMD Milan CPU implementations.[2] Numenta's work demonstrates the capacity to dramatically reduce the overall cost of running language models in production on Intel, unlocking entirely new natural language processing (NLP) capabilities for customers.

Products and Solutions
Intel® Xeon® CPU Max Series
4th Gen Intel® Xeon® Scalable processors
Intel® Advanced Matrix Extensions

Industry
Software Development

Organization Size
11–50

Country
United States

Learn more
Article

1, 2 For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/numenta-hbm-customer-story.html