

OpenAI makes GPT-5.6 Sol, Terra and Luna generally available

Source copy kept by wheresthe.ai for the story "OpenAI makes GPT-5.6 Sol, Terra and Luna generally available" (<https://wheresthe.ai/d/openai-gpt-5-6-general-release-2026-07/>).

Source: <https://openai.com/index/gpt-5-6/>

Text saved 18 Oct 2026 (the page text through a reader, because the site refused a direct fetch).

SHA-256 of the saved text: cd4cc3b87d9afdac692f5dcd424abc8588db2b365e9a2ab12514eda3d840f721

****_Update on August 21, 2026:_**_OpenAI dropped the API and credit pricing of GPT-5.6 Sol by over 20% for the next 3 months._**

****_Update on July 30, 2026:_**_OpenAI reduced the price of GPT-5.6 Luna by 80% and GPT-5.6 Terra by 20%.__Learn more here__._**

We're launching the GPT-5.6 family of models for general availability following our limited preview: our new flagship, ****Sol****, alongside ****Terra****, a balanced model for everyday work, and ****Luna****, our most cost-efficient model.

GPT-5.6 Sol sets a new standard for both intelligence and efficiency, achieving state-of-the-art results across coding, knowledge work, cybersecurity, and science while outperforming previous and competing frontier models with fewer tokens and at lower estimated cost. The result is stronger performance per dollar: more successful work for the same spend, or comparable results at a lower total cost. We also introduce a new way to accelerate the most demanding work: ``ultra`` is our highest-capability setting, coordinating multiple agents across parallel workstreams to finish complex tasks faster. Stronger computer use and design judgment make GPT-5.6 Sol our most polished collaborator yet, helping it inspect, refine, and deliver ready-to-use results.

We trained GPT-5.6 to get more useful work from every token. On ****Agents' Last Exam****(opens in a new window), an evaluation of long-running professional workflows across 55 fields, GPT-5.6 Sol sets a new high of 53.6, eclipsing Claude Fable 5 (adaptive reasoning) by 13.1 points. Even at medium reasoning, it beats Fable 5 by 11.4 points at roughly one-quarter the estimated cost. That efficiency extends to smaller models, which are essential to making intelligence more abundant and affordable: GPT-5.6 Terra and GPT-5.6 Luna outperform Fable 5 at around one-sixteenth the cost. On the ****Artificial Analysis Intelligence Index****(opens in a new window), a broad measure of intelligence spanning agentic work, coding, scientific reasoning, and general capabilities, GPT-5.6 Sol with max reasoning comes within one point of Fable 5 while completing tasks in 61% less time at roughly half the estimated cost.

GPT-5.6 launches with our most robust safeguards to date, designed to be resilient against determined and adaptive misuse without broadly limiting legitimate work. Before general availability, we put the models and safeguards through our most extensive evaluation period yet, combining human red teaming with large-scale automated testing. During the preview, we worked closely with expert organizations and with trusted partners to pressure-test defenses and strengthen safeguards before broader launch. The resulting system layers protections trained into the model with real-time checks, monitoring, and access calibrated to trust and risk.

GPT-5.6 Sol is our best coding model yet. On the ****Artificial Analysis Coding Agent Index,**** GPT-5.6 Sol with max reasoning sets a new state of the art at 80, 2.8 points above Fable 5, while using less than half the output tokens, taking less than half the time, and costing about one-third less. That advantage extends across the family: Terra performs just above Fable 5, while Luna outperforms Opus 4.8; each does so in roughly one-third of the time, with about half as many output tokens, and at approximately one-quarter the estimated cost. It also sets new state-of-the-art results on Terminal-Bench 2.1 and DeepSWE, which test complex command-line workflows and long-horizon engineering in real codebases.

