



22 Verified AI-Tool Changes That Mattered in August 2026

VERSION 1.0

Change Type	Count
Capability change	6
Shutdown	5
Model or product launch	4
Availability or access change	4
Plan change	1
Policy change	1
Price change	1

Key findings

August did not produce one clean story such as "models got better" or "AI got cheaper." The month was more operational than that. Choose.ly verified 22 material changes across the monitored August dataset. Capability changes were the largest primary class at six events, followed by five shutdowns. Model or product launches and availability or access changes each contributed four events. The remaining three were one plan change, one price change, and one policy change.

The clearest pattern was that AI systems gained more ability to act while the infrastructure around those systems became less safe to treat as permanent. Grok Bot introduced persistent delegated agents. Anthropic moved computer use, reusable skills, file handling, cross-workspace memory, authenticated browser action, and an isolated Cowork browser further into production use. ChatGPT scheduled tasks gained event triggers from supported external apps. GitHub published an agent-plugin standard designed to package reusable agent capabilities across compatible clients.

At the same time, five products or production surfaces reached shutdown states during August: GitHub Spark's workbench, ChatGPT Atlas, Imagen 4 API endpoints, OpenAI's Assistants API, and the official DALL·E GPT. OpenAI o3 also reached its published ChatGPT retirement date, although that event is classified as an access change because the retirement applied to ChatGPT rather than the underlying API.

Pricing and plan movement also became harder to summarize with a single direction. ChatGPT for PowerPoint moved beyond its temporary free-use window into published paid or flexible usage treatment. Anthropic did the opposite with Claude Sonnet 5, cancelling a planned September API price increase and making its lower introductory rate permanent. Google launched Gemini 3.7 Flash at an introductory rate that was half Gemini 3.6 Flash's July launch rate, but that discount is explicitly time-limited through December 31, 2026.

The practical August lesson is therefore not "buy more agents" or "switch to the cheapest model." It is to treat AI adoption as an operating-system decision. Teams need to understand permissions, billing, model routing, endpoint durability, migration paths, memory, browser trust boundaries, and which dependencies are allowed to change automatically.

August at a glance

Primary change class	Changes	Share
Model or product launch	4	18.2%
Availability or access change	4	18.2%
Capability change	6	27.3%
Plan change	1	4.5%
Price change	1	4.5%
Policy change	1	4.5%
Shutdown	5	22.7%

Coverage statement

This report covers material AI-tool and model changes that were released, became effective, or produced a material user consequence from August 1 through August 31, 2026, and were verified through September 2, 2026.

At the cutoff, Choose.ly's canonical monitored catalog contained 332 records. The normal month-end Catalog Intelligence scan generated August 31 covered 329 tools and 678 official URLs. These are different system scopes and are intentionally reported separately.

A change qualified only when Choose.ly could establish reliable first-party evidence, a defensible August state transition, the affected users or workflows, and a practical decision consequence. A prior state is published only when evidence supports it.

This is not a census of every AI product worldwide. It does not cover every vendor, geography, private beta, negotiated enterprise arrangement, unpublished change, or lower-consequence interface update. The figures describe Choose.ly's verified August dataset and should be cited with the reporting period, verification cutoff, and monitored-catalog scope intact.

The 22 verified changes

AUG-2026-001 | SHUTDOWN | EFFECTIVE 2026-08-31 | IMPACT CRITICAL HIGH

GitHub Spark's editable workbench reached retirement

GitHub began Spark's terminal wind-down on August 4 by stopping new-user access and new app creation. Existing users retained the workbench through August 31 so they could export apps, after which the editable Spark experience retired. GitHub said already-deployed apps would continue working. The event also exposed a dependency on July's GitHub Models retirement. Spark apps using the GitHub Models llm() function had already lost that capability after July 30.

PREVIOUS STATE Spark accepted new users, allowed new app creation and provided an editable workbench.	NEW STATE New users and new app creation stopped August 4. Existing users could access the workbench through August 31 to export apps. GitHub states already-deployed apps continue after Spark retires; apps using GitHub Models llm() had already been affected by the July 30 GitHub Models retirement.
WHO IS AFFECTED Spark builders, app maintainers and apps using llm().	DECISION EFFECT Export editable apps, replace llm() dependencies and move future development to a supported environment.

SOURCES GitHub announcement: <https://github.blog/changelog/2026-08-04-upcoming-deprecation-of-github-spark-on-github-com/>

AUG-2026-002 | AVAILABILITY OR ACCESS CHANGE | EFFECTIVE 2026-08-05 | IMPACT HIGH

Grok Voice's latest alias moved to Think Fast 2.0

The grok-voice-latest alias moved from Think Fast 1.0 to Think Fast 2.0 on August 5. Production users following an unpinned "latest" alias therefore experienced a model transition without changing the alias in their own code. SpaceXAI provided a path for teams that wanted to remain on the earlier model and listed Think Fast 2.0 at US\$0.08 per minute of audio.

PREVIOUS STATE grok-voice-latest routed to Think Fast 1.0.	NEW STATE The alias moved to Think Fast 2.0; users wanting stable 1.0 behavior were directed to pin the prior version. Think Fast 2.0 was listed at US\$0.08 per minute of audio.
WHO IS AFFECTED Developers using the unpinned alias in production voice agents.	DECISION EFFECT Re-test behavior after alias migration or pin a dated version when consistency matters.

