

Mistral releases Medium 3.5, a 128-billion-parameter open model, and cloud coding agents

Source copy kept by wheresthe.ai for the story "Mistral releases Medium 3.5, a 128-billion-parameter open model, and cloud coding agents" (<https://wheresthe.ai/d/eu-mistral-medium-3-5-remote-agents-2026-05/>).

Source: <https://mistral.ai/news/vibe-remote-agents-mistral-medium-3-5>

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

SHA-256 of the saved text: 450ae794206505ed35934ebac10ad2b0ab60ae3fb6f6f17e966215b5f04499ba

Product Remote agents in Vibe. Powered by Mistral Medium 3.5. May 22, 2026 By Mistral

[Back to Blog](#)

6 min read

[Share this post](#)

[Copy url to clipboard](#) Copied

Thinking Summary

Mistral has moved coding agents to the cloud, enabling parallel, asynchronous task execution via Mistral Medium 3.5—a new 128B open-weight model optimized for long-horizon coding and productivity work. The update introduces remote agents in Mistral Vibe and Le Chat, along with a new Work mode for multi-step tasks like research and cross-tool workflows, all while maintaining human oversight for sensitive actions. Available today on Pro, Team, and Enterprise plans, Mistral Medium 3.5 is priced at \$1.5 per million input tokens and \$7.5 per million output tokens.

Coding agents have mostly lived on your laptop. Today we're moving them to the cloud, where they run on their own, in parallel, and notify you when they're done. You can start them from the Mistral Vibe CLI or directly in Le Chat, offloading a coding task without leaving the conversation. Powering this is Mistral Medium 3.5 in public preview, our new default model in Mistral Vibe and Le Chat, built to run for long stretches on coding and productivity work. The new Work mode in Le Chat (Preview) extends this with a powerful agent for complex, multi-step tasks like research, analysis, and cross-tool actions. Highlights. Mistral Medium 3.5, a new flagship model that merges instruction-following, reasoning, and coding into a single 128B dense model. Released as open weights, under a modified MIT license.

Strong real-world performance at a size that runs self-hosted on as few as four GPUs.

Mistral Vibe remote agents for async coding: sessions run in the cloud, can be spawned from the CLI or Le Chat, and a local CLI session can be teleported up to the cloud.

Start Mistral Vibe coding tasks in Le Chat. Sessions run on the same remote runtime and keep going while you step away.

Work mode in Le Chat runs on a new agent, powered by Mistral Medium 3.5, that works through multi-step tasks, calling tools in parallel until the job is done.

Mistral Medium 3.5. Mistral Medium 3.5 is our first flagship merged model, available in public preview. It is a dense 128B model with a 256k context window, handling instruction-following, reasoning, and coding in a single set of weights. It performs strongly in real-world use, with self-hosting possible on as few as four GPUs. Reasoning effort is now configurable per request, so the same model can answer a quick chat reply or work through a complex agentic run. We trained the vision encoder from scratch to handle variable image sizes and aspect ratios. Mistral Medium 3.5 scores 77.6% on SWE-Bench Verified, ahead of Devstral 2 and models like Qwen3.5 397B A17B. It also has strong agentic capabilities and scores 91.4 on T³-Telecom.

The model was built for long-horizon tasks, calling multiple tools reliably, and producing structured output that downstream code can consume. It is the model that made async cloud agents in Vibe practical to ship. Mistral Medium 3.5 becomes the default model in Le Chat. It also replaces Devstral 2 in our coding agent, Vibe CLI. Vibe remote agents. From today, coding sessions can work through long tasks while you're away. Many can run in

parallel, and you stop being the bottleneck on every step the agent takes. You can start the cloud agents from the Mistral Vibe CLI or from Le Chat. While they run, you can inspect what the agent is doing, with file diffs, tool calls, progress states, and questions surfaced as you go. Ongoing local CLI sessions can be teleported up to the cloud when you want to leave them running, with session history, task state, and approvals carrying across.

Vibe sits between the systems engineering teams already use, with humans in the loop wherever they're needed. It plugs into GitHub for code and pull requests, Linear and Jira for issues, Sentry for incidents, and apps like Slack or Teams for reporting. Each coding session runs in an isolated sandbox, including broad edits and installs. When the work is done, the agent can open a pull request on GitHub and notify you, so you review the result instead of every keystroke that produced it. It fits the high-volume, well-defined work that takes a developer's time without taking their judgment: module refactors, test generation, dependency upgrades, CI investigations, as well as bug fixes.

We use Workflows orchestrated in Mistral Studio to bring Mistral Vibe into Le Chat. We originally built this for our own in-house coding environment, then for our enterprise customers . Today the capability opens up to everyone, who can now launch coding tasks from the web. And without being tied to a local terminal, a developer can run several in parallel. You can start coding sessions directly in Le Chat, so a task described in chat runs on the same remote runtime as the CLI and the web, and comes back later as a finished branch or a draft PR.

New Work mode in Le Chat (Preview). Work mode is a powerful new agentic mode for complex tasks in Le Chat, powered by a new harness and Mistral Medium 3.5. The agent becomes the execution backend for the assistant itself, so Le Chat can read and write, use several tools at once, and work through multi-step projects until it completes what you've asked.

Here's what Work mode enables you to do today. Cross-tool workflows: catch up across email, messages, and calendar in a single run; prepare for a meeting with attendee context, latest news, and talking points pulled from your sources.

Research and synthesis: dive into a topic across the web, internal docs, and connected tools, then produce a structured brief or report you can edit before exporting or sending.

Triage your inbox and draft replies; create issues in Jira from your team and customer discussions; send a summary to your team on Slack.

Sessions persist longer than a typical chat reply, so an agent can keep going across many turns, through trial-and-error, and through to completion. In Work mode, connectors are on by default rather than chosen manually, which lets the agent reach into documents, mailboxes, calendars, and other systems for the rich context it needs to take correct action. Every action the agent takes is visible: you see each tool call and the thinking rationale. Le Chat will ask for explicit approval—based on your permissions—before proceeding with sensitive tasks like sending a message, writing a document, or modifying data. Get started. Mistral Medium 3.5 is available today in Mistral Vibe and Le Chat , and powers remote coding agents and Work mode in Le Chat on the Pro, Team, and Enterprise plans . Through API, it's priced at \$1.5 per million input tokens and \$7.5 per million output tokens. Open weights are on Hugging Face under a modified MIT license. It is also available for prototyping, hosted on NVIDIA GPU-accelerated endpoints on build.nvidia.com and as a scalable containerized inference microservice, NVIDIA NIM . Build the future of agentic systems with us. We're hiring across research, engineering, and product to push agentic systems further. See our open roles .

Mistral Medium 3.5

Modified MIT

Merged dense model unifying instruction, reasoning, and coding; configurable reasoning effort with multimodal input. Text-to-text Agentic Multimodal Coding Reasoning

Input (/M tokens)

\$1.5

Output (/M tokens)

\$7.5

mistral-medium-latest

Copy to clipboard Copied

0%