GPT-5.6 can write and run lightweight programs that coordinate tools, process intermediate results, monitor progress, and choose the next action as work unfolds. This lets tool-heavy tasks advance with fewer tokens, fewer model round trips, and less guidance. Instead of requiring developers to script every step or passing every tool response back through the model, Programmatic Tool Calling(opens in a new window) in the Responses API can filter large amounts of intermediate data, retain only what matters, and adapt its workflow along the way.

For problems that reward a greater investment of time and compute, GPT-5.6 can push beyond this efficient

default. ``max`` gives GPT-5.6 even more time than ``xhigh`` to reason and explore alternatives, run checks, and revise its approach. ``ultra`` goes further by coordinating four agents in parallel by default, trading higher token use for stronger results and faster time-to-result on demanding tasks. The charts below compare ``ultra``'s default four-agent setup with a one-agent baseline across BrowseComp, SEC-Bench Pro, and Terminal-Bench 2.1; BrowseComp and SEC-Bench Pro also show 16-agent configurations. Across all three evaluations, adding parallel agents shifts the score-latency frontier upward and to the left, reaching stronger results in less time. In the API, developers can build ``ultra``-like experiences using the multi-agent([opens in a new window](#)) beta in the Responses API. , ,

1 of 11

GPT-5.6 delivers a step change in design judgment. With only high-level direction, GPT-5.6 creates tasteful, ergonomic, and functional interfaces. Its stronger computer-use capabilities let it inspect and refine the rendered result—not just generate the underlying code or content—so it can catch visual and functional issues and apply finishing touches before handing the work back.

GPT-5.6's frontend capabilities also turn natural-language requests into polished, interactive explanations and visualizations within ChatGPT Work.

GPT-5.6 delivers better results for professional tasks. It takes messy context from your documents and everyday workflows like Slack, Notion, Microsoft 365, and Google Drive, and converts it into expert-level, shareable artifacts.

GPT-5.6's strength on knowledge work shows up in evaluations spanning long-horizon professional analysis, browsing, tool use, and computer use. GPT-5.6 Sol sets new state-of-the-art results on **BrowseComp** at 92.2% and **OSWorld 2.0** at 62.6%; on OSWorld, it surpasses Opus 4.8 while using 85% fewer output tokens. Here, the performance-per-dollar gains extend across the GPT-5.6 family. Luna nearly matches GPT-5.5's peak performance at less than half the estimated cost, while Terra surpasses it at a lower cost.

GPT-5.6 Sol improves quality in presentations, documents, and spreadsheets, producing outputs that are more polished and accurate. It can create fully editable presentations from scratch, translating a prompt and source material into a coherent visual narrative with strong layouts, hierarchy, and design.

The improvement is especially pronounced when following templates and reference decks. **GPT-5.6** can infer a deck's design system—layouts, typography, spacing, colors, and recurring content patterns, including rules embedded in the Slide Master—and apply those conventions consistently to new material. In this example, when asked to update numbers based on a reference file, the GPT-5.5 output is missing key components from the master slide, while GPT-5.6 follows the reference structure more faithfully.

GPT-5.6 also creates more visually refined documents and spreadsheets. It follows complex reference formats more faithfully, which is important for repeatable knowledge work activities. It handles equations and financial models with greater precision, and makes better use of typography, spacing, hierarchy, and page or worksheet layout.

Early customers testing GPT-5.6 saw improvements to knowledge work outputs across domains.

1 of 9

GPT-5.6 is our strongest cybersecurity model yet, achieving frontier performance with significantly fewer tokens. On **ExploitBench**, which measures progress from reaching vulnerable code through arbitrary code execution, it scores 73.5% versus GPT-5.5's 47.9% at a comparable output-token budget. On **ExploitGym**, which asks agents to turn real-world vulnerabilities into working exploits, it almost doubles GPT-5.5's peak pass rate, from 15.1% to 24.9% under the two-hour cap; with six hours, it reaches 33.7%. On **SEC-Bench Pro**, which tests proof-of-concept generation on complex software, it scores 71.2% versus GPT-5.5's 45.8% at an improved latency.

