

Paris start-up Mistral AI releases Mixtral 8x7B under Apache 2.0

Source copy kept by wheresthe.ai for the story "Paris start-up Mistral AI releases Mixtral 8x7B under Apache 2.0" (<https://wheresthe.ai/d/fr-mistral-mixtral-8x7b-release-2023-12/>).

Source: <https://mistral.ai/news/mixtral-of-experts>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 829c5560120882ebd41b5f71a40b7c57a08d1047b2c130c830a2f4ab53f4efa6

Research Mixtral of experts December 11, 2023 By Mistral team

[Back to Blog](#)

4 min read

[Share this post](#)

[Copy url to clipboard](#) Copied

Mistral AI continues its mission to deliver the best open models to the developer community. Moving forward in AI requires taking new technological turns beyond reusing well-known architectures and training paradigms. Most importantly, it requires making the community benefit from original models to foster new inventions and usages. Today, the team is proud to release Mixtral 8x7B, a high-quality sparse mixture of experts model (SMoE) with open weights. Licensed under Apache 2.0. Mixtral outperforms Llama 2 70B on most benchmarks with 6x faster inference. It is the strongest open-weight model with a permissive license and the best model overall regarding cost/performance trade-offs. In particular, it matches or outperforms GPT3.5 on most standard benchmarks. Mixtral has the following capabilities. It gracefully handles a context of 32k tokens.

It handles English, French, Italian, German and Spanish.

It shows strong performance in code generation.

It can be finetuned into an instruction-following model that achieves a score of 8.3 on MT-Bench.

Pushing the frontier of open models with sparse architectures Mixtral is a sparse mixture-of-experts network. It is a decoder-only model where the feedforward block picks from a set of 8 distinct groups of parameters. At every layer, for every token, a router network chooses two of these groups (the "experts") to process the token and combine their output additively. This technique increases the number of parameters of a model while controlling cost and latency, as the model only uses a fraction of the total set of parameters per token. Concretely, Mixtral has 46.7B total parameters but only uses 12.9B parameters per token. It, therefore, processes input and generates output at the same speed and for the same cost as a 12.9B model. Mixtral is pre-trained on data extracted from the open Web – we train experts and routers simultaneously. Performance We compare Mixtral to the Llama 2 family and the GPT3.5 base model. Mixtral matches or outperforms Llama 2 70B, as well as GPT3.5, on most benchmarks.

On the following figure, we measure the quality versus inference budget tradeoff. Mistral 7B and Mixtral 8x7B belong to a family of highly efficient models compared to Llama 2 models.

The following table give detailed results on the figure above.

Hallucination and biases. To identify possible flaws to be corrected by fine-tuning / preference modelling, we measure the base model performance on BBQ/BOLD.

Compared to Llama 2, Mixtral presents less bias on the BBQ benchmark. Overall, Mixtral displays more positive sentiments than Llama 2 on BOLD, with similar variances within each dimension. Language. Mixtral 8x7B masters French, German, Spanish, Italian, and English.

Instructed models We release Mixtral 8x7B Instruct alongside Mixtral 8x7B. This model has been optimised through supervised fine-tuning and direct preference optimisation (DPO) for careful instruction following. On MT-Bench, it reaches a score of 8.30, making it the best open-source model, with a performance comparable to GPT3.5. Note: Mixtral can be gracefully prompted to ban some outputs from constructing applications that require a strong level of moderation, as exemplified here . A proper preference tuning can also serve this purpose. Bear in

mind that without such a prompt, the model will just follow whatever instructions are given. Deploy Mixtral with an open-source deployment stack To enable the community to run Mixtral with a fully open-source stack, we have submitted changes to the vLLM project, which integrates Megablocks CUDA kernels for efficient inference. Skypilot allows the deployment of vLLM endpoints on any instance in the cloud. Use Mixtral on our platform. We're currently using Mixtral 8x7B behind our endpoint mistral-small , which is available in beta . Register to get early access to all generative and embedding endpoints. Acknowledgement We thank CoreWeave and Scaleway teams for technical support as we trained our models.

0%