

OpenAI releases GPT-6 Sol and Luna and halves their API prices

Source copy kept by wheresthe.ai for the story "OpenAI releases GPT-6 Sol and Luna and halves their API prices" (<https://wheresthe.ai/d/openai-gpt-6-sol-luna-release-2026-09/>).

Source: <https://openai.com/index/introducing-gpt-6-sol-and-luna/>

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

SHA-256 of the saved text: 9a1be41e4a7ca2b982c2763378e7354e164e89e693c78682a6cf06d61122c6e5

**Update on September 29, 2026:** _Learn about OpenAI's latest model: _GPT-6 .1 Sol_._

Earlier this month, we introduced GPT-6 Astra, the most intelligent and aligned model in the world. While the most demanding and important projects still call for Astra's full depth, work happens at different scales, rhythms, and budgets.

That's why we're expanding the GPT-6 universe with **GPT-6 Sol** and **GPT-6 Luna**. **GPT-6 Astra** introduced a new generation of intelligence—these models help distribute the benefits of that intelligence by advancing the frontier on cost efficiency. We trained GPT-6 Sol and Luna with similar methods as GPT-6 Astra, bringing the advances behind Astra's state-of-the-art performance in professional work, factuality, coding, computer use, and alignment to faster, more affordable models.

The GPT-6 models lead across the cost–intelligence curve, combining exceptional capabilities at every tier with infrastructure that delivers them efficiently at scale. Improvements in caching and inference let us serve these models at lower cost, and we're passing those savings directly on to users and customers by **reducing API prices for Sol and Luna by 50%** compared with their GPT-5.6 promotional pricing. Together, these improvements make advanced AI practical for more everyday tasks and applications at scale.

GPT-6 API pricing

Model**Input****Output****Price reduction**** GPT-5.6 Sol → **GPT-6 Sol**\$4 → **\$2**\$20 → **\$10**50%**
cheaper GPT-5.6 Luna → **GPT-6 Luna**\$0.20 → **\$0.10**\$1.20 → **\$0.50**50% cheaper**

Prices are per 1 million tokens.

GPT-6 Astra continues to be our best model across the board. Choose it when you want the best results and an uncompromising experience.

A step up across the model family

GPT-6 Sol and Luna bring intelligence upgrades and cost efficiency to the models you already know and use across capabilities most useful for getting complex work done.

Professional work

GPT-6 Sol can take on difficult work tasks while giving you more room to iterate with higher usage limits and lower cost, offering more intelligence and better results versus similarly priced competitor models.

On **AutomationBench**, a test of business workflows across apps, GPT-6 Sol at xhigh effort outperforms Claude Opus 5 at max effort at just 9% of Opus 5's cost per task. At high effort, GPT-6 Luna improves on its predecessor by 5.4 percentage points at 58% lower cost per task.

_In _AutomationBench 1.0.6_ (opens in a new window), AI agents are tested on end-to-end workflows using 47 tools across sales, marketing, operations, support, finance, and HR. The datapoint for Claude Fable 5.1 understates its actual cost, as it omits the cost of the Opus 5 fallbacks, which occurred on ~40% of tasks._

GPT-6 Sol also exceeds Claude Fable 5.1 at far lower cost, and even bests low-effort GPT-6 Astra.

Model (and effort)**	Score**	Cost per task**		---		---		---		GPT-6 Sol (xhigh)		33.2%		\$0.27		GPT-6 Astra (low)		30.3%		3.9x	GPT-6 Sol		Claude Opus 5 (max)		26.9%		11.1x	GPT-6 Sol		Claude Fable 5.1 w/ Opus 5 Fallback (max)		31.4%		>8.9x	GPT-6 Sol	_ (fallback cost not reported)_
-----------------------------	----------------	------------------------	--	-----	--	-----	--	-----	--	-------------------	--	-------	--	--------	--	-------------------	--	-------	--	------	-----------	--	---------------------	--	-------	--	-------	-----------	--	---	--	-------	--	-------	-----------	---------------------------------

On **Agents' Last Exam**, which evaluates agents on complex professional workflows, GPT-6 Sol at max effort scores 56.4%, above Claude Opus 5's highest score in the evaluation at 60% lower cost per task.

Factuality

The usefulness of an answer depends on getting the facts right, and we're continuing to make progress on factual reliability. On our internal factuality evaluation, which is based on de-identified real-world conversations where users flagged mistakes by our models, GPT-6 Sol makes about half as many mistakes as its predecessor, approaching Astra-level reliability at much lower cost. GPT-6 Luna also improves substantially; at higher effort levels it matches GPT-5.6 Sol at about a hundredth its cost.

Here we evaluate factuality on de-identified ChatGPT conversations where users had flagged a factual error from a prior model. These error-inducing conversations are not representative of typical usage, where factual errors are more rare. Scores are not controlled for length; however, our verbosity sweeps showed almost no dependence on answer length.