GPT-5.6 supports important defensive tasks such as secure code review, patching, threat modeling, and blue teaming. Qualified individuals and organizations in OpenAI Daybreak's Trusted Access for Cyber program can access more of its defensive capability through more precise safeguards for verified work in authorized environments, including vulnerability triage and validation, malware analysis, detection engineering, and patch validation.

Individuals can verify their identity and request trusted access(opens in a new window), and organizations can apply for their teams. Individual members will need to enable Advanced Account Security(opens in a new window) with hardware-backed passkeys by September 1 to retain access to our most cyber-capable frontier models; those who do not will return to default access. Users who do not already have hardware-backed passkeys can receive preferred pricing(opens in a new window) from our partner, Yubico. We are also taking additional steps to restrict access to high-risk entities and in high-risk jurisdictions.

GPT-5.6 Sol also shows broad gains across ****scientific research****. On life sciences evaluations, GPT-5.6 demonstrates Pareto improvements over GPT-5.5 on real-world biology, life science research workflows, and chemistry.

GPT-5.6 is our strongest model yet for accelerating AI research. Inside OpenAI, researchers use it across the development loop: diagnosing failures, optimizing training systems, running experiments, and interpreting results. We already saw that acceleration and stronger adoption during the internal testing period of GPT-5.6, as average daily output tokens per active researcher were more than twice the highest level observed for GPT-5.5.

This way of working is quickly becoming standard. Over the past six months, the share of research compute devoted to internal coding inference grew 100-fold, while internal agentic token usage increased approximately 22-fold. These adoption metrics do not measure research progress on their own, but they show how rapidly AI assistance is increasing for research and across other teams like sales, marketing, user ops, finance, and more.

To measure this capability directly, we developed an internal suite of evaluations based on real AI research tasks, including debugging research systems, optimizing kernels and training recipes, running machine-learning experiments, and improving another model.

As model capabilities increase, we strengthen our safety stack so advanced intelligence can remain broadly useful while applying greater scrutiny to the highest-risk uses. For GPT-5.6, we built our most robust safety system to date, calibrated to each model's capabilities and powered by more compute than ever before.

The GPT-5.6 models are more capable than our earlier models in both biology and cybersecurity but do not cross the Critical threshold in either category. In cybersecurity, our testing suggests GPT-5.6 is better at finding and fixing vulnerabilities than at reliably carrying out autonomous, end-to-end attacks against hardened targets—giving defenders an opportunity to strengthen systems before weaknesses are exploited. In biology, our testing suggests GPT-5.6 can support legitimate research but does not provide the end-to-end capability needed to create, engineer, or synthesize a highly dangerous novel threat.

Both domains are inherently dual-use. In cybersecurity, the same capabilities that could help an attacker exploit a vulnerability can help a defender find it, reproduce it, and build a reliable fix. Overblocking therefore creates a security risk of its own. It can prevent defenders from testing systems and deploying patches while malicious actors continue using other models, including increasingly capable open-source models, as well as established tools. Effective safeguards account for the context and likely consequences of a request, preserving legitimate defensive work while applying stronger controls where the evidence indicates a serious risk of harm.

GPT-5.6's safeguards are layered for greater accuracy and redundancy, and designed to adapt quickly as new attacks emerge. Protections trained into the model work alongside real-time checks, continuous monitoring, and account-level enforcement, to help the system remain safe even when a particular layer does not work as intended. In many systems, classifier flags alone decide what to block, relying on lower intelligence models that are harder to change in order to prevent harm. Our approach adds a reasoning monitor that reviews the conversation to determine if there is a potential for harm. This design is intended to enable defensive work while blocking serious misuse, with the most sensitive capabilities reserved for verified users through Trusted Access. Because some protections use test-time reasoning, we can rapidly update them to close gaps without retraining classifiers from scratch.

