AI Field Guide

CHAPTER ONE

Sol, Terra & Luna

OpenAI's GPT-5.6 family, taken apart tier by tier — what the benchmarks say, what OpenAI didn't publish, and which one you should actually point your agents at.

Compiled 14 July 2026 · GPT-5.6 released 9 July 2026 · every figure sourced

CHAPTER ONE

The three tiers

THE SHORT VERSION

Default to Terra. It lands within 2–3 points of the flagship on almost everything, at half the price, with essentially the same long-context recall.

Never point Luna at a long document. It has a 1M-token window it cannot actually use — see the cliff below.

Sol is the flagship, but it is not the workhorse. That distinction matters and OpenAI's own guidance agrees.

You remembered two of the three names. The full family is Luna (budget), Terra (mid), and Sol (flagship). They arrived on **9 July 2026** — five days before this document was written — after an unusual limited preview in which OpenAI shipped only to trusted partners first, at the request of a US-government safety review.

The naming scheme is the actual news

OpenAI has thrown away `mini` and `nano`. In their words: *the number identifies a model's generation, while Sol, Terra, and Luna identify durable capability tiers that can advance on their own cadence.*

That last clause is the point. Under the old scheme, `gpt-5.4-mini` could not ship until `gpt-5.4` shipped — the tiers were welded to the generation. Now Terra can get a mid-cycle upgrade without forcing a rename of the whole family. The suffixes also described *size* rather than *capability*, which flattered nobody: "nano" implied a toy, and Luna is not a toy.

The comparison table

	LUNA	TERRA	SOL
API model ID	`gpt-5.6-luna`	`gpt-5.6-terra`	`gpt-5.6-sol`
Alias	—	—	bare `gpt-5.6` routes here
Context window	1.05M	1.05M	1.05M
Max output	128K	128K	128K

	LUNA	TERRA	SOL
Input / 1M tokens	$1.00	$2.50	$5.00
Output / 1M tokens	$6.00	$15.00	$30.00
Cache read	90% discount across all three · cache write 1.25× input · 30-min minimum cache life		
Reasoning effort	`none · low · medium · high · xhigh · max` — six levels, all tiers		
Modalities	text + image in → text out		
Knowledge cutoff	16 February 2026		
Intelligence Index	51	55	**59**
Cost per task	$0.21	$0.55	$1.04
Coding Agent Index	74.6	77.4	**80.0**
Terminal-Bench 2.1	84.7%	87.4%	**88.8%** (91.9% Ultra)
Agents' Last Exam	50.3	50.4	**53.6**
MRCR long-context	**41.3%**	89.6%	91.5%
Built for	High-volume: classification, summarising, drafting, routine automation	**The default.** Everyday interactive work and agentic coding	Long-horizon agents, large-codebase reasoning, security research

THE ONE NUMBER THAT WILL BITE YOU

Luna scores 41.3% on MRCR long-context recall. Sol scores 91.5%.

Luna advertises the same 1.05M-token context window as its siblings — and cannot use it. Feed it a long document and it will confidently lose the middle of your prompt. This is the sharpest capability cliff anywhere in the family, and OpenAI does not foreground it. Luna is for *short*, *well-scoped*, *high-volume* work. Nothing else.

"Max" and "Ultra" are not the same knob

GPT-5.6 adds a sixth reasoning level, `max`, sitting above `xhigh`. Separately — and confusingly — there is Ultra, which is **not a reasoning-effort value at all**. Ultra is an execution mode that spawns four subagents in parallel and merges their results.

- `max` = *think harder, in one chain.*
- **Ultra** = *fan out, then synthesise.*

Cost scales the way you'd fear: Ultra runs roughly **3× a single Max call**, because token spend stacks across every subagent, and it buys about three points on Terminal-Bench. Reach for it rarely.

> COUNTERINTUITIVE MIGRATION ADVICE — FROM OPENAI
>
> When you move off 5.5, **keep your current effort as a baseline and then test one level *lower***. GPT-5.6 reaches frontier quality using fewer output tokens, so whatever effort level you tuned for 5.5 is probably now overkill — you'd be paying for reasoning you no longer need.

A caution on the speed numbers

Artificial Analysis measured Sol at **max** effort at 67.7 output tokens/sec with a **~131-second time-to-first-token**. That is not network latency — it is reasoning time, and it is what `max` costs you in wall-clock. If you are running Sol at max inside an interactive loop, that pause is the model thinking, and you should expect it.

Per-tier speed figures for **Terra and Luna have not been published** by anyone. Various blogs quote confident numbers ("Sol ~180ms TTFT, Terra ~320ms") — these are internally contradictory and almost certainly fabricated. Ignore them. A Cerebras-hosted Sol at up to 750 tok/s launched in July 2026 if you need a genuinely low-latency tier.

