# OpenAI returns old models to ChatGPT as Sam Altman admits ‘bumpy’ GPT-5 rollout

Want smarter insights in your inbox? Sign up for our weekly newsletters to get only what matters to enterprise AI, data, and security leaders. Subscribe Now  
  
OpenAI co-founder and CEO Sam Altman is publicly acknowledging major hiccups in yesterday’s rollout of GPT-5, the company’s new, flagship large language model (LLM) — advertised as its most powerful and capable yet.  
  
Answering user questions in a Reddit AMA (Ask Me Anything) thread and in a post on X this afternoon, Altman admitted to a range of issues that have disrupted the launch of GPT-5, including faulty model switching, poor performance, and user confusion — prompting OpenAI to partially walk back some of its platform changes and reinstate user access to earlier models like GPT-4o.  
  
“It was a little more bumpy than we hoped for,” Altman wrote in reply to a question on Reddit regarding the big GPT-5 launch.  
  
As for erroneous model performance charts shown off during OpenAI’s GPT-5 livestream, Altman said: “People were working late and were very tired, and human error got in the way. A lot comes together for a livestream in the last hours.”  
  
AI Scaling Hits Its Limits  
  
Power caps, rising token costs, and inference delays are reshaping enterprise AI. Join our exclusive salon to discover how top teams are:  
  
Secure your spot to stay ahead: https://bit.ly/4mwGngO  
  
While he noted the accompanying blog post and system card were accurate, the missteps further muddied a launch already facing scrutiny from early users and developers.  
  
GPT-5 rollout updates:\*We are going to double GPT-5 rate limits for ChatGPT Plus users as we finish rollout.\*We will let Plus users choose to continue to use 4o. We will watch usage as we think about how long to offer legacy models for.\*GPT-5 will seem smarter starting…  
  
One key reason for the trouble according to Altman stems from OpenAI’s new automatic “router” that assigns user prompts to one of four GPT-5 variants — regular, mini, nano, and pro — with an optional “thinking” mode for heavier reasoning tasks.  
  
On X, Altman revealed that a key part of that system — the autoswitcher — was “out of commission for a chunk of the day,” causing GPT-5 to appear “way dumber” than intended.  
  
In response, OpenAI says it’s implementing changes to the model decision boundary and will make it more transparent which model is responding to a given query.  
  
A UI update is also on the way to help users manually trigger thinking mode.  
  
Additionally, Altman confirmed that OpenAI will now allow ChatGPT Plus users to continue using GPT-4o — the prior default model — after a wave of complaints about GPT-5’s inconsistent performance. He said on Reddit the company is “trying to gather more data on the tradeoffs” before deciding how long to offer legacy models.  
  
they brought back the old models!!!! (it's in settings under "legacy models") pic.twitter.com/0MN85jmicG  
  
Yet many users including OpenAI beta testers like Wharton School of Business professor Ethan Mollick expressed confused and dismay at OpenAI unilaterally upgrading their ChatGPT experiences to GPT-5 and initially taking away access to the older models.  
  
OpenAI’s internal benchmarks may show GPT-5 leading the pack of LLMs, but real-world users are sharing a different experience.  
  
Since the launch, users have posted numerous examples of GPT-5 making basic errors in math, logic, and coding tasks.  
  
Data scientist Colin Fraser posted screenshots of GPT-5 incorrectly solving whether 8.888 repeating equals 9 (it does not, obviously), while another user showed it flubbing a simple algebra problem: 5.9 = x + 5.11.  
  
And still other users reported trouble getting accurate answers to math word problems or using GPT-5 to debug its own presentation charts.  
  
Developer feedback hasn’t been much better, with users posting images of GPT faring worse at “one-shot” certain programming tasks — completing them well with a single-prompt — compared to rival AI lab Anthropic’s new model Claude Opus 4.1.  
  
And security firm SPLX found GPT-5 still suffers from serious vulnerabilities to prompt injection and obfuscated logic attacks unless its safety layer is hardened.  
  
With 700 million weekly users on ChatGPT, OpenAI remains the largest player in generative AI by audience.  
  
But that scale has brought growing pains. Altman noted in his X post that API traffic doubled over 24 hours following the GPT-5 launch, contributing to platform instability.  
  
