AI Field Guide

CHAPTER TWO

The Whole Board

Claude, GPT, Gemini, DeepSeek, Kimi, Qwen, Grok and the open-weight insurgency — every model that matters in July 2026, what it costs, and what it's actually for.

Compiled 14 July 2026 · prices in USD per million tokens · every figure sourced

CHAPTER TWO

Before the table

THE FOUR THINGS THAT ACTUALLY CHANGED

1. Chinese open-weight models now process **the majority of tokens on OpenRouter** — north of 20 trillion a week. This stopped being a hobbyist story.

2. The frontier still belongs to Anthropic. **Claude Fable 5 leads SWE-bench Pro at 80.3%**, and nothing Chinese is within 18 points of it.

3. **Alibaba closed its frontier tier.** "Qwen = open weights" is no longer true, and Meta has effectively left the race entirely.

4. Self-hosting to save money is **almost always a mistake** at your volume. The maths is brutal and it's in this chapter.

One warning before the numbers. **SWE-bench Verified and SWE-bench Pro are not the same benchmark** and are not comparable. Pro is far harder — roughly, 80% on Verified corresponds to about 60% on Pro. Labs have started quoting whichever one flatters them, so any table that mixes the two without saying so is lying to you. This one says so.

CHAPTER TWO

Anthropic — and the naming mess

Start here, because you use it daily and because the lineup genuinely confuses people. **There is no "Opus 5."** A new tier name — Fable — now sits *above* Opus, while Sonnet has jumped a whole generation number and Opus is still on 4.x. Ranked by capability:

TIER	MODEL	WHAT IT'S FOR
Frontier	**Claude Fable 5**	The most capable widely-released model in the world. Long-horizon agents, hardest reasoning
Frontier (gated)	**Claude Mythos 5**	Fable 5 *without safety classifiers*. Invitation-only, defensive cyber. Not self-serve
Workhorse	**Claude Opus 4.8**	Anthropic's own recommended default for complex agentic coding. *This is what you're running now.*
Default	**Claude Sonnet 5**	Best speed/intelligence blend. Default on Free and Pro

TIER	MODEL	WHAT IT'S FOR
Cheap & fast	**Claude Haiku 4.5**	Low-latency, cost-sensitive. Noticeably older — Feb 2025 cutoff, 200K context

	FABLE 5	OPUS 4.8	SONNET 5	HAIKU 4.5
Context	1M	1M	1M	200K
Price in / out	**$10 / $50**	**$5 / $25**	**$3 / $15** intro $2/$10 to 31 Aug	**$1 / $5**
Knowledge cutoff	Jan 2026	Jan 2026	Jan 2026	Feb 2025
SWE-bench **Pro**	**80.3% ✱**	69.2%	63.2%	—

✱ Best score of any model tested, open or closed.

Three things about Claude that will bite you

> THE TOKENIZER CHANGED — YOUR CONTEXT IS SMALLER THAN YOU THINK
>
> Opus 4.7 introduced a new tokenizer, now used by Fable 5. **The same text produces roughly 30% more tokens** than on pre-4.7 models. Your 1M-token window is not as big as the number suggests, and your bill is not comparable to last year's. Nobody advertises this.

Refusals return HTTP 200, not an error. Fable 5 can come back with `stop_reason: "refusal"` — a successful response containing a refusal. If your code only handles error status codes, you will silently treat a refusal as a valid completion. You are not billed for a pre-output refusal, and there is a server-side fallback parameter, but you have to actually wire it up.

Fable 5 sometimes silently routes to Opus 4.8. When a query touches cyber, bio, chem, or model-distillation topics, you may get the weaker model without being told. Under 5% of sessions — but you paid Fable prices.

