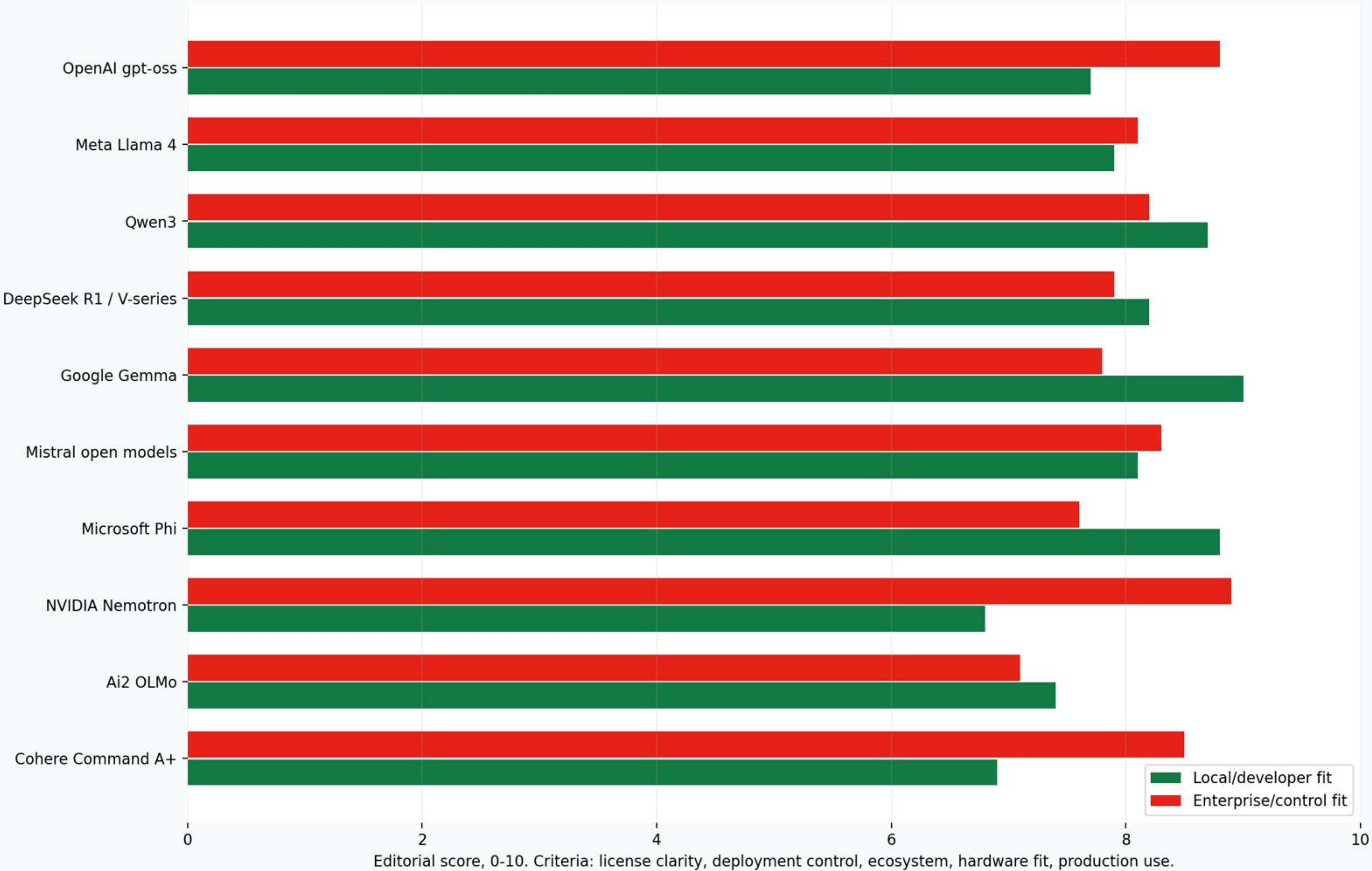


State of Open-Weight AI Models



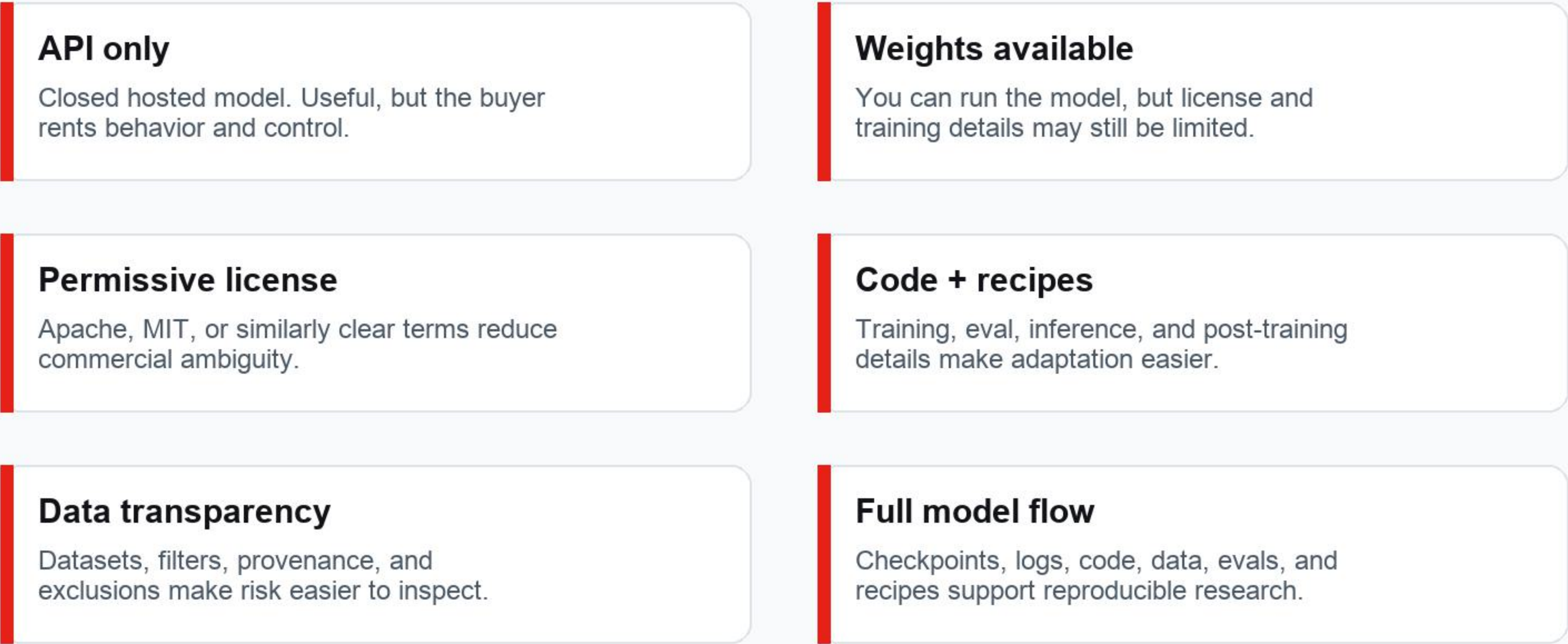
State of Open-Weight AI Models

Kingy Open-Weight Utility Scores: 2026



The Open-Weight Openness Ladder

Weights alone are useful. Weights plus license clarity, data, code, and recipes are stronger.



Open-Weight Deployment Map

The right model depends on where it runs and what can go wrong.

Laptop / edge

Small models, quantization, privacy, offline work, low cost, constrained quality.

Single GPU

Developer inference, prototypes, internal tools, model routers, and eval harnesses.

GPU server

Team workloads, batch jobs, private RAG, coding agents, and fine-tuning experiments.

Private cloud

Data residency, logging, access control, compliance, and cost optimization.

Hosted provider

Open weights with API convenience, but less direct operational control.

Hybrid router

Use open models for routine work and closed frontier models for hardest tasks.

Buyer Checklist

Do not choose open weights because they sound virtuous. Choose them when control changes the outcome.

License

Can the business use, modify, distribute, and commercialize the model safely?

Hardware

Can the model run at the required latency, quality, and cost on available compute?

Evals

Does it win on your private task, not just public benchmarks?

Safety

Can you monitor, filter, red-team, and roll back behavior in production?

Ecosystem

Does it work with vLLM, Ollama, Hugging Face, quantization, and your stack?

Fallbacks

What closed or open model takes over when quality is not good enough?