

WIKIPEDIA

The Free Encyclopedia

Mistral AI

**Mistral AI** is a French company selling artificial intelligence (AI) products. It was founded in April 2023 by previous employees of Meta Platforms and Google DeepMind.<sup>[1]</sup> The company raised €385 million in October 2023<sup>[2]</sup> and in December 2023 it was valued at more than \$2 billion.<sup>[3][4][5]</sup>

It produces open source large language models,<sup>[6]</sup> citing the foundational importance of open-source software, and as a response to proprietary models.<sup>[7]</sup>

As of March 2024, two models have been published and are available as weights.<sup>[8]</sup> Three more models, Small, Medium and Large, are available via API only.<sup>[9][10]</sup>

## History

Mistral AI was co-founded in April 2023 by Arthur Mensch, Guillaume Lample and Timothée Lacroix. Prior to co-founding Mistral AI, Arthur Mensch worked at Google DeepMind which is Google's artificial intelligence laboratory, while Guillaume Lample and Timothée Lacroix worked at Meta Platforms.<sup>[11]</sup> The co-founders met while students at École polytechnique. Mistral is named for a strong wind that blows in France.<sup>[12]</sup>


In June 2023, the start-up carried out a first fundraising of €105 million (\$117 million) with investors including the American fund Lightspeed Venture Partners, Eric Schmidt, Xavier Niel and JCDecaux. The valuation is then estimated by the Financial Times at €240 million (\$267 million).

On 27 September 2023, the company made its language processing model “Mistral 7B” available under the free Apache 2.0 license. This model has 7 billion parameters, a small size compared to its competitors.

On 10 December 2023, Mistral AI announced that it had raised €385 million (\$428 million) as part of its second fundraising. This round of financing notably involves the Californian fund Andreessen Horowitz, BNP Paribas and the software publisher Salesforce.<sup>[13]</sup>

On 11 December 2023, the company released the “Mixtral 8x7B” model with 46.7 billion parameters but using only 12.9 billion per token thanks to the mixture of experts architecture. The model masters 5 languages (French, Spanish, Italian, English and German) and outperforms, according to its developers' tests, the "LLama 2 70B" model from Meta. A version trained to follow instructions and called “Mixtral 8x7B Instruct” is also offered.<sup>[14]</sup>

Mistral AI



MISTRA  
AI

Company type	Private
Industry	Artificial intelligence
Founded	28 April 2023
Founders	Arthur Mensch (Co-Founder & CEO) Guillaume Lample (Co-Founder & Chief Scientist) Timothée Lacroix (Co-Founder & CTO)
Headquarters	Paris, France
Products	Mistral 7B Mixtral 8x7B Mistral Medium Mistral Large
Website	<a href="https://mistral.ai/">mistral.ai</a> ( <a href="https://mistral.ai/">https://mistral.ai/</a> )

On 26 February 2024, [Microsoft](#) announced a new partnership with the company to expand its presence in the rapidly evolving [artificial intelligence](#) industry. Under the agreement, Mistral's rich language models will be available on [Microsoft's Azure cloud](#), while the multilingual conversational assistant "Le Chat" will be launched in the style of [ChatGPT](#).<sup>[15]</sup>

## Models

### Open Weight Models

#### Mistral 7B

Mistral 7B is a 7.3B parameter language model using the transformers architecture. Officially released on September 27, 2023, via a [BitTorrent magnet link](#),<sup>[16]</sup> and [Hugging Face](#).<sup>[17]</sup> The model was released under the [Apache 2.0](#) license. The release blog post claimed the model outperforms [LLaMA 2 13B](#) on all benchmarks tested, and is on par with [LLaMA 34B](#) on many benchmarks tested.<sup>[18]</sup>

Mistral 7B uses a similar architecture to LLaMA, but with some changes to the attention mechanism. In particular it uses Grouped-query attention (GQA) intended for faster inference and Sliding Window Attention (SWA) intended to handle longer sequences.

Sliding Window Attention (SWA) reduces the computational cost and memory requirement for longer sequences. In sliding window attention, each token can only attend to a fixed number of tokens from the previous layer in a "sliding window" of 4096 tokens, with a total context length of 32768 tokens. At inference time, this reduces the cache availability, leading to higher latency and smaller throughput. To alleviate this issue, Mistral 7B uses a rolling buffer cache.