The export-control episode, briefly

Worth knowing because it explains behaviour you may have noticed. On **12 June 2026**, US export controls hit Fable and Mythos after a research report demonstrated bypassing safeguards to find software vulnerabilities. Anthropic suspended access for roughly nineteen days. Controls lifted on 30 June; the model returned globally on 1 July with a new cyber classifier — and **Anthropic itself acknowledged that the new classifier flags benign coding and debugging requests more often than before**. If Claude has felt twitchier about security-adjacent code in the last two weeks, that is why, and it is not your imagination.

CHAPTER TWO

Google — where the Flash beats the Pro

The Gemini line is in an odd state: **the newest Flash is a generation ahead of the newest Pro.** There is no Gemini 3.5 Pro.

MODEL	IN / OUT PER 1M	NOTE
Gemini 3.5 Flash (May 2026)	$1.50 / $9.00	**Beats 3.1 Pro on most coding and agentic benchmarks, ~4× faster**
Gemini 3.1 Pro (Feb 2026)	$2 / $12 (≤200K) $4 / $18 (>200K)	80.6% SWE-bench Verified · 94.3 GPQA-D · 1M context
Gemini 3 Flash	$0.50 / $3.00	Cheapest credible Google tier
Gemini Omni Flash	$1.50 / $9.00 $17.50 video out	See below — it is not what you'd assume

For agent work, **reach for the Flash, not the Pro**. That inverts the usual intuition and it is the single most useful thing to know about Google's lineup right now.

> "GEMINI OMNI" IS NOT A CHAT MODEL
>
> It got a lot of headlines at I/O 2026 and it is *real* — but it is a **generative world model** for any-input-to-any-output media, starting with video, edited conversationally instead of on a timeline. Demis Hassabis framed it as a world model with physics understanding, not a video generator. It ships in the Gemini app, Google Flow, and YouTube Shorts Remix.
>
> **Do not plan a coding-agent migration around it.** It is not competing with Sol or Opus.

CHAPTER TWO

The open-weight insurgency

DeepSeek — the price floor

DeepSeek V4 (March 2026) ships two models, both **1M context**, both **MIT-licensed open weights**:

MODEL	INPUT	OUTPUT	NOTE
`deepseek-v4-pro`	$0.435	**$0.87**	1.6T params MoE · **80.6% SWE-bench Verified** — ties Gemini 3.1 Pro

MODEL	INPUT	OUTPUT	NOTE
`deepseek-v4-flash`	$0.14	**$0.28**	The price floor of the entire industry

Sit with those numbers for a second. **DeepSeek V4-Pro's output costs $0.87 per million tokens. Claude Fable 5's costs $50.00.** That is a **57× ratio**, for a model roughly fifteen SWE-bench points behind. V4-Flash against GPT-5.6 Sol is **107×**.

> DEADLINE: 24 JULY 2026 — TEN DAYS OUT
>
> `deepseek-chat` and `deepseek-reasoner` are **deprecated on 24 July**. They map to the non-thinking and thinking modes of `deepseek-v4-flash`. If either string appears anywhere in your code, that is a this-week problem.

(The long-awaited R2 reasoning model **never shipped** — reporting says the CEO was unsatisfied with it. The reasoning capability was folded into V4's thinking mode instead.)

Kimi — the one built for overnight runs

Kimi K2.6 (April 2026), 1T params MoE, effectively MIT-licensed, **$0.95 / $4.00**, 256K context, **80.2% SWE-bench Verified**.

> AGENT SWARM — GENUINELY UNMATCHED
>
> Kimi ships an architecture that runs **up to 300 sub-agents across ~4,000 coordinated steps for 12+ hours of continuous autonomous execution.** Nothing else — open or closed — ships an equivalent.
>
> If you ever seriously pursue the "let it run overnight on the repo" pattern you sketched for Cinefolio, this is the only model in the world explicitly built for it. Worth a look on that project specifically.

