CHOOSELY.AI

MONTHLY AI-TOOL CHANGES REPORT

8 Verified AI-Tool Changes That Mattered in July 2026

Evidence-based analysis of the launches, access changes, pricing movement, and shutdown that materially affected AI-tool users during July.

VERSION 1.0

July 1 to July 31, 2026 | Verified through August 1, 2026 | Choosely Editorial | Published August 1, 2026

8	319	87.5%	80%
material changes verified	tools in monitored catalog	of verified changes were model-related	largest verified price cut

The HTML report is the canonical record. Material corrections are dated and reflected across the HTML, PDF, and CSV.

Key findings

8	319	87.5%	80%
material changes verified	tools in the monitored catalog	of verified changes were model-related	largest verified price cut

Most monitored tools recorded no public change that met both the report's materiality and evidence standards during July.

July accelerated the shift from model scarcity to model competition. Major providers expanded access, launched more specialized tiers, moved deeper into coding and agent workflows, and sharpened price-performance competition. Choosely verified eight changes significant enough to affect user, buyer, developer, or administrator decisions.

The strongest movement came from OpenAI's GPT-5.6 launch and rapid repricing, Google's expanded Flash family, Kimi K3's service and open-weight release, Grok 4.5's expansion into API and work surfaces, and Anthropic's Opus 5 launch. Fable 5 returned after an access suspension, while GitHub Models demonstrated the opposite risk by completing a full retirement path.

The practical conclusion is not to rebuild a stack around every announcement. It is to update routing assumptions, compare completed-work economics, test the relevant alternatives, and remove access dependencies that no longer have a stable future.

July at a glance

PRIMARY CHANGE CLASS	CHANGES	SHARE
Model or product launch	5	62.5%
Availability or access change	1	12.5%
Price change	1	12.5%
Shutdown	1	12.5%

One primary class is assigned to each change for headline counting. Individual changes may also carry secondary pricing, capability, billing, or distribution consequences.

Coverage statement

This report covers material AI-tool and model changes that were effective or publicly released from July 1 through July 31, 2026 and verified by August 1, 2026. At the cutoff, Choosely's monitored catalog contained 319 AI tools.

A change qualified only when Choosely could establish a practical consequence, reliable first-party evidence, a relevant July date, the affected users or workflows, and a defensible new state. A prior state is published only when reliable evidence supports it.

This is not a census of every AI product worldwide. It does not cover every vendor, region, private beta, enterprise contract, or unpublished change. The figures describe Choosely's verified July dataset and should be cited with the reporting period, verification cutoff, and monitored-catalog scope intact.

The eight verified changes

JUL-2026-001 AVAILABILITY OR ACCESS CHANGE JULY 1, 2026 IMPACT: HIGH

Claude Fable 5 returned to global availability

Anthropic announced the restoration on June 30, with Fable 5 access effective July 1. Mythos 5 also returned for a restricted set of approved US organizations. The operational detail matters: the restored model carried revised cyber safeguards, fallback behavior, and temporary usage terms rather than simply returning unchanged.

PREVIOUS STATE

Access to Fable 5 and Mythos 5 had been suspended for all customers on June 12 after a US export-control directive took immediate effect.

NEW STATE

Fable 5 returned globally on July 1 across Claude Platform, Claude.ai, Claude Code, and Claude Cowork. Eligible plans received up to 50% of weekly usage limits through July 7 before access moved to usage credits. Blocked Fable 5 requests can route to Opus 4.8, and Anthropic warned of additional false positives on routine coding and debugging requests.

WHO IS AFFECTED

Claude users, Claude Code and Cowork users, API developers, and teams that had removed Fable 5 from production routing during the suspension.

DECISION EFFECT

Reassess Fable 5 for high-complexity work, but test benign coding and security-adjacent requests for false positives and account for usage-credit consumption after the temporary allowance ended.

Sources: Anthropic redeployment announcement Anthropic suspension statement
Choosely: Tool page Read Choosely's Fable 5 provider-risk analysis

JUL-2026-002 MODEL OR PRODUCT LAUNCH JULY 8, 2026 IMPACT: HIGH

Grok 4.5 entered API, coding, and office workflows

The official source trail contains three relevant dates, so the report records them rather than forcing one disputed launch date. The material change was broader than another chat release: Grok 4.5 was positioned for coding, agentic tasks, engineering, and office-document workflows, while long-context pricing introduced a meaningful cost threshold.