Mistral 7B uses grouped-query attention (GQA), which is a variant of the standard attention mechanism. Instead of computing attention over all the hidden states, it computes attention over groups of hidden states.<sup>[19]</sup>

Both a base model and "instruct" model were released with the later receiving additional tuning to follow chat-style prompts. The fine-tuned model is only intended for demonstration purposes, and does not have guardrails or moderation built-in.<sup>[18]</sup>

#### Mixtral 8x7B

Much like Mistral's first model, Mixtral 8x7B was released via BitTorrent on December 9, 2023,<sup>[6]</sup> and later Hugging Face and a blog post were released two days later.<sup>[14]</sup>

Unlike the previous Mistral model, Mixtral 8x7B uses a sparse [mixture of experts](#) architecture. The model has 8 distinct groups of "experts", giving the model a total of 46.7B usable parameters.<sup>[20][21]</sup> Each single token can only use 12.9B parameters, therefore giving the speed and cost that a 12.9B parameter model would incur.<sup>[14]</sup>

Mistral AI's testing shows the model beats both LLaMA 70B, and [GPT-3.5](#) in most [benchmarks](#).<sup>[22]</sup>

In March 2024, research conducted by Patronus AI comparing performance of LLMs on a 100-question test with prompts to generate text from books protected under [U.S. copyright law](#) found that [Open AI's GPT-4](#), Mixtral, [Meta AI's LLaMA-2](#), and [Anthropic's Claude2](#) generated

copyrighted text verbatim in 44%, 22%, 10%, and 8% of responses respectively.<sup>[23][24]</sup>

## API-Only Models

Unlike Mistral 7B and Mixtral 8x7B, the following models are closed-source and only available through the Mistral API.<sup>[25]</sup>

### Mistral Large

Mistral Large was launched on February 26, 2024, and Mistral claims it is second in the world to OpenAI's GPT-4.

It is fluent in English, French, Spanish, German, and Italian, with Mistral claiming understan of both grammar and cultural context, and provides coding capabilities. As of early 2024, Mistral's flagship AI.<sup>[26]</sup> It is also available on Microsoft Azure.

### Mistral Medium

Mistral Medium is trained in various languages including English, French, Italian, German, Spanish and code with a score of 8.6 on MT-Bench.<sup>[27]</sup> It is ranked in performance above Claude and below GPT-4 on the LMSys ELO Arena benchmark.<sup>[28]</sup>

The number of parameters, and architecture of Mistral Medium is not known as Mistral has not published public information about it.

### Mistral Small

Like the Large model, Small was launched on February 26, 2024. It is intended to be a light-weight model for low latency, with better performance than Mixtral 8x7B.<sup>[29]</sup>

## References

1. "France's unicorn start-up Mistral AI embodies its artificial intelligence hopes" ([https://www.lemonde.fr/en/economy/article/2023/12/12/french-unicorn-start-up-mistral-ai-embodies-its-artificial-intelligence-hopes\\_6337125\\_19.html](https://www.lemonde.fr/en/economy/article/2023/12/12/french-unicorn-start-up-mistral-ai-embodies-its-artificial-intelligence-hopes_6337125_19.html)). Le Monde.fr. 2023-12-12. Retrieved 2023-12-16.
2. Metz, Cade (10 December 2023). "Mistral, French A.I. Start-Up, Is Valued at \$2 Billion in Funding Round" (<https://www.nytimes.com/2023/12/10/technology/mistral-ai-funding.html>). The New York Times.
3. Fink, Charlie. "This Week In XR: Epic Triumphs Over Google, Mistral AI Raises \$415 Million, \$56.5 Million For Essential AI" (<https://www.forbes.com/sites/chariefink/2023/12/14/this-week-in-xr-epic-triumphs-over-google-mistral-ai-raises-415-million-565-million-for-essential-ai/>). Forbes. Retrieved 2023-12-16.
4. "A French AI start-up may have commenced an AI revolution, silently" (<https://www.hindustantimes.com/business/a-french-ai-start-up-may-have-commenced-an-ai-revolution-silently-101702370816617.html>). Hindustan Times. December 12, 2023.
5. "French AI start-up Mistral secures €2bn valuation" (<https://www.ft.com/content/ea29ddf8-91cb-45e8-86a0-f501ab7ad9bb>). ft.com Financial Times.
6. "Buzzy Startup Just Dumps AI Model That Beats GPT-3.5 Into a Torrent Link" (<https://gizmodo.com/mistral-artificial-intelligence-gpt-3-openai-1851091217>). Gizmodo. 2023-12-12. Retrieved 2023-12-16.
7. "Bringing open AI models to the frontier" (<https://mistral.ai/news/about-mistral-ai/>). Mistral AI. 27 September 2023. Retrieved 4 January 2024.

