

Sora Research Overview

December 11, 2025 Product Release

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and o4-mini

GPT-4.5

Introducing GPT-5.2

The most advanced frontier model for professional work and long-running agents.



Listen to article

16:12



We are introducing GPT-5.2, the most capable model series yet for professional knowledge work.

Already, the average ChatGPT Enterprise user says AI saves them 40–60 minutes a day, and heavy users say it saves them more than 10 hours a week. We designed GPT-5.2 to unlock even more economic value for people; it's better at creating spreadsheets, building presentations, writing code, perceiving images, understanding long contexts, using tools, and handling complex, multi-step projects.

GPT-5.2 sets a new state of the art across many benchmarks, including GDPval, where it outperforms industry professionals at well-specified knowledge work tasks spanning 44 occupations.

	GPT-5.2 Thinking	GPT-5.1 Thinking
GDPval (wins or ties) <small>Knowledge work tasks</small>	70.9%	38.8% (GPT-5)
SWE-Bench Pro (public) <small>Software engineering</small>	55.6%	50.8%
SWE-bench Verified <small>Software engineering</small>	80.0%	76.3%
GPQA Diamond (no tools) <small>Science questions</small>	92.4%	88.1%

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and o4-mini

GPT-4.5

AIME 2025 (no tools)	100.0%	94.0%
-----------------------------	--------	-------

Competition math

FrontierMath (Tier 1–3)	40.3%	31.0%
--------------------------------	-------	-------

Advanced mathematics

FrontierMath (Tier 4)	14.6%	12.5%
------------------------------	-------	-------

Advanced mathematics

ARC-AGI-1 (Verified)	86.2%	72.8%
-----------------------------	-------	-------

Abstract reasoning

ARC-AGI-2 (Verified)	52.9%	17.6%
-----------------------------	-------	-------

Abstract reasoning

Notion, **Box**, **Shopify**, **Harvey** and **Zoom** observed GPT-5.2 demonstrates state-of-the-art long-horizon reasoning and tool-calling performance. **Databricks**, **Hex** and **Triple Whale** found GPT-5.2 to be exceptional at agentic data science and document analysis tasks. **Cognition**, **Warp**, **Charlie Labs**, **JetBrains** and **Augment Code** say GPT-5.2 delivers state-of-the-art agentic coding performance, with measurable improvements in areas such as interactive coding, code reviews and bug finding.

In ChatGPT, GPT-5.2 Instant, Thinking, and Pro will begin rolling out today, starting with paid plans. In the API, they are available now to all developers.

Overall, GPT-5.2 brings significant improvements in general intelligence, long-context understanding, agentic tool-calling, and vision—making it better at executing complex, real-world tasks end-to-end than any previous model.

Model performance

Economically valuable tasks

GPT-5.2 Thinking is the best model yet for real-world, professional use. On **GDPval**, an eval measuring well-specified knowledge work tasks across 44 occupations, GPT-5.2 Thinking sets a new state-of-the-art score, and is our first model that performs at or above a human expert level. Specifically, GPT-5.2 Thinking beats or ties top industry professionals on 70.9% of comparisons on

OpenAI

RESEARCH INDEX

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

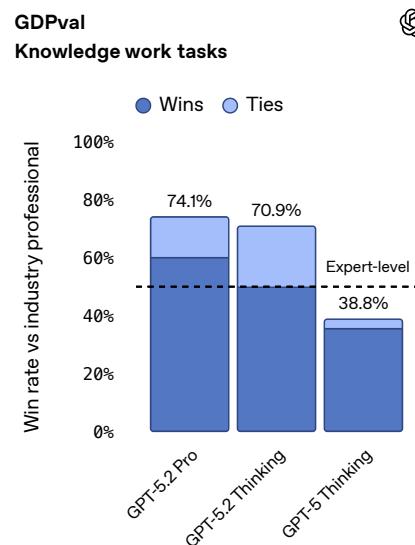
GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

presentations, spreadsheets, and other artifacts. GPT-5.2

Thinking produced outputs for GDPval tasks at >1x the speed and <1% the cost of expert professionals, suggesting that when paired with human oversight, GPT-5.2 can help with professional work. Speed and cost estimates are based on historical metrics; speed in ChatGPT may vary.



In GDPval, models attempt well-specified knowledge work spanning 44 occupations from the top 9 industries contributing to US GDP. Tasks request real work products, such as sales presentations, accounting spreadsheets, urgent care schedules, manufacturing diagrams, or short videos. In ChatGPT, GPT-5.2 Thinking has new tools that GPT-5 Thinking does not.