In response, OpenAI says it will double rate limits for ChatGPT Plus users, and continue to tweak infrastructure as it gathers feedback.  
  
But the early missteps — compounded by confusing UX changes and errors in a high-profile launch — have opened a window for rivals to gain ground.  
  
The pressure is on for OpenAI to prove that GPT-5 isn’t just an incremental update, but a true step forward. Based on the initial rollout, many users aren’t convinced — yet.  
  
If you want to impress your boss, VB Daily has you covered. We give you the inside scoop on what companies are doing with generative AI, from regulatory shifts to practical deployments, so you can share insights for maximum ROI.  
  
Read our Privacy Policy  
  
Thanks for subscribing. Check out more VB newsletters here.  
  
An error occured.

# OpenAI’s GPT-5 rollout is not going smoothly

Want smarter insights in your inbox? Sign up for our weekly newsletters to get only what matters to enterprise AI, data, and security leaders. Subscribe Now  
  
Updated Friday August 8, 5:21 pm ET: shortly after this post’s publication, OpenAI co-founder and CEO Sam Altman announced the company would restore access to GPT-4o and other old models for selected users, admitting the GPT-5 launch was “more bumpy than we hoped for.”  
  
The launch of OpenAI’s long anticipated new model, GPT-5, is off to a rocky start to say the least.  
  
Even forgiving errors in charts and voice demos during yesterday’s livestreamed presentation of the new model (actually four separate models, and a ‘Thinking’ mode that can be engaged for three of them), a number of user reports have emerged since GPT-5’s release showing it erring badly when solving relatively simple problems that preceding OpenAI models — and rivals from competing AI labs — answer correctly.  
  
For example, data scientist Colin Fraser posted screenshots showing GPT-5 getting a math proof wrong (whether 8.888 repeating is equal to 9 — it is of course, not).  
  
AI Scaling Hits Its Limits  
  
Power caps, rising token costs, and inference delays are reshaping enterprise AI. Join our exclusive salon to discover how top teams are:  
  
Secure your spot to stay ahead: https://bit.ly/4mwGngO  
  
Wow, I was just playing around before but it actually is stupid pic.twitter.com/ao51nOH0Ui  
  
It also failed on a simple algebra arithmetic problem that elementary schoolers could probably nail, 5.9 = x + 5.11.  
  
This is concerning. https://t.co/PUbeCSgtRV  
  
Using GPT-5 to judge OpenAI’s own erroneous presentation charts also did not yield helpful or correct responses.  
  
Q. Prove using an LLM-as-a-judge still doesn't workA. pic.twitter.com/KnCK5Xs9ja  
  
It also failed on this trickier math word problem below (which, to be fair, stumped this human at first…though Elon Musk’s Grok 4 AI answered it correctly. For a hint, think of the fact that flagstones in this case can’t be divided into smaller portions. They must remain in tact as 80 separate units, so no halves or quarters).  
  
Careful not to cut yourself on the jagged frontier pic.twitter.com/buJGgJ6baI  
  
The older 4o model performed better for me on at least one of these math problems. Unfortunately, OpenAI is slowly deprecating those older models — including the former default GPT-4o and the powerful reasoning model o3 — for users of ChatGPT, though they’ll continue to be available in the application programming interface (API) for developers for the foreseeable future.  
  
Even though OpenAI’s internal benchmarks and some third-party external ones have shown GPT-5 to outperform all other models at coding, it appears that in real world usage, Anthropic’s recently updated Claude Opus 4.1 seems to do a better job at “one-shotting” certain tasks, that is, completing the user’s desired application or software build to their specifications. See an example below from developer Justin Sun posted to X :  
  
Opus 4.1's one-shot attempt at "create a 3d capybara petting zoo" – 8 minutes totalThis was honestly pretty insane, not only are the capybaras way cuter and moving, there are individual pet affinity levels, a day/night switcher, feeding, and even a screenshot feature pic.twitter.com/FiKTO3FKK4  
  
In addition, a report from security firm SPLX found that OpenAI’s internal safety layer left major gaps in areas like business alignment and vulnerability to prompt injection and obfuscated logic attacks.  
  
While anecdotal, the checking the temperature on how the model is faring with early AI adopters seems to indicate a chilly reception.  
  
