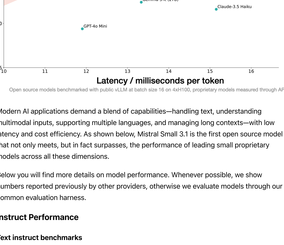
Mistral Small 3.1 | Mistral AI

Today we announce Mistral Small 3.1: the best model in its weight class.

Building on Mistral Small 3, this new model comes with improved text performance, multimodal understanding, and an expanded context window of up to 128k tokens. The model outperforms comparable models like Gemma 3 and GPT-4o Mini, while delivering inference speeds of 150 tokens per second.

Mistral Small 3.1 is released under an Apache 2.0 license.



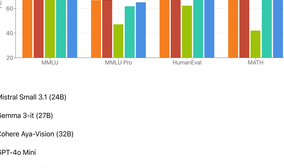
Open source models benchmarked with public vLLM at batch size 16 on 4xH100s, proprietary models measured through API

Modern AI applications demand a blend of capabilities—handling text, understanding multimodal inputs, supporting multiple languages, and managing long contexts—with low latency and cost efficiency. As shown below, Mistral Small 3.1 is the first open source model that not only meets, but in fact surpasses, the performance of leading small proprietary models across all these dimensions.

Below you will find more details on model performance. Whenever possible, we show numbers reported previously by other providers, otherwise we evaluate models through our common evaluation harness.

Instruct Performance

Text instruct benchmarks



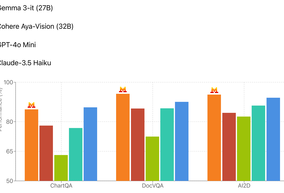
Mistral Small 3.1 (24B)

Gemma 3-it (27B)

Cohere Aya-Vision (32B)

GPT-4o Mini

Claude-3.5 Haiku



Mistral Small 3.1 (24B)

Gemma 3-it (27B)

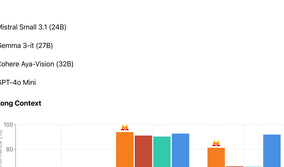
Cohere Aya-Vision (32B)

GPT-4o Mini

Claude-3.5 Haiku

Multimodal Instruct Benchmarks

MM-MT-Bench scaled to between 0 and 100.



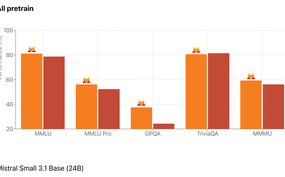
Mistral Small 3.1 (24B)

Gemma 3-it (27B)

Cohere Aya-Vision (32B)

GPT-4o Mini

Claude-3.5 Haiku



Mistral Small 3.1 (24B)

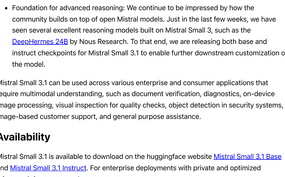
Gemma 3-it (27B)

Cohere Aya-Vision (32B)

GPT-4o Mini

Claude-3.5 Haiku

Multilingual



Mistral Small 3.1 (24B)

Gemma 3-it (27B)

Cohere Aya-Vision (32B)

GPT-4o Mini

Long Context

Mistral Small 3.1 (24B)

Gemma 3-it (27B)

GPT-4o Mini

Claude-3.5 Haiku

Pretrained Performance

We also release the pretrained base model for Mistral Small 3.1.

All pretrain

Mistral Small 3.1 Base (24B)

Gemma 3-pt (27B)

Use cases

Mistral Small 3.1 is a versatile model designed to handle a wide range of generative AI tasks, including instruction following, conversational assistance, image understanding, and function calling. It provides a solid foundation for both enterprise and consumer-grade AI applications.

Key Features and Capabilities

- Lightweight: Mistral Small 3.1 can run on a single RTX 4090 or a Mac with 32GB RAM. This makes it a great fit for on-device use cases.
- Fast-response conversational assistance: Ideal for virtual assistants and other applications where quick, accurate responses are essential.
- Low-latency function calling: Capable of rapid function execution within automated or agentic workflows
- Fine-tuning for specialized domains: Mistral Small 3.1 can be fine-tuned to specialize in specific domains, creating accurate subject matter experts. This is particularly useful in fields like legal advice, medical diagnostics, and technical support.
- Foundation for advanced reasoning: We continue to be impressed by how the community builds on top of open Mistral models. Just in the last few weeks, we have seen several excellent reasoning models built on Mistral Small 3, such as the DeepHermes 24B by Nous Research. To that end, we are releasing both base and instruct checkpoints for Mistral Small 3.1 to enable further downstream customization of the model.

Mistral Small 3.1 can be used across various enterprise and consumer applications that require multimodal understanding, such as document verification, diagnostics, on-device image processing, visual inspection for quality checks, object detection in security systems, image-based customer support, and general purpose assistance.

Availability

Mistral Small 3.1 is available to download on the huggingface website Mistral Small 3.1 Base and Mistral Small 3.1 Instruct. For enterprise deployments with private and optimized inference infrastructure, please contact us.

You can also try the model via API on Mistral AI's developer playground La Plateforme starting today. The model is also available on Google Cloud Vertex AI. Mistral Small 3.1 will be available on NVIDIA NIM and Microsoft Azure AI Foundry in the coming weeks.

Happy building!