GLM-5.2 — the quiet overachiever

Z.ai's GLM-5.2 (June 2026), ~750B MoE, **MIT**, 1M context, **$1.40 / $4.40**. It scores **62.1 on SWE-bench Pro — beating GPT-5.5's 58.6** — and is the **top open model on the Artificial Analysis Intelligence Index**. It landed within one percentage point of Claude Opus 4.8 on a closely-watched agentic benchmark at roughly a fifth of the cost. There is an $18/month coding plan.

Two changes of allegiance you may have missed

Alibaba closed its frontier tier. Qwen 3.5 and 3.6 remain Apache 2.0 — but **Qwen3.7 Max is closed, API-only**. The biggest Chinese open-weights champion took its best model private. Stop assuming "Qwen" means "open."

Meta has effectively left. Llama 4 Behemoth was previewed, delayed, and **shelved — never released**. In April 2026 Meta Superintelligence Labs shipped **Muse Spark**, which is *closed-weight* and consumer-facing (meta.ai, WhatsApp, Instagram, the glasses). There is no serious first-party developer API story. The company that started the open-weights movement is no longer in it.

The one nobody covered: NVIDIA

Nemotron 3 Ultra — 550B hybrid Mamba-2/Transformer MoE, OpenMDW license, **$0.42 / $2.61**, and *the strongest US open-weight model in existence*. NVIDIA also launched a "Nemotron Coalition" of global labs for open frontier models. If you want frontier-adjacent open weights that are **not from China** — for a client with procurement rules, say, which given your ERP work is a real scenario — this is now the answer, and it got almost no press.

CHAPTER TWO

The master table

USD per million tokens, standard tier. SWE-bench **Verified** unless marked **Pro** — again, they are not comparable.

MODEL	COMPANY	OPEN?	CTX	IN	OUT	SWE-BENCH	STANDOUT STRENGTH
Claude Fable 5	Anthropic	Closed	1M	10.00	50.00	**80.3% Pro ✱**	Best coding model, period
Claude Opus 4.8	Anthropic	Closed	1M	5.00	25.00	69.2% Pro	Anthropic's agentic default — *yours*
Claude Sonnet 5	Anthropic	Closed	1M	3.00	15.00	63.2% Pro	Best price/perf in the Claude line
Claude Haiku 4.5	Anthropic	Closed	200K	1.00	5.00	—	Fastest — but stale (Feb 2025 cutoff)
GPT-5.6 Sol	OpenAI	Closed	1.05M	5.00	30.00	64.6% Pro	ARC-AGI-3 SOTA; terminal/browser agents
GPT-5.6 Terra	OpenAI	Closed	1.05M	2.50	15.00	n/p	Best all-round default; ~Sol at ½ price
GPT-5.6 Luna	OpenAI	Closed	1.05M	1.00	6.00	n/p	Cheap. ⚠️ 41.3% long-context recall
Gemini 3.5 Flash	Google	Closed	1M	1.50	9.00	n/p	Beats its own Pro at coding, 4× faster
Gemini 3.1 Pro	Google	Closed	1M	2.00	12.00	80.6%	Long-context; 94.3 GPQA-D

MODEL	COMPANY	OPEN?	CTX	IN	OUT	SWE-BENCH	STANDOUT STRENGTH
DeepSeek V4-Pro	DeepSeek	**MIT**	1M	**0.435**	**0.87**	**80.6%**	Frontier-adjacent coding at 1/57th Fable's output cost
DeepSeek V4-Flash	DeepSeek	**MIT**	1M	**0.14**	**0.28**	n/p	The industry price floor
Kimi K2.6	Moonshot	**~MIT**	256K	0.95	4.00	**80.2%**	**Agent Swarm** — 300 sub-agents, 12hr runs
GLM-5.2	Z.ai	**MIT**	1M	1.40	4.40	62.1% Pro	Top open model on AA index; beats GPT-5.5
Qwen3.7 Max	Alibaba	**Closed**	1M	2.50	7.50	n/p	Multi-hour autonomous agent; GPQA 92.3
Qwen 3.5	Alibaba	Apache 2.0	—	self-host		n/p	Best *truly* open Qwen; 91.3 AIME
Grok 4.5	xAI	Closed	500K	2.00	6.00	n/p	**#1 agentic tool use**; 2× token efficiency. No EU
Mistral Large 3	Mistral	Apache 2.0	262K	self-host		n/p	Best multilingual; EU-sovereign
Nemotron 3 Ultra	NVIDIA	**OpenMDW**	—	0.42	2.61	n/p	Strongest **US** open-weight model
Muse Spark	Meta	Closed	—	consumer only		n/p	No real developer API. Meta is out