SOURCES SpaceXAI documentation: <https://docs.x.ai/developers/release-notes> | SpaceXAI announcement: <https://x.ai/news/grok-voice-think-fast-2>

AUG-2026-003 | CAPABILITY CHANGE | EFFECTIVE 2026-08-06 | IMPACT HIGH

ChatGPT changed GPT-5.6 reasoning controls and the Free/Go access ladder

OpenAI changed how users route reasoning inside ChatGPT. Plus and Pro gained the updated GPT-5.6 Sol reasoning control, while Free and Go moved to Luna as the default with expanded text access and a Think control for harder prompts. The boundary matters: the change applied to ChatGPT conversations and did not silently alter the Sol version used by Work or Codex.

PREVIOUS STATE The August 6 Sol reasoning-effort control and the announced Free/Go Luna access position were not yet present.	NEW STATE Plus and Pro gained the updated GPT-5.6 Sol reasoning control; Free and Go moved to Luna as the default with expanded text access and a Think control. The release did not change Work or Codex.
WHO IS AFFECTED ChatGPT Plus, Pro, Free and Go users.	DECISION EFFECT Route reasoning effort by task consequence and distinguish text access from separate tool limits.

SOURCES OpenAI announcement: <https://openai.com/index/improving-gpt-5-6-sol-in-chatgpt/> | OpenAI Help Center: <https://help.openai.com/en/articles/20001354-gpt-5-6>

AUG-2026-004 | MODEL OR PRODUCT LAUNCH | EFFECTIVE 2026-08-06 | IMPACT HIGH

Adobe launched one unified ChatGPT plugin across its creative stack

Adobe launched a unified ChatGPT plugin spanning more than 70 tools across multiple Adobe products. The material change was not merely another integration. It reduced the number of handoffs between chat, creative work, and document production while still leaving advanced finishing available inside native Adobe applications.

PREVIOUS STATE Adobe's professional creative and document capabilities were not available through one unified ChatGPT plugin.	NEW STATE Adobe launched a unified plugin spanning more than 70 tools across multiple creative and document products.
WHO IS AFFECTED Creative professionals, marketers, document teams and ChatGPT users.	DECISION EFFECT Test a real end-to-end job before consolidating handoffs; retain native Adobe surfaces where advanced finishing is required.

SOURCES Adobe announcement: <https://blog.adobe.com/en/publish/2026/08/06/introducing-adobe-chatgpt-create-edit-get-work-done-all-in-chatgpt>

AUG-2026-005 | AVAILABILITY OR ACCESS CHANGE | EFFECTIVE 2026-08-06 | IMPACT HIGH

Kimi K3 began its general-availability rollout inside GitHub Copilot

Kimi K3 launched in July, but August changed where users could access it. GitHub announced the model as generally available through a gradual rollout to eligible Copilot plans. Business and Enterprise access remained off by default until administrator enablement. GitHub temporarily paused the rollout during a GitHub Actions incident and later resumed it. Published rates were US\$3 input, US\$15 output, and US\$0.30 cached input per million tokens.

PREVIOUS STATE Kimi K3 was a July-launched model but was not available through GitHub Copilot.	NEW STATE GitHub announced Kimi K3 as generally available through a gradual rollout across eligible Copilot plans. Business and Enterprise remained off by default until administrator enablement. Rates were US\$3 input, US\$15 output and US\$0.30 cached input per million tokens.
WHO IS AFFECTED Copilot developers and administrators.	DECISION EFFECT Evaluate Kimi K3 in the actual Copilot workflow and review administrator and data-governance controls before organization-wide enablement.

SOURCES GitHub announcement: <https://github.blog/changelog/2026-08-06-kimi-k3-is-now-available-in-github-copilot/>

AUG-2026-006 | PLAN CHANGE | EFFECTIVE 2026-08-07 | IMPACT HIGH

ChatGPT for PowerPoint moved beyond its free-use window

OpenAI's published schedule kept ChatGPT for PowerPoint free for Business and Enterprise usage through August 6, then moved the product into paid or flexible usage treatment after that date. For token-based Enterprise usage, OpenAI's rate card uses token billing rather than a fixed per-task charge and lists GPT-5.6 Sol at US\$4 input, US\$0.40 cached input, and US\$20 output per million tokens. This is best classified as a plan or billing treatment change rather than a clean continuing-SKU price increase.

PREVIOUS STATE Published Business and Enterprise usage remained free through August 6.	NEW STATE OpenAI's published schedule places PowerPoint usage into paid/flexible usage after August 6. For token-based Enterprise, pricing moves from fixed credits per task to token-calculated usage and lists GPT-5.6 Sol at US\$4 input, US\$0.40 cached input and US\$20 output per million tokens.
WHO IS AFFECTED Teams using ChatGPT for PowerPoint and administrators forecasting shared agentic usage.	DECISION EFFECT Add PowerPoint to cost controls and compare accepted-deck cost rather than assuming the temporary free window persists.

SOURCES OpenAI Help Center: <https://help.openai.com/en/articles/20001415> | OpenAI Help Center: <https://help.openai.com/en/articles/10128477>

AUG-2026-007 | SHUTDOWN | EFFECTIVE 2026-08-09 | IMPACT CRITICAL HIGH

ChatGPT Atlas reached its published shutdown date

OpenAI had already announced that Atlas would stop working on August 9 while browser-based agent capability moved into other ChatGPT and Codex surfaces. The migration problem was not only feature replacement. OpenAI warned that browser state such as bookmarks, open tabs, and history would not transfer automatically.