AI influencer and former Googler Bilawal Sidhu posted a poll on X asking for a “vibe check” from his followers and the wider userbase, and so far, with 172 votes in, the overwhelming response is “Kinda mid.”  
  
Alright, GPT-5 vibe check  
  
And as the pseudonymous AI Leaks and News account wrote, “The overwhelming consensus on GPT-5 from both X and the Reddit AMA are overwhelmingly negative.”  
  
The overwhelming consensus on GPT-5 from both X and the Reddit AMA are overwhelmingly negativeMost users are disgruntled about the broken model picker and non-pro users not having access to legacy modelsWhat are your initial thoughts on GPT-5?  
  
Tibor Blaho, lead engineer at AIPRM and a popular AI leaks and news poster on X, summarized the many problems with the ChatGPT-5 rollout in an excellent post, highlighting that one of the new marquee features — an automatic “router” in ChatGPT that chooses a thinking or non-thinking mode for the underlying GPT-5 model depending on the difficulty of the query — has become one of the chief complaints, given the model seemed to default to non-thinking mode for many users.  
  
A bit sad how the GPT-5 launch is going so far, especially after the long wait and high expectations– The automatic switching between models (the router) seems partly broken/unreliable– It's unclear exactly which model you're actually interacting with (standard or mini,…  
  
Thus, the sentiment toward ChatGPT-5 is far from universally positive, highlighting a major problem for OpenAI as it faces increasing competition from major U.S. rivals like Google and Anthropic, and a growing list of free, open source and powerful Chinese LLMs offering features that many U.S. models lack.  
  
Take the Alibaba Qwen Team of AI researchers, who just today updated their highly performant Qwen 3 model to have 1 million token context — giving users the ability to exchange nearly 4x as much information with the model in a single back/forth interaction as GPT-5 offers.  
  
Given OpenAI’s other big release this week — that of new open source gpt-oss models — also received a mixed reception from early users, things are not looking up for the number one dedicated AI company by users right now (700 million weekly active users of ChatGPT as of this month).  
  
Indeed, this is also exemplified by users of the betting marketplace Polymarket overwhelmingly deciding following the release of GPT-5 that Google would likely have the best AI model by the end of this month, August 2025.  
  
Other power users like Otherside AI co-founder and CEO Matt Shumer, who received early access to GPT-5 and blogged about it favorably in a review here, opined that views would shift as more people figured out the best ways to use the new model and adjusted their integration approaches:  
  
A lot of folks who are having a bad experience are using GPT-5 in agent harnesses that aren't yet optimized for it.For every new model release, there's a time lag between release + when companies that integrate the model have it truly working well.Agent companies rush to…  
  
While it’s still early days for GPT-5 — and the sentiment could change dramatically as more users get their hands on it and try it for different tasks — the early indications are not looking like this is a “home run” release for OpenAI in the same way that prior releases such as GPT-4, or even the newer 4o and o3, were. And that’s a concerning indicator for a company that just raised yet another funding round, yet remains unprofitable due to its high costs of research and development.  
  
If you want to impress your boss, VB Daily has you covered. We give you the inside scoop on what companies are doing with generative AI, from regulatory shifts to practical deployments, so you can share insights for maximum ROI.  
  
Read our Privacy Policy  
  
Thanks for subscribing. Check out more VB newsletters here.  
  
An error occured.

# OpenAI launches GPT-5, nano, mini and Pro — not AGI, but capable of generating ‘software-on-demand’

Want smarter insights in your inbox? Sign up for our weekly newsletters to get only what matters to enterprise AI, data, and security leaders. Subscribe Now  
  
After years of hype and speculation, OpenAI has officially launched a new lineup of large language models (LLMs), all different-sized variants of GPT-5, the long-awaited successor to its GPT-4 model from March 2023, nearly 2.5 years ago.  
  
The company is rolling out four distinct versions of the model — GPT-5, GPT-5 Mini, GPT-5 Nano and GPT-5 Pro — to meet varying needs for speed, cost, and computational depth.  
  
GPT-5 will soon be powering ChatGPT exclusively and replace all other models going forward for its 700 million weekly users, though ChatGPT Pro subscribers ($200) month can still select older models for the next 60 days.  
  
