NVIDIA & MISTRAL AI RELEASE MISTRAL NEMO 12B LANGUAGE MODEL

STEVE MCDOWELL, CHIEF ANALYST
7/18/24

CONTEXT

Mistral AI and NVIDIA launched Mistral NeMo 12B, a state-of-the-art language model for enterprise applications such as chatbots, multilingual tasks, coding, and summarization. The collaboration combines Mistral AI's training data expertise with NVIDIA's optimized hardware and software ecosystem, offering high performance across diverse applications.

Key Highlights of the Mistral NeMo 12B model announcement include:

- **Collaboration and Training**: Mistral NeMo was trained on the NVIDIA DGX Cloud AI platform using 3,072 H100 80GB Tensor Core GPUs, leveraging NVIDIA's TensorRT-LLM and NeMo development platform for accelerated and optimized performance.
- **Capabilities**: The model excels in multi-turn conversations, math, common sense reasoning, world knowledge, and coding. It features a 128K context length for processing extensive information coherently and accurately.
- **Efficiency and Deployment**: Using the FP8 data format, Mistral NeMo reduces memory size and speeds deployment without loss of accuracy. It is available as an NVIDIA NIM inference microservice, allowing quick deployment in minutes and optimized inference with TensorRT-LLM engines.
- **Enterprise-Grade Support**: Part of NVIDIA AI Enterprise, the model includes comprehensive support, direct access to NVIDIA AI experts, and defined service-level agreements, ensuring reliable performance.

- **Open Source and Accessibility**: Released under the Apache 2.0 license, Mistral NeMo fosters innovation and supports the broader AI community. It is designed to fit on the memory of NVIDIA L40S, GeForce RTX 4090, or RTX 4500 GPUs, offering high efficiency, low compute cost, and enhanced security and privacy.

BACKGROUND: WHO IS MISTRAL AI?

Mistral AI is a cutting-edge technology company developing advanced artificial intelligence models and solutions. Known for its innovative approaches and high-performance AI systems, Mistral AI has established itself as a significant player in the AI landscape.

Mistral AI's mission is to democratize access to advanced AI technologies, making powerful and versatile AI models available to researchers, developers, and enterprises worldwide. The company envisions a future where AI can seamlessly integrate into various applications, driving innovation and efficiency across industries.

KEY MILESTONES

- **Formation and Early Development**: Mistral AI was established by a team of AI researchers and engineers with extensive experience in machine learning, natural language processing, and large-scale data analysis. The founders aimed to address the growing demand for more powerful and versatile AI models.

- **Collaborations and Partnerships**: Early in its history, Mistral AI formed strategic partnerships with major tech companies, including NVIDIA. These collaborations have been instrumental in leveraging the latest advancements in hardware and software to develop highly efficient AI models.

- **Release of Mistral 7B**: One of its significant early achievements was releasing the Mistral 7B model, which set a benchmark for performance and integration ease in various applications.

NEW: MISTRAL NEMO 12B

The Mistral NeMo 12B model was developed through a strategic partnership between Mistral AI and NVIDIA. This collaboration utilized:

- **NVIDIA DGX Cloud AI Platform**: Providing dedicated, scalable access to the latest NVIDIA architecture.
- **NVIDIA TensorRT-LLM**: Accelerating inference performance on large language models.
- **NVIDIA NeMo Development Platform**: Facilitating the building of custom generative AI models.

Mistral NeMo 12B is designed to handle large context windows of up to 128,000 tokens, offering unprecedented accuracy in reasoning, world knowledge, and coding within its size category. Built on a standard architecture, it ensures seamless integration, serving as a drop-in replacement for systems currently using the Mistral 7B model.

KEY FEATURES & ADVANCEMENTS

1. **Large 128K Token Context Window**: Mistral NeMo supports an extensive context window of up to 128,000 tokens, enabling it to handle large and complex datasets easily.
2. **Compatibility**: Designed with a standard architecture, Mistral NeMo can be seamlessly integrated into systems currently using the Mistral 7B model, serving as a direct drop-in replacement.
3. **Open-Source Licensing**: The pre-trained base and instruction-tuned checkpoints are released under the Apache 2.0 license, encouraging widespread use and adaptation by researchers and enterprises.
4. **Quantization Awareness**: The model is trained with quantization awareness, allowing for efficient FP8 inference without any performance loss, making it both powerful and efficient.
5. **Multilingual Capabilities**: Mistral NeMo is optimized for multilingual applications, excelling in languages such as English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi.
6. **Efficient Tekken Tokenizer**: The new Tekken tokenizer, based on Tiktoken, is highly efficient, particularly in compressing source code and various languages. It offers:
    - ~30% more efficiency in compressing source code, Chinese, Italian, French, German, Spanish, and Russian languages.