PREVIOUS STATE Atlas remained available during its published wind-down window.	NEW STATE OpenAI scheduled Atlas to stop working August 9 while moving browser capability to other ChatGPT and Codex surfaces. Browser state did not transfer automatically.
WHO IS AFFECTED Atlas users and teams with stored browser workflows/state.	DECISION EFFECT Migrate workflows and preserve important browser state.

SOURCES OpenAI Help Center: <https://help.openai.com/en/articles/20001371-evolving-atlas-into-chatgpt-for-browser-based-agentic-work>

AUG-2026-008 | MODEL OR PRODUCT LAUNCH | EFFECTIVE 2026-08-11 | IMPACT HIGH

Grok Bot launched persistent delegated agents

SpaceXAI launched Grok Bot in early beta and described persistent agents that use managed computers, work across applications, continue while the user's own device is offline, and return work for approval. That changes the operating model. The user is no longer simply waiting for a response inside one conversation. The agent can continue work across time and applications.

PREVIOUS STATE SpaceXAI did not publicly offer Grok Bot.	NEW STATE SpaceXAI launched Grok Bot in early beta and described persistent agents that operate managed computers, work across apps and can continue while the user's own device is offline.
WHO IS AFFECTED Operators and teams considering persistent delegated work.	DECISION EFFECT Pilot bounded, reversible work and review credentials, permissions and approvals before broad deployment.

SOURCES SpaceXAI announcement: <https://x.ai/news/introducing-grok-bot>

AUG-2026-009 | MODEL OR PRODUCT LAUNCH | EFFECTIVE 2026-08-12 | IMPACT HIGH

Grok 4.6 launched and quickly spread across managed platforms

SpaceXAI launched Grok 4.6 with a 500,000-token context window and configurable reasoning effort. The same-period appearances of Grok 4.6 in third-party managed platforms are treated as distribution consequences of this August launch rather than additional headline events. That avoids inflating one model release into several counts.

PREVIOUS STATE Grok 4.5 was the preceding flagship model.	NEW STATE SpaceXAI launched Grok 4.6 with a 500,000-token context window and configurable reasoning effort; later August third-party availability is treated as same-period distribution consequence.
WHO IS AFFECTED API developers, coding teams and agent builders.	DECISION EFFECT Test 4.6 against 4.5 on long-running workloads and compare delivery surfaces separately.

SOURCES SpaceXAI announcement: <https://x.ai/news/grok-4-6>

AUG-2026-010 | MODEL OR PRODUCT LAUNCH | EFFECTIVE 2026-08-13 | IMPACT HIGH

Gemini 3.7 Flash arrived only 23 days after Gemini 3.6 Flash

Google launched Gemini 3.7 Flash on August 13, only 23 days after the July 21 Gemini 3.6 Flash launch. The new model arrived at US\$0.75 input and US\$3.75 output per million tokens through December 31, 2026, versus 3.6 Flash's original July launch rate of US\$1.50 input and US\$7.50 output. Google also applied the same introductory lower rate to 3.6 Flash. The rate is explicitly temporary and returns to US\$1.50 / US\$7.50 on January 1, 2027.

PREVIOUS STATE Gemini 3.6 Flash launched July 21 at US\$1.50 input / US\$7.50 output per million tokens.	NEW STATE Gemini 3.7 Flash launched at US\$0.75 input / US\$3.75 output per million tokens through December 31, 2026, reverting to US\$1.50 / US\$7.50 January 1, 2027. The same introductory rate was applied to 3.6 Flash.
WHO IS AFFECTED Developers using Gemini Flash for coding, agents and high-volume production work.	DECISION EFFECT Re-test routing while treating the 50% lower rate as an introductory condition rather than a permanent price cut.

SOURCES Google announcement: <https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/> | Google AI documentation: <https://ai.google.dev/gemini-api/docs/pricing>

AUG-2026-011 | CAPABILITY CHANGE | EFFECTIVE 2026-08-20 | IMPACT HIGH

Claude's production-agent building blocks moved further into GA

Anthropic made computer use, Skills API, and Files API generally available and also launched a browser-use tool for structured web interaction. The combination matters because production agents increasingly need more than model reasoning. They need an environment they can operate, reusable procedures, file movement, and controlled web interaction.

PREVIOUS STATE Computer use, Skills API and Files API were not all in the same GA production state.	NEW STATE Anthropic made computer use, Skills API and Files API generally available and also launched a browser-use tool.
WHO IS AFFECTED Developers building production agents that operate software and files.	DECISION EFFECT Move bounded workflows into production with versioned skills, narrow permissions and review for consequential actions.

SOURCES Anthropic announcement: <https://claude.com/blog/computer-use-skills-api-files-api>

AUG-2026-012 | CAPABILITY CHANGE | EFFECTIVE 2026-08-25 | IMPACT HIGH

ChatGPT scheduled tasks became event-driven

OpenAI expanded scheduled tasks beyond a purely clock-based model. Eligible users gained triggers connected to supported Gmail, Slack, and GitHub activity, with approval pauses where the workflow requires them. That moves automation closer to "when something happens, do this" rather than "at 9:00 AM, run this prompt."

PREVIOUS STATE Scheduled tasks were primarily time-triggered.	NEW STATE Eligible users gained event-triggered tasks connected to supported Gmail, Slack and GitHub activity, with approval pauses where required.
WHO IS AFFECTED Eligible Plus, Pro, Business, Enterprise, Edu and Healthcare users, subject to plan/admin controls.	DECISION EFFECT Use event triggers for bounded automations while keeping consequential actions approval-gated.

