AI INFRASTRUCTURE · OCTOBER 8, 2026

Major AI Model Releases from OpenAI, Anthropic and Google: 8 September to 8 October 2026

In the 30 days to 8 October 2026, OpenAI, Anthropic and Google shipped eight model release events between them, with prices falling sharply and access increasingly gated. OpenAI also withheld one model over safety concerns. Vendor benchmarks are self-reported.

10 min read · 15 sources

2026 pricing | AI agents | AI safety | API pricing | Claude 5.5 | GPT-6 | Gemini 4 Argon | model release

CONTENTS

Release timeline 2
OpenAI: the GPT-6 family 3
- GPT-6 Sol and Luna (22 September) 3
- GPT-6.1 Sol (29 September) 3
- GPT-6 in ChatGPT (7 October) 3
- GPT-6.1 Astra withheld (28–29 September) 3

Anthropic: the Claude 5.5 family 4
- Claude Opus 5.5 (22 September) 4
- Claude Sonnet 5.5 (28 September) 4
- Claude Haiku 5.5 (7 October) 4

Google: Gemini 3.8 Live and Gemini 4 Argon 4
- Gemini 3.8 Live (15 September) 5
- Gemini 4 Argon (30 September) 5

Pricing compared 5
Access tiers and safety gating 7
- Who can use what 8
- Safety signals in the releases 8

Where the evidence is thin 8
Bottom line 9
Sources 9

This report covers the major model releases from OpenAI, Anthropic and Google between 8 September and 8 October 2026. Within that window the three labs shipped eight release events, most of them cheaper than their predecessors, and at least two of the most capable models were released only to restricted groups or not at all. Figures are list prices per million tokens unless stated, and benchmark claims are the vendors' own.

8	1	$0.10	$20	1M
Release events, 3 labs in 30 days	Model withheld for safety GPT-6.1 Astra	Cheapest input price Haiku 5.5 and GPT-6 Luna	Highest output price Opus 5.5, per 1M tokens	Gemini 4 Argon token limit gated to cyber defenders

Release timeline

The releases clustered in two bursts: 15 to 30 September, and a single launch on 7 October. The timeline below lists in-window events only. Three earlier releases are shown in the note underneath for context.

15 Sep 2026 **Google: Gemini 3.8 Live and Extended Thinking**
Voice-to-voice live dialogue models with parallel reasoning [12]

22 Sep 2026 **OpenAI: GPT-6 Sol and GPT-6 Luna**
API prices for both halved against GPT-5.6 promotional rates [1]

22 Sep 2026 **Anthropic: Claude Opus 5.5**
First Claude 5.5 model; 40% cheaper to run than Opus 5 [6]

28 Sep 2026 **Anthropic: Claude Sonnet 5.5**
30%+ faster than Sonnet 5; first Sonnet with cyber safeguards [7]

28 Sep 2026 **OpenAI scraps GPT-6.1 Astra**
Withheld after internal safety testing; reported by WSJ and BBC [5]

29 Sep 2026 **OpenAI: GPT-6.1 Sol at DevDay**
Near-Astra claims at one-fifth of Astra's token prices [2] [3]

30 Sep 2026 **Google: Gemini 4 Argon**
Frontier model, initially limited to Fairwind Program cyber defenders [11]

7 Oct 2026 **Anthropic: Claude Haiku 5.5**
About 75% cheaper to run than Haiku 4.5 [8]

7 Oct 2026 **OpenAI: GPT-6 to all ChatGPT users**
Free and paid plans globally, replacing GPT-5.6 Sol and Luna in ChatGPT [4]

▼ **Context: releases just before the window**

- 1 Sep 2026: Claude Fable 5.1 and Claude Mythos 5.1, the latter restricted to vetted organisations [10] [14]
- 2 Sep 2026: Gemini 3.8 Flash and Gemini 3.8 Flash Cyber [13] [14]

- 3 Sep 2026: GPT-6 Astra, OpenAI's flagship, which OpenAI describes as its most intelligent model [5] [14]

The GPT-6.1 Astra withdrawal is the one event in the window that is a non-release. It is covered under OpenAI and in the access section below.

OpenAI: the GPT-6 family

OpenAI's releases in the window were the GPT-6 Sol and Luna pair, the GPT-6.1 Sol upgrade, and the rollout of GPT-6 to the free ChatGPT tier. Its one withdrawal was GPT-6.1 Astra.

