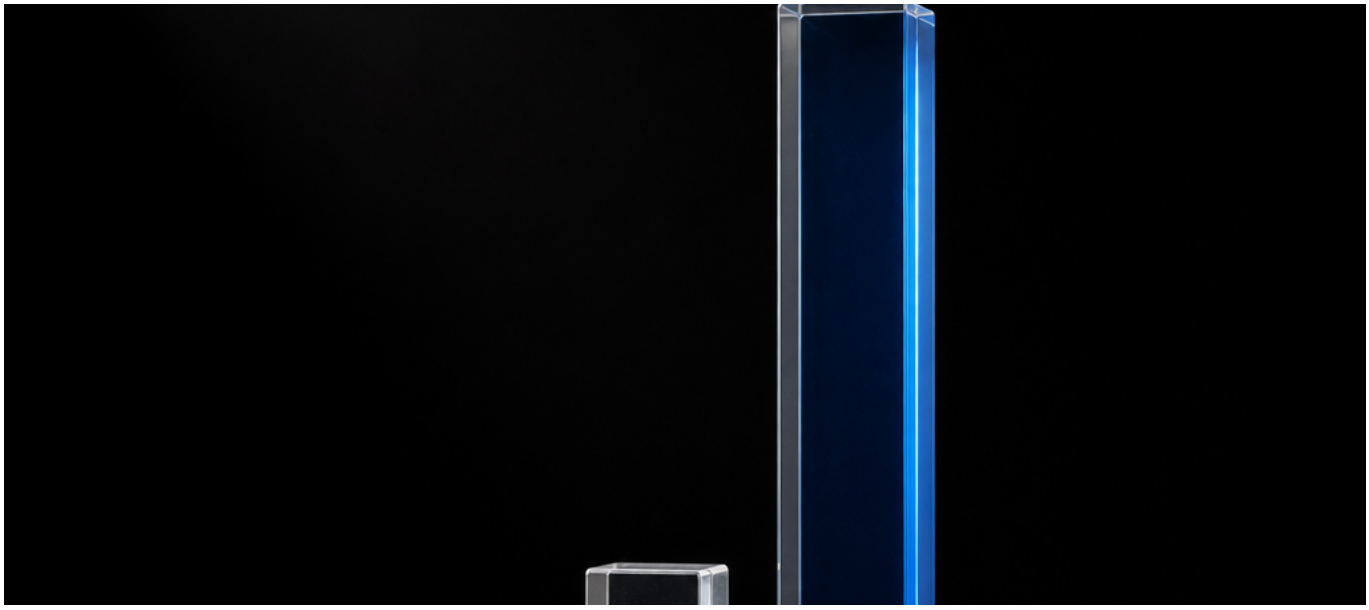


Mistral Large 4 vs DeepSeek V4.1 Flash vs Kimi K3.1 Pro: the honest open-weight comparison for CTOs

Three trillion-class open-weight models shipped in five weeks, all advertising 1M context. What decides your choice is a licence clause and a 3.7x spread in VRAM.

THE SHORT VERSION

For most teams, DeepSeek V4.1 Flash (updated October 17, 2026) is the default: MIT licence, ~429 GB in NVFP4, and the cheapest reported API at $0.15/M input. Mistral Large 4, open-weighted October 29, 2026, wins on agentic coding and cyber work (79% CyberGym-E2E against DeepSeek's 76%) but forbids Model-as-a-Service resale without a separate commercial contract.

ATTRIBUTE	MISTRAL LARGE 4	DEEPSEEK V4.1 FLASH	KIMI K3.1 PRO
Open-weight date	Oct 29, 2026 (upd. Nov 5)	Oct 17, 2026	Nov 2, 2026
Licence	Mistral Ecosystem Open-Weight (modified MIT)	**MIT**	Custom "Kimi K3.1-Pro License"
MaaS resale	Separate commercial licence required	**Permitted**	n/a
Total / active params	1T / ~52B	552B backbone + 196B Engram / 8B prefill, 16B decode	2.8T / 104B
Context	1M	1M	1M
Serving footprint	~610 GB NVFP4 (~233 GB heavily quantised)	**~429 GB NVFP4**	~1,560 GB native MXFP4
API input / output per M	$0.70 / $2.10	**$0.15 / $0.60 off-peak (~2x at peak)**	n/a
CyberGym-E2E	**79%**	76%	n/a
Terminal-Bench 4.0	**41% (up from 28% after post-preview RL)**	n/a	n/a

Two things about that table. Every "n/a" is a real gap in the public record rather than a zero: Kimi K3.1 Pro's benchmark position and API pricing are not in the material I have, and I will not guess them. The licence row is also the only one that cannot be fixed with more hardware budget.

The clause that costs more than the GPUs

Mistral describes the Mistral Ecosystem Open-Weight License as a modified MIT licence. It permits use, modification, fine-tuning and commercialisation. Then the announcement adds the carve-out: a business running a Model-as-a-Service offering must obtain a separate commercial licence first.

That single sentence splits the audience cleanly. If you are embedding the model inside your own product, in your own cloud, serving your own users, you are fine. If your product is

"customers send us prompts, we return completions, we bill per token," you are in negotiation rather than in deployment.