We are taking a more conservative approach as we continue to strengthen the system against adaptive attacks. Compared with previous models, our GPT-5.6 Sol cyber safeguards block roughly ten times more potentially harmful activity. Because these measures can create friction for benign use, we provide an option in ChatGPT and Codex to easily retry prompts on lower-capability models, and we will continue reducing the impact of our safeguards on benign use while maintaining a high robustness bar. This reflects our iterative deployment approach: starting conservatively and improving based on what we learn from real-world use.

Before general availability, we ran our most intensive safety evaluations to date, including extensive red teaming,

robust capability and safeguard testing with external experts, and approximately 700,000 NVIDIA A100 Tensor Core GPU-equivalent hours of black-box automated red teaming. This enabled us to systematically probe likely weak points, surface jailbreaks, and help us strengthen the system before launch.

There is no such thing as perfect security, and our work to secure increasingly capable models continues. New weaknesses will be discovered, as will new jailbreaks that circumvent existing safeguards. Each new generation of model will also create new avenues for attack and misuse. We build for that reality through layered safeguards, continuous monitoring, rapid remediation, and collaboration across the defensive community. For GPT-5.6, we have paired our existing security (opens in a new window) and biology bug bounty programs with a new rapid-remediation process and our strongest monitoring effort to date. Findings from researchers, monitoring, and real-world misuse will feed into new evaluations and stronger safeguards on an ongoing basis.

GPT-5.6 spans three model tiers: Sol, our flagship; Terra, a lower-cost model with performance competitive with GPT-5.5; and Luna, our fastest and most affordable model. The number identifies the generation, while Sol, Terra, and Luna are durable capability tiers that can advance on their own cadence.

GPT-5.6 is available starting today across ChatGPT, Codex, and the OpenAI API. **The rollout is starting globally now and will continue gradually toward full availability over the next 24 hours.**

* **Chat:** Plus, Pro, Business, and Enterprise users access GPT-5.6 Sol through medium and higher effort settings. Pro and Enterprise users can also select GPT-5.6 Sol Pro for the highest-quality results on complex tasks.

* **ChatGPT Work and Codex:** Free and Go users access GPT-5.6 Terra. Plus, Pro, Business, and Enterprise users can choose among GPT-5.6 Sol, Terra, and Luna and set an effort level for each. `max` is available to all users with access to GPT-5.6 in ChatGPT Work and Codex and can be toggled on in settings. In ChatGPT Work, `ultra` is available to Pro and Enterprise users. In Codex, it is available to Plus and higher plans.

* **API:** Developers can access Sol, Terra, and Luna through the OpenAI API. In the Responses API, Programmatic Tool Calling lets GPT-5.6 write and run programs in-memory that coordinate tools and process intermediate results, making it Zero Data Retention (ZDR) compatible. Multi-agent, initially available in beta, lets GPT-5.6 run concurrent subagents and synthesize their work in a single request.

GPT-5.6 is priced per 1M tokens across three model sizes: Sol is \$5 input / \$30 output; Terra is \$2.50 input / \$15 output; and Luna is \$1 input / \$6 output. GPT-5.6 also introduces more predictable prompt caching, including support for explicit cache breakpoints (opens in a new window) and a 30-minute minimum cache life. For GPT-5.6 and later models, cache writes are billed at 1.25x the model's uncached input rate, while cache reads continue to receive the 90% cached-input discount.