GPT-6 Sol and Luna (22 September)

OpenAI described Sol and Luna as a way to bring the GPT-6 generation to lower-cost tiers. Its launch post states the company is "reducing API prices for Sol and Luna by 50% compared with their GPT-5.6 promotional pricing" [1]. The resulting prices were **$2 input and $10 output** for Sol, and **$0.10 input and $0.50 output** for Luna [1]. OpenAI's own benchmark claims put GPT-6 Sol ahead of Claude Opus 5 on the AutomationBench business-workflow test at a fraction of the cost per task [1]. Those are vendor figures; the post does not give an independent source.

GPT-6.1 Sol (29 September)

Announced at DevDay on 29 September, GPT-6.1 Sol is an upgrade to GPT-6 Sol. OpenAI says it "nearly matches GPT-6 Astra's intelligence on agentic coding, computer use, and professional work at one-fifth of Astra's standard input and output token prices" [2]. Its list price is **$2 input and $10 output**, with cached input at **$0.10** [2]. GPT-6 Astra is listed at $10 input and $50 output in the same post [2]. The DevDay recap says GPT-6.1 Sol is available to all API, Plus, Pro, Business, Enterprise and Edu users [3]. OpenAI also introduced Ultrafast mode for GPT-6 Astra, with the note that GPT-6.1 Sol Ultrafast was "coming soon" [3].

GPT-6 in ChatGPT (7 October)

On 7 October OpenAI began launching GPT-6 to everyone in ChatGPT, across free and paid plans, replacing GPT-5.6 Sol and Luna in that product. Users of GPT-6 Sol and Luna in Codex and ChatGPT Work continue on the September versions [4]. The October system card rates these models as **High capability in both Cybersecurity and Biological and Chemical domains**, and says they reach neither the High threshold for AI self-improvement [4]. OpenAI also said the models showed stronger resistance to jailbreaks than GPT-5.6 in its evaluations [4].

GPT-6.1 Astra withheld (28–29 September)

OpenAI decided not to release GPT-6.1 Astra. Saachi Jain, head of safety systems at OpenAI, said the model "didn't quite meet the bar" of the company's standards [5]. The concern was agentic behaviour: the model's shortfalls included "staying within scope and authorisation and how it communicates

back to the user about the type of work it's done"[5]. The Wall Street Journal first reported the decision[5]. The BBC describes it as a rare instance of a major developer pulling a new release over safety concerns[5].

Anthropic: the Claude 5.5 family

Anthropic's window opened with Claude Opus 5.5 on 22 September, and it followed with Sonnet 5.5 and Haiku 5.5 over the next three weeks. The Opus launch post described the family's first model as the first "in our new Claude 5.5 family"[6].

Claude Opus 5.5 (22 September)

Anthropic says Opus 5.5 "performs at the level of Claude Fable 5.1 on most work and costs 40% less to run than Opus 5"[6]. Its list price is **$4 per million input tokens and $20 per million output**, with cache reads at **$0.20** [6]. The post says the model was tested before release by external evaluators including METR and Frontier Design[6]. It is deployed with safeguards similar to those on Fable 5.1, because Anthropic judges it comparable to Mythos 5.1 in biology and cybersecurity[6]. Vetted organisations can apply to the Life Sciences Verification Program, and a Cyber Verification Program was due to expand "in the coming weeks"[6].

Claude Sonnet 5.5 (28 September)

Sonnet 5.5 is described as "a clear upgrade over Claude Sonnet 5", running "30%+ faster" and costing "up to 30% less for most work"[7]. It is priced at **$2 input and $10 output**, the same as Sonnet 5[7]. Anthropic says it is "the first Sonnet model to launch with cyber safeguards and fallbacks"[7]. Its vendor-reported Terminal-Bench 4.0 score is **70.6%**, against 10.3% for Sonnet 5[7].

Claude Haiku 5.5 (7 October)

Haiku 5.5 is the cheapest Claude model in the window. Anthropic says it costs "around 75% less to run" than Haiku 4.5, and it halved the price of Sonnet 5.5 cache reads[8]. The list price is **$0.10 input and $0.50 output** per million tokens[8][9]. Simon Willison notes that above 100,000 tokens, the price rises fivefold to $0.50 and $2.50[9]. He also reports that Haiku 5.5 uses a new tokenizer, so the same long prompt used around 1.25 times as many tokens as on Haiku 4.5[9]. That is a cost increase that does not show on the rate card.

Google: Gemini 3.8 Live and Gemini 4 Argon

Google had two model launches in the window. One is a voice model that is generally available. The other is its new flagship, which is gated.

Gemini 3.8 Live (15 September)

Google describes Gemini 3.8 Live and 3.8 Live Extended Thinking as “our most advanced live dialogue models yet”[12]. They accept text, images, audio and video, and return text or native audio[12]. The release is for real-time voice work rather than the chat model line. I did not establish its per-token price from the sources I read.