CHAPTER ONE

The benchmark that OpenAI attacked

This is the most important section in the chapter, because it is the one place where the marketing and the evidence come apart — and it happens to land directly on what you do all day, which is agents working inside real repositories.

What OpenAI published

BENCHMARK	SOL	TERRA	LUNA	CLAUDE FABLE 5
Terminal-Bench 2.1	88.8% (Ultra 91.9%)	87.4%	84.7%	86.0%

BENCHMARK	SOL	TERRA	LUNA	CLAUDE FABLE 5
Coding Agent Index	80.0	77.4	74.6	77.2
Agents' Last Exam	53.6	50.4	50.3	40.5
ExploitBench 1	73.5%	—	—	—
SWE-Bench Pro	**64.6%**	—	—	**80.0%**

What OpenAI did *not* publish

SWE-bench Verified. GPQA Diamond. AIME. MMLU. ARC-AGI-2. FrontierMath. All of them were omitted from the launch table — every one of the traditional academic benchmarks that defined frontier comparison right up through GPT-5.5.

That omission is the tell. OpenAI leads on the *agentic* benchmarks and trails on several *academic* ones, so the former were foregrounded and the latter quietly dropped.

> THE SWE-BENCH PRO INCIDENT — READ THIS ONE TWICE
>
> Sol scores **64.6%** on SWE-Bench Pro. Claude Fable 5 scores **80%**. That is a fifteen-point loss on the single benchmark that most closely resembles "can this thing work inside my actual repository."
>
> **The day before announcing GPT-5.6, OpenAI published a critique of SWE-Bench Pro** arguing that roughly 30% of its tasks are broken and advising developers to scrutinise results from it.
>
> As Vellum put it, and it is hard to improve on: *when a lab leads on a benchmark, they cite it; when they lose on it, they question its methodology.*

The independent corroboration matters here. Simon Willison — who has no stake in either company — reports that Sol "*hasn't struck me as better than Fable*" on complex coding tasks, despite Sol's wins on the benchmarks OpenAI chose to show. His hands-on impression matches SWE-Bench Pro, not the launch table.

> WHAT THIS MEANS FOR YOU SPECIFICALLY
>
> You build web apps with coding agents all day. On *repo-level coding* — the thing you actually do — the independent evidence still favours **Claude Fable 5** over Sol, and OpenAI's response to that evidence was to attack the ruler rather than the result. That is worth knowing before you switch anything.

Where Sol genuinely is the best model in the world

None of the above means Sol is weak. On **ARC-AGI-3** it is the only performant model in existence as of July 2026, and the first ever to win a public game (87% on ft09). Its verified ARC-AGI-2 score at max effort is **92.5%**. It hits **94.6% on GPQA Diamond** and **89.0% on FrontierMath v2**.

And on Artificial Analysis's Intelligence Index, Sol at max scores **59** — one point behind Claude Fable 5 at 60, but at roughly **a third of the cost per task** ($1.04 vs ~$3.00). Treat the top of that leaderboard as a statistical tie with a large price gap in Sol's favour.

> A DETAIL WORTH INTERNALISING
>
> On ARC-AGI-1, Sol peaks at `xhigh` (97.5%) and gets *slightly worse* at `max` (96.5%). **More reasoning is not monotonically better.** If you reflexively crank effort to maximum, you are sometimes paying more for a worse answer.

CHAPTER ONE

The lineage, and what happened to 5.3

You didn't ask about this, but it resolves a genuine confusion: **GPT-5.3 does exist.** I assumed it didn't, and I was wrong.

What actually happened is that the Instant / Thinking / Codex lines stopped versioning in lock-step. `GPT-5.3-Codex` shipped in February 2026 and `GPT-5.3 Instant` in March — but **no 5.3 flagship "Thinking" model was ever released**. The March launch paired 5.3 *Instant* with 5.4 *Thinking*. So the version number is very much in use; it just never had a frontier model attached to it.

VERSION	DATE	WHAT CHANGED
GPT-5	Aug 2025	Unified system — fast base model plus a reasoning layer behind a real-time router
5.1	Nov 2025	**Adaptive reasoning**: dynamic compute allocation, 2–3× faster on easy queries. Warmer default tone
5.2	Dec 2025	Aimed at professional knowledge work. First model past 90% on ARC-AGI-1
5.3	Feb–Mar 2026	Codex + Instant only, **no flagship**. Instant jumped to 400K context, 26.8% fewer hallucinations with web search
5.4	Mar 2026	**Native computer use**, **1M-token context**, tool search
5.5	Apr 2026	Big jump in agentic coding and long-horizon work. More expensive, but far more token-efficient

VERSION	DATE	WHAT CHANGED
5.6	**Jul 2026**	Luna / Terra / Sol. Programmatic tool calling, multi-agent beta, `max` effort, Ultra mode

The o-series is dead — and the migration path is a trap

There is no longer a separate reasoning line. Every active ChatGPT model belongs to the GPT-5 family; `reasoning_effort` *is* the o-series now, absorbed into the parameter. (There was never an o4 flagship — only `o4-mini`.)