As per rumors and reports, OpenAI has replaced the previous system of having users switch the underlying model powering ChatGPT with an automatic router that decides to engage a special “GPT-5 thinking” mode with “deeper reasoning” that takes longer to respond on harder queries, or uses the regular GPT-5 or mini models for simpler queries.  
  
AI Scaling Hits Its Limits  
  
Power caps, rising token costs, and inference delays are reshaping enterprise AI. Join our exclusive salon to discover how top teams are:  
  
Secure your spot to stay ahead: https://bit.ly/4mwGngO  
  
In the API, the three reasoning-focused models — GPT-5, GPT-5 mini, and GPT-5 nano — are available as gpt-5, gpt-5-mini, and gpt-5-nano, respectively. GPT-5 Pro is not currently accessible via API, as it is used only to power ChatGPT for Pro-tier subscribers.  
  
GPT-5’s release comes just days after OpenAI launched a set of free, new open source LLMs under the name GPT-oss, which can be downloaded, customized and used offline by individuals and developers on consumer devices like PCs/Mac desktops and laptops.  
  
The biggest takeaway, though, is likely not what GPT-5 is, but what it isn’t: AGI, artificial general intelligence, OpenAI’s stated goal of an autonomous AI system that outperforms humans at most economically valuable work.  
  
Whether or not you, the reader, personally believe such a system is possible or desirable, OpenAI declaring AGI would have material business impacts. Wired reported previously that there is a clause in OpenAI’s contract with Microsoft that permits OpenAI to begin charging Microsoft for access to its newest models or cut off access to OpenAI models if OpenAI’s board determines that the company has achieved AGI or generates more than $100 billion in profit.  
  
But apparently, that is not the case today. As co-founder and CEO Sam Altman said, flanked by other OpenAI staffers on an embargoed video call with reporters last night, “the way that most of us define AGI, we’re still missing something quite important — many things that are quite important, actually — but one big one is a model that continuously learns as its deployed, and GPT-5 does not.”  
  
I also asked OpenAI the following question directly: “Is OpenAI considering GPT-5 AGI? Will it trigger any changes regarding Microsoft negotiations?”  
  
To which an OpenAI spokesperson responded over email:  
  
“GPT-5 is a significant step toward AGI in that it shows substantial improvements in reasoning and generalization, bringing us closer to systems that can perform a wide range of tasks with human-level capability. However, AGI is still a weakly defined term and means different things to different people. While GPT-5 meets some early criteria for AGI, it doesn’t yet reach the threshold of fully human-level AGI. There are still key limitations in areas like persistent memory, autonomy, and adaptability across tasks. Our focus remains on advancing these capabilities safely, rather than speculating on specific timelines.“  
  
Yet benchmark results shared by OpenAI show GPT-5 is nearing the threshold of performing as well as, and is close to exceeding, the average human expert performance at various tasks across law, logistics, sales, and engineering.  
  
As OpenAI writes: “When using reasoning, GPT-5 is comparable to or better than experts in roughly half the cases, while outperforming OpenAI o3 and ChatGPT Agent.”  
  
With numerous alternative models now available from OpenAI and a growing list of competitors, including Chinese startups offering powerful open-source models, what does GPT-5 bring to the table?  
  
Altman described the leap in capability as more than incremental. He compared the experience of using GPT-5 to upgrading from a pixelated display to a retina screen — something users simply don’t want to go back from.  
  
“GPT-3 felt like talking to a high school student,” Altman said. “GPT-4 was like a college student. GPT-5 is the first time it feels like talking to a PhD-level expert in your pocket.”  
  
Among the most impressive capabilities demoed for reporters during the embargoed call was the ability to generate the code for a fully working web application from a single prompt, in this case, a French language learning app with built-in game where English-to-French phrases were shown every time the user guided a virtual mouse to collect slices of cheese, with fully working emoji-inspired characters, backdrop/setting, and clickable interactive menus. The given prompt was also only a single paragraph.  
  
As Altman stated: “This idea of software on demand will be a defining part of the new GPT-5 era.”  
  