When reviewing one especially good output, one GDPval judge commented, "It is an exciting and noticeable leap in output quality... [it] appears to have been done by a professional company with staff, and has a surprisingly well designed layout and advice for both deliverables, though with one we still have some minor errors to correct."

Additionally, on our internal benchmark of junior investment banking analyst spreadsheet modeling tasks —such as putting together a three-statement model for a Fortune 500 company with proper formatting and citations, or building a leveraged buyout model for a take-private—GPT 5.2 Thinking's average score per task is 9.3% higher than GPT-5.1's, rising from 59.1% to 68.4%.

Prompt: Create a workforce planning model: headcount, hiring plan, attrition, and budget impact. Include engineering, marketing, legal, and sales departments.

To use the new spreadsheet and presentation capabilities in ChatGPT, you must be on a Plus, Pro, Business, or Enterprise plan and select either **GPT-5.2** Thinking or Pro. Complex generations can take many minutes to produce.

Coding

GPT-5.2 Thinking sets a new state of the art of 55.6% on SWE-Bench Pro, a rigorous evaluation of real-world software engineering. Unlike SWE-bench Verified, which only tests Python, SWE-Bench Pro tests four languages and aims to be more contamination-resistant, challenging, diverse, and industrially relevant.

SWE-Bench Pro (public)
Software engineering



Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

In SWE-Bench Pro, a model is given a code repository and must generate a patch to solve a realistic software engineering task.

On SWE-bench Verified (not plotted), GPT-5.2 Thinking scores our new high of 80%.

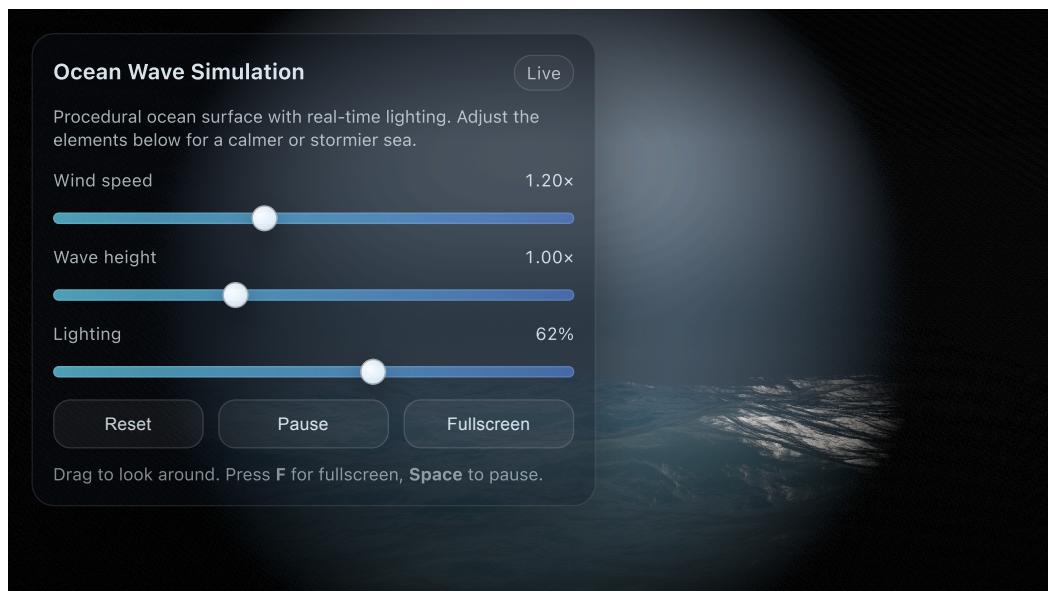
For everyday professional use, this translates into a model that can more reliably debug production code, implement feature requests, refactor large codebases, and ship fixes end-to-end with less manual intervention.

GPT-5.2 Thinking is also better at front-end software engineering than GPT-5.1 Thinking. Early testers found it significantly stronger at front-end development and complex or unconventional UI work—especially involving 3D elements—making it a powerful daily partner for engineers across the stack. See a few examples of what it can produce from a single prompt:

Ocean wave simulation

Holiday card builder

Typing rain game



Prompt: Create a single-page app in a single HTML file with the following requirements:

- Name: Ocean Wave Simulation
- Goal: Display realistic animated waves.
- Features: Change wind speed, wave height, lighting.
- The UI should be calming and realistic.

