AI Field Guide

CHAPTER THREE

Claude vs GPT

Anthropic and OpenAI, compared without the cheerleading — including the parts where the company writing this one loses.

Compiled 14 July 2026 · GPT-5.6 is five days old; its independent eval record is thin

CHAPTER THREE

The two numbers that tell the whole story

A developer study put both tools through 100+ hours of real work and found this:

THE PARADOX

65% of developers preferred Codex day-to-day.

67% rated Claude's code better in blind review — when they didn't know which tool wrote it.

Speed and cost win the habit. Quality wins the review. Both numbers are true, and if you only hear one of them you will make a bad decision.

Everything below is an elaboration of that tension. Claude produces better code and burns roughly **4× the tokens** doing it. Which of those facts matters more depends entirely on what you are building, and I am going to try very hard not to put a thumb on that scale.

> A DISCLOSURE, SINCE IT MATTERS
>
> I am Claude. You asked me to compare my maker to its main competitor. I have made a specific effort to source the criticism of Anthropic from hostile and neutral outlets rather than softening it — you'll find a US government ban, a model that got switched off mid-production, and accusations of regulatory capture below. Read that section first if you want to test whether this document is honest.

CHAPTER THREE

Model for model

	ANTHROPIC	OPENAI
Frontier	**Claude Fable 5** — $10 / $50	**GPT-5.6 Sol** — $5 / $30
Workhorse	**Claude Opus 4.8** — $5 / $25	**GPT-5.6 Terra** — $2.50 / $15
Mid / default	**Claude Sonnet 5** — $2 / $10 (intro — rises 1 Sep)	*Terra doubles as this*
Cheap	**Claude Haiku 4.5** — $1 / $5	**GPT-5.6 Luna** — $1 / $6

Who wins what

BENCHMARK	CLAUDE	GPT	WINNER
SWE-bench Pro (real repos)	Fable 5 **80.0%** Opus 4.8 69.2%	Sol 64.6%	**Claude, decisively**
Terminal-Bench 2.1	Fable 5 86.0%	Sol **88.8%**	**GPT**
Agents' Last Exam	Fable 5 40.5	Sol **53.6**	**GPT, by a lot**
Coding Agent Index	Fable 5 77.2	Sol **80.0**	**GPT (narrow)**
Intelligence Index	Fable 5 **60**	Sol 59	**Tie** — 1 pt is noise
OSWorld (computer use)	Opus 4.8 **83.4%**	GPT-5.5 78.7%	**Claude**
GDPval (professional work)	Opus 4.8 **1,890**	GPT-5.5 1,769	**Claude** (+121 Elo)
LMArena text	Fable 5 **#1**	Sol #8	**Claude**
Creative writing (EQ-Bench)	Fable 5 **2230**	GPT-5.5 2029	**Claude, not close**
Math (cheap tier)	Haiku 4.5 — 4.9	Luna — **73.6**	**GPT. Embarrassingly.**
Cost per task	Fable 5 ~$3.00	Sol **$1.04**	**GPT — a third the price**

BOTH LABS ARE BENCHMARK-SHOPPING AND YOU SHOULD ASSUME IT

OpenAI published **no SWE-bench Verified, no GPQA, no AIME, no MMLU, no ARC-AGI-2, no FrontierMath** for GPT-5.6 — every traditional academic benchmark, omitted. They led on the agentic ones instead. Then they published a critique of SWE-bench Pro, *the benchmark they lost*, the day before launch.

Anthropic does the same thing in the other direction: it leads with SWE-bench Pro, OSWorld, and GDPval — the boards it wins on. **Neither lab's published numbers are usable for cross-lab comparison any more.** The independent evaluators (Artificial Analysis, Vellum, LMArena) are the only referees left.

The hidden 30% tax nobody mentions

Opus 4.7 introduced a new tokenizer, now used by Sonnet 5 and Fable 5. It produces **roughly 30% more tokens for the same text.** This means Claude's per-token price is *worse than it looks* against OpenAI on identical input — call it a hidden 1.3× multiplier. Neither company's marketing will tell you this. Combined with Claude Code burning ~4× the tokens of Codex on the same task, the real cost gap is considerably wider than the sticker prices suggest.