However, this basic capability — prompt to working software — has been available already from prior OpenAI models such as o3 and o4-mini, o4-high, and rival services like Anthropic’s Claude Artifacts, which I (and many others) have used for many months to create interactive first-person and clickable games as well.  
  
The advantage GPT-5 seems to offer in making games, apps, and other software from prompts seems to be in speed — it produced this demo app in a matter of mere minutes — and completeness, with very few discernible bugs and a completely playable experience in “one-shot,” or from a single prompt without back-and-forth conversation, as the developers like to say.  
  
GPT-5 is not restricted to premium subscribers. OpenAI has made the model available across all ChatGPT tiers, including free users — a deliberate move aligned with the company’s mission to ensure broad benefits from AI.  
  
Free-tier users can access GPT-5 and GPT-5 Mini, with usage limits — though exactly what those usage limits are remains undefined for now, and I’d guess will likely change on an irregular cadence depending on demand.  
  
Subscribers to the ChatGPT Plus ($20 per month) tier receive higher usage allowances, while subscribers to the ChatGPT Pro ($200 monthly), Team ($30 per month or $240 annually), and Enterprise (variable pricing depending on company size and usage) customers get unlimited or prioritized access.  
  
GPT-5 Pro will become available to Team, Enterprise, and EDU customers in the coming days.  
  
The new unified ChatGPT experience eliminates the need to manually select a model. Once users reach usage limits on GPT-5, the system automatically shifts to GPT-5 mini — a more lightweight but still highly capable fallback.  
  
According to OpenAI, GPT-5 offers the most accurate, responsive and context-aware AI system the company has ever shipped.  
  
It reduces hallucinations, handles multi-step reasoning more reliably and generates better-quality code, content, and responses across diverse domains.  
  
The GPT-5 system delivers ~45% fewer factual errors than GPT-4o in real-world traffic, and up to ~80% fewer when using its “thinking” mode.  
  
This mode, which users can trigger by explicitly asking the model to take its time, enables more complex and robust responses — powered by GPT-5 Pro in certain configurations. In tests, GPT-5 Pro sets new state-of-the-art scores on benchmarks like GPQA (88.4%), AIME 2025 math (100% when using Python to answer the questions) and HealthBench Hard (46.2%).  
  
Performance improvements show up across key academic and real-world benchmarks. In coding, GPT-5 sets new state-of-the-art results on SWE-Bench Verified (74.9%) and Aider Polyglot (88%).  
  
Perhaps most incredibly, on Humanity’s Last Exam — a newish benchmark of 2,500 extremely difficult tasks for programs — GPT-5 pro achieves a record-high 42%, blowing away the competition and all prior OpenAI models except the new ChatGPT agent unveiled last month that controls its own computer and cursor like a human.  
  
On writing tasks, GPT-5 adapts more smoothly to tone, context and user intent. It is better at maintaining coherence, structuring information clearly and completing complex writing assignments.  
  
The improvements are not just technical — OpenAI’s team emphasized how GPT-5 feels more natural and humanlike in conversation.  
  
Health-related use cases have also been enhanced. While OpenAI continues to caution that ChatGPT is not a replacement for medical professionals, GPT-5 is more proactive about flagging concerns, helping users interpret medical results and guiding them through preparing for appointments or evaluating options. The system also adjusts answers based on user location, background knowledge and context — leading to safer and more personalized assistance.  
  
One of the most significant updates is in safe completions, a new system that helps GPT-5 avoid abrupt refusals or unsafe outputs.  
  
Instead of declining queries outright, GPT-5 aims to provide the most helpful response within its safety boundaries and explains when it cannot assist — a change that dramatically reduces unnecessary denials while maintaining trustworthiness.  
  
GPT-5 is also a major upgrade for developers working on agentic systems and tool-assisted workflows. OpenAI has introduced a suite of developer-friendly controls in the GPT-5 API, including:  
  
For the first time, developers can also enable a new parameter option for reasoning effort, called ‘minimal’. This setting allows for the model to operate in reasoning mode, optimized for speed. “This is so that you can use these reasoning models, but with minimalization,” one OpenAI researcher explained during the company’s announcement livestream on YouTube earlier today, “so that they can slot into the very fastest and most latency sensitive applications.”  
  