SOURCES OpenAI Help Center: <https://help.openai.com/en/articles/10291617> | OpenAI Help Center: <https://help.openai.com/en/articles/6825453>

AUG-2026-013 | AVAILABILITY OR ACCESS CHANGE | EFFECTIVE 2026-08-26 | IMPACT HIGH

OpenAI o3 reached its published ChatGPT retirement date

OpenAI had announced a 90-day ChatGPT sunset for o3. That published retirement date arrived on August 26. The important boundary is scope. OpenAI's notice applied to ChatGPT model selection and explicitly said the retirement notice did not change API availability.

PREVIOUS STATE Paid ChatGPT users could select o3 during the announced sunset period.	NEW STATE OpenAI scheduled o3 to retire from ChatGPT August 26. The notice applies to ChatGPT and explicitly states no API change from that retirement notice.
WHO IS AFFECTED Paid ChatGPT users with o3-specific prompts, guidance or routing.	DECISION EFFECT Remove o3-specific ChatGPT assumptions and verify successor behavior; treat API status separately.

SOURCES OpenAI Help Center: <https://help.openai.com/en/articles/9624314-model-release-notes>

AUG-2026-014 | POLICY CHANGE | EFFECTIVE 2026-08-26 | IMPACT HIGH

GitHub Copilot's global model policy began enforcement

GitHub began staged enforcement of its global model policy for Copilot Business and Enterprise on August 26, continuing through September 1. New and previously unconfigured generally available models inherit the global policy, while explicit per-model decisions remain in force. That makes model access partly a governance question rather than a sequence of one-off model decisions.

PREVIOUS STATE Global default model policy could be configured but had not started controlling future/unconfigured model access.	NEW STATE Staged enforcement began August 26 through September 1. New and previously unconfigured GA models inherit the global policy while explicit model decisions remain.
WHO IS AFFECTED Copilot Business/Enterprise administrators and governed teams.	DECISION EFFECT Review global defaults and explicit model settings before future models arrive.

SOURCES GitHub announcement: <https://github.blog/changelog/2026-07-29-default-model-enablement-for-copilot-business-and-enterprise/> | GitHub announcement: <https://github.blog/changelog/2026-08-26-global-model-policy-generally-available/>

AUG-2026-015 | AVAILABILITY OR ACCESS CHANGE | EFFECTIVE 2026-08-26 | IMPACT HIGH

Claude in Chrome reached general availability

Anthropic made Claude in Chrome generally available on paid Claude plans. The primary state transition is access. The secondary consequence is capability: Anthropic says a safety classifier now evaluates browser actions against the user's request while Claude can take more actions without individual approval.

PREVIOUS STATE Claude in Chrome was pre-GA with a more explicit approval posture.	NEW STATE Anthropic made Claude in Chrome generally available on paid plans. Anthropic says its safety classifier evaluates actions while Claude can perform more browser actions without individual approval.
WHO IS AFFECTED Paid Claude users working in authenticated browser sessions.	DECISION EFFECT Use bounded permissions and prompt-injection awareness; do not treat vendor safety controls as guarantees.

SOURCES Anthropic announcement: <https://claude.com/blog/claude-in-chrome-generally-available>

AUG-2026-016 | SHUTDOWN | EFFECTIVE 2026-08-30 | IMPACT HIGH

The official DALL·E GPT retired from ChatGPT

The official DALL·E GPT reached retirement at the end of August. OpenAI's current Images guidance identifies ChatGPT Images as the successor path. The change applies to the official DALL·E GPT. User-created GPTs with image generation enabled remain a separate supported path.

PREVIOUS STATE The official DALL·E GPT remained available during the published retirement window.	NEW STATE OpenAI's current Images FAQ says the official DALL·E GPT has been retired. ChatGPT Images is the successor surface; user-created GPTs with image generation enabled remain supported.
WHO IS AFFECTED Users relying on the official DALL·E GPT and documentation pointing to it.	DECISION EFFECT Move workflows and internal guidance to ChatGPT Images.

SOURCES OpenAI Help Center: <https://help.openai.com/en/articles/6825453> | OpenAI Help Center: <https://help.openai.com/en/articles/11084440-chatgpt-images>

AUG-2026-017 | PRICE CHANGE | EFFECTIVE 2026-08-10 | IMPACT HIGH

Anthropic cancelled Claude Sonnet 5's planned September price increase

Claude Sonnet 5 launched with US\$2 input and US\$10 output per million tokens through August 31, with a published move to US\$3 input and US\$15 output scheduled for September 1. On August 10, Anthropic changed that future price state. The lower US\$2 / US\$10 rate became permanent and the scheduled increase was cancelled. This is a real price event even though the transaction users experienced in August was the removal of a future increase rather than a same-day bill reduction.

PREVIOUS STATE US\$2 input / US\$10 output per million tokens was introductory through August 31, with US\$3 / US\$15 scheduled from September 1.	NEW STATE Anthropic made US\$2 / US\$10 permanent and cancelled the scheduled US\$3 / US\$15 rate.
WHO IS AFFECTED Claude API customers and teams forecasting Sonnet 5 production cost.	DECISION EFFECT Remove the planned September increase from forecasts and compare Sonnet 5 at the permanent rate.