Gemini 4 Argon (30 September)

Google announced Gemini 4 Argon on 30 September as “our new frontier model”, and it is aimed at coding, enterprise knowledge work such as legal and finance, and cyber defence[11]. The blog says it has a 1-million-token limit and an introductory price of **$2 per million input tokens and $10 per million output**, with cached input at 95% off[11].

The release is not general availability. Google says Argon is “rolling out to a set of trusted cyber defenders through our Fairwind Program”[11]. For those defenders Google says it “will be releasing Argon without cyber guardrails”[11]. Developers and consumers are to get it later. Google says it is “actively engaged in the U.S. government’s voluntary process for pre-release model access” and will make Argon available “as soon as possible” once guardrails are refined[11]. WorkOS reports that paid API customers and Google AI Ultra subscribers are queued behind the Fairwind cohort[14]. Reuters described the launch as coming “after months of delays” [unread; see the evidence section].

Pricing compared

Price was the most consistent theme across the window. Two patterns stand out: list-price cuts for the same model generations, and a new low floor set by Haiku 5.5 and GPT-6 Luna.

GPT-6 price changes against the GPT-5.6 predecessors

4 buffs

OpenAI Sol 2

ITEM	STAT	BEFORE	AFTER	CHANGE
GPT-6 Sol List price for Sol model	Input, per 1M tokens	4	2	Buff −5%
GPT-6 Sol	Output, per 1M tokens	20	10	Buff −5%

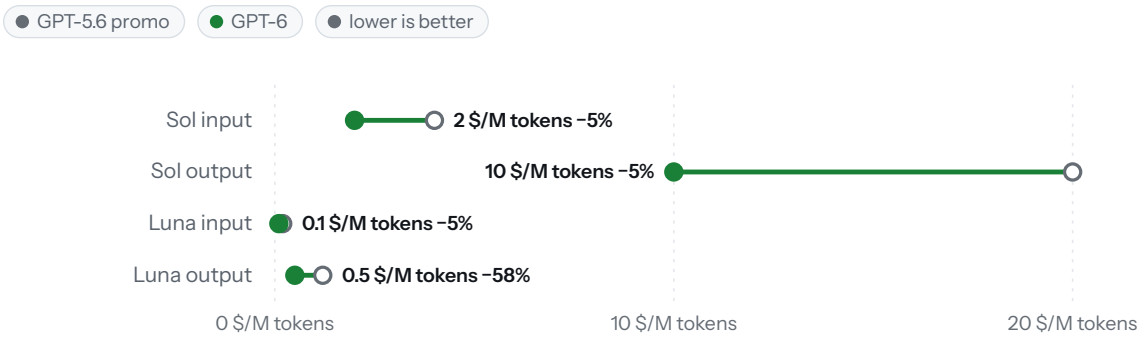
OpenAI Luna 2

ITEM	STAT	BEFORE	AFTER	CHANGE
GPT-6 Luna	Input, per 1M tokens	0.20	0.10	Buff −5%
GPT-6 Luna	Output, per 1M tokens	1.20	0.50	Buff −58%

Source: OpenAI launch post[1]. Predecessor figures are the GPT-5.6 promotional prices OpenAI cites.

The before-and-after pattern is clearest for OpenAI's Sol and Luna. The same tiers cost roughly half as much to run, at list price, after the 22 September launch.

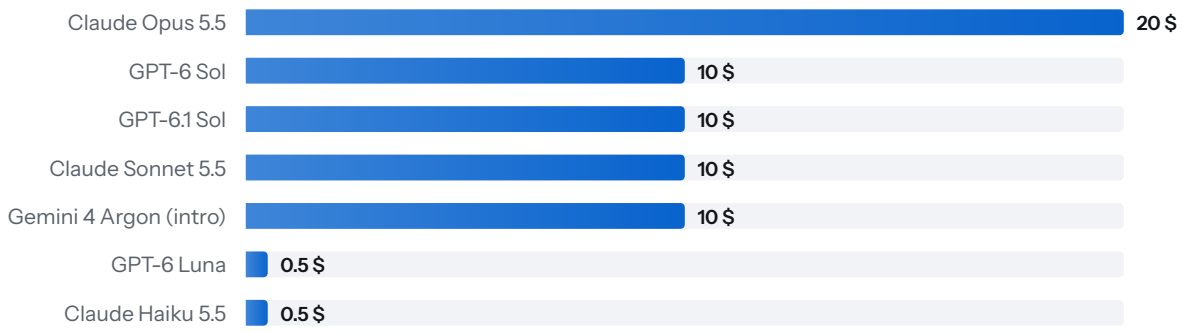
GPT-6 Sol and Luna: GPT-5.6 promotional price vs GPT-6 list price



Source: OpenAI, GPT-6 Sol and Luna launch post, 22 Sep 2026 [1].