> IF YOU HAVE ANYTHING RUNNING ON O3
>
> OpenAI maps `o3` → `gpt-5.5` and `o3-pro` → `gpt-5.5-pro`. But **o3 was a dedicated reasoning model with its own thinking paradigm, and gpt-5.5 is a chat-completion model with configurable reasoning.** A drop-in ID swap is *not* behaviourally equivalent. Validate the behaviour; don't just diff the cost and latency and assume you're done.

Deprecations — one of these is nine days out

23 July 2026 — `gpt-5-chat-latest`, `gpt-5.1-chat-latest`, all `gpt-5.1-codex` variants, legacy audio/realtime models, and the deep-research variants. If any of those strings appear in your code, that is a this-week problem.

23 October 2026 — the GPT-4 family, `gpt-4-turbo`, `o1`, `o3-mini`, `o4-mini`, `gpt-3.5-turbo`.

11 December 2026 — the original GPT-5 snapshots (`gpt-5-2025-08-07` and its mini/nano/pro siblings) and `o3` itself.

CHAPTER ONE

Codex ate the workspace

The structural change worth knowing: **there is no `gpt-5.6-codex` model.** The dedicated Codex fork is dead. Codex ran its own models from `gpt-5-codex` (Sept 2025) through `5.3-codex` (Feb 2026), and then stopped — from GPT-5.5 onward, the mainline models ship straight into Codex. You now pick Sol, Terra, or Luna inside Codex directly.

On the same day GPT-5.6 launched, Codex was **merged into the ChatGPT desktop app** — it is no longer a separate application. PR review moved inside it, inline diff editing arrived, and Codex Remote went GA (start work on a connected Mac, approve actions from your phone). OpenAI's own framing is that Codex has gone from a coding tool to *"a broader workspace for getting work done with AI."*

If that sounds familiar, it should — it is a direct answer to the Claude Code desktop app, and we will compare them properly in **Chapter Four**.

CHAPTER ONE

What I could not verify

Stated plainly, because a guide that hides its gaps is worse than useless:

- **Tokens/sec and time-to-first-token for Terra and Luna.** Nobody has published them. The confident figures circulating on SEO blogs contradict each other.
- **AIME scores for GPT-5.6.** Not published — almost certainly because the benchmark is saturated (GPT-5.2 reportedly hit 100% on AIME 2025).
- **SWE-bench Verified for GPT-5.6.** Not published for any tier. The last known figure in the family is GPT-5.1's 76.3%.
- **GPQA / ARC-AGI for Terra and Luna specifically.** Only Sol has been independently tested.
- **Whether "Sol Ultra" has its own API model ID** or is a request-level flag on `gpt-5.6-sol`. The evidence points to a mode rather than a model, but I could not confirm the parameter name.
- **The exact retirement date for GPT-5.2 / 5.2-Codex / 5.3-Codex.** Sources conflict; they do not appear on OpenAI's official deprecations page. GitHub Copilot dropped GPT-5.2 on 1 June 2026, which is the one corroborated data point.

CHAPTER ONE

Six things to actually do

1. **Default your agents to `gpt-5.6-terra`**, not Sol. Half the price, near-identical long-context recall, within a few points on everything else.
2. **Keep Claude in the loop for repo-level coding.** SWE-Bench Pro 80% vs 64.6% is not a rounding error, and the independent hands-on reports agree with the benchmark rather than the launch table.
3. **Never point Luna at a long context.** 41.3% recall. It has a window it cannot use.
4. **When migrating off 5.5, try one reasoning level *lower*.** OpenAI recommends this themselves.
5. **Grep your code for `gpt-5.1-codex`, `gpt-5-chat-latest`, `gpt-5.1-chat-latest` before 23 July.**
6. **Spike on programmatic tool calling.** The model writes JavaScript that runs in a sandboxed V8 with no network access, orchestrating tool calls in code instead of one round-trip per call. Of everything in 5.6, this is the feature most likely to change how you actually build agents.

Sources. OpenAI model docs and deprecations pages · OpenAI GPT-5.6 launch · Artificial Analysis (Intelligence Index, cost/task, measured throughput) · Vellum (benchmark analysis) · ARC Prize verified results · Simon Willison ·

TechCrunch · GitHub Copilot changelog. Figures current as of 14 July 2026 and will move; the live version of this chapter is kept updated.