It also has a regulatory tail. The European Commission's AI Act FAQ states that the open-source GPAI exemption requires a free and open-source licence permitting use, modification and distribution without monetisation restrictions, plus public weights and architecture information. A MaaS restriction is a monetisation restriction. My read, and I am a consultant and not your counsel: Mistral Large 4 looks unlikely to qualify for that exemption on the licence limb, while the MIT-licensed DeepSeek weights have no such problem. If your GPAI obligations analysis leans on the open-source carve-out, that distinction is worth a lawyer's hour before it is worth a procurement decision. I made the broader version of this argument in why MIT weights are not open source.

Kimi K3.1 Pro ships under its own custom "Kimi K3.1-Pro License", neither MIT nor Apache. The roundup that reports it does not quote the terms, so I will say only this: a custom licence from a vendor you have not negotiated with before is a legal review, scheduled, with a named owner, before anyone downloads 1.56 TB of weights.

Can we actually serve these on hardware we already have?

Mostly, no. The gap between the three is larger than the gap between any two benchmark scores here.

429 GB
DeepSeek V4.1 Flash, NVFP4
gpuwerk open-weight roundup, autumn 2026

610 GB
Mistral Large 4, NVFP4
gpuwerk, autumn 2026

1,560 GB
Kimi K3.1 Pro, native MXFP4
gpuwerk, autumn 2026

DeepSeek's number is the one that changes plans. At 8B active parameters on prefill and 16B on decode, with a 552B backbone plus a 196B Engram component, it is architected to be cheap at inference in a way the other two are not. Mistral's ~52B active on a 1T total body sits in the middle, and the FP8 and NVFP4 checkpoints on Hugging Face run on vLLM, SGLang, NVIDIA NIM, NeMo AutoModel and TensorRT-LLM, with community GGUF

quantisations as well. There is a reported ~233 GB heavily quantised Mistral variant, the only configuration in this comparison that a single well-specced node might swallow. I have not seen quality numbers at that quantisation level, so treat 233 GB as a hardware fact and not a capability claim.

Kimi K3.1 Pro at 1.56 TB is a multi-node commitment with everything multi-node brings: interconnect, scheduling, failure modes that only appear under load. 2.8T total parameters with 104B active is the most capable-looking configuration on paper. On paper is where it stays until someone publishes benchmarks I can cite.

Three points on CyberGym is not a procurement decision

Mistral is specific about its target. It positions Large 4 as its most capable model for agentic coding and cybersecurity, and the open-weight drop carries the line "Built for Coders, Trusted by Defenders". The evidence behind that: 79% on CyberGym-E2E, matching MiMo-V2.6-Pro and ahead of DeepSeek V4.1 Flash at 76%, and a Terminal-Bench 4.0 jump from 28% to 41% after post-preview reinforcement learning, reportedly tying GLM-5.3, per Fortune.

Three points between 79% and 76% will not justify restructuring a procurement decision. The Terminal-Bench movement interests me more than the absolute figure: a 13-point lift from post-training RL on an already-shipped preview suggests the capability came from targeted work on agentic trajectories rather than from scale. That kind of gain tends to be narrow. It should hold up on tasks that look like the training target (terminal-driven agent loops) and generalise less well elsewhere. Unconfirmed, but it is how I would budget my expectations.

WHERE I LAND

My take: the $0.07/M cached input rate on Mistral Large 4 is the underrated number in this whole comparison, a 10x discount on the $0.70/M standard rate. For agentic coding against a large, stable repository context, cache hit rate drives your bill more than the headline price does. That makes Mistral's $0.70/$2.10 look much closer to DeepSeek's $0.15/$0.60 off-peak than the sticker comparison suggests, and DeepSeek's roughly 2x peak-hour pricing cuts the other way.

Pick DeepSeek unless you have a reason not to

Pick DeepSeek V4.1 Flash if you are building a product others pay to use, if your legal position benefits from a clean MIT licence, or if your serving budget is the binding constraint. It has the smallest footprint and the cheapest tokens, and it is the only one of the three with no resale question to answer.

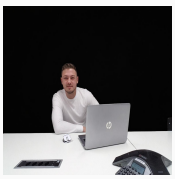
Pick Mistral Large 4 if the workload is agentic coding or security analysis inside your own org, if your context reuse is high enough to live on the cached rate, and if you are not reselling inference. Running it on EU infrastructure under a licence from an EU vendor matters to some procurement processes, which is a judgment about your buyers rather than about the model.

Pick Kimi K3.1 Pro if you have 1.56 TB of VRAM already standing idle and the appetite to evaluate a custom licence. That is a short list of companies. I would not start there on the evidence currently public.

One more date to note before you ship anything: Mistral's open-weight announcement is dated October 29 and was updated November 5, 2026, while the original model page is dated October 6. A licence text that moved within a week of publication is a licence text to re-read, with a date stamp.

WHAT TO DO WITH THIS

Benchmark the two you can actually serve, on your own tasks, and settle the licence question first, because it is the one no amount of GPU spend will solve. If you want a second opinion on which side of the MaaS clause your product sits, that is a short conversation. Book a call →



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →