IBM Newsroom

IBM Announces Availability of Open-Source Mistral AI Model on watsonx, Expands Model Choice to Help Enterprises Scale AI with Trust and Flexibility

- IBM offers an optimized version of Mixtral-8x7B that showed potential to cut latency by up to 75%

- Adds to growing catalogue of IBM, third-party and open-source models to give clients choice and flexibility

- Latest open-source model available on watsonx AI and data platform with enterprise-ready AI studio, data store and governance capabilities

ARMONK, N.Y., Feb. 29, 2024 /PRNewswire/ -- IBM (NYSE: IBM) today announced the availability of the popular open-source Mixtral-8x7B large language model (LLM), developed by Mistral AI, on its watsonx AI and data platform, as it continues to expand capabilities to help clients innovate with IBM's own foundation models and those from a range of open-source providers.

IBM offers an optimized version of Mixtral-8x7B that, in internal testing, was able to increase throughput — or the amount of data that can be processed in a given time period — by 50 percent when compared to the regular model.[1] This could potentially cut latency by 35-75 percent, depending on batch size — speeding time to insights. This is achieved through a process called quantization, which reduces model size and memory requirements for LLMs and, in turn, can speed up processing to help lower costs and energy consumption.

The addition of Mixtral-8x7B expands IBM's open,multi-model strategy to meet clients where they are and give them choice and flexibility to scale enterprise AI solutions across their businesses. Through decades-long AI research and development, open collaboration with Meta and Hugging Face, and partnerships with model leaders, IBM is expanding its watsonx.ai model catalog and bringing in new capabilities, languages, and modalities.

IBM's enterprise-ready foundation model choices and its watsonx AI and data platform can empower clients to usegenerative AI to gain new insights and efficiencies, and create new business models based on principles of trust. IBM enables clients to select the right model for the right use cases and price-performance goals for targeted business domains like finance.

Mixtral-8x7B was built using a combination of Sparse modeling — an innovative technique that finds and uses only the most

essential parts of data to create more efficient models — and the Mixture-of-Experts technique, which combines different models ("experts") that specialize in and solve different parts of a problem. The Mixtral-8x7B model is widely known for its ability to rapidly process and analyze vast amounts of data to provide context-relevant insights.

"Clients are asking for choice and flexibility to deploy models that best suit their unique use cases and business requirements," said Kareem Yusuf, Ph.D, Senior Vice President, Product Management & Growth, IBM Software. "By offering Mixtral-8x7B and other models on watsonx, we're not only giving them optionality in how they deploy AI — we're empowering a robust ecosystem of AI builders and business leaders with tools and technologies to drive innovation across diverse industries and domains."

This week, IBM also announced the availability of ELYZA-japanese-Llama-2-7b, a Japanese LLM model open-sourced by ELYZA Corporation, on watsonx. IBM also offers Meta's open-source models Llama-2-13B-chat and Llama-2-70B-chat and other third-party models on watsonx, with more to come in the next few months.

Statements regarding IBM's future direction and intent are subject to change or withdrawal without notice and represent goals and objectives only.

Media Contact:

Amy Angelini
alangeli@us.ibm.com

[1] Based on IBM testing over two days using internal workloads captured on an instance of watsonx for IBM use.

SOURCE IBM