SOURCES anthropic.com source: <https://www.anthropic.com/news/claude-sonnet-5>

AUG-2026-018 | CAPABILITY CHANGE | EFFECTIVE 2026-08-12 | IMPACT HIGH

Agent Plugins 1.0 gave reusable agent extensions a cross-client packaging path

The Agent Plugins 1.0 standard packages agent skills and MCP servers into one installable plugin. GitHub announced support across VS Code, Copilot CLI, the Copilot SDK, and the Copilot app. The material change is not another isolated extension format. It creates a path for reusable agent capabilities to be packaged once and delivered across multiple compatible clients.

PREVIOUS STATE Skills and MCP servers did not have the announced Agent Plugins 1.0 package plus compatible-client support.	NEW STATE Agent Plugins 1.0 packages agent skills and MCP servers into one installable plugin; GitHub announced support in VS Code, Copilot CLI, Copilot SDK and the Copilot app.
WHO IS AFFECTED Developers maintaining reusable agent extensions and organizations governing compatible clients.	DECISION EFFECT Evaluate whether one plugin package can replace vendor-specific duplicated agent extension work.

SOURCES GitHub announcement: <https://github.blog/changelog/2026-08-12-agent-plugins-1-0-in-vs-code-copilot-cli-and-the-copilot-app/>

AUG-2026-019 | SHUTDOWN | EFFECTIVE 2026-08-17 | IMPACT CRITICAL HIGH

Imagen 4 production API endpoints shut down

Google listed August 17 as the shutdown date for the Imagen 4 standard, ultra, and fast production endpoints in the Gemini API. This is a classic production dependency event. A model can remain historically important while its exact endpoint becomes unusable.

PREVIOUS STATE Imagen 4 standard, ultra and fast production endpoints were callable through the Gemini API.	NEW STATE Google lists all three Imagen 4 production endpoints with an August 17 shutdown date and directs developers to newer Gemini image-generation endpoints.
WHO IS AFFECTED Developers and applications using Imagen 4 API endpoints.	DECISION EFFECT Migrate endpoint names and response handling before retired calls break production.

SOURCES Google AI documentation: <https://ai.google.dev/gemini-api/docs/deprecations> | Google AI documentation: <https://ai.google.dev/gemini-api/docs/changelog>

AUG-2026-020 | CAPABILITY CHANGE | EFFECTIVE 2026-08-25 | IMPACT HIGH

Claude introduced one memory across chat and Cowork

Anthropic introduced a unified memory that can carry context across Claude chat and Cowork. Users can inspect remembered topics and edit or delete them. Anthropic also says sensitive-topic memory is off by default unless a user enables it. This is more than convenience. Shared persistent context changes how organizations should think about workspace continuity, data governance, and the boundaries between conversational and task-execution surfaces.

PREVIOUS STATE Cowork did not share the same persistent memory state as Claude chat.	NEW STATE Anthropic introduced one memory across chat and Cowork with user-visible topic controls to inspect, edit or delete remembered information. Anthropic says sensitive topics remain off by default unless enabled.
WHO IS AFFECTED Claude users moving between chat and Cowork and teams governing persistent context.	DECISION EFFECT Treat memory settings as part of workspace governance and assume enabled memory can persist context across chat and Cowork.

SOURCES Anthropic announcement: <https://claude.com/blog/claudes-memory-works-everywhere-and-you-decide-whats-in-it>

AUG-2026-021 | SHUTDOWN | EFFECTIVE 2026-08-26 | IMPACT CRITICAL HIGH

OpenAI Assistants API reached its shutdown date

OpenAI's legacy Assistants API reached its published August 26 shutdown date after a one-year deprecation window. OpenAI directs developers toward the Responses API plus Conversations API as the replacement path. This is one of August's highest-consequence changes because production integrations can fail outright when a retired API is still embedded in an application.

PREVIOUS STATE Developers could continue using Assistants API during its one-year deprecation window.	NEW STATE OpenAI's deprecations page lists August 26, 2026 as the shutdown date and directs developers to Responses API plus Conversations API.
WHO IS AFFECTED Developers and products built on Assistants API.	DECISION EFFECT Complete migration to Responses plus Conversations and remove production reliance on the retired API.

SOURCES [developers.openai.com](https://developers.openai.com/api/docs/deprecations) source: <https://developers.openai.com/api/docs/deprecations>

AUG-2026-022 | CAPABILITY CHANGE | EFFECTIVE 2026-08-26 | IMPACT HIGH

Claude Cowork gained an isolated built-in browser

Anthropic added a browser directly inside the Cowork desktop app. It runs separately from the user's own browser session and can navigate sites, read pages, click, type, and fill forms without requiring the Claude in Chrome extension. That creates a different trust model from Claude in Chrome. One surface operates inside the user's authenticated browser session. The other uses an isolated browser inside Cowork.

PREVIOUS STATE Cowork web interaction relied on other browser access paths such as Claude in Chrome.	NEW STATE Anthropic added a browser inside Cowork that runs separately from the user's own browser and can navigate, read, click, type and fill forms without the Chrome extension.
WHO IS AFFECTED Cowork users and organizations choosing between isolated-browser and authenticated-browser agent access.	DECISION EFFECT Prefer the isolated browser when the workflow does not need the user's existing authenticated browser session.

SOURCES Anthropic announcement: <https://claude.com/blog/cowork-built-in-browser>

Pricing and billing changes

August does not support a broad claim that AI pricing moved consistently up or down.

The month contains three different pricing patterns.

First, ChatGPT for PowerPoint moved beyond a temporary free-use period into published paid or flexible usage treatment. That is a plan and billing transition, not a clean continuing-SKU price increase.

