03: The model landscape

Who makes the models, and how to pick one when the names keep changing

Lineups and prices as of November 2026

Why this matters

- Names change monthly. The tier ladder doesn't.
- Benchmarks and daily use often disagree
- Access can vanish overnight, for reasons unrelated to you
- The right model depends on the job, not the launch chart

Part 1

Reading the landscape

Labs, models, and products

- A frontier lab trains its own models at the edge of capability
- It takes tens of thousands of GPUs and months of work
- A model is the trained network. A product is what you use.
- One model powers many products; a product can switch models under you

"ChatGPT got worse" could mean the model, the system prompt, the router, or the harness.

Reading a model name

Part	Example	What it tells you
Family	Claude, GPT, Gemini, Grok, Kimi, Qwen	Which lab, which product line
Generation	5, 5.5, 6, 6.1, 4	Roughly when it was trained
Tier	Haiku / Sonnet / Opus; Luna / Terra / Sol / Astra; Flash-Lite / Flash / Pro	Size, price, and speed
Variant	Pro, Code, Flash, Cyber, Open, -Thinking	A specialization or serving mode
Snapshot / ID	`claude-opus-5-5`, `gpt-6.1-sol`	The exact string for an API call

The tier matters most.

The tier ladder (as of November 2026)

Tier	For	List price per million tokens (in / out)
Top / flagship	Hardest reasoning, long autonomous work	$4–$10 / $20–$50
Workhorse	Most coding and knowledge work	$1–$2 / $5–$10
Fast / cheap	Bulk processing, classification, sub-agents	$0.10–$0.30 / $0.50–$2.50

On Anthropic's own numbers, Sonnet 5.5 beats Opus 5.5 on Terminal-Bench 4.0 (70.6% vs 66.4%) at half the price.

Anthropic's top tier ships as two releases

Haiku → Sonnet → Opus → Mythos-class (new in 2026)

Release	Who gets it	Safeguards
Claude Fable	General release	Classifiers for cyber, bio, chem, and distillation
Claude Mythos	Vetted organizations only	Some lifted

Version numbers don't move in lockstep. Fable 5.1 sits above Opus 5.5.

Closed lineups (as of November 2026)

Lab	Top	Workhorse	Small / cheap
Anthropic	Fable 5.1 ($10 / $50), Opus 5.5 ($4 / $20)	Sonnet 5.5 ($2 / $10)	Haiku 5.5 ($0.10 / $0.50)
OpenAI	GPT-6 Astra ($10 / $50)	GPT-6.1 Sol ($2 / $10), GPT-6.1 Terra ($1 / $5)	GPT-6 Luna ($0.10 / $0.50)
Google	Gemini 4 Argon ($2 / $10 intro, then $4 / $20)	Gemini 3.8 Flash	Gemini 3.5 Flash-Lite ($0.30 / $2.50)

Prices per million tokens, input / output. Most have about 1M tokens of context.

Leading open-weight models (as of November 2026)

Model	Lab	Size (total / active)	License
Kimi K3.1 Pro	Moonshot	2.8T / 104B	Custom
GLM-5.3	Z.ai	753B MoE	Custom
DeepSeek V4-Pro	DeepSeek	1.6T / 49B	MIT
Qwen3.8-Max	Alibaba	2.4T / 95B	Custom
Mistral Large 4	Mistral	~1T / ~49–52B	Custom
MiniMax M3	MiniMax	428B / 23B	Custom (community)

They trail the closed frontier by roughly six months to a year.

Where is Meta's Muse? (as of November 8, 2026)

2026	Model	Weights
Apr 8	Muse Spark, the first MSL model	Closed
Jul 9	Muse Spark 1.1, plus the Meta Model API (preview)	Closed
Aug 10	Muse Glimmer 30B	Open, Apache 2.0
Oct 12	Muse Wonder ($4 / $20 on OpenRouter)	Closed
Oct 15	Muse Spark Open (816B)	Open, custom license

Meta's best models are closed. Its open ones don't lead the open field.

Gated tiers

In 2026 every top lab started holding back its most cyber-capable model.

Lab	Gated model or rollout
Anthropic	Mythos Preview, then Mythos 5 and 5.1, through Project Glasswing
OpenAI	GPT-6-Cyber
Google	Argon went to cyber defenders first; Gemini 4 Carbon is still limited to some of them

Public versions route or refuse some security and biology requests.

