

Mistral Medium 3.5: A 128B Open-Weight Model with Cloud Coding Agents

Kabui, Charles

2026-05-11

[Read at ToKnow.ai](#)

Mistral Medium 3.5:
128B Open Weights with
Cloud Coding Agents

Vibe remote agents run async in the cloud on four GPUs

- 128B**
Dense parameters, open-weight model
- 77.6%**
SWE-Bench Verified coding benchmark
- 4 GPUs**
Minimum to self-host the full model

May 11, 2026

ToKnow.ai

Mistral AI released Mistral Medium 3.5, a 128-billion parameter dense model with a 256k context window, alongside Vibe remote agents for cloud-based async coding. Medium 3.5 merges instruction-following, reasoning, and coding into a single set of open weights under a Modified MIT license. It scores 77.6% on [SWE-Bench Verified](#), a standard benchmark for AI agents resolving real GitHub issues. Reasoning effort is adjustable per request, so the same

model handles quick replies and complex multi-step runs. API pricing is \$1.5 per million input tokens and \$7.5 per million output. It self-hosts on as few as four GPUs. Vibe remote agents move coding sessions to the cloud: they run in isolated sandboxes, execute in parallel, and can open pull requests when finished. Local terminal sessions can be “teleported” to the cloud mid-task, with full session history carrying over. Vibe plugs into GitHub for code, Linear and Jira for issues, Sentry for incidents, and Slack or Teams for notifications.

For developers, Vibe remote agents mean you can kick off five refactoring jobs, close your laptop, and come back to finished pull requests. For organizations with data sovereignty needs, Medium 3.5 on four GPUs in your own data center delivers frontier-class coding AI without sending code to external APIs. At \$1.5/\$7.5 per million tokens, Mistral undercuts larger competing models while targeting teams that want strong coding results without the compute cost of 400B+ alternatives.

Medium 3.5’s unified dense architecture, handling chat, reasoning, and code in one set of weights, challenges the trend of shipping separate specialized models for each task. If one 128B model can match larger competitors on coding benchmarks while also powering a general assistant, the case for maintaining multiple models weakens.

Read more: [The Complete Coding Agent Landscape in May 2026](#)

Sources:

- [Mistral AI: Remote Agents in Vibe, Powered by Medium 3.5](#)
- [Mistral Medium 3.5 on HuggingFace](#)
- [Mistral Medium 3.5 Model Card](#)
- [Mistral Vibe Product Page](#)

Disclaimer: For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation. Read more: [/terms-of-service](#)