Second, Anthropic cancelled a planned price increase for Claude Sonnet 5. The US\$2 input / US\$10 output per million token rate, originally presented as temporary through August, became permanent instead of rising to US\$3 / US\$15 on September 1.

Third, Gemini 3.7 Flash launched at a temporary introductory rate equal to half Gemini 3.6 Flash's July launch rate, and Google applied the same lower rate to 3.6 Flash. That pricing advantage is explicitly time-limited through December 31, 2026.

Grok Voice Think Fast 2.0 also carried a published US\$0.08-per-minute rate, but the material August event was the alias migration rather than a clean repricing of the same continuing product state.

What the pricing evidence means

1. "Cheaper" depends on which layer moved. A subscription can stay unchanged while one agentic feature becomes metered, or a model can become cheaper while the surrounding workflow becomes more expensive to run.
2. Avoid treating promotional rates as structural economics. Gemini's current Flash pricing is materially attractive, but the published expiry date belongs in every forecast.
3. Cancelled future increases matter. Sonnet 5's August price event changed forward budgets even though users did not receive a same-day invoice cut from a higher live rate.
4. Cost per accepted outcome remains the better comparison. Token rates, audio minutes, agent steps, retries, tool calls, human correction, and review time all contribute to completed-work economics.
5. No market-wide price index should be inferred from this month. August gives useful vendor-level evidence, not a claim that the AI market as a whole became cheaper or more expensive.

Category analysis

August's volatility spread outward from the model itself.

Capability changes were the largest primary class at 27.3%, but five shutdowns represented another 22.7% of the dataset. Launches and access changes each accounted for 18.2%. That makes August less launch-dominated than July and more dependent on the operating layer around AI products.

1. AI authority is expanding across more surfaces

Grok Bot introduced persistent delegated work. Claude's production-agent stack moved further into GA. ChatGPT gained event-triggered tasks. Claude in Chrome reached general availability. Cowork gained its own isolated browser. Unified Claude memory connected conversational and task-execution surfaces.

The relevant question is increasingly not only what a model knows. It is what environment the system can enter, what state it can remember, what event can wake it up, what it can change, and which action still requires a person.

2. Browser architecture is becoming a trust-boundary decision

Claude in Chrome and Cowork's built-in browser are useful because they are not the same product architecture.

Claude in Chrome acts inside the user's authenticated browser environment. Cowork's built-in browser is separate from that environment. Both can be useful, but they create different exposure to logged-in accounts, saved sessions, credentials, and malicious instructions embedded in webpages.

Teams evaluating browser agents should therefore ask which browser context the agent actually controls, not simply whether the product "can browse."

3. Reusable agent infrastructure is becoming a product layer

Agent Plugins 1.0, Claude Skills API, event-triggered ChatGPT tasks, and persistent Grok Bot execution all point toward reusable agent systems rather than one-off prompts.

That raises the value of versioned procedures, permission templates, auditable tools, and portable extension formats. The durable asset may increasingly be the workflow definition around the model rather than the model itself.

4. Model competition remains fast enough to disrupt routing assumptions

Grok moved from 4.5 to 4.6. Gemini moved from 3.6 Flash to 3.7 Flash in 23 days. Kimi K3 expanded from its July launch into GitHub Copilot during August.

A model decision made once per quarter can therefore become stale quickly. Production teams need explicit retest triggers and should separate model evaluation from the surface used to deliver it.

5. Shutdown risk is now an operating concern, not an edge case

Spark, Atlas, Imagen 4 API, Assistants API, and the official DALL·E GPT all hit shutdown states in the same month.

These were different types of products: a workbench, a browser surface, production image endpoints, a developer API, and a ChatGPT GPT. The common lesson is that operational dependency risk exists at every layer.

A workflow can remain conceptually useful while the exact endpoint, workbench, model selector, or interface that carries it disappears.

6. Governance is becoming part of product behavior

GitHub's global Copilot model policy is the cleanest August example. When new models can inherit an administrator's global default, policy becomes part of how the product behaves.

Claude memory controls, browser approval models, and ChatGPT event-trigger approvals reinforce the same direction. Governance is no longer only an after-the-fact enterprise wrapper. It increasingly shapes what the AI system can do at runtime.

7. July watchlists are proving their value

Two material August changes were already visible from July: the Grok Voice alias migration and GitHub Copilot's model-policy transition.

That validates a core reporting rule. A future-effective vendor announcement should remain on the watchlist until the actual user consequence arrives. The announcement date and effective date are not interchangeable.

Stack implications

1. Treat agent adoption as a permissions design exercise

Before giving an AI system broader authority, identify what it can read, what it can change, which accounts it can enter, which actions require approval, how much it can spend, and what happens if it encounters a malicious instruction inside something it reads.

The practical default is straightforward: delegate reversible work aggressively and irreversible authority reluctantly.

2. Choose browser context deliberately

An isolated browser and an authenticated personal browser solve different problems.

Use isolated browser execution when the workflow can run without broad access to existing sessions. Use authenticated browser access only when the task genuinely requires it. The browser context itself should be treated as a permission boundary.

3. Pin production dependencies when consistency matters

The Grok Voice alias migration shows why "latest" is a product choice, not a neutral default.

Record which models, aliases, endpoints, and delivery surfaces are pinned and which are intentionally allowed to float. Automatic upgrades are useful when experimentation matters more than behavioral stability.