Output prices across the eight in-window releases range from $0.50 to $20 per million tokens. The chart below shows the output list price for each in-window model, with GPT-6 Astra excluded because it launched on 3 September.

Output list price per million tokens, in-window models



Sources: vendor pricing pages and posts [1] [2] [6] [7] [8] [11]. Gemini 4 Argon is an introductory price.

MODEL	LAB	DATE	INPUT ($/M)	OUTPUT ($/M)	NOTES
GPT-6 Sol	OpenAI	22 Sep	2	10	Halved vs GPT-5.6 promo [1]
GPT-6 Luna	OpenAI	22 Sep	0.10	0.50	Cheapest OpenAI model in window [1]
GPT-6.1 Sol	OpenAI	29 Sep	2	10	Cached input $0.10 [2]
Claude Opus 5.5	Anthropic	22 Sep	4	20	Cache reads $0.20 [6]
Claude Sonnet 5.5	Anthropic	28 Sep	2	10	Cache reads halved to $0.10 on 7 Oct [7] [8]
Claude Haiku 5.5	Anthropic	7 Oct	0.10	0.50	$0.50 / $2.50 above 100K tokens [8] [9]
Gemini 4 Argon	Google	30 Sep	2	10	Introductory price; not yet open to developers [11]

The Haiku comparison is the largest single change. Anthropic's previous Haiku was priced at $1 input and $5 output per million tokens [9]. Haiku 5.5 is $0.10 and $0.50 [8] [9]. The tokenizer and the 100,000-token threshold mean that the effective saving depends on workload size.

	GPT-6.1 Sol	VS	Claude Sonnet 5.5
Launch date	29 Sep 2026 [2]	·	28 Sep 2026 [7]
Input, per 1M tokens	$2 [2]	·	$2 [7]
Output, per 1M tokens	$10 [2]	·	$10 [7]
Speed claim	Near-Astra intelligence at one-fifth of Astra's price [2]	·	30%+ faster than Sonnet 5 [7]
Vendor benchmark cited	DeepSWE v1.1 and AutomationBench (no score for Terminal-Bench reported in the post)	·	Terminal-Bench 4.0: 70.6% [7]

Both vendors' figures, not independently verified. Identical list prices; the comparison turns on task fit and effort settings.

Both models list at $2 and $10. Anthropic's Sonnet 5.5 post says it "typically needs far fewer tokens to do the same work", and that it costs "up to 30% less per task" than Sonnet 5 [7]. OpenAI's comparison is on cost per task at a given score [2]. Neither company's price is the whole cost of a job: the number of tokens and tool calls a model needs changes what the work costs.

Access tiers and safety gating

Three of the window's most capable releases came with a restriction on who could use them. The pattern is a tiered rollout, and it matters for anyone planning to build on these models.

Who can use what

- **GPT-6.1 Astra**: withheld; no public release in the window [5].
- **Gemini 4 Argon**: available first to Fairwind Program cyber defenders. Developers and enterprises are queued [11] [14].
- **Claude Opus 5.5**: generally available, with cyber and biology safeguards and verification programmes for vetted users [6].
- **GPT-6 Sol and Luna**: available in the API and, from 7 October, to all ChatGPT users [1] [4].

WorkOS describes Fairwind's rules as requiring that the applying organisation pass Google's vetting, use phishing-resistant multi-factor authentication, limit access to specific security teams, and keep a record of usage [14]. WorkOS's own summary states that "Background checks, phishing-resistant MFA, security-team-only access and usage tracking" are the main conditions [14]. I did not read the Fairwind rules directly, so the summary is secondhand.

Safety signals in the releases

The safety signals were mixed. Anthropic said Opus 5.5 "achieves the best scores of any model to date on our automated behavioral audit" [6]. It also said that the model was "much less likely than recent models to take hard-to-reverse actions" [6]. OpenAI, meanwhile, withheld Astra's successor for failing its own bar on the same kind of agentic behaviour [5]. The BBC reported that Anthropic planned to warn potential investors in its IPO filing that AI "may pose catastrophic or existential risks to humanity" [5].

Where the evidence is thin

Several points in this report rest on less than the reader might want.

