

If you've recently noticed a jump in your AI subscription costs and encountered the message that "**grok-4-fast was retired on May 15, 2026**", you're not alone. Many longtime users of **Grok** tools through **xAI** are scratching their heads trying to understand why this popular model vanished, why there's no longer a paid-plan free trial available, and how this affects their monthly billing.

In this article, we'll unpack the key changes around **grok-4-fast retired**, why switching storefronts from *grok.com* to xAI's *Premium bundles* matters, what the free tiers really look like now, and how different versions like **SuperGrok Lite** versus Heavy models factor into your wallet — not just your results.

Two Storefronts: grok.com vs X Premium Bundles

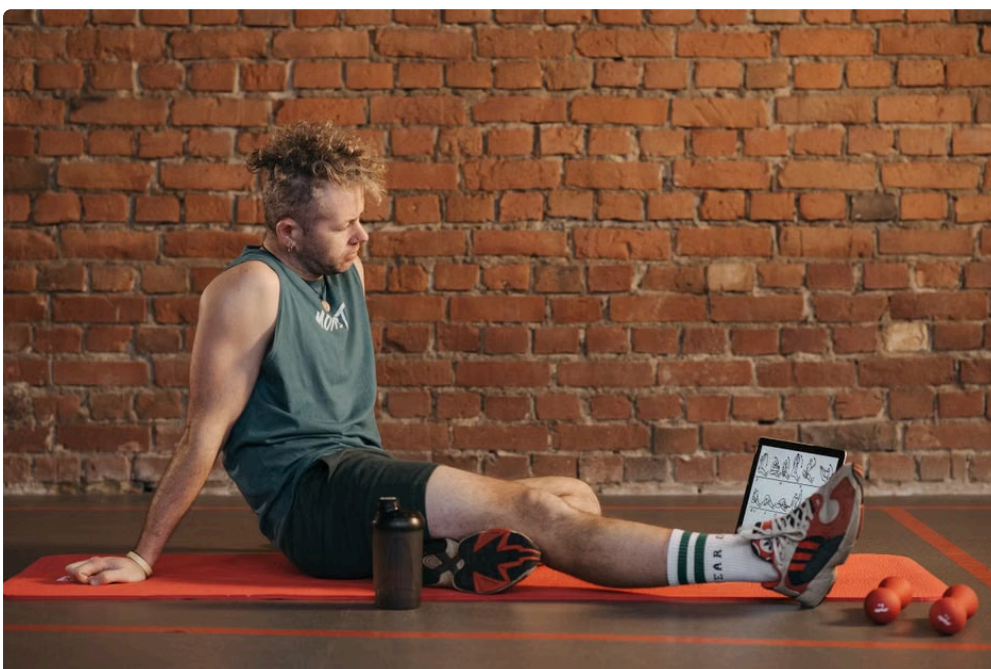
Originally, Grok's AI services were accessed via *grok.com*, offering a straightforward pay-as-you-go plan including the *grok-4-fast* variant favored for quick, low-latency workloads. However, since early 2026, xAI has transitioned much of this lineup exclusively to their *X Premium bundles*, consolidating AI tool offerings under a unified subscription strategy.

- **Grok.com storefront:** Featured mostly usage-based pricing with a <\$0 Free tier offering very limited demo access, alongside paid plans generally suitable for smaller or experimental teams.
- **X Premium bundles:** More integrated subscriptions combining access to Grok models plus enhanced tools like DeepSearch and Big Brain, but typically at higher base costs without a paid-plan free trial.

Because two different checkout systems handle billing, workflows, and model access differently, users migrating to xAI Premium bundles often face a steeper entry cost while gaining access to a richer suite of capabilities.

What Changed on May 15, 2026?

xAI officially retired the **grok-4-fast** model on **May 15, 2026**. Any requests to that model endpoint are now redirected automatically to **grok-4.3**, a cousin model that emphasizes more thorough, but computationally heavier processing.



Model Status Redirect Compute & Priority grok-4-fast Retired (May 15, 2026) Redirect to grok-4.3 Lower compute, lower priority grok-4.3 Active N/A Higher compute, higher priority SuperGrok Lite Active N/A Moderate compute, mid-tier priority

The key impact on your bill relates to compute usage and prioritization. While grok-4-fast favored speed and cost-efficiency, grok-4.3 demands more compute power, which [grok trial](#) translates into higher costs when billed on a subscription or usage basis.

No Paid-Plan Free Trial — Only Free Tier Demo

One of the most frustrating shifts for users is the removal of a **paid-plan free trial**. In the old grok.com storefront, users often enjoyed a limited trial for small-scale projects at no charge. This gave IT and finance teams the ability to assess real-world costs before committing to a subscription.

In the xAI storefront, the only free access level available is a **\$0 Free tier**, which functions more like a demo with strict caps:

- Limited daily token usage
- Restricted model selections, often defaulting to lower-priority or Demo-only modes
- No access to premium or experimental features found in paid plans

While it's great to have a zero-cost level, this *free tier hardly mimics real paid usage*, making it harder to estimate post-trial expenses.