4. Maintain an exit path for every operationally important AI surface

Five shutdowns in one month are enough evidence to make migration planning a standard operating requirement.

Export important data. Keep source files outside proprietary workbenches where practical. Document replacement paths. Avoid creating a workflow that can only be reconstructed inside one vendor surface.

5. Track model retirement separately from API retirement

OpenAI o3 is a useful example. A model can leave a consumer surface without the same announcement changing API availability.

Do not treat one product-surface retirement as proof that every access path for the underlying model has disappeared.

6. Measure completed-work economics, not only model rates

The cheapest advertised token rate does not automatically produce the cheapest outcome.

Track retries, agent duration, tool calls, human correction, review time, and whether the final output is accepted. Temporary pricing should also carry an expiry date in internal forecasts.

7. Review global governance before model choice expands

GitHub's Copilot policy means future model access can inherit a global setting.

Organizations should know whether new models appear automatically, require explicit enablement, or inherit another policy state before users discover the answer through production behavior.

8. Treat memory as part of workspace governance

Persistent memory is operational context.

Teams using shared or persistent work surfaces should understand what can be remembered, who can inspect or delete it, whether sensitive categories are enabled, and how memory crosses between chat and task-execution environments.

9. Separate the model from the surface that delivers it

Kimi K3 inside Copilot and Grok 4.6 across managed platforms demonstrate why distribution needs its own evaluation.

A model may be attractive while one delivery surface has the wrong billing, data, governance, or workflow fit. Test the full route, not only the model name.

What to watch in September

The September watchlist is not included in August event totals.

- GitHub Copilot model deprecations effective September 1. These belong in September if the scheduled state transitions complete and clear the materiality threshold.
- Completion of GitHub's staged global-model-policy enforcement. August records the start of enforcement. September should only count a separate event if a distinct material consequence emerges.
- Post-shutdown migration outcomes. Watch Spark, Atlas, Imagen 4 API, Assistants API, and DALL·E GPT successor paths for materially different breakage, replacement guidance, or access changes.
- Claude browser trust boundaries. Claude in Chrome and Cowork's isolated browser now create different operating models that may diverge further in permissions, authentication, and safety controls.
- Persistent memory governance. Watch for administrator controls, enterprise defaults, export behavior, and clearer cross-workspace memory boundaries.
- Grok Bot reliability and economics. Persistent-agent launches become materially more useful once there is stable evidence about completed jobs, failure recovery, cost, and permission handling.
- Gemini 3.7 Flash introductory pricing. Keep the December 31 expiry visible in routing and budget decisions.
- Any August-announced change with a September effective date. Do not pull it into August simply because the announcement happened earlier.

Methodology

Primary event taxonomy

Choosely assigns one stable primary class to each event:

- Model or product launch
- Availability or access change
- Price change
- Plan change
- Capability change
- Policy change
- Deprecation
- Shutdown

An event can carry secondary pricing, billing, capability, distribution, or migration consequences, but it is counted once.

Source standard

Material claims prioritize first-party vendor announcements, pricing pages, release notes, changelogs, support pages, policy pages, and product documentation.

Choosely AI Radar can interpret a verified change but does not replace the first-party source. Third-party reporting can identify a candidate or explain a source conflict, but it does not independently qualify an event when first-party evidence should exist.

Vendor performance, safety, and capability claims are attributed to the vendor unless Choosely has run controlled independent testing.

Timing rule

An event qualifies for August when the release, effective date, or material user consequence occurred from August 1 through August 31, 2026.

An earlier announcement can qualify when the state transition takes effect in August. This is why Grok Voice, ChatGPT for PowerPoint, ChatGPT Atlas, o3 in ChatGPT, GitHub Copilot policy enforcement, Imagen 4 API, Assistants API, and the official DALL·E GPT belong in August despite earlier announcement dates.

An August announcement with a material consequence scheduled for September normally remains on the September watchlist.

Distribution rule

Distribution that occurs in the same reporting period as a parent model or product launch is normally folded into the parent event unless it creates an independently material billing, migration, governance, or access consequence.

Distribution that occurs in a later reporting period is assessed as a follow-on event on its own materiality.

Kimi K3's August Copilot rollout therefore qualifies as a follow-on to its July model launch. Grok 4.6's same-month third-party distributions remain inside the August Grok 4.6 launch record.

Materiality rule

A change enters the public dataset only when it affects a user, buyer, developer, administrator, or workflow decision strongly enough to justify durable reporting.

Lower-consequence interface changes, ambiguous rollouts, scanner noise, temporary retrieval failures, and facts that cannot be tied to a defensible August state transition are excluded from the headline count.

Vendor representation is not an inclusion or exclusion criterion.

Limitations

The monitored catalog is a defined Choosely population, not the entire global AI-tool market.

Regional rollouts, negotiated enterprise arrangements, private betas, unpublished changes, and products outside Choosely's monitored scope may not be represented.

A product can have several distinct surfaces under one catalog slug. August's public dataset therefore adds a `product_or_surface` field so model, API, browser, workbench, and plugin events do not collapse into misleading tool-level aggregation.

Vendor capability, performance, efficiency, and safety claims are attributed to vendor sources unless Choosely has run controlled independent testing.

A launch price is not necessarily a continuing-SKU price change. A temporary introductory price should not be treated as permanent. A third-party distribution event is not automatically a new model launch. A product retirement may move functionality into another surface rather than eliminating the underlying capability.

Some vendor support and release-note pages are mutable. Choosely records the verification date and uses the publication evidence available at the reporting cutoff.

