Mistral NeMo

Mistral NeMo: our new best small model. A state-of-the-art 12B model with 128k context length, built in collaboration with NVIDIA, and released under the Apache 2.0 license.

July 18, 2024 Mistral AI team

Today, we are excited to release Mistral NeMo, a 12B model built in collaboration with NVIDIA. Mistral NeMo offers a large context window of up to 128k tokens. Its reasoning, world knowledge, and coding accuracy are state-of-the-art in its size category. As it relies on standard architecture, Mistral NeMo is easy to use and a drop-in replacement in any system using Mistral 7B.

We have released pre-trained base and instruction-tuned checkpoints checkpoints under the Apache 2.0 license to promote adoption for researchers and enterprises. Mistral NeMo was trained with quantisation awareness, enabling FP8 inference without any performance loss.

The following table compares the accuracy of the Mistral NeMo base model with two recent open-source pre-trained models, Gemma 2 9B, and Llama 3 8B.

	Context Window	HellaSwag (0-shot)	Winogrande (0-shot)	NaturalQ (5-shot)	TriviaQA (5-shot)	MMLU (5-shot)	OpenBookQA (0-shot)	CommonSense QA (0-shot)	TruthfulQA (0-shot)
Mistral NeMo 12B	**128k**	**83.5%**	**76.8%**	**31.2%**	**73.8%**	68.0%	**60.6%**	**70.4%**	**50.3%**
Gemma 2 9B	8k	80.1%	74.0%	29.8%	71.3%	**71.5%**	50.8%	60.8%	46.6%
Llama 3 8B	8k	80.6%	73.5%	28.2%	61.0%	62.3%	56.4%	66.7%	43.0%

Table 1: Mistral NeMo base model performance compared to Gemma 2 9B and Llama 3 8B.

Multilingual Model for the Masses

The model is designed for global, multilingual applications. It is trained on function calling, has a large context window, and is particularly strong in English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, and Hindi. This is a new step toward bringing frontier AI models to everyone's hands in all languages that form human culture.

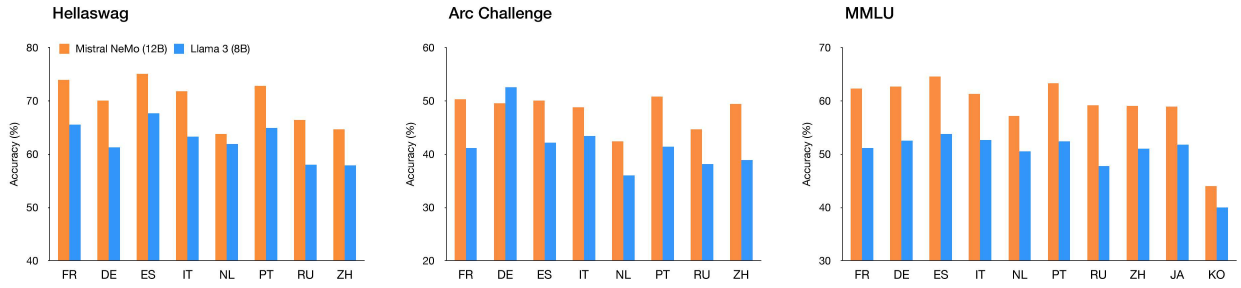

Figure 1: Mistral NeMo performance on multilingual benchmarks.

Tekken, a more efficient tokenizer

Mistral NeMo uses a new tokenizer, Tekken, based on Tiktoken, that was trained on over more than 100 languages, and compresses natural language text and source code more efficiently than the SentencePiece tokenizer used in previous Mistral models. In particular, it is ~30% more efficient at compressing source code, Chinese, Italian, French, German, Spanish, and Russian. It is also 2x and 3x more efficient at compressing Korean and Arabic, respectively. Compared to the Llama 3 tokenizer, Tekken proved to be more proficient in compressing text for approximately 85% of all languages.

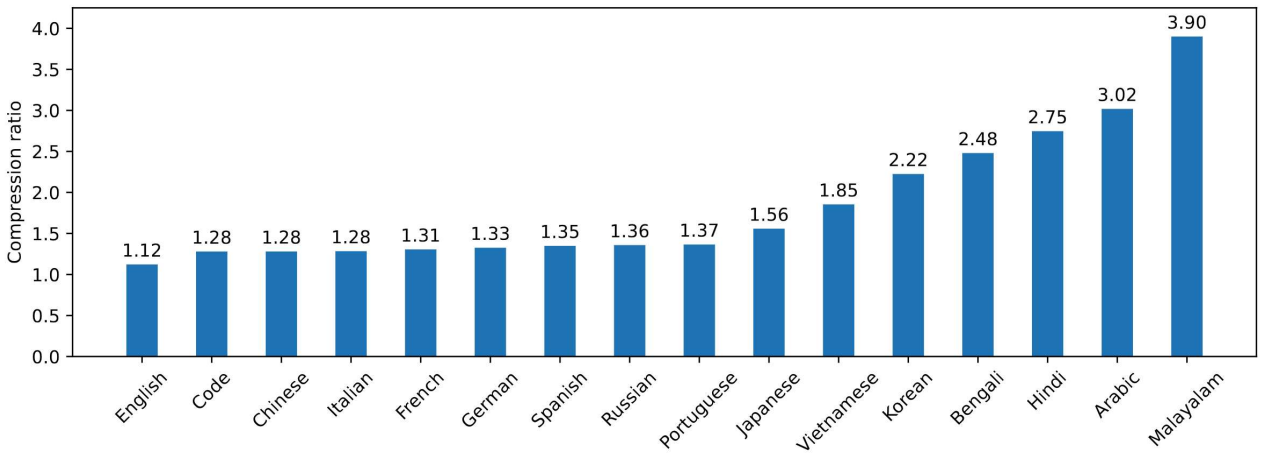

Figure 2: Tekken compression rate.

Instruction fine-tuning

Mistral NeMO underwent an advanced fine-tuning and alignment phase. Compared to Mistral 7B, it is much better at following precise instructions, reasoning, handling multi-turn conversations, and generating code.

	MT Bench	WildBench
Mistral 7B	6.48	25.55
Llama 3 8B	6.85	28.77
Mistral NeMo	**7.84**	**42.57**

Table 2: Mistral NeMo instruction-tuned model accuracy. Evals done with GPT4o as judge on official references.

Links

Weights are hosted on HuggingFace both for the base and for the instruct models. You can try Mistral NeMo now with mistral-inference and adapt it with mistral-finetune. Mistral NeMo is exposed on la Plateforme under the name `open-mistral-nemo-2407`. This model is also packaged in a container as NVIDIA NIM inference microservice and available from ai.nvidia.com.

Careers

Terms of Use

Privacy Policy

Data Processing Agreement