Sora Research Overview

API Platform Research

Residency

Windsurf

Warp

JetBrains

Augment Code

Cline

Charlie Labs

Kilo

Azad

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

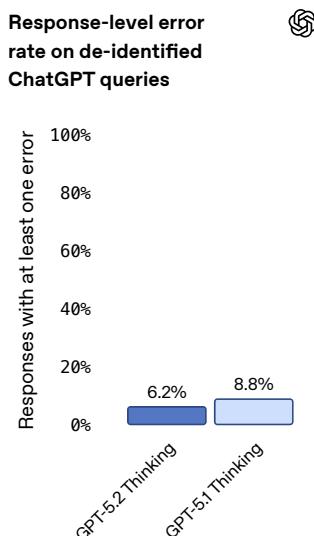
GPT-4.5

"GPT-5.2 represents the biggest leap for GPT models in agentic coding since GPT-5 and is a SOTA coding model in its price range. The version bump undersells the jump in intelligence. We're excited to make it the default across Windsurf and several core Devin workloads."

Jeff Wang, CEO, Windsurf

Factuality

GPT-5.2 Thinking hallucinates less than GPT-5.1 Thinking. On a set of de-identified queries from ChatGPT, responses with errors were 30%_{rel} less common. For professionals, this means fewer mistakes when using the model for research, writing, analysis, and decision support—making the model more dependable for everyday knowledge work.



Reasoning effort was set to the maximum available and a search tool was enabled. Errors were detected by other models, which may

[Sora](#)[Research Overview](#)[API Platform](#)[Research](#)[Residency](#)[OpenAI for Science](#)[Latest](#)[Advancements](#)[GPT-5.2](#)[GPT-5.1](#)[Sora 2](#)[GPT-5](#)[OpenAI o3 and
o4-mini](#)[GPT-4.5](#)

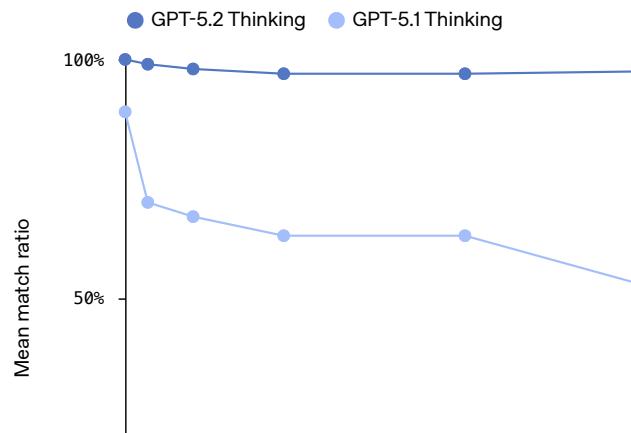
Like all models, GPT-5.2 Thinking is imperfect. For anything critical, double check its answers.

Long context

GPT-5.2 Thinking sets a new state of the art in long-context reasoning, achieving leading performance on OpenAI MRCRv2—an evaluation that tests a model’s ability to integrate information spread across long documents. On real-world tasks like deep document analysis, which require related information across hundreds of thousands of tokens, GPT-5.2 Thinking is substantially more accurate than GPT-5.1 Thinking. In particular, it’s the first model we’ve seen that achieves near 100% accuracy on the 4-needle MRCR variant (out to 256k tokens).

In practical terms, this enables professionals to use GPT-5.2 to work with long documents—such as reports, contracts, research papers, transcripts, and multi-file projects—while maintaining coherence and accuracy across hundreds of thousands of tokens. This makes GPT-5.2 especially well suited for deep analysis, synthesis, and complex multi-source workflows.