✱ Best of any model tested. `n/p` = not published by the vendor — if someone quotes you one, they're guessing.

CHAPTER TWO

Should you self-host? Almost certainly not

This question comes up constantly and the honest answer is unpopular, so here is the arithmetic.

BREAK-EVEN AGAINST…	YOU'D NEED TO BE RUNNING…
A premium API (~$5/Mtok)	**~256 million tokens/month**, at 60–70% sustained GPU utilisation
A cheap open API like DeepSeek V4-Flash	**~5.7 billion tokens/month** on a single H100

Read the second row again. **Self-hosting DeepSeek to beat DeepSeek's own API price is essentially never worth it.** And raw GPU cost is only 30–40% of true total cost of ownership —

budget a 2.5–3× multiplier, because engineer time is another 25–30%, and a senior inference engineer runs $250–360K loaded.

> THE VERDICT FOR YOU
>
> Self-host for **data residency, compliance, or fine-tuning** — never to save money. At the volumes any of your projects run, the open-weight APIs (DeepSeek, Moonshot, Z.ai, Together, Fireworks) capture essentially 100% of the cost advantage with 0% of the operations burden.
>
> The one place this could genuinely matter to you: **an Iraqi client with data-residency requirements.** That is a compliance argument, not a cost one — and it's the right reason to reach for it.

CHAPTER TWO

How to actually route your work

The strategic move in 2026 is not picking *a* model. It is routing by task, because the cost spread is now 50× while the quality spread on *easy* tasks is nearly zero.

WHEN YOU'RE DOING…	REACH FOR…	WHY
The gnarly refactor you can't afford to get wrong	**Fable 5 / Opus 4.8**	80.3% SWE-bench Pro is not matched anywhere, at any price
Everyday agentic coding	**Opus 4.8** or **GPT-5.6 Terra**	Both land near the frontier at a fraction of flagship cost
Bulk codegen, test generation, classification, anything in a loop	**DeepSeek V4-Flash**, **GLM-5.2**, **Kimi K2.6**	The 20–60× cost delta dwarfs the quality delta on easy tasks
Overnight autonomous runs on a big repo	**Kimi K2.6**	Agent Swarm — nothing else does 12-hour continuous execution
A client who cannot use Chinese weights	**NVIDIA Nemotron 3 Ultra**	Strongest US open-weight option

> THE HONEST CAVEAT ON THE CHINA STORY
>
> The top of the frontier is still American. **Nothing Chinese matches Fable 5's 80.3% on SWE-bench Pro** — the best is GLM-5.2 at 62.1, an eighteen-point gap on the hardest agentic coding benchmark. That gap is worth $50 per million tokens to some teams and precisely nothing to others. Know which one you are before you switch.

Sources. Anthropic platform docs & Fable 5 announcement · OpenAI developer docs · Google Gemini API pricing & I/O 2026 announcements · DeepSeek API docs & HuggingFace · Moonshot Kimi platform docs · Z.ai / VentureBeat ·

Artificial Analysis · Vellum · OpenRouter open-weights report (June 2026) · CNBC on Chinese model adoption · NVIDIA Nemotron Coalition · Mistral · x.ai. Prices and scores current as of 14 July 2026 and moving fast; the live version of this chapter is kept updated.