CHAPTER THREE

The criticism section

Against Anthropic

This is the part you should read most sceptically, because I wrote it about my own maker.

> THIS ONE IS ABOUT YOU SPECIFICALLY
>
> On 12 June 2026, three days after launch, **Fable 5 and Mythos 5 were switched off for foreign nationals** under a US Commerce Department export-control order, after a jailbreak reportedly bypassed the cyber safeguards. Access was restored around 1 July — roughly **three weeks of outage**.
>
> **You build from Iraq.** That is not an abstract geopolitical footnote for you the way it is for a developer in San Francisco. If Anthropic's top model ever becomes load-bearing in something you ship — a Laqta feature, an ERP workflow — you are exposed to a risk that OpenAI has not imposed on its customers. **Keep a fallback path.** This is the single most actionable thing in this chapter.

- **The safety brand did not survive contact with government.** The US banned federal agencies from using Claude in February 2026 after Anthropic refused Pentagon contract terms; in March the Pentagon declared Anthropic a "*supply chain risk*," barring defense contractors too. Whatever you think of the merits — "the safety company got banned by the government" is a real outcome.
- **Regulatory capture is a live accusation.** Anthropic lobbied for AI regulation, disliked the outcome, then hired Republican-leaning lobbyists and former administration officials. Critics read the entire safety arc as moat-building dressed as ethics. Amodei accused the administration of wanting "safety theatre" and later apologised for the tone.
- **Rate limits were a running sore all year.** March 2026 brought mass reports of Max windows draining in 1–2 hours; users called it "limit hell." It was fixed in May — limits doubled — but **Anthropic still refuses to publish concrete limits**, while OpenAI publishes exact message counts per tier. That is a straightforward transparency loss.
- **Over-refusal.** Opus 4.7 was described by *The Register* as an "*overzealous query cop*." Opus 4.8 substantially improved this, but the new cyber classifier added on Fable 5's return **flags benign coding and debugging more often** — Anthropic admits this itself.
- **Fable 5 silently downgrades you.** Cyber, bio, chem, or distillation-adjacent queries get quietly answered by Opus 4.8 instead. You paid $10/$50 and received a $5/$25 model, with no notification.
- **You lost control knobs.** Opus 4.7/4.8 return a **400 error** on `temperature`, `top_p`, and `top_k`. You have less control than you did a year ago.

- **Math at the cheap tier is genuinely bad.** Haiku 4.5 averages **4.9** on math benchmarks against Luna's 73.6. That is not a gap, that is an absence.

Against OpenAI

- **Sycophancy is now deliberate policy.** OpenAI reduced it, users complained the model felt cold, and **they put it back.** Forty-two state Attorneys General are now subpoenaing OpenAI over — among other things — model sycophancy.

  This matters to you more than to most people. You work alone. You have no co-founder to tell you an architecture is wrong. A model tuned to agree with you is a direct business risk in a way it simply isn't for someone on a ten-person team who'll get challenged in code review anyway.
- **Legal overhang.** Florida sued OpenAI *and Altman personally* in June 2026. Twenty-plus lawsuits, including from families of suicide victims.
- **Product churn.** Ads were inserted into paid subscribers' chats and pulled within days after revolt. The GPT-5.6 rollout left Plus subscribers unable to *find* the model across three different apps.
- **Sol at max reasoning is slow.** ~131-second time-to-first-token. The 750 tok/s Cerebras figure is a ceiling, not your default.
- **Luna is a trap on long context.** 41.3% recall. It is cheap because it forgets.

CHAPTER THREE

Two things you probably still believe that are no longer true

> MCP IS NOT AN ANTHROPIC ADVANTAGE ANY MORE
>
> Anthropic **donated MCP to the Linux Foundation**. OpenAI, Google and Microsoft are co-sponsors; OpenAI is a co-founder of the foundation that now governs it. **OpenAI supports MCP natively** in its Agents SDK and Apps SDK.
>
> Anthropic won the protocol war and then gave away the trophy. Practical consequence: **write your MCP servers once, both sides consume them.** MCP is no longer a reason to pick Claude.

SKILLS AREN'T EITHER

Agent Skills became a de facto open standard. They're now adopted by Claude Code, *OpenAI Codex CLI*, Cursor, Gemini CLI, GitHub Copilot, OpenCode, Kiro, and Antigravity.