PREVIOUS STATE

Grok 4.5 was not available through the SpaceXAI API or its later distribution surfaces.

NEW STATE

SpaceXAI release notes record API availability on July 8 at $2 input and $6 output per million tokens. A broader launch announcement followed on July 16 across Grok Build, Cursor, and the API, with EU API-console availability on July 17. Requests using at least 200,000 context tokens are priced at $4 input and $12 output per million tokens.

WHO IS AFFECTED

API developers, Cursor users, Grok Build users, engineering teams, and people applying AI to spreadsheets, documents, and long-context workflows.

DECISION EFFECT

Evaluate Grok 4.5 on the actual coding or document workflow, and include the higher long-context rates when requests may exceed 200,000 tokens.

Sources: SpaceXAI release notes SpaceXAI launch announcement SpaceXAI pricing
Choosely: Tool page Read Choosely's Grok 4.5 analysis

JUL-2026-003 MODEL OR PRODUCT LAUNCH JULY 9, 2026 IMPACT: HIGH

OpenAI moved GPT-5.6 to general availability

The launch created a clearer routing ladder between frontier work, everyday production tasks, and lower-cost high-volume execution. It also introduced a new billing mechanic for prompt-cache writes, which means launch economics cannot be summarized by input and output token rates alone.

PREVIOUS STATE

GPT-5.6 Sol, Terra, and Luna were available only through a limited preview for selected partners and organizations.

NEW STATE

The family became generally available across ChatGPT, Codex, and the OpenAI API with a staged global rollout. Launch rates were $5/$30 for Sol, $2.50/$15 for Terra, and $1/$6 for Luna per million input/output tokens. GPT-5.6 also introduced cache-write billing at 1.25 times uncached input rates while retaining a 90% discount for cache reads.

WHO IS AFFECTED

ChatGPT paid users, ChatGPT Work users, Codex users, API developers, and teams routing tasks across model tiers.

DECISION EFFECT

Re-evaluate OpenAI model routing by task complexity, latency, completed-work quality, cache behavior, and total cost. Use the later July 30 Terra and Luna rates for current comparisons.

Sources: OpenAI GPT-5.6 announcement
Choosely: Tool page Compare GPT-5.6, Grok 4.5, and Claude Fable 5

JUL-2026-004 MODEL OR PRODUCT LAUNCH JULY 16, 2026 IMPACT: HIGH

Kimi K3 made open-weight, long-context evaluation harder to ignore

The correct description is open-weight, not open-source in the OSI sense. The weights and code are governed by the Kimi K3 License, which includes commercial conditions for certain model-as-a-service businesses and very large products. The timing also matters: service access began before the full weights were published.

PREVIOUS STATE

Kimi K3 was unavailable across Kimi products and the API.

NEW STATE

Kimi K3 launched on July 16 across Kimi products and the API with native multimodality, a one-million-token context window, and pricing of $0.30 cached input, $3 uncached input, and $15 output per million tokens. Full weights followed on July 27 under the bespoke Kimi K3 License.

WHO IS AFFECTED

Developers, coding-agent users, open-weight adopters, and teams handling large codebases, long documents, or repeated cached context.

DECISION EFFECT

Add Kimi K3 to evaluations where deployment control, open weights, cache reuse, or long context have practical value. Review the license before commercial deployment and keep output cost and task reliability in the comparison.

Sources: Kimi product timeline Kimi K3 repository Kimi K3 License Kimi K3 pricing
Choosely: Tool page Read Choosely's Kimi K3 analysis

JUL-2026-005 MODEL OR PRODUCT LAUNCH JULY 21, 2026 IMPACT: HIGH

Google expanded the Flash family for production agents

This was a three-model announcement, although only 3.6 Flash and 3.5 Flash-Lite were broadly available at launch. The key operator signal was routing: a higher-value workhorse, a throughput-focused tier, and a restricted cyber model, with lower output cost for 3.6 Flash but higher generation-to-generation pricing for Flash-Lite than 3.1 Flash-Lite.

PREVIOUS STATE

Gemini 3.5 Flash and Gemini 3.1 Flash-Lite were the prior production baselines for these roles.

NEW STATE

Google introduced Gemini 3.6 Flash at $1.50 input and $7.50 output per million tokens, Gemini 3.5 Flash-Lite at $0.30/$2.50, and the restricted Gemini 3.5 Flash Cyber model for a trusted-partner CodeMender pilot. Google reported that 3.6 Flash used 17% fewer output tokens than 3.5 Flash on the Artificial Analysis Index while lowering output price from $9 to $7.50.

