Mistral 7B

The best 7B model to date, Apache 2.0

September 27, 2023 Mistral AI team

Mistral AI team is proud to release Mistral 7B, the most powerful language model for its size to date.

Mistral 7B in short

Mistral 7B is a 7.3B parameter model that:

- Outperforms Llama 2 13B on all benchmarks
- Outperforms Llama 1 34B on many benchmarks
- Approaches CodeLlama 7B performance on code, while remaining good at English tasks
- Uses Grouped-query attention (GQA) for faster inference
- Uses Sliding Window Attention (SWA) to handle longer sequences at smaller cost

We're releasing Mistral 7B under the Apache 2.0 license, it can be used without restrictions.

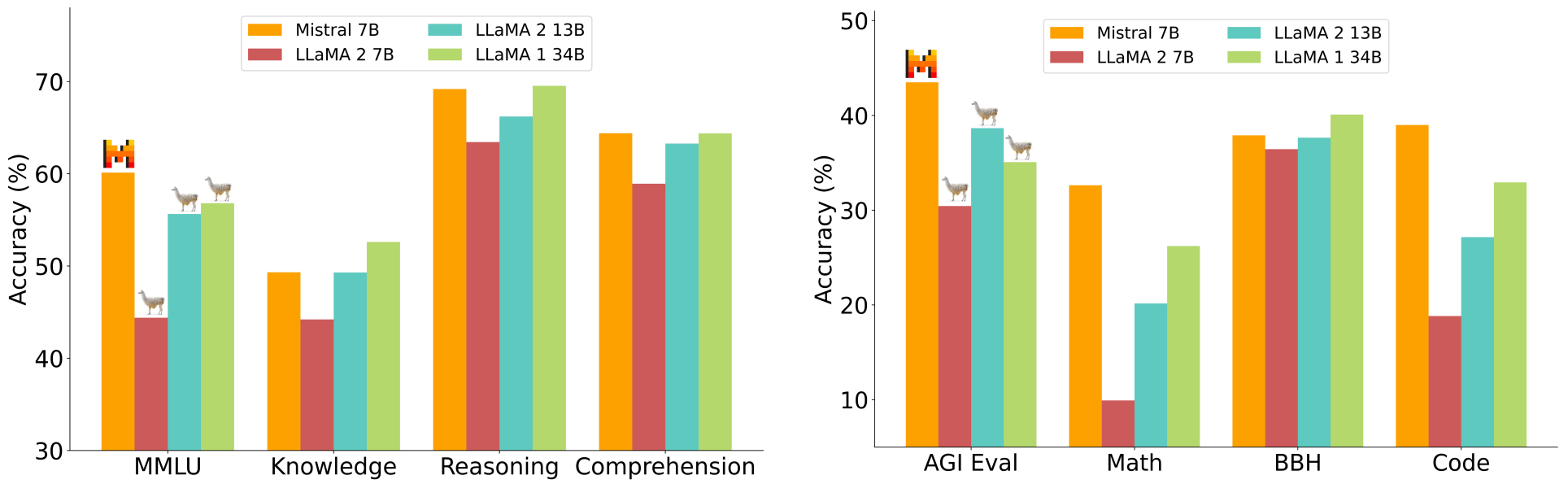
- Download it and use it anywhere (including locally) with our reference implementation,

- Deploy it on any cloud (AWS/GCP/Azure), using vLLM inference server and skypilot,
- Use it on HuggingFace.

Mistral 7B is easy to fine-tune on any task. As a demonstration, we're providing a model fine-tuned for chat, which outperforms Llama 2 13B chat.

Performance in details

We compared Mistral 7B to the Llama 2 family, and re-run all model evaluations ourselves for fair comparison.