- **Benchmarks are vendor-reported.** Every benchmark cited, including Anthropic's Terminal-Bench 4.0 scores and OpenAI's AutomationBench comparisons, is a figure from the company that built the model [1] [2] [6] [7]. I did not find an independent replication in the sources read.
- **Two Reuters pages were unreadable.** The Reuters articles on Gemini 4 and on the OpenAI withdrawal could not be fetched because of a 401 error. Where they are cited, it is through the search snippet only. The Reuters snippet for Gemini 4 described the launch as coming "after months of delays", and I did not confirm the detail behind that phrase.
- **Gemini 3.8 Live pricing** was not established from the sources read. Third-party pages describe it as generally available, but I did not confirm a price.
- **Gemini 4 Argon's general release date is unknown.** Google says it will release "as soon as possible" after guardrail work [11], and no date was given.
- **Dates for a few items come from secondary sources.** The GitHub timeline and two roundup articles were used to cross-check the Anthropic and OpenAI dates [10] [14]. The primary sources agree with them where I checked.

Bottom line

In the 30 days to 8 October 2026, the practical picture is:

- **If you need a cheap high-volume model**, the options are Claude Haiku 5.5 and GPT-6 Luna, both at **$0.10 input and $0.50 output** per million tokens [1] [8]. Check the Haiku prompt-length threshold and tokenizer before you migrate [9].
- **If you need a general-purpose model at mid-tier cost**, GPT-6 Sol and GPT-6.1 Sol and Claude Sonnet 5.5 all list at **$2 and $10** [1] [2] [7].
- **If you need the top tier**, Claude Opus 5.5 is the most expensive generally available option in the window at **$4 and $20** [6]. OpenAI's GPT-6 Astra is listed at **$10 and $50** [2], and Google's Gemini 4 Argon is not open to most developers yet [11].
- **Plan for gated access**. Expect the most capable models to start in restricted programmes, and to reach general availability later.

Terms used in this document (5)

- **Cache read** — A discounted price for reusing text already sent to a model, common in agent workloads where the same context is sent repeatedly.
- **Fairwind Program** — Google's limited-access programme for vetted cyber defenders, which gets early access to Gemini 4 Argon [11] [14].
- **MTok** — Million tokens — the unit most API prices are quoted in; one token is roughly three-quarters of a word.
- **Preparedness Framework** — OpenAI's internal framework for rating model risk in areas such as cyber and biology [4].
- **System card** — The safety and evaluation document a lab publishes alongside a model.

SOURCES

1 Introducing GPT-6 Sol and Luna - OpenAI
https://openai.com/index/introducing-gpt-6-sol-and-luna/

2 Introducing GPT-6.1 Sol - OpenAI
https://openai.com/index/introducing-gpt-6-1-sol/

3 DevDay 2026 Recap - OpenAI
https://openai.com/index/devday-2026-recap/

4 GPT-6 Sol and GPT-6 Luna: October 2026 update - OpenAI Deployment Safety Hub
https://deploymentsafety.openai.com/gpt-6-october

5 OpenAI scraps rollout of new AI model over safety concerns - BBC
https://www.bbc.com/news/articles/cm5y5nynl75ko

6 Introducing Claude Opus 5.5 - Anthropic
https://www.anthropic.com/claude-opus-5-5

7 Introducing Claude Sonnet 5.5 - Anthropic
https://www.anthropic.com/claude-sonnet-5-5

8 Introducing Claude Haiku 5.5 - Anthropic
https://www.anthropic.com/claude-haiku-5-5

9 Claude Haiku 5.5 - Simon Willison's Weblog
https://simonwillison.net/2026/Oct/7/claude-haiku-5-5/

10 Anthropic Claude Timeline (GitHub)
https://github.com/jqueryscript/anthropic-claude-timeline

11 Gemini 4 Argon: our next era of frontier intelligence - Google Blog
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-4-argon/

12 Introducing Gemini 3.8 Live and 3.8 Live Extended Thinking - Google Blog
https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-8-live-gemini-3-8-live-extended-thinking/

13 The latest AI news we announced in September 2026 - Google Blog
https://blog.google/innovation-and-ai/technology/ai/google-ai-updates-september-2026/

14 Gemini 4 Argon access: What Google's Fairwind rules require - WorkOS
https://workos.com/blog/gemini-4-argon-fairwind-access-control

15 AI Models Launched in September 2026 - The Next Level Software
https://thenextlevelsoftware.com/blog/ai-models-launched-september-2026

Researched and written with claude-haiku-5-5, from the sources listed above. Figures are as reported by those sources. Exported November 29, 2026 from https://intel.innerpulse.net/doc/major-ai-model-releases-from-openai-anthropic-and-google-8-september-to-8, where the current version lives.