8. "Open-weight models and Mistral AI Large Language Models" (<https://docs.mistral.ai/models/>). *docs.mistral.ai*. Retrieved 2024-01-04.
9. "Endpoints and Mistral AI Large Language Models" (<https://docs.mistral.ai/platform/endpoints/#medium>). *docs.mistral.ai*.
10. "Endpoints and benchmarks | Mistral AI Large Language Models" (<https://docs.mistral.ai/platform/endpoints/>). *docs.mistral.ai*. Retrieved 2024-03-06.
11. "France's unicorn start-up Mistral AI embodies its artificial intelligence hopes" ([https://www.lemonde.fr/en/economy/article/2023/12/12/french-unicorn-start-up-mistral-ai-embodies-its-artificial-intelligence-hopes\\_6337125\\_19.html](https://www.lemonde.fr/en/economy/article/2023/12/12/french-unicorn-start-up-mistral-ai-embodies-its-artificial-intelligence-hopes_6337125_19.html)). *Le Monde.fr*. 12 December 2023.
12. Journal, Sam Schechner | Photographs by Edouard Jacquinet for The Wall Street. "The 9-Month AI Startup Challenging Silicon Valley's Giants" (<https://www.wsj.com/tech/ai/the-9-month-old-artup-challenging-silicon-valleys-giants-ee2e4c48>). *WSJ*. Retrieved 2024-03-31.
13. "Mistral lève 385 M€ et devient une licorne française - le Monde Informatique" (<https://www.ledeinformatique.fr/actualites/lire-mistral-leve-385-meteuro-et-devient-une-licorne-francaise-97202.html>). 11 December 2023.
14. "Mixtral of experts" (<https://mistral.ai/news/mixtral-of-experts/>). *mistral.ai*. 2023-12-11. Retrieved 2024-01-04.
15. Bableshwar (2024-02-26). "Mistral Large, Mistral AI's flagship LLM, debuts on Azure AI Models-as-a-Service" (<https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/mistral-large-mistral-ai-s-flagship-llm-debuts-on-azure-ai/ba-p/4066996>). *techcommunity.microsoft.com*. Retrieved 2024-02-26.
16. Goldman, Sharon (2023-12-08). "Mistral AI bucks release trend by dropping torrent link to new open source LLM" (<https://venturebeat.com/ai/mistral-ai-bucks-release-trend-by-dropping-torrent-link-to-new-open-source-llm/>). *VentureBeat*. Retrieved 2024-01-04.
17. Coldewey, Devin (27 September 2023). "Mistral AI makes its first large language model free for everyone" (<https://techcrunch.com/2023/09/27/mistral-ai-makes-its-first-large-language-model-free-for-everyone/>). *TechCrunch*. Retrieved 4 January 2024.
18. "Mistral 7B" (<https://mistral.ai/news/announcing-mistral-7b/>). *mistral.ai*. Mistral AI. 27 September 2023. Retrieved 4 January 2024.
19. Jiang, Albert Q.; Sablayrolles, Alexandre; Mensch, Arthur; Bamford, Chris; Chaplot, Devendra Singh; Casas, Diego de las; Bressand, Florian; Lengyel, Gianna; Lample, Guillaume (2023-10-10). "Mistral 7B". arXiv:2310.06825v1 (<https://arxiv.org/abs/2310.06825v1>) [cs.CL (<https://arxiv.org/archive/cs>.CL)].
20. "Mixture of Experts Explained" (<https://huggingface.co/blog/moe>). *huggingface.co*. Retrieved 2024-01-04.
21. Marie, Benjamin (2023-12-15). "Mixtral-8x7B: Understanding and Running the Sparse Mixture of Experts" (<https://towardsdatascience.com/mixtral-8x7b-understanding-and-running-the-sparse-mixture-of-experts-0e3fc7fde818>). *Medium*. Retrieved 2024-01-04.
22. Franzen, Carl (2023-12-11). "Mistral shocks AI community as latest open source model eclipses GPT-3.5 performance" (<https://venturebeat.com/ai/mistral-shocks-ai-community-as-latest-open-source-model-eclipses-gpt-3-5-performance/>). *VentureBeat*. Retrieved 2024-01-04.
23. Field, Hayden (March 6, 2024). "Researchers tested leading AI models for copyright infringement using popular books, and GPT-4 performed worst" (<https://www.cnbc.com/2024/03/06/gpt-4-researchers-tested-leading-ai-models-for-copyright-infringement.html>). *CNBC*. Retrieved March 6, 2024.
24. "Introducing CopyrightCatcher, the first Copyright Detection API for LLMs" (<https://www.patronus.ai/blog/introducing-copyright-catcher>). Patronus AI. March 6, 2024. Retrieved March 6, 2024.
25. "Pricing and rate limits | Mistral AI Large Language Models" (<https://docs.mistral.ai/platform/pricing/>). *docs.mistral.ai*. Retrieved 2024-01-22.
26. AI, Mistral (2024-02-26). "Au Large" (<https://mistral.ai/news/mistral-large/>). *mistral.ai*. Retrieved 2024-03-06.

27. AI, Mistral (2023-12-11). "La plateforme" (<https://mistral.ai/news/la-plateforme/>). *mistral.ai*. Retrieved 2024-01-22.
28. "LMSys Chatbot Arena Leaderboard - a Hugging Face Space by lmsys" (<https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>). *huggingface.co*. Retrieved 2024-01-22.
29. AI, Mistral (2024-02-26). "Au Large" (<https://mistral.ai/news/mistral-large/>). *mistral.ai*. Retrieved 2024-03-06.

---

Retrieved from "[https://en.wikipedia.org/w/index.php?title=Mistral\\_AI&oldid=1216522388](https://en.wikipedia.org/w/index.php?title=Mistral_AI&oldid=1216522388)"

■

