Models, providers, and the frontier

The market map behind every model choice.

Same question. Four different answers.

All four useful. None wrong. Four different stances on being helpful.

Claude

Careful, long-form, nuanced

GPT

Generalist, broad ecosystem

Gemini

Multimodal, reasoning-first

Llama

Open, sovereign, tunable

The differences aren't noise. They're design.

Two dimensions.

Cross them and you get the market map.

DIMENSION 01

Closed ↔ Open-weight

Closed models live behind an API you don't control. Weights stay on the provider's servers. Open-weight models ship as files — you can download, run, fine-tune.

Closed: Claude, GPT, Gemini · Open: Llama, Mistral, Qwen, DeepSeek

DIMENSION 02

Frontier ↔ Commodity

Frontier models push capability forward — biggest training runs, flagship prices. Commodity models deliver proven capability cheaply, tuned for volume.

Frontier moves. What was flagship 18 months ago is commodity today.

The 2×2 map.

SNAPSHOT · APRIL 2026

Refresh annually.

	CLOSED	OPEN-WEIGHT
FRONTIER	THE FLAGSHIPS Claude Opus 4.7 GPT-5.4 Gemini 3.1 Pro	SOVEREIGN CAPABILITY Llama 4 Maverick DeepSeek V3.2 Qwen 3.5 · GLM-5.1
COMMODITY	THE WORKHORSES Claude Sonnet 4.6 Gemini 3.1 Flash-Lite GPT-5 Turbo · mini	THE EDGE Llama 4 Scout Mistral Small 4 · Phi Qwen 3.5 Small · Gemma 4

THE REFRAME

~~*"Which model is best?"*~~

A question with no single answer. Benchmark compression hides the real differences.

"Which model for which task, at what stake, with what constraint?"

An answerable question. A different answer for every combination.

No single model wins every task. Design for that.

A decision sequence.

Instead of picking a model, follow the order.

01

Constraints

Data residency, latency, regulation. These eliminate most of the map before capability matters.

02

Stakes

Rung 1-2 work → commodity model. Rung 3-4 work → frontier model as default.

03

Fit

Private probe suite (5-6 tasks from real work). Winner is who performs best on your tasks.

04

Lock-in check

Can you route, swap, and move? If no, you're buying a dependency, not a model.

Multi-model is the new default.

Three common shapes. Each saves 50-90% vs. flagship-for-everything.

ROUTER

Right model per call

A classifier decides. Easy cases go to commodity. Hard cases go to frontier. Sensitive data routes to on-prem. 70-90% savings vs. single-flagship.

ENSEMBLE

Two answer, one judges

Higher cost per call. Higher quality at the top of the curve. Reserved for tasks where that quality matters — and they exist.

AGENT LOOP

Different model per step

Planner on frontier. Executors on commodity. Verifier on a third. Dramatically better cost and latency than one-model-for-everything.

Provider risk, named.

Single-vendor AI strategy = single-vendor cloud strategy + a few new ones.

01	**PRICING**	*Per-token costs in 2027 don't have to look like 2026's. Thin unit economics are exposed.*
02	**POLICY**	*A use case that worked in January can quietly stop in March. Policies change.*
03	**CAPABILITY**	*Model deprecation. New version, similar benchmarks, different behavior in the ten places your prompts relied on subtle quirks.*
04	**AVAILABILITY**	*Outages. Rate limits. Geographies added to restricted lists. Rare individually, common in aggregate.*

A worked example: a real stack.

A typical production workload distribution in April 2026.

70% — **GPT-5.4 or Gemini 3.1 Pro**

Bulk generation · chat · simple extraction

20% — **Claude Sonnet 4.6**

Mid-tier coding · everyday developer tasks

10% — **Claude Opus 4.7**

Quality-sensitive agentic reasoning · hardest coding

This routing typically cuts monthly spend 50-65% vs. running everything through the flagship.

THE MAINTENANCE NOTE

The quadrants endure. The model names don't.

This deck names Claude Opus 4.7, GPT-5.4, Llama 4 Maverick, Gemini 3.1 Pro. In a year those names will shift one or two versions. The structure — closed/open, frontier/commodity — will still be the map.

Refresh the model names. Keep the framework.

Procurement gets granular.

One sentence changes everything downstream.

THE SHORTCUT

"We're standardizing on X."

Convenient. Locks your cost structure, risk profile, and capability ceiling to one provider's roadmap.

THE ARCHITECTURE

"We're standardizing on an architecture, with X as the default and open routing to others."

Same sentence length. Radically different downstream flexibility.

The analogy worth holding.

The model market is starting to look like the cloud market.

CLOSED FLAGSHIPS

For premium workloads where capability matters more than cost — same logic as dedicated hyperscaler services.

OPEN + COMMODITY

For price-sensitive ones, on-device, sovereignty — same logic as on-prem and edge infrastructure.

MULTI-PROVIDER

The default for anything production. A handful of hyperscalers plus a long tail. Familiar pattern.

The analogy is imperfect. It's also the best one available. Plan accordingly.

THE CORE IDEA

Pick the architecture. Not the model.

Most serious stacks mix at least two quadrants.

Takeaways

01

Two dimensions, four quadrants.

Closed/Open × Frontier/Commodity. Each quadrant has a different logic and a different use case.

02

"Best model" is the wrong question.

Best for which task, at what stake, with what constraint. A different answer for every combination.

03

Multi-model is the new default.

Router, ensemble, or agent loop. Each saves 50-90% vs. running everything through the flagship.