

The Brief Life of Frontier AI

The AI industry's rapid release cycle is making flagship models increasingly short-lived. Instead of supporting multiple generations in parallel, leading AI companies are retiring or winding down older models within months of their launch. This strategy helps reduce infrastructure costs and concentrate computing resources on newer frontier models, but it also forces developers and enterprises to migrate applications more frequently.

ET's **Saliha Nasline** takes a look at the AI models being phased out this year, how quickly flagship models are being replaced and what comes next.

MODELS BEING WOUND DOWN

Company Status Replacement

Amazon

Nova Premier, Nova Omni, Canvas, Reel

Shifted to maintenance mode; active development halted

New frontier model under Frontier Model Research (FMR); expected later this year

OpenAI

GPT-4o, GPT-4.1, GPT-4.1 mini, o4-mini (ChatGPT)

Retired from ChatGPT

GPT-5.x family; models remain available via API for now

Google

Older Gemini preview models and Gemini 3.1 Flash Lite (scheduled)

Preview models deprecated; some stable models have announced shut-down timelines

Gemini 3.5 Flash Lite and newer Gemini releases

SHELF LIFE OF FLAGSHIP AI MODELS

Company	Model	Launch date	Retirement/wind-down	Lifespan
Amazon	Nova Premier	Dec 2024	Aug 2026	20 months
OpenAI	GPT-4o	May 2024	Feb 2026	21 months
Google	Gemini 3.1 Flash Lite	Mid-2026	Shutdown scheduled for 2027	12 months
Anthropic	Claude Sonnet 4	May 2025	Aug 2026	15 months

The lifespan of flagship AI models is typically around 15-21 months, much shorter than traditional software cycles

Anthropic*

Claude Sonnet 4 (GitHub Copilot deployment)

Deprecated in Copilot

Claude Sonnet 4.6

*The deprecation announced is specific to GitHub Copilot

LIFE AFTER THE RETIREMENT FOR AI MODELS

- Removed from consumer products
- Replaced by newer models with better reasoning, speed, multimodal capabilities
- May remain available via API for developers for a transition period
- Shifted to maintenance mode, receive only critical bug fixes and security updates
- Given end-of-life timelines, allowing customers time to migrate

CAPABILITY OF THE MODELS

Model	Context window
GPT-4o	128K tokens
Nova Premier	1m tokens
Claude Sonnet 4	1m tokens
Gemini 3.1 Flash Lite	1m tokens

The retirement of AI models shows just how quickly the market is moving. Today's leading model might be surpassed tomorrow on cost or reasoning capability. Enterprises should, therefore, avoid building their strategy around any single model. Firms that lead in AI will build flexible, governed foundations that adopt the best model for each task as the market evolves

Baris Gultekin,
Vice President – AI,
Snowflake

Foundation models are likely to be replaced every year or two; enterprises need to mitigate risks like vendor lock-in, cost escalation, intelligence sovereignty and evolving regulatory compliance needs. The goal isn't to solve for model selection but for model evolution

Mukund Kalmanker,
Global head-Data,
Analytics & AI, Apexon

Source: Company announcements, media reports and ET Research