Event counts depend on the reporting period, verification cutoff, definitions, materiality threshold, and evidence available at that cutoff.

No broad market-wide pricing trend is inferred from August's one primary price event, one primary plan event, and secondary pricing consequences.

Downloadable data and reuse terms

Canonical HTML: <https://choosely.ai/reports/ai-tool-changes/2026-08>

PDF: </reports/ai-tool-changes/2026-08/report.pdf>

CSV: </reports/ai-tool-changes/2026-08/data.csv>

The HTML report is the canonical publication. The PDF is a convenience copy, and the CSV contains the 22 public event-level records used in the report.

The August CSV extends the existing public schema with `product_or_surface` while retaining `canonical_slug` for catalog joins.

© 2026 Choosely. All rights reserved.

The dataset may be quoted and referenced with attribution. Redistribution, republication, or commercial reuse of the complete dataset requires written permission from Choosely.

The public CSV excludes detection timestamps, monitoring cadence, scanner logic, source inventories, scoring rules, rejected candidates, and unpublished review material.

Cite this report

Recommended citation

Choosely Editorial. 22 Verified AI-Tool Changes That Mattered in August 2026. Choosely, September 2, 2026.
<https://choosely.ai/reports/ai-tool-changes/2026-08>

Corrections and version history

Version 1.0 - September 2, 2026. Initial publication containing 22 verified material changes. Material corrections will be dated and reflected across the HTML, PDF, and CSV.

Final August verdict

August 2026 expanded the operational authority of AI assistants while making the surrounding access, billing, governance, and retirement layers harder to treat as stable infrastructure.

The model race remained fast. Grok 4.6 arrived. Gemini 3.7 Flash followed 3.6 Flash after only 23 days. Kimi K3 moved into Copilot. ChatGPT changed how users route reasoning through GPT-5.6.

But the more consequential movement happened around the models.

Persistent agents gained managed computers. Production agent building blocks reached broader GA status. Browser agents gained more authority. Cowork gained its own isolated browser. ChatGPT tasks became event-driven. Claude memory crossed between chat and Cowork. Agent Plugins 1.0 created a more reusable packaging layer for agent capabilities. Enterprise policy began determining how future Copilot models inherit access.

Then the durability warning arrived in parallel.

GitHub Spark's workbench retired. Atlas reached its published shutdown date. Imagen 4 production endpoints shut down. Assistants API reached retirement. The official DALL·E GPT left ChatGPT. o3 reached its ChatGPT retirement date.

The useful response is not to resist AI becoming more capable. It is to operate that capability deliberately.

Build bounded workflows. Choose browser context intentionally. Keep consequential actions visible. Pin dependencies that need stable behavior. Track cost per accepted outcome. Put expiry dates on promotional economics. Export important state. Know which policy determines model access. Treat memory as governance. Maintain a replacement path before an AI surface becomes too important to lose.

The next competitive advantage is not simply choosing the smartest model.

It is building a stack that can absorb rapid AI change without losing control of cost, access, state, or workflow continuity.

Primary source register

- 1 <https://github.blog/changelog/2026-08-04-upcoming-deprecation-of-github-spark-on-github-com/>
- 2 <https://docs.x.ai/developers/release-notes>
- 3 <https://x.ai/news/grok-voice-think-fast-2>
- 4 <https://openai.com/index/improving-gpt-5-6-sol-in-chatgpt/>
- 5 <https://help.openai.com/en/articles/20001354-gpt-5-6>
- 6 <https://blog.adobe.com/en/publish/2026/08/06/introducing-adobe-chatgpt-create-edit-get-work-done-all-in-chatgpt>
- 7 <https://github.blog/changelog/2026-08-06-kimi-k3-is-now-available-in-github-copilot/>
- 8 <https://help.openai.com/en/articles/20001415>
- 9 <https://help.openai.com/en/articles/10128477>
- 10 <https://help.openai.com/en/articles/20001371-evolving-atlas-into-chatgpt-for-browser-based-agentic-work>
- 11 <https://x.ai/news/introducing-grok-bot>
- 12 <https://x.ai/news/grok-4-6>
- 13 <https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/>
- 14 <https://ai.google.dev/gemini-api/docs/pricing>
- 15 <https://claude.com/blog/computer-use-skills-api-files-api>
- 16 <https://help.openai.com/en/articles/10291617>
- 17 <https://help.openai.com/en/articles/6825453>
- 18 <https://help.openai.com/en/articles/9624314-model-release-notes>
- 19 <https://github.blog/changelog/2026-07-29-default-model-enablement-for-copilot-business-and-enterprise/>
- 20 <https://github.blog/changelog/2026-08-26-global-model-policy-generally-available/>
- 21 <https://claude.com/blog/claude-in-chrome-generally-available>
- 22 <https://help.openai.com/en/articles/11084440-chatgpt-images>
- 23 <https://www.anthropic.com/news/claude-sonnet-5>
- 24 <https://github.blog/changelog/2026-08-12-agent-plugins-1-0-in-vs-code-copilot-cli-and-the-copilot-app/>
- 25 <https://ai.google.dev/gemini-api/docs/deprecations>
- 26 <https://ai.google.dev/gemini-api/docs/changelog>
- 27 <https://claude.com/blog/claude-memory-works-everywhere-and-you-decide-whats-in-it>
- 28 <https://developers.openai.com/api/docs/deprecations>
- 29 <https://claude.com/blog/cowork-built-in-browser>