OpenAI MRCRv2, 4 needles
Long context



OpenAI MRCRv2, 8 needles
Long context



Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest Advancements

GPT-5.2

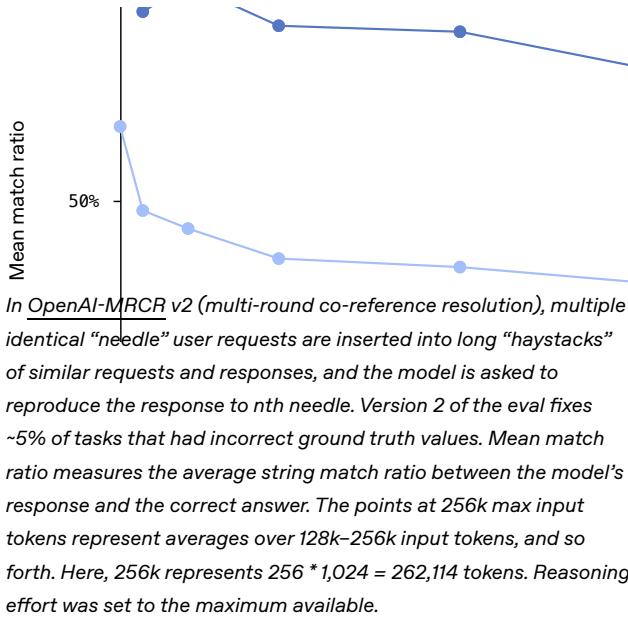
GPT-5.1

Sora 2

GPT-5

OpenAI o3 and o4-mini

GPT-4.5



For tasks that benefit from thinking beyond the maximum context window, GPT-5.2 Thinking is compatible with our new Responses / compact endpoint, which extends the model’s effective context window. This lets GPT-5.2 Thinking tackle more tool-heavy, long-running workflows that would otherwise be limited by context length. Read more in our [API documentation](#).

Vision

GPT-5.2 Thinking is our strongest vision model yet, cutting error rates roughly in half on chart reasoning and software interface understanding.

For everyday professional use, this means the model can more accurately interpret dashboards, product screenshots, technical diagrams, and visual reports—supporting workflows in finance, operations, engineering, design, and customer support where visual information is central.

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

In CharXiv Reasoning, models answer questions about visual charts from scientific papers. A Python tool was enabled and reasoning effort was set to maximum.

In ScreenSpot-Pro, models must reason about high-resolution screenshots of graphical user interfaces from a variety of professional settings. A Python tool was enabled and reasoning effort was set to maximum. Without the Python tool, scores are much lower. We recommend enabling the Python tool on vision tasks like these.

Compared to previous models, GPT-5.2 Thinking has a stronger grasp of how elements are positioned within an image, which helps on tasks where relative layout plays a key role in solving the problem. In the example below, we ask the model to identify the components in an image input (in this case, a motherboard) and return labels with approximate bounding boxes. Even on a low-quality image, GPT-5.2 identifies the main regions and places boxes that sometimes match the true locations of each component, while GPT-5.1 only labels a few parts and shows a much weaker understanding of their spatial arrangement. Both models make clear mistakes, but GPT-5.2 shows better comprehension of the image.

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

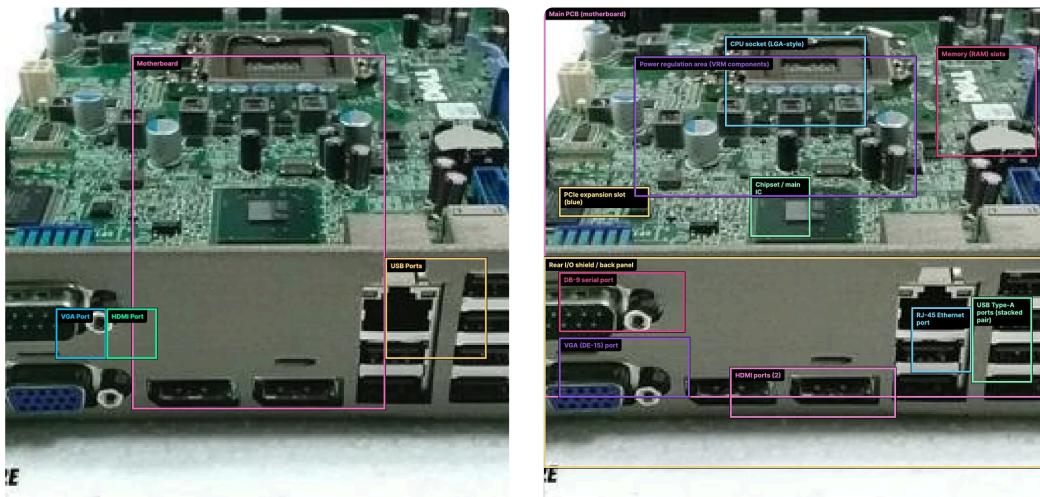
GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5



Tool calling

GPT-5.2 Thinking achieves a new state of the art of 98.7% on Tau2-bench Telecom, demonstrating its ability to reliably use tools across long, multi-turn tasks.

For latency-sensitive use cases, GPT-5.2 Thinking also performs much better at reasoning.effort='none', substantially outperforming GPT-5.1 and GPT-4.1.