Professional

Eval	**GPT-5.6 Sol**	**GPT-5.6 Terra**	**GPT-5.6 Luna**	**GPT-5.5**	**Claude Fable 5**	**Claude Opus 4.8**	**Gemini 3.1 Pro Preview**	**Gemini 3.5 Flash**		---	---	---	---	---	---	---	---	---	Agents' Last Exam	52.7%	50.4%	50.3%	46.9%	40.5%	45.2%	32.1%	—	—	GDPval-AA v2	1,747.8 Elo	1,593 Elo	1,591.8 Elo	1,493.7 Elo	1,759.6 Elo	1,600.1 Elo	962.3 Elo	1,348.8 Elo	—	—	Big Finance Bench	53%	51%	36%	49%	—	44%	—	—	Artificial Analysis Intelligence Index v4.1	58.9 Index score	55 Index score	51.2 Index score	54.8 Index score	59.9 Index score	55.7 Index score	46.5 Index score	50.2 Index score
----------	-----------------	-------------------	------------------	-------------	--------------------	---------------------	----------------------------	----------------------	--	-----	-----	-----	-----	-----	-----	-----	-----	-----	-------------------	-------	-------	-------	-------	-------	-------	-------	---	---	--------------	-------------	-----------	-------------	-------------	-------------	-------------	-----------	-------------	---	---	-------------------	-----	-----	-----	-----	---	-----	---	---	---	------------------	----------------	------------------	------------------	------------------	------------------	------------------	------------------

Coding

Eval	**GPT-5.6 Sol**	**GPT-5.6 Sol Ultra**	**GPT-5.6 Terra**	**GPT-5.6 Luna**	**GPT-5.5**	**Claude Mythos 5**	**Claude Mythos Preview**	**Claude Fable 5**	**Claude Opus 4.8**	**Gemini 3.1 Pro Preview**		---	---	---	---	---	---	---	---	Artificial Analysis Coding Agent Index v1.1	80 Index score	—	77.4 Index score	74.6 Index score	76.4 Index score	—	—	77.2 Index score	72.5 Index score	42.7 Index score	—	SWE-Bench Pro	64.6%	—	63.4%	62.7%	59.4%	80.3%	77.8%	80%	69.2%	54.2%	—	DeepSWE v1.1	72.7%	—	69.6%	67.2%	67%	—	—	69.7%	59%	11.8%	—	Terminal-Bench 2.1	88.8%	91.9%	87.4%	84.7%	85.6%	88%	—	83.1%	78.9%	70.7%
----------	-----------------	-----------------------	-------------------	------------------	-------------	---------------------	---------------------------	--------------------	---------------------	----------------------------	--	-----	-----	-----	-----	-----	-----	-----	-----	---	----------------	---	------------------	------------------	------------------	---	---	------------------	------------------	------------------	---	---------------	-------	---	-------	-------	-------	-------	-------	-----	-------	-------	---	--------------	-------	---	-------	-------	-----	---	---	-------	-----	-------	---	--------------------	-------	-------	-------	-------	-------	-----	---	-------	-------	-------

Science and health

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Fable 5** | **Claude Opus 4.8** | **Gemini 3.1 Pro Preview** | **Gemini 3.5 Flash** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | GeneBench Pro | 28.7% | 23.3% | 10.8% | 12% | — | 16% | 3.1% | 8.14% | | LifeSciBench | 59.9% | 56% | 51.2% | 50.4% | — | 53.6% | — | — | | MedChemBench (Internal) | 48.3% | 35% | 30.4% | 35.5% | — | — | — | — | | HealthBench Professional⁶ | 60.5% | 57.7% | 55.7% | 49.5% | 60.9% | 53% | — | — |

Computer use

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Sol Ultra** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Mythos 5** | **Claude Mythos Preview** | **Claude Opus 4.8** | **Gemini 3.1 Pro Preview** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | OSWorld 2.0 | 62.6% | — | 50.2% | 45.6% | 47.5% | — | — | 54.8% | — | | BrowseComp | 90.4% | 92.2% | 87.5% | 83.3% | 84.4% | 88% | 87.9% | 84.3% | 85.9% | | BenchCAD | 70.6% | — | 62.3% | 63.1% | 44.4% | 38.4% | 35.5% | 27.3% | — | | BenchCAD (python tool) | 83.4% | — | 78.2% | 73.9% | 55.8% | 65% | 61% | 51.8% | — |