The Free Tier Caps and What They Feel Like

The \$0 Free tier might look attractive on paper, but the usage limits are so restrictive that most productive teams feel handcuffed within hours. For example:

- Daily token limit might be 2,000 tokens, which translates roughly to 1-2 short documents or conversations.
- Queue priority for free-tier requests is at the bottom, often incurring noticeable latency spikes during peak times.
- No access to optimized compute settings means slow processing times for larger or more complex tasks.

Many users seeking to run critical workloads or multi-user team projects quickly bump into those caps and find the free tier unusable beyond initial testing.

SuperGrok Lite vs Heavy: Compute and Priority, Not Just Features

Another important nuance is how **SuperGrok Lite** differs from “Heavy” models like the retired *grok-4-fast* or its successor *grok-4.3*. While feature lists are often similar — all can process natural language, perform summarization, or provide coding assistance — the real differences lie in compute needs and operational priority:

- **SuperGrok Lite:** Designed to be a mid-tier subscription option with moderate compute allocation. It offers faster throughput than heavy models but with slightly less accuracy or contextual depth, ideal for startups or small teams balancing cost and performance.
- **Heavy Models (grok-4.3, previous grok-4-fast):** Prioritize accuracy and processing depth, requiring more expensive compute infrastructure and higher subscription fees. They enjoy priority routing on xAI’s backend but can rack up bills quickly as usage scales.

In other words, the upgrade from *grok-4-fast* to *grok-4.3* didn’t just swap your model — it suddenly increased the backend compute intensiveness, which directly affects your subscription or usage-based costs.

Integrations Matter: DeepSearch and Big Brain

xAI’s Premium bundles don’t just include AI language models. They bundle in powerful tools like **DeepSearch** and **Big Brain** for enhanced data querying and knowledge management. These tools can be extremely valuable but also contribute to your overall compute consumption if used heavily.

- **DeepSearch:** Provides semantic search capabilities that demand ongoing indexing and fast retrieval, drawing on compute resources beyond standard query volumes.
- **Big Brain:** Offers advanced reasoning and context retention over larger datasets, which further ramp up compute and memory requirements.

If your billing ticked higher post-retirement of *grok-4-fast*, it might be because xAI is encouraging users to adopt these higher-value tools that naturally require more resources — instead of relying solely on the earlier, simpler *grok* models.

How to Sanity Check Your Bill: The Team-of-Five Example

It’s always smart to sanity-check per-seat or per-user charges by imagining a typical small team scenario. For example, here’s a quick hypothetical monthly billing breakdown using xAI’s Premium bundles for a team of five users who each:

- Run about 20,000 tokens/day through *grok-4.3* or SuperGrok Lite
- Tap into DeepSearch for 10 queries/day
- Use Big Brain for knowledge base operations

Charge Category	Monthly Cost / User	Team of 5	Cost Base Premium Bundle	Subscription
\$50	\$250	Compute (Model Usage)	\$30	\$150
DeepSearch Queries	\$10	\$50	Big Brain Usage	\$20
\$100	Total Monthly	\$110	\$550	

If your bills are creeping toward these levels, it might be a reflection of the new bundling and model computations rather than a billing error. The retired *grok-4-fast* was a lightweight, low-cost option that disappeared — leaving the more powerful, more costly *grok-4.3* or SuperGrok Lite as the baseline.

What to Do Next: Your Quick Checklist

1. **Audit your active subscriptions:** Check whether you're still on grok.com plans versus xAI Premium bundles and understand your renewal terms.
2. **Track model usage:** Identify which models your workflows call. Are you inadvertently using grok-4.3 instead of lighter options?
3. **Review free tier limits:** If you're trying to stay \$0, confirm how daily token caps are affecting your productivity and consider if paid plans with trials can provide better cost control.
4. **Explore SuperGrok Lite:** Test SuperGrok Lite to see if it offers a better balance between cost, speed, and accuracy compared to grok-4.3 Heavy models.
5. **Limit extra tool usage:** DeepSearch and Big Brain can add up; ensure usage aligns with your needs and negotiate team seat counts accordingly.
6. **Contact xAI support:** Clarify billing details and consider requesting usage reports for better forecasting.

Summary

The retirement of **grok-4-fast** on May 15, 2026, marked not just the end of a popular lightweight AI model, but a billing paradigm shift for many users. Moving from grok.com to the new xAI Premium bundles introduces no paid-plan free trials, tighter free tier caps, and a push toward more compute-heavy models like grok-4.3 and SuperGrok Lite.

Understanding these changes, the difference between model variants in terms of compute and priority, and the hidden costs of integrated tools like DeepSearch and Big Brain helps finance and IT teams forecast and manage your bills better.

Keep calm, analyze your usage patterns, and use the checklist above to prevent your monthly AI costs from sneaking up on you — even as xAI evolves your AI experience.