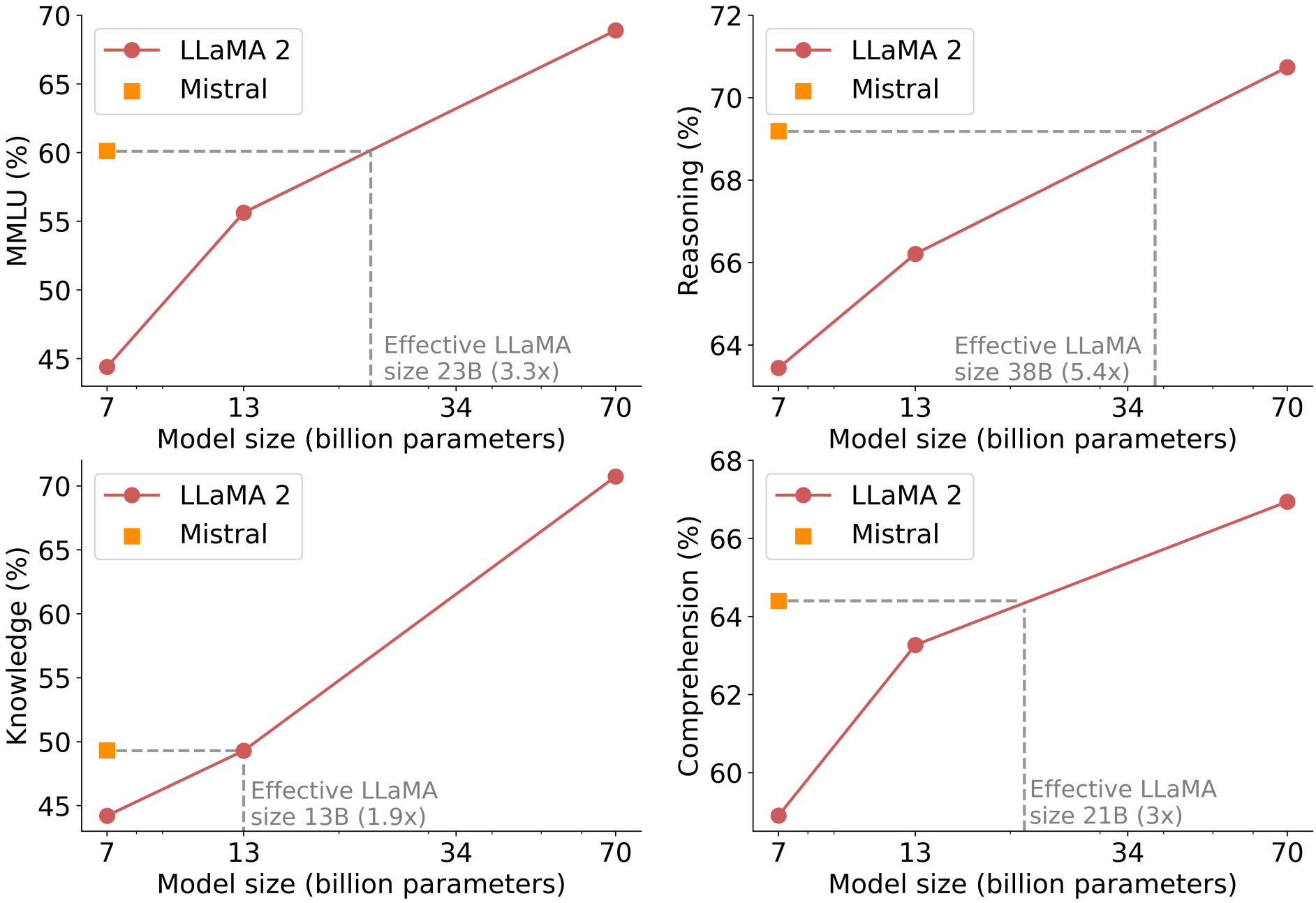
Performance of Mistral 7B and different Llama models on a wide range of benchmarks. For all metrics, all models were re-evaluated with our evaluation pipeline for accurate comparison. Mistral 7B significantly outperforms Llama 2 13B on all metrics, and is on par with Llama 34B (since Llama 2 34B was not released, we report results on Llama 34B). It is also vastly superior in code and reasoning benchmarks.

The benchmarks are categorized by their themes:

- Commonsense Reasoning: 0-shot average of Hellaswag, Winogrande, PIQA, SIQA, OpenbookQA, ARC-Easy, ARC-Challenge, and CommonsenseQA.
- World Knowledge: 5-shot average of NaturalQuestions and TriviaQA.
- Reading Comprehension: 0-shot average of BoolQ and QuAC.
- Math: Average of 8-shot GSM8K with maj@8 and 4-shot MATH with maj@4
- Code: Average of 0-shot Humaneval and 3-shot MBPP
- Popular aggregated results: 5-shot MMLU, 3-shot BBH, and 3-5-shot AGI Eval (English multiple-choice questions only)

Model	Modality	MMLU	HellaSwag	WinoGrande	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	**80.7%**	72.9%	80.8%	75.2%	48.8%	**29.0%**	**69.6%**	18.9%	35.4%	6.0%	34.3%
Code LLaMA 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	**31.1%**	**52.5%**	5.2%	20.8%
Mistral 7B	Pretrained	**60.1%**	**81.3%**	**75.3%**	**83.0%**	**80.0%**	**55.5%**	**28.8%**	**69.9%**	**30.5%**	47.5%	**13.1%**	**52.1%**

An interesting metric to compare how models fare in the cost/performance plane is to compute "equivalent model sizes". On reasoning, comprehension and STEM reasoning (MMLU), Mistral 7B performs equivalently to a Llama 2 that would be more than 3x its size. This is as much saved in memory and gained in throughput.