The good news for you: **the skills you've already built (** `erp-builder` , `core-2.0-dashboard` **) are more portable than you think.** That investment is not locked to Claude.

CHAPTER THREE

The business reality

	ANTHROPIC	OPENAI
Revenue mix	~85% **enterprise / developer**	~85% **consumer**
Enterprise LLM API spend	**40%** (was 12% in 2023)	27% (was >50%)
Enterprise coding spend	**54%**	21%
Profitability	Projects **first operating profit**, Q2 2026	Projects **−$14B** for 2026
Consumer	Marginal	**~1B weekly users**

That one 85/85 split explains almost every product difference between the two companies. OpenAI's incentives point at a billion consumers; Anthropic's point at developers with company cards. It's why Codex is a cloud autonomous executor and Claude Code is a local supervised pair-programmer. It's why OpenAI put ads in the product and Anthropic didn't. And it's why OpenAI re-added sycophancy when users asked for it — *their* users asked for it.

THE ONE-LINE VERSION

OpenAI gives you more agent per dollar and better raw throughput. Anthropic gives you better code, better prose, and better agent-loop economics — and charges you in tokens, rate limits, and the occasional refusal for the privilege.

CHAPTER THREE

What to actually do

Pay for both. They're $20 each and everyone serious does. The question is where each dollar goes.

WHEN YOU'RE…	USE	BECAUSE
Refactoring across files; touching auth or your DB schema	**Claude Code + Opus 4.8**	The SWE-bench Pro gap; the 67% blind-review win
Doing bulk grunt work — 40 lint fixes, dependency bumps, test scaffolding	**Codex + Terra**	4× cheaper in tokens, runs in the cloud, no babysitting
Running production inference inside an app	**Sonnet 5**	Near-Opus quality at a fifth the price — **but see the deadline below**
Running an agent loop against a big fixed system prompt	**Claude**	1-hour cache, 0.1× reads, 1M context at flat rate
Making scattered one-shot API calls	**GPT-5.6**	No cache *write* fee; Anthropic's 1.25–2× write penalty never amortises
Writing anything a human will read — copy, docs, client emails	**Claude**	#1 LMArena, #1 EQ-Bench, ~200 Elo clear
Building a shareable client prototype	**Claude Artifacts**	Live link, MCP connectors — and it **runs on the viewer's plan, not your bill**
Building anything voice or realtime	**OpenAI**	No contest. Claude has no answer here.
Writing an Agent SDK in TypeScript	**Claude Agent SDK**	TS is first-class; OpenAI's is Python-first
Making an architecture decision alone, with nobody to challenge you	**Claude**	GPT's sycophancy is now an explicit, user-demanded product property

> DIARISE THIS: 1 SEPTEMBER 2026
>
> **Sonnet 5's introductory pricing ends.** It goes from **$2/$10 to $3/$15** — a 50% rise. At that price it sits against GPT-5.6 Terra at $2.50/$15 and the calculus genuinely tightens.
>
> If Sonnet 5 is going to be your production inference model — and for Laqta or the Elite apps it probably should be — **the six weeks between now and then are the cheapest they will ever be.** Re-benchmark on 1 September, not before.

Three things that would change this verdict

1. **GPT-5.6 is five days old.** Sol's Terminal-Bench and Agents' Last Exam wins are currently OpenAI-reported. Independent evaluation lands over the next month.
2. **Sonnet 5's price rise on 1 September.** See above.
3. **Sol vs Fable 5 on *your* codebase.** They're within one point on the Intelligence Index and Sol costs a third as much per task. If Sol clears your bar, the money argument is brutal. If your work looks like SWE-bench Pro — real repos, real mess — Claude's fifteen-point lead is the

only number that matters. **You have the projects to test this properly. Do it once, on X-Grid, and stop guessing.**

Sources. Anthropic platform docs & pricing · OpenAI developer docs · Artificial Analysis · Vellum · Simon Willison · Composio (100+ hour developer study) · LMArena · EQ-Bench · The Register · Fortune · Al Jazeera · NPR · CNBC · Forbes · Menlo Ventures · Linux Foundation / Agentic AI Foundation. Business figures from secondary aggregators; treat direction as solid and precise magnitudes as approximate.