The researcher stressed that minimal mode means developers don’t have to choose between accuracy and responsiveness: “Now you don’t actually have to choose between a bunch of models… you can use GPT-5 for all of your use cases, and just dilute reasoning effort.”  
  
This approach aims to make GPT-5 viable for ultra-low-latency scenarios like live customer interactions, fast-refresh dashboards, and real-time tool integrations, while still leveraging the reasoning capabilities that differentiate it from smaller or older models.  
  
The API itself is getting major upgrades. A researcher explained that the new custom tools feature moves beyond JSON-only outputs: “Custom tools are just free form plain text,” with the option to enforce formats using “a regular expression or even a context-free grammar… super useful if you have your own SQL fork and specify that the models always follow that format.”  
  
Developers also gain tool call preambles, so “the model’s ability to output explanation of what it’s about to do before it calls tools” can be switched on or tailored, and a verbosity parameter to set responses to “low, medium and high.”  
  
OpenAI also touted GPT-5’s leap in coding performance. On SWE-Bench, a benchmark for Python, the model scored 74.9%, beating GPT-4’s best of 69.1%, and it hit 88% on Polyglot, which covers multiple programming languages. Human testers preferred its code “70% of the time for its improved aesthetic abilities, but also better capabilities overall.”  
  
Developers can access GPT-5 through OpenAI’s platform for the following prices:  
  
The context window now spans 256,000 tokens (about the length of a 600-800 page book of text), allowing GPT-5 to handle substantially larger documents and more extensive conversations than its predecessor, GPT-4 Turbo.  
  
For those who require even more, GPT-4.1 (which supports 1 million-token context windows) remains available.  
  
Compared to the primary competitors — Anthropic and Google — OpenAI’s GPT-5 models are on par or cheaper for developers to access through the API, placing more downward pressure on the cost of intelligence.  
  
Several high-profile companies have already adopted GPT-5 in early trials. JetBrains is using it to power intelligent developer tools, and Notion has integrated GPT-5 to improve document generation and productivity workflows.  
  
At AI developer tool startup Cursor, co-founder and CEO Michael Truell said in a quote provided to reporters by OpenAI: “Our team has found GPT-5 to be remarkably intelligent, easy to steer, and even to have a personality we haven’t seen in any other model. It not only catches tricky, deeply-hidden bugs but can also run long, multi-turn background agents to see complex tasks through to the finish—the kinds of problems that used to leave other models stuck. It’s become our daily driver for everything from scoping and planning PRs to completing end-to-end builds. ”  
  
Other customers report major gains: GitLab cites a drop in tool call volume, GitHub notes improvements in reasoning across large codebases, and Uber is testing GPT-5 for real-time, domain-aware service applications. At Amgen, the model has already improved output quality and reduced ambiguity in scientific tasks.  
  
GPT-5’s launch coincides with several new features coming now and soon to ChatGPT.  
  
Users can now personalize the interface with chat colors (with exclusive options for paid users) and experiment with preset personalities like Cynic, Robot, Listener, and Nerd — designed to match different communication styles.  
  
ChatGPT will also soon support seamless integration with Gmail, Google Calendar, and Google Contacts. Once enabled, these services will be automatically referenced during chats, eliminating the need for manual toggling. These connectors launch for Pro subscribers next week, with broader availability to follow.  
  
A new Advanced Voice mode understands instructions more effectively and enables users to adjust tone and delivery. Voice will be available across all user tiers and included in custom GPTs.  
  
In 30 days, OpenAI will retire the older “Standard Voice Mode” and fully transition to this unified experience.  
  
With safer design, more robust reasoning, expanded developer tooling, and broad user access, GPT-5 reflects a maturing AI ecosystem that’s inching closer to real-world utility on a global scale.  
  
OpenAI’s approach this time is less about flash and more about integration. GPT-5 isn’t a separate offering that users have to seek out — it’s simply there, powering the tools millions already use, making them smarter and more capable and unlocking a whole new raft of use cases for developers.  
  
If you want to impress your boss, VB Daily has you covered. We give you the inside scoop on what companies are doing with generative AI, from regulatory shifts to practical deployments, so you can share insights for maximum ROI.  
  
Read our Privacy Policy  
  
Thanks for subscribing. Check out more VB newsletters here.  
  
An error occured.