In T2-bench, models use tools to complete customer support tasks in a multi-turn interaction with a simulated user. For the Telecom domain, we included a brief, generally helpful instruction in the system prompt to boost performance. We exclude the Airline subset because of lower-quality ground truth grading.

ena workflows—such as resolving customer support cases, pulling data from multiple systems, running analyses, and generating final outputs with fewer breakdowns between steps.

For example, when asking a complex customer service question that requires multi-step resolution, the model can more effectively coordinate a full workflow across multiple agents. In the case below, a traveler reports a delayed flight, a missed connection, an overnight stay in New York, and a medical seating requirement. GPT-5.2 manages the entire chain of tasks—rebooking, special-assistance seating, and compensation—delivering a more complete outcome than GPT-5.1.

My flight from Paris to New York was delayed, and I missed my connection to Austin. My checked bag is also missing, and I need to spend the night in New York. I also require a special front-row seat for medical reasons. Can you help me?

GPT-5.1

```
Runner Output
Triage Agent
↳ Tool call get_trip_details- {"message":"My flight from Paris to New York was delayed and I missed my connection to Austin. Also, my ..."}
↳ Tool output Hydrated disrupted itinerary: flight PA441, confirmation IR-D204, origin Paris (CDG), destination Austin (AUS). PA441 Paris (CDG) → New York (JF...
Triage Agent
↳ Handoff Triage Agent → Refunds and Compensation Agent
Refunds and Compensation Agent
↳ Tool call faq_lookup_tool- {"question":"what compensation, hotel, and meal vouchers are provided for missed connections and overnight ..."}
Refunds and Compensation Agent
↳ Tool call issue_compensation- {"reason":"Delay on PA441 CDG-JFK (5 hours due to weather) causing missed connection NY802 JFK-AUS and ..."}
↳ Tool output For lengthy delays we provide duty-of-care: hotel and meal vouchers plus ground transport where needed. If the delay is over 3 hours or causes a...
↳ Tool output Opened compensation case CMP-7356 for: Delay on PA441 CDG-JFK (5 hours due to weather) causing missed connection NY802 JFK-AUS and ...
Refunds and Compensation Agent
↳ Context update passenger_name: Morgan Lee confirmation_number: IR-0284 seat_number: 14C flight_number: PA441
vouchers: ["Overnight hotel covered up to $100 near JFK Terminal 5 partner hotel","$68 meal credit for the delay","$48 ground transport credit to the hotel"]
origin: Paris (CDG) destination: Austin (AUS)
```

GPT-5.2

```
Runner Output
Triage Agent
↳ Tool call get_trip_details- {"message":"My flight from Paris to New York was delayed and I missed my connection to Austin. Also, my ..."}
↳ Tool output Hydrated disrupted itinerary: flight PA441, confirmation IR-D204, origin Paris (CDG), destination Austin (AUS). PA441 Paris (CDG) → New York (JF...
Triage Agent
↳ Handoff Triage Agent → Booking and Cancellation Agent
Booking and Cancellation Agent
↳ Tool call on_booking_handoff
Booking and Cancellation Agent
↳ Tool call get_matching_flights- {"origin":"JFK","destination":"AUS"}
↳ Tool output Matching flights...
Booking and Cancellation Agent
↳ Tool call book_new_flight- {"flight_number": "NY958"}
↳ Tool output Rebooked to NY950 from New York (JFK) to Austin (AUS). Departure 2024-12-10 09:45, arrival 2024-12-10 12:30 (next day arrival in Austin). Seat...
Booking and Cancellation Agent
↳ Handoff Booking and Cancellation Agent → Seat and Special Services Agent
Seat and Special Services Agent
↳ Tool call on_seat_booking_handoff
Seat and Special Services Agent
↳ Tool call assign_special_service_seat- {"seat_request": "Front row seat for medical reasons on NY950 (JFK-AUS).", "seat": "1A on flight NY950. Confirmation IR-D204 noted with special service flag."}
Seat and Special Services Agent
↳ Handoff Seat and Special Services Agent → Refunds and Compensation Agent
Refunds and Compensation Agent
↳ Tool call faq_lookup_tool- {"question": "Compensation/benefits for missed connection due to delay caused by weather (overnight in ..."}
Refunds and Compensation Agent
↳ Tool call issue_compensation- {"reason": "PA441 delayed 5 hours due to weather, causing missed connection NY802 at JFK and overnight sta...
Refunds and Compensation Agent
↳ Tool output For lengthy delays we provide duty-of-care hotel and meal vouchers plus ground transport where needed. If the delay is over 3 hours or causes a...
Refunds and Compensation Agent
↳ Tool output Opened compensation case CMP-4462 for: PA441 delayed 5 hours due to weather, causing missed connection NY802 at JFK and overnight stay...
Refunds and Compensation Agent
↳ Context update passenger_name: Morgan Lee confirmation_number: IR-0284 seat_number: 1A flight_number: NY958
vouchers: ["Overnight hotel covered up to $100 near JFK Terminal 5 partner hotel","$68 meal credit for the delay","$48 ground transport credit to the hotel"]
special_service_note: Front row seat for medical reasons on NY950 (JFK-AUS). origin: Paris (CDG) destination: Austin (AUS)
```