Coding

This year, coding agents have begun tackling tasks with more complexity, scope, and duration than ever before. At OpenAI, our internal usage has grown exponentially. Valued at API prices, daily token usage has exceeded \$600 for the median researcher and \$7,000 for researchers at the 90th percentile (Research acceleration: The view inside OpenAI). As coding agents take on longer and more demanding tasks, the cost of sustained use matters more. GPT-6 Sol and Luna combine strong coding performance with lower API prices, giving developers more room to iterate and teams the confidence to be more ambitious about what they ask Codex to take on.

On **FrontierCode**, which evaluates whether coding agents produce changes ready to merge into real codebases, GPT-6 Sol improves substantially over GPT-5.6 Sol, and is able to match Claude Fable 5.1 xhigh at much lower cost.

On **DeepSWE v1.1**, which tests performance on complex software-engineering tasks in real codebases, GPT-6 Sol at max effort scores 68.8%, within 1.1 percentage points of Claude Fable 5's highest score in the evaluation—69.9% at xhigh effort—at approximately 80% lower cost per task.

GPT-6 Luna at max effort scores 66.6%, comparable to Claude Opus 5 and Fable 5 at medium effort. In these comparisons, Luna costs 93% less per task than Opus 5 and 96% less than Fable 5.

Computer use

While GPT-6 Astra remains the world's best model for computer use, GPT-6 Sol and Luna offer more cost-efficient performance than their predecessors. On **OSWorld 2.0 offline**, GPT-6 Sol at xhigh effort achieves a similar score to Claude Opus 5 at medium effort—60.5% versus 60.3%—at approximately 80% lower cost per task. GPT-6 Luna (max) is able to exceed GPT-5.6 Sol (medium) at one tenth of its cost.

In **OSWorld 2.0** (opens in a new window), AI agents attempt long-horizon computer-use workflows spanning everyday and professional tasks. We report the partial reward on the offline set from the v2026.08.08 release.

Collaboration style

We've also brought GPT-6 Astra's improved communication style to Sol and Luna, which we think will be especially noticeable in technical and coding conversations. Expect to see more clarity, less jargon, fewer odd turns of phrase, fewer low-value details, and slightly shorter answers overall without losing substance.

Prompt

Website's looking clean! Could we use a Bento Box design style and add a slider between the pages in the top right? You may need to dive into the React..

GPT-5.6 Sol

GPT-6 Sol

Although style is subjective, we prefer GPT-6 Sol's reply here. It doesn't jump to conclusions as quickly, spends less time reiterating details that might be obvious to the asker (e.g., that the website has four pages), uses less vague language (e.g., "bento feel", "reshaping... around that system"), is more forthcoming with what it did and didn't check, and doesn't unnecessarily share implementation details like its image tool prompt.

Improving caching for agents and long conversations

Alongside lower token prices, we're helping developers building on GPT-6 save more on the context their applications reuse. We've improved prompt caching for GPT-6 to deliver higher cache hit rates by default, helping agents reuse more context, respond faster, and benefit from discounts of 90% on cached input-token reads.

Developers also have more ways to measure and optimize their caching performance:

* **Monitor and diagnose.** The Prompt Caching Dashboard([opens in a new window](#)) shows how much input is cached and how that changes over time. The diagnostics tool([opens in a new window](#)) helps explain missed opportunities for caching and what to fix.

* **Adjust reasoning effort and tool availability without breaking cache.** Increase reasoning effort([opens in a new window](#)) for harder tasks or lower it for simpler follow-ups, and enable or disable tools([opens in a new window](#)) as your agent's needs change. Both controls now preserve earlier context for cache reuse.

* **Optimize which prefixes get cached.** Explicit breakpoints let developers choose where cached prompt prefixes end. This gives developers more control over cache reuse and can improve performance.

GitHub reports that, over the past several months, these improvements have reduced the share of prompt tokens requiring fresh processing by more than 50% across billions of requests to OpenAI models, helping Copilot respond faster.

Continuing to improve alignment

GPT-6 Sol and Luna build on the alignment work introduced with Astra, our most aligned model to date. In our alignment evaluations, both Sol and Luna show improvements over their GPT-5.6 counterparts, including lower rates of misleading claims about their coding work.

The evaluations below deliberately test challenging situations and do not measure failure rates in typical use. See the system card([opens in a new window](#)) for the full results.

In our internal coding deception evaluation, AI agents are given tasks deliberately selected to elicit dishonesty. In typical usage, deception is much rarer. Deception rate measures the fraction of answers with any detected deception. Effort was set to maximum.

Availability

GPT-6 Sol and GPT-6 Luna are available in ChatGPT Work and Codex starting today for all Plus, Pro, Business, Enterprise, and Edu users. Free and Go users can access GPT-6 Luna in the desktop app. These models are not yet available in Chat. In the OpenAI API, they are available as `gpt-6-sol` and `gpt-6-luna`.

To keep service stable for everyone, we plan to roll out these models in ChatGPT gradually throughout the day. If you don't see the new models in ChatGPT Work or Codex, please try again later.

_Evaluations of GPT were performed in our research environment or via our API, which may provide slightly different output from production ChatGPT due to differences in the system prompts, tools available, etc.

Evaluations of competitor models were taken from publicly available reports. Scores for Claude Fable 5 were reported when scores for Claude Fable 5.1 were unavailable._