

Mistral releases Small 4 under Apache 2.0 and joins NVIDIA's Nemotron Coalition

Source copy kept by wheresthe.ai for the story "Mistral releases Small 4 under Apache 2.0 and joins NVIDIA's Nemotron Coalition" (<https://wheresthe.ai/d/eu-mistral-small-4-nvidia-coalition-2026-03/>).

Source: <https://mistral.ai/news/mistral-small-4>

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

SHA-256 of the saved text: b9b8dddc7c7e6fba973a2102773cbbd1554b81156557b3da76b209f5e6f00aa5

Research Introducing Mistral Small 4 March 16, 2026 By Mistral

[Back to Blog](#)

5 min read

[Share this post](#)

[Copy url to clipboard](#) Copied

Today, we are announcing Mistral Small 4. This model is the next major release in the Mistral Small family. Mistral Small 4 is the first Mistral model to unify the capabilities of our flagship models, Magistral for reasoning, Pixtral for multimodal, and Devstral for agentic coding, into a single, versatile model. With Small 4, users no longer need to choose between a fast instruct model, a powerful reasoning engine, or a multimodal assistant: one model now delivers all three, with configurable reasoning effort and best-in-class efficiency. Mistral Small 4 is released under the Apache 2.0 license, continuing our commitment to open, accessible, and customizable AI. A new standard for multimodal, reasoning-optimized models Mistral Small 4 is a hybrid model optimized for general chat, coding, agentic tasks, and complex reasoning. Its architecture supports both text and image inputs, making it versatile for a wide range of applications. With Mistral Small 4, we reaffirm our commitment to open-source models and are proud to join the NVIDIA Nemotron Coalition as a founding member, advancing collaboration and innovation in AI development. Key architectural details Mixture of Experts (MoE): 128 experts, with 4 active per token, enabling efficient scaling and specialization.

119B total parameters, with 6B active parameters per token (8B including embedding and output layers).

256k context window, supporting long-form interactions and document analysis.

Configurable reasoning effort: Toggle between fast, low-latency responses and deep, reasoning-intensive outputs.

Native multimodality: Accepts both text and image inputs, unlocking use cases from document parsing to visual analysis.

Performance highlights 40% reduction in end-to-end completion time (latency-optimized setup).

3x more requests per second (throughput-optimized setup) compared to Mistral Small 3.

Why Mistral Small 4? Unified capabilities Mistral Small 4 consolidates the strengths of Magistral (reasoning), Devstral (coding agents), and Mistral Small (instruct) into a single model. Whether you need a chat assistant, a research partner, or a coding agent, Small 4 adapts to your task, no need to switch between specialized models. Reasoning on demand With the new `reasoning_effort` parameter, users can dynamically adjust the model's behavior: `reasoning_effort="none"` : Fast, lightweight responses for everyday tasks, equivalent to the same chat style of Mistral Small 3.2.

`reasoning_effort="high"` : Deep, step-by-step reasoning for complex problems, with equivalent verbosity to previous Magistral models.

Enterprise-grade efficiency Minimum infrastructure: 4x NVIDIA HGX H100 , 2x NVIDIA HGX H200 , or 1x NVIDIA DGX B200 .

Recommended setup: 4x NVIDIA HGX H100, 4x NVIDIA HGX H200, or 2x NVIDIA DGX B200 for optimal performance.

Mistral Small 4 is fully open source. Fine-tune it for specialized tasks or deploy it out of the box for general-

purpose use. Thanks to collaboration with the community, it's now available on vLLM, llama.cpp, SGLang, Transformers, and more.

Delivering advanced open-source AI models requires broad optimization. Through close collaboration with NVIDIA, inference has been optimized for both open source vLLM and SGLang, ensuring efficient, high-throughput serving across deployment scenarios.

Figure: Score vs. Output Length across three benchmarks. Top: accuracy scores (higher is better). Bottom: average output length in thousands of characters (shorter is better). Mistral Small 4 with reasoning achieves competitive scores, matching or surpassing GPT-OSS 120B on all three benchmarks, while generating significantly shorter outputs. On AA LCR, Mistral Small 4 scores 0.72 with just 1.6K characters, whereas Qwen models need 3.5-4x more output (5.8-6.1K) for comparable performance. On LiveCodeBench, Mistral Small 4 outperforms GPT-OSS 120B while producing 20% less output. This efficiency gap matters in practice: shorter outputs mean lower latency, reduced inference costs, and a better user experience. For enterprise buyers:

Efficiency per token directly impacts cost and scalability. Models that maintain or improve performance as responses grow longer reduce the need for manual intervention, lower operational costs, and ensure consistent quality, even for complex, high-stakes tasks like report generation, customer support, or decision-making workflows. Hybrid reasoning models deliver better value by maximizing accuracy without proportional increases in resource use, making them ideal for large-scale deployments where both performance and cost-efficiency are critical. For technical teams and data scientists:

Performance per token is a key metric for model selection and optimization. Models that scale efficiently allow teams to deploy solutions for longer, more nuanced tasks (e.g., detailed analytics, multi-step reasoning) without sacrificing accuracy or inflating computational costs. This means fewer trade-offs between quality and resource allocation, enabling more innovative and reliable AI-driven applications. It also simplifies fine-tuning and integration, as the model's robustness reduces the need for constant adjustments or fallback systems. Intended use cases Mistral Small 4 is designed for: Developers: Coding automation, codebase exploration, and code agentic workflows.

Enterprises: General chat assistants, document understanding, and multimodal analysis.

Researchers: Math, research, and complex reasoning tasks.

Its open-source license and customizable architecture make it ideal for fine-tuning and specialization. Availability
Mistral API and AI Studio

Hugging Face Repository

Developers can prototype Mistral Small 4 for free on NVIDIA accelerated computing at build.nvidia.com, and for production deployment, Mistral Small 4 is available day-0 as an NVIDIA NIM, delivering optimized, containerized inference out of the box. It can also be customized with NVIDIA NeMo for domain-specific fine-tuning.

Technical documentation for customers is available on our AI Governance Hub

For enterprise deployments, custom fine-tuning, or on-premises solutions, contact our team. The future of AI is open. By unifying instruct, reasoning, and multimodal capabilities, Mistral Small 4 simplifies AI integration and empowers users to tackle a wider range of tasks with a single, adaptable tool, bringing the benefits of open source AI to real-world use cases.

Mistral Small 4

Apache 2.0

Hybrid instruct and reasoning MoE; unifies Instruct, Magistral, and Devstral into one model with multimodal input.
Text-to-text Agentic Multimodal Lightweight

Input (/M tokens)

\$0.15

Output (/M tokens)

\$0.6

mistral-small-latest

Copy to clipboard Copied

0%