Part 2

Choosing in a moving market

Different places in the chain

```
Closed frontier (Anthropic, OpenAI, Google)
  ↓  months later: similar capability
  ↓  outputs used for training (legitimate or not)
Open-weight leaders (Kimi, GLM, DeepSeek, Qwen)
  ↓  fine-tunes, quantizations, distillations
Community and enterprise derivatives

Open leaders → cheap hosting → price floor → pressure on the closed frontier
```

Open models trail by six months to a year

Evidence	Gap
Epoch AI, November 2024 (benchmarks)	About one year (90% range: 5–22 months)
Epoch AI, October 2026 (cyber)	6–7 months; top open model was GLM-5.3
Simon Willison on Mistral Large 4, October 2026	"maybe about 6 months behind the frontier"

The real gap may be larger than benchmarks show.

Open models set the price floor

The September 2026 releases started "a new price war" (Simon Willison):

- GPT-6 Sol and Luna cost half as much as the GPT-5.6 models they replaced
- Anthropic cut Opus from $5 / $25 to $4 / $20
- Haiku 5.5 at $0.10 / $0.50 is a tenth of Haiku 4.5's list price
- Kimi K3.1 Pro, open weights, at a reported $2 / $10

Best model for the job (as of November 2026)

Job	Good choices	Basis
Agentic coding	Opus 5.5 or Sonnet 5.5 in Claude Code; GPT-6.1 Sol or Astra in Codex	Opinion + vendor benchmarks
Code review	A *different* model from the one that wrote the code	Opinion
Long documents and video	Gemini 4 Argon, Gemini 3.8 Flash	Fact + opinion
Cheap bulk work	Haiku 5.5, GPT-6 Luna, Flash-Lite, DeepSeek V4.1-Flash	Fact (price)
Local and private	Qwen3.8-27B, Gemma 4, gpt-oss-20b, Muse Glimmer 30B	Fact + opinion
Security research	Mythos 5.1 or GPT-6-Cyber, via verification	Fact (access)

Which models for coding?

> **Bryan's take:** For coding, Claude and Codex are the best. For certain other tasks you might use Grok or Gemini for very specific cases, but not for coding, because they fall far behind there. Different models are better for different use cases.

Launches don't always land

Launch	What happened
GPT-5 (Aug 2025)	Router down on launch day; a "chart crime"; users wanted GPT-4o back
Llama 4 (Apr 2025)	Its top Arena score came from an unreleased experimental version
Opus 5 (Jul 2026)	Led leaderboards; its communication style drew common complaints
Fable 5 (Jun 2026)	Suspended for all users three days after launch
Gemini 3.5 Pro (2026)	Promised for June; never shipped

On Opus 5 and Opus 5.5

> **Bryan's take:** Opus 5 felt benchmark-maxed. Great numbers, and it kind of sucked in daily use. People didn't like it. Opus 5.5, people really liked.

On trusting labs

> **Bryan's take:** I don't trust some labs much. Google in particular tends to overhype its models, and they underperform in my use.

The counterpoint

- Gemini 3 Pro launched at the top of LMArena (1501 Elo), November 2025
- Scored 87.6% on Video-MMMU at Gemini 3 Pro's launch
- Its Flash models are among the cheapest capable models
- Offers Nano Banana images, Veo video, Live voice
- Gemini 4 Argon tops the Vals Index, by Google's account

The critique fits Google's 2026 Pro line best, and multimodal and price-per-token work worst.

How to pick a model next month

1. Define the job, then pick 2–3 candidates from different labs
2. Check access rules for retention, region, safety routing, approved vendors
3. Read the launch post skeptically: harness, effort, who ran the numbers
4. Run 5–10 of your own real tasks in your own harness
5. Compare cost per finished task, and try the effort settings
6. Wait a week if you can. Write down the decision and the date.

Try it

On your own: a mini bake-off (60–90 minutes)

- Three real tasks: a bug fix, a refactor, "explain this module"
- Two models from different labs, in the same harness
- Record pass/fail, tokens or cost, time, and how readable the final message was
- Did the benchmark ranking predict your result?

As a class, each group builds a living "model card" page for one lab.

Discussion

1. Opus 5 led benchmarks and drew complaints. What does each kind of evidence miss?
2. If open models trail by 6–12 months, when is that good enough for you?
3. Should a lab reroute your request without asking? What disclosure would you want?
4. Google cancelled a promised model. How much weight should future-model announcements carry?

Read more

- Module: 03 The model landscape
- Epoch AI, How far behind are open models? (2024)
- Simon Willison, Opus 5.5, GPT-6 Sol, GPT-6 Luna and a new price war (September 2026)
- OpenRouter models: dates and prices across providers