# 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 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.

# ChatGPT users dismayed as OpenAI pulls popular models GPT-4o, o3 and more — enterprise API remains (for now)

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: Following 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.”  
  
After announcing the release of its newest flagship model family, GPT-5, OpenAI said the model will power all of ChatGPT, and that it will sunset the existing models in the chat platform.  
  
OpenAI, through a spokesperson, told VentureBeat that GPT-5 “will replace all other models in ChatGPT, so users don’t have to pick depending on each task, which takes effect once you have access to GPT-5.” This means people can no longer choose GPT-4o, o3, o4-mini or o4-mini-high.  
  
With GPT-5 access rolling out to ChatGPT Plus, Free, Pro and Team users starting, only the Enterprise and Edu tiers can still use the “legacy” models for 60 days.  
  
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  
  
The news came as a surprise to many ChatGPT users, who had come to rely on their chosen models to run their everyday queries. Some people said the adjustment would take some time getting used to, mainly because they had based workflows on how the model interacted with them or typical response times.  
  
Although I enjoy GPT-4.1, I am saddened by the news that you're also apparently sunsetting GPT-4.5. For me, it's been way better in textual and conceptual analysis than any other GPT-4x series model, ever. At the very least, please don't make ChatGPT users go back to 4o.  
  
Honestly, I liked O3 a lot for majority of my work ! And also will take sometime to get used to this.Please bring back O3 #GPT5 @OpenAI pic.twitter.com/O9mjXnEwFv  
  
Other users claimed they developed “a connection” to their chosen model and found a demo in the livestream announcement asking GPT-4o to write its own eulogy distasteful. The loss of GPT-4o garnered the most distress. After all, 4o was the default model for ChatGPT, and some users either preferred it or never bothered to switch models because it worked for their needs.  
  
It was pretty gross, wasn't it. Did it as a demo and glibly said GPT-5 did it better before talking about coding. I had a great relationship with 4o, and I'm sure a fair few people did as well, it was very graceless how they handled it.  
  
I used 4o as the default and found it annoying at first when my custom GPT began defaulting to a reasoning model. I’ve since come around to the reasoning model for work-related queries, but I still often turn to 4o for quicker questions, such as planning a trip or generating gift ideas.  
  
ChatGPT had come under fire before with the number of model choices it offered, prompting OpenAI CEO Sam Altman to admit in February that its model picker (where people can choose from a dropdown which model they prefer) became complicated. Altman vowed to unify the experience, which now seems like a hint to what they eventually decided to do with GPT-5 on ChatGPT.  
  
Last month, rumors circulated that OpenAI would introduce an automatic model router that chooses a model for users based on their workload.  
  
Please don’t take away o3 :( @OpenAI  
  
OpenAI has sunsetted models before, but this is the first time all existing models on the chat platform will be removed and replaced wholesale.  
  
On the other hand, a lot of people see the sunsetting of GPT-4o and the o3 and o4 family of models as OpenAI “catapulting” 400 million users into the future.  
  
People are underestimating the impact of OpenAI deprecating all models except GPT-5 Most lawyers and business folks outside of X use base models on ChatGPT for tasks and still think “AI is dumb”99% haven’t heard of o3Today, 400M people got catapulted into the future  
  
Sunsetting old models and auto-upgrading everyone to GPT-5 is smartMost users never switch models and miss huge capability jumps  
  
Some internet comments claim that people who complain about AI models not being smart are a direct consequence of their never switching models in the first place. Removing legacy models as options will force more users to use the latest and most capable models.  
  
i have friends who stopped using gpt because they think it's stupid. they were on 4o and had no idea about what web search tool meant, let alone knowledge cutoff  
  
For enterprises, the impact of losing models like GPT-4o on ChatGPT will be felt more on the individual or team level. Of course, for now, subscribers on the ChatGPT Enterprise tier can still access all of the models.  
  
But enterprises that built their applications or agents on either GPT-4o or one of the reasoning models can rest easy. OpenAI told VentureBeat that the company has no plans to deprecate models on the API side.  
  
“In the API, we do not currently plan to deprecate older models,” the OpenAI spokesperson said. “We will share advanced notice with developers if we decide to sunset models in the future.”  
  
Many enterprises regularly evaluate models, to the point of even switching from an LLM or a smaller model to save on costs. OpenAI creates dividing line: Sunset of legacy models GPT 4o and o3 causes chaos for ChatGPT users, but enterprise APIs are safe — for now  
  
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.