Cybersecurity

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Sol Ultra** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Mythos 5** | **Claude Mythos Preview** | **Claude Opus 4.8** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Capture-the-Flag Challenges | 96.7% | — | 91.8% | 85.2% | 88.1% | — | — | — | | SEC-Bench Pro | 71.2% | 74.3% | 57.7% | 48.9% | 45.8% | — | — | — | | ExploitBench | 73.5% | — | 52.9% | 33.2% | 47.9% | 78% | 74.2% | 40% | | ExploitGym | 33.7% | — | 23.2% | 12.4% | 15.1% | — | — | — |

Self-improvement

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | | --- | --- | --- | --- | --- | | Internal Research Debugging Evaluation | 68.3% | 67.8% | 50.8% | 50% | | KernelGen 1P | 61.1% | 49.2% | 22.4% | 29.3% | | NanoGPT | 9.69% | 14.5% | 1.66% | 2.65% | | PostTrainBench Lite | 50.3% | 51.5% | 29.6% | 38.8% | | RSI Index | 57.9% | 56.3% | 41.9% | 41.7% |

Multimodal

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Fable 5** | **Claude Opus 4.8** | **Gemini 3.1 Pro Preview** | | --- | --- | --- | --- | --- | --- | --- | --- | | MMMU Pro (no tools) | 83% | 80.7% | 78.4% | 81.2% | — | — | 80.5% | | MMMU Pro (with tools) | 84.6% | 82% | 79.5% | 83.2% | — | — | — | | gdp.pdf | 30.7% | 24.7% | 22.7% | 26% | 29.8% | 22.5% | 16.7% |

Academic

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Mythos 5** | **Claude Mythos Preview** | **Claude Fable 5** | **Claude Opus 4.8** | **Gemini 3.1 Pro Preview** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | GPQA Diamond | 94.6% | 92.9% | 92.3% | 93.6% | 94.1% | 94.6% | 92.6% | 92% | 94.3% | | FrontierMath Tier 1-3 (v2) | 89% | 84.9% | 78.6% | 85.3% | — | — | 87% | 80% | 59.6% | | FrontierMath Tier 4 (v2) | 83% | 68.3% | 58.5% | 72.5% | — | — | 87.8% | 56.1% | — |

Tool use

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Mythos 5** | **Claude Mythos Preview** | **Claude Fable 5** | **Claude Opus 4.8** | **Gemini 3.1 Pro Preview** | **Gemini 3.5 Flash** | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | AutomationBench | 18.1% | 15.2% | 14.9% | 12.9% | — | — | 17.4% | 15.5% | — | 14.5% | | Toolathlon | 58% | 53.1% | 53.4% | 55.6% | 61.7% | 61.1% | 61.7% | 59.9% | 48.8% | — |

Long context

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Mythos 5** | **Claude Mythos Preview** | **Claude Opus 4.8** | | --- | --- | --- | --- | --- | --- | --- | --- | | OpenAI MRCC v2 8-needle 256K-512K | 91.5% | 89.6% | 41.3% | 81.5% | — | — | — | | OpenAI MRCC v2 8-needle 512K-1M | 73.8% | 72.5% | 41.3% | 74% | — | — | — | | GraphWalks BFS 256k f1 | 90.7% | 76.9% | 81.3% | 73.7% | 91.1% | 85.7% | 85.9% | | GraphWalks BFS 1mil f1 | 77.1% | 71.2% | 51.2% | 45.4% | 79.4% | 74.3% | 68.1% |

Abstract reasoning

| **Eval** | **GPT-5.6 Sol** | **GPT-5.6 Terra** | **GPT-5.6 Luna** | **GPT-5.5** | **Claude Opus 4.8** | **Gemini

3.1 Pro Preview** | | --- | --- | --- | --- | --- | --- | | ARC-AGI-3⁷ | 7.78% | 0.8% | 0.18% | 0.43% | 1.5% | 0.42% |