Results on MMLU, Commonsense Reasoning, World Knowledge and Reading comprehension for Mistral 7B and Llama 2 (7B/13/70B). Mistral 7B largely outperforms Llama 2 13B on all evaluations, except on knowledge benchmarks, where it is on par (this is likely due to its limited parameter count, which restricts the amount of knowledge it can compress).

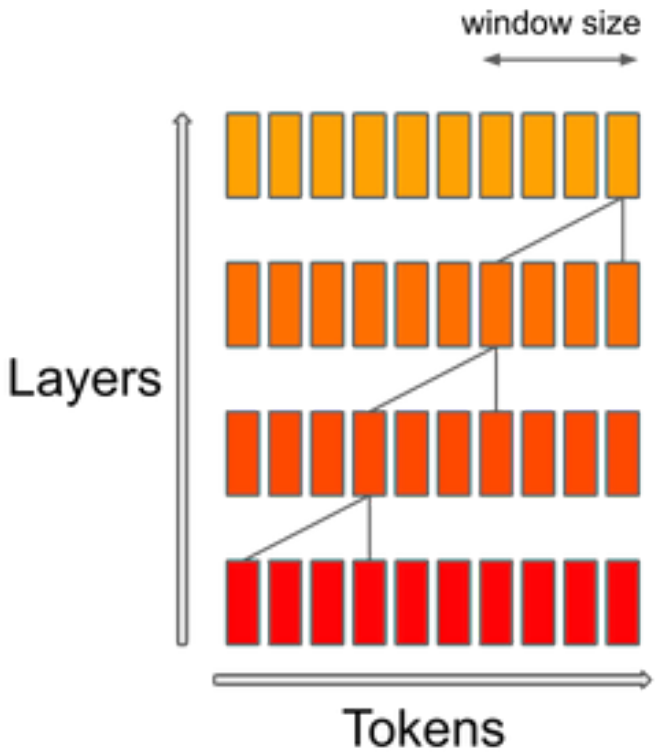
Note: Important differences between our evaluation and the LLaMA2 paper's:

- For MBPP, we use the hand-verified subset
- For TriviaQA, we do not provide Wikipedia contexts

Flash and Furious: Attention drift

Mistral 7B uses a sliding window attention (SWA) mechanism (Child et al., Beltagy et al.), in which each layer attends to the previous `4,096` hidden states. The main improvement, and reason for which this was initially investigated, is a linear compute cost of O(sliding_window.seq_len). In practice, changes made to FlashAttention and xFormers yield a 2x speed improvement for sequence length of 16k with a window of 4k. A huge thanks to Tri Dao and Daniel Haziza for helping include these changes on a tight schedule.

Sliding window attention exploits the stacked layers of a transformer to attend in the past beyond the window size: A token `i` at layer `k` attends to tokens `[i-sliding_window, i]` at layer `k-1`. These tokens attended to tokens `[i-2*sliding_window, i]`. Higher layers have access to information further in the past than what the attention patterns seems to entail.



Finally, a fixed attention span means we can limit our cache to a size of `sliding_window` tokens, using rotating buffers (read more in our reference implementation repo). This saves half of the cache memory for inference on sequence length of `8192`, without impacting model quality.

Fine-tuning Mistral 7B for chat

To show the generalization capabilities of Mistral 7B, we fine-tuned it on instruction datasets publicly available on HuggingFace. No tricks, no proprietary data. The resulting model, Mistral 7B Instruct, outperforms all 7B models on MT-Bench, and is comparable to 13B chat models.

Model	MT Bench
WizardLM-13b-v1.2	7.2
Vicuna-13B-16k	6.92
Mistral 7B Instruct	**6.84 ± 0.065**
WizardLM-13B-v1.1	6.76
Llama-2-13b-chat	6.65
Llama-2-7b-chat	6.27
Vicuna-7B-16k	6.22
Alpaca-13B	4.53

> ⓘ Note
>
> The Mistral 7B Instruct model is a quick demonstration that the base model can be easily fine-tuned to achieve compelling performance. It does not have any moderation mechanism. We're looking forward to engaging with the community on ways to make the model finely respect guardrails, allowing for deployment in environments requiring moderated outputs.

Acknowledgements

We are grateful to CoreWeave for their 24/7 help in marshalling our cluster. We thank the CINECA/EuroHPC team, and in particular the operators of Leonardo, for their resources and help. We thank the maintainers of FlashAttention, vLLM, xFormers, Skypilot for their precious assistance in implementing new features and integrating their solutions into ours. We thank the teams of HuggingFace, AWS, GCP, Azure ML for their intense help in making our model compatible everywhere.

MISTRAL AI_

LINKS

Developers

Technology

Business

About Us

News

ABOUT

Contact Us

Careers

Terms of Use

Privacy Policy

Data Processing Agreement