    - 2x efficiency for Korean and 3x for Arabic compared to previous models.
    - Outperforms the Llama 3 tokenizer for approximately 85% of all languages.

7. **Advanced Instruction Fine-Tuning**: Mistral NeMo has undergone comprehensive fine-tuning, which significantly improves its abilities in:
    - **Instruction Following**: Enhanced ability to follow precise instructions.
    - **Reasoning**: Improved reasoning capabilities.
    - **Multi-turn Conversations**: Better handling of extended conversations.
    - **Code Generation**: Superior accuracy in generating code.
8. **Superior Benchmark Performance**: The model demonstrates state-of-the-art reasoning, world knowledge, and coding accuracy within its size category, outperforming recent open-source models like Gemma 2 9B and Llama 3 8B.

VERSATILITY AND ENTERPRISE READINESS

The model is packaged as an Nvidia NIM inference microservice, offering performance-optimized inference with TensorRT-LLM engines, allowing for deployment anywhere in minutes. This containerized format ensures enhanced flexibility and ease of use for various applications.

ENTERPRISE-GRADE SUPPORT AND SECURITY

As part of Nvidia AI Enterprise, Mistral NeMo 12B also includes comprehensive support features from Nvidia:

- **Dedicated Feature Branches**: Ensuring specialized and reliable performance.
- **Rigorous Validation Processes**: Maintaining high standards of accuracy and efficiency.
- **Enterprise-Grade Security**: Protecting data integrity and privacy.

This allows direct access to Nvidia AI experts and defined service-level agreements, delivering consistent and reliable performance for enterprise users.

OPEN-SOURCE OPTIONS

While the new model is available as part of Nvidia AI Enterprise, its availability is much broader, including availability on Hugging Face. Mistral released NeMo under the Apache 2.0 license, where anyone interested can use the technology.

As a small language model, Mistral NeMo is designed to fit on the memory of affordable accelerators like Nvidias L40S, GeForce RTX 4090, or RTX 4500 GPUs, offering high efficiency, low compute cost, and enhanced security and privacy.

ANALYSIS

In an era where artificial intelligence is rapidly advancing, the role of small language models has become increasingly important. Despite their size, these models offer significant advantages:

- **Efficiency**: Smaller models are computationally less demanding, making them ideal for applications where resources are limited. This efficiency is crucial for real-time applications and edge computing.
- **Accessibility**: With lower resource requirements, small language models can be deployed more widely, democratizing access to advanced AI capabilities.
- **Specialization**: They can be fine-tuned for specific tasks or industries, providing highly specialized performance without the overhead of larger models.

Mistral NeMo 12B embodies these benefits, delivering high performance in a compact form. Its ability to handle large context windows and support multilingual applications makes it a versatile tool for various industries, from tech and finance to healthcare and education.

The AI market is one of the most competitive markets in technology, with giants like OpenAI, IBM, Anthropic, Cohere, and nearly every public cloud provider all working to find the right solutions to bring the value of generative AI into the enterprise. The line between competitor and partner is often blurry, such as Mistral's relationship with Microsoft, which has its internal efforts and a deep relationship with OpenAI. This is a world that continues to evolve.

Mistral AI performs very well in the AI model space, showing the necessary blend of technical competence and execution. Its execution is impressive. In August alone, Mistral released its NeMo model with Nvidia, its new Codestral Mamba model for code generation, and Mathstral for math reasoning and scientific discovery. It has strong relationships with companies like Nvidia, Microsoft, Google Cloud, and Hugging Face but faces equally fierce competition.

Mistral AI was founded with the mission to push the boundaries of AI capabilities. Thanks to its innovative approaches and strategic partnerships, the company has no difficultly adhering to that mission while growing its importance in the field. The release of Mistral NeMo 12B continues that momentum forward. We can't wait to see what's next.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nand-research.com or visit our website at nand-research.com.