WHO IS AFFECTED

Developers running production agents, coding workflows, multimodal document processing, high-volume classification, and computer-use tasks.

DECISION EFFECT

Re-test Google routing using completion quality, latency, retries, tool calls, and total task cost. Treat the 3.5 Flash-Lite launch price as a new-generation rate, not a repricing of the continuing 3.1 SKU.

Sources: Google Flash family announcement Gemini API pricing Gemini 3.1 Flash-Lite launch pricing
Choosely: Tool page Read Choosely's Gemini Flash analysis

JUL-2026-006 MODEL OR PRODUCT LAUNCH JULY 24, 2026 IMPACT: HIGH

Anthropic launched Claude Opus 5 as a lower-cost frontier workhorse

The launch is the primary event. Copilot distribution is recorded as a consequence rather than counted separately, which keeps the report consistent with other model integrations. GitHub bills Opus 5 at provider API list price under usage-based billing and warns that enhanced cyber safeguards may block some security-adjacent requests.

PREVIOUS STATE

Claude Opus 4.8 was Anthropic's current Opus model for complex agentic coding and enterprise work.

NEW STATE

Claude Opus 5 launched at $5 input and $25 output per million tokens, matching Opus 4.8 and sitting at half Fable 5 list pricing. It supports a one-million-token context window, up to 128,000 output tokens, and a May 2026 reliable knowledge cutoff. It became available through the Claude API and major cloud platforms, and GitHub began a gradual Copilot rollout on the same day.

WHO IS AFFECTED

Claude API developers, enterprise teams, complex coding users, GitHub Copilot users, and administrators controlling model access and usage-based billing.

DECISION EFFECT

Test Opus 5 against Opus 4.8 and Fable 5 on complex, long-running work. In Copilot, review administrator enablement, provider-list billing, and cyber safeguard behavior before broad rollout.

Sources: Claude Platform release notes Claude models overview GitHub Copilot availability
Choosely: Tool page Read Choosely's Claude Opus 5 analysis

JUL-2026-007 SHUTDOWN JULY 30, 2026 IMPACT: CRITICAL HIGH

GitHub Models completed its retirement path

The July event was not the original decision to retire the service. It was the publication of the final timeline, two user-impacting brownouts, and the effective shutdown. That distinction matters because it shows both the earlier warning and the July migration deadline.

PREVIOUS STATE

GitHub closed Models to new customers on June 16 while existing customers retained the playground, API, model catalog, and BYOK access.

NEW STATE

On July 1, GitHub set July 30 as the full retirement date, scheduled brownouts on July 16 and July 23, and stated that the playground, model catalog, inference API, BYOK endpoints, and related UI would become unavailable to all customers.

WHO IS AFFECTED

Developers using GitHub Models for experimentation, model discovery, inference, or bring-your-own-key workflows.

DECISION EFFECT

Remove GitHub Models dependencies, confirm replacement endpoints and authentication, and compare Microsoft Foundry, GitHub Copilot, or another provider path for the specific workflow.

Sources: GitHub June retirement step GitHub final retirement timeline
Choosely: Read the July 30 Change Brief

JUL-2026-008 PRICE CHANGE JULY 30, 2026 IMPACT: HIGH

OpenAI cut Terra and Luna prices and replaced Priority Processing

This is the report's only continuing-SKU repricing event. OpenAI also reduced how Terra and Luna usage counts against paid Codex and ChatGPT Work subscriptions, while subscription prices and quota budgets remained unchanged.

PREVIOUS STATE

Terra cost $2.50 input and $15 output per million tokens. Luna cost $1 input and $6 output. Priority Processing was the premium API service tier.

NEW STATE

Terra fell to $2 input and $12 output, a 20% reduction. Luna fell to $0.20 input and $1.20 output, an 80% reduction. Fast mode replaced Priority Processing for GPT-5.6 Sol, offering up to 2.5 times Standard speed at twice the price while remaining backward compatible with priority-tagged requests.

WHO IS AFFECTED

API developers, Codex users, ChatGPT Work users, and teams routing high-volume or latency-sensitive workloads.

DECISION EFFECT

Update every GPT-5.6 cost model based on launch rates. Retest Luna for background agent loops, Terra for everyday production work, and Fast mode only where lower latency justifies the 2x Sol rate.

Sources: OpenAI July 30 pricing update
Choosely: Tool page Read Choosely's GPT-5.6 price-cut analysis

Pricing and billing changes

What counts as a price event

A price event is a change to the published rate of a continuing model, product, or plan. Launch pricing and generation-to-generation differences are recorded inside launch events but are not counted as repricing events.

Continuing-SKU repricing

MODEL	OLD INPUT	NEW INPUT	OLD OUTPUT	NEW OUTPUT	CHANGE
GPT-5.6 Terra	$2.50	$2.00	$15.00	$12.00	-20%
GPT-5.6 Luna	$1.00	$0.20	$6.00	$1.20	-80%

At a fixed monthly workload of 100 million input tokens and 100 million output tokens, Terra falls from $1,750 to $1,400 and Luna from $700 to $140. Actual completed-task economics depend on caching, context length, retries, tool calls, and correction time.

Generation-to-generation and billing context

CHANGE	OPERATOR INTERPRETATION
Gemini 3.6 Flash output pricing fell from Gemini 3.5 Flash's $9 to $7.50 while input remained $1.50. Google also reported 17% fewer output tokens on the Artificial Analysis Index.	A lower generation-level output rate and lower token use can improve task economics, but this is launch pricing for a new model rather than a repricing of the continuing 3.5 SKU.
Gemini 3.5 Flash-Lite launched at $0.30 input and $2.50 output, above 3.1 Flash-Lite's $0.25 and $1.50.	The new model costs more per token than the earlier generation. The decision depends on speed, quality, retries, and completed-work cost.
Claude Opus 5 launched at $5/$25, matching Opus 4.8 and sitting at half Fable 5's $10/$50 list price.	Opus 5 is a new lower-cost frontier option, not a Fable 5 price cut.
GPT-5.6 introduced cache-write billing at 1.25x uncached input rates, and Fast mode later replaced Priority Processing for Sol at 2x Standard pricing.	Routing and latency choices now require attention to billing mechanics beyond headline input and output rates.

The July sample is too small to support a market-wide claim that AI-tool prices rose or fell.

Category analysis

Seven of eight verified changes were model launches, model access, or model pricing events. The remaining change was the shutdown of a developer model-access platform.

Model choice is becoming more granular. Providers increasingly offer families and tiers for different combinations of capability, speed, latency, context, and cost. One default model is less likely to suit an entire workflow.

Distribution surfaces matter, but they should not be double-counted. Opus 5 becoming available inside GitHub Copilot was operationally important, but it is recorded inside the parent launch. Third-party distribution becomes a separate event only when it creates a distinct material migration, billing, or access consequence.

Open weights are entering practical comparisons. Kimi K3 adds pressure in coding and long-context work, but its bespoke license and delayed weight release matter as much as the headline capability.

Access layers remain fragile. GitHub Models' retirement shows that a convenient platform can disappear even while underlying model choice expands.

Stack implications

1 Re-run model routing.

Separate frontier reasoning, everyday production work, background agent loops, coding, and long-context processing instead of applying one model to every step.

2 Update cost assumptions.

Any GPT-5.6 comparison using July 9 launch rates is already stale for Terra and Luna. New-model comparisons should also include cache writes, long-context thresholds, and premium speed tiers.

3 Compare completed-work economics.

A lower token rate can still cost more when a model needs extra tool calls, retries, longer outputs, or manual correction.

4 Use distribution as a consolidation test.

Opus 5 inside Copilot may reduce the need for a separate coding surface, but usage-based billing, administrator controls, and safeguard behavior still matter.

5 Audit access dependencies.

Record which workflows depend on a provider, alias, endpoint, playground, or third-party access layer, then identify replacements before retirement becomes urgent.

6 Review licenses before adopting open-weight models.

Kimi K3 may be attractive for deployment control and long context, but its commercial terms need to be part of the decision.

What to watch in August

The watchlist is not included in July's event total.

- Model aliases and endpoint migrations announced in July but effective in August
- Enterprise default-model policies that alter access without an individual user decision
- Whether Luna's lower price changes high-volume agent routing in production
- Whether Gemini 3.6 Flash lowers completed-task cost rather than only token cost
- Whether Kimi K3's open-weight release translates into reliable commercial adoption
- Migration outcomes after the GitHub Models retirement
- Verified free-tier or plan-limit changes that were incomplete at the July cutoff
- How premium speed tiers affect total cost for latency-sensitive workflows

Methodology

Primary event taxonomy

Choosely uses one stable primary class per event: model or product launch, availability or access change, price change, plan change, capability change, deprecation, shutdown, or policy change. A record can carry secondary effects, but it is counted once.

Source standard

Material claims prioritize first-party vendor announcements, pricing pages, release notes, changelogs, support pages, and product documentation. Third-party reporting can identify a candidate but does not independently qualify an event when first-party evidence should exist.

Timing rule

An event qualifies for July when the release, effective date, or material user consequence occurred in July. A June announcement can qualify when the change took effect in July, as with Fable 5. A July announcement that takes effect later normally belongs in the watchlist.

Distribution rule

Third-party model availability is normally folded into the parent launch unless it creates a separate material migration, billing, or access consequence. This prevents the same model release from being counted once for every downstream platform.

Exclusions

The report excludes redirects, blocked pages, temporary retrieval failures, non-material site changes, rumors, current facts that cannot be tied to July, and signals without sufficient evidence for the old state, new state, timing, or affected-user scope. One free-plan signal was excluded because those fields could not be fully established by the cutoff.

Limitations

- The monitored catalog is a defined Choosely population, not the entire global AI-tool market.
- Regional, negotiated enterprise, private beta, and unpublished changes may not be represented.
- Vendor performance claims are attributed to vendor sources and are not presented as controlled Choosely testing.
- Event counts depend on the reporting period, verification cutoff, definitions, and evidence available at the cutoff.
- No broad pricing trend is inferred from one continuing-SKU repricing event.

Corrections and version history

Material corrections will be dated and explained. Stable event IDs remain attached to records that are corrected or superseded. Factual correction requests can be sent to info@choosely.ai.

VERSION	DATE	CHANGE
1.0	August 1, 2026	Initial publication containing eight verified material changes.

Downloadable data and reuse terms

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

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. *8 Verified AI-Tool Changes That Mattered in July 2026*. Choosely, August 1, 2026. https://choosely.ai/reports/ai-tool-changes/2026-07

Dataset attribution

Choosely Editorial, *Choosely Verified AI-Tool Changes - July 2026*, covering July 1 to July 31, 2026 and verified through August 1, 2026.

Preserve the monitored-catalog scope and verification cutoff when quoting an aggregate statistic. Vendor-specific facts should continue to cite the relevant first-party source. Choosely is the source for the cross-market aggregation, classification, and interpretation.

Final July verdict

July 2026 was a model-market month.

New model families, wider access, and sharper price competition created the strongest movement. OpenAI's rapid GPT-5.6 repricing showed how quickly cost assumptions can become stale. Google's Flash launch, Anthropic's Opus 5 launch, Kimi's open-weight release, and Grok's expansion widened the production choice set.

At the same time, Fable 5's restoration showed that model access can be shaped by policy and safeguards, while GitHub Models' retirement showed that access infrastructure can disappear even when model choice expands.

The useful response is to review routing, update costs, test the most relevant alternatives, and remove dependencies that no longer have a stable future. Rebuilding the entire stack around every announcement would be a considerably less useful use of August.

Primary source register

1. https://ai.google.dev/gemini-api/docs/pricing
2. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-flash-lite/
3. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-6-flash-3-5-flash-lite-3-5-flash-cyber/
4. https://docs.x.ai/developers/pricing
5. https://docs.x.ai/developers/release-notes
6. https://github.blog/changelog/2026-06-16-github-models-is-no-longer-available-to-new-customers/
7. https://github.blog/changelog/2026-07-01-github-models-is-being-fully-retired-on-july-30-2026/
8. https://github.blog/changelog/2026-07-24-claude-opus-5-is-now-available-in-github-copilot/
9. https://github.com/MoonshotAI/Kimi-K3
10. https://github.com/MoonshotAI/Kimi-K3/blob/main/LICENSE
11. https://openai.com/index/advancing-the-price-performance-frontier-with-gpt-5-6/
12. https://openai.com/index/gpt-5-6/
13. https://platform.claude.com/docs/en/about-claude/models/overview
14. https://platform.claude.com/docs/en/release-notes/overview
15. https://www.anthropic.com/news/fable-mythos-access
16. https://www.anthropic.com/news/redeploying-fable-5
17. https://www.kimi.com/help/agent/agent-overview
18. https://www.kimi.com/resources/kimi-k3-pricing
19. https://x.ai/news/grok-4-5