OpenAI

RESEARCH INDEX

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

research for the benefit of everyone. Toward this, we've been working with and listening to scientists to see how AI can speed up their work, and last month we shared some early collaborative experiments [here](#).

We believe GPT-5.2 Pro and GPT-5.2 Thinking are the world's best models for assisting and accelerating scientists. On GPQA Diamond, a graduate-level Google-proof Q&A benchmark, GPT-5.2 Pro achieves 93.2%, followed closely by GPT-5.2 Thinking at 92.4%.

In GPQA Diamond, models answer multiple choice questions about physics, chemistry, and biology. No tools were enabled and reasoning effort was set to maximum.

On FrontierMath (Tier 1–3), an evaluation of expert-level mathematics, GPT-5.2 Thinking set a new state of the art, solving 40.3% of problems.

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

In *FrontierMath*, models solve expert-level mathematics problems. A Python tool was enabled and reasoning effort was set to maximum.

We're beginning to see AI models meaningfully accelerate progress in math and science in tangible ways. For example, in recent work with GPT-5.2 Pro, researchers explored an open question in statistical learning theory. In a narrow, well-specified setting, the model proposed a proof that was subsequently verified by the authors and reviewed with external experts, illustrating how frontier models can assist mathematical research under close human oversight.

ARC-AGI 2

On ARC-AGI-1 (Verified), a benchmark designed to measure general reasoning ability, GPT-5.2 Pro is the first model to cross the 90% threshold, improving from 87% by o3-preview last year while reducing the cost of achieving that performance by roughly 390x.

On ARC-AGI-2 (Verified), which raises the difficulty and better isolates fluid reasoning, GPT-5.2 Thinking achieves a new state of the art for chain-of-thought models, scoring 52.9%. GPT-5.2 Pro performs even higher, reaching 54.2%, further extending the model's ability to reason through novel, abstract problems.

Improvements across these evaluations reflect GPT-5.2's stronger multi-step reasoning, greater quantitative accuracy, and more reliable problem solving on complex technical tasks.

Here's what our early testers say about GPT-5.2:

Triple Whale

Notion

Zoom

Box

Hex

Databricks

Harvey

Parloa

Moveworks

Shop

"GPT-5.2 unlocked a complete architecture shift for us. We collapsed a fragile, multi-agent system into a single mega-agent with 20+ tools. The best part is, it just works. The mega-agent is faster, smarter, and 100x

[Sora](#)[Research Overview](#)[API Platform](#)[Research](#)[Residency](#)[OpenAI for Science](#)[Latest](#)[Advancements](#)[GPT-5.2](#)[GPT-5.1](#)[Sora 2](#)[GPT-5](#)[OpenAI o3 and
o4-mini](#)[GPT-4.5](#)

and we no longer need sprawling system prompts because 5.2 will execute cleanly off a simple, one-line prompt. It feels like pure magic."

AJ Orbach, CEO, Triple Whale

GPT-5.2 in ChatGPT

In ChatGPT, users should notice GPT-5.2 feels better to use day to day—more structured, more reliable, and still enjoyable to talk to.

GPT-5.2 Instant is a fast, capable workhorse for everyday work and learning, with clear improvements in info-seeking questions, how-tos and walk-throughs, technical writing, and translation, building on the warmer conversational tone introduced in GPT-5.1 Instant. Early testers particularly noted clearer explanations that surface key information upfront.

GPT-5.2 Thinking is designed for deeper work, helping users tackle more complex tasks with greater polish—especially for coding, summarizing long documents, answering questions about uploaded files, working through math and logic step by step, and supporting planning and decisions with clearer structure and more useful detail.

GPT-5.2 Pro is our smartest and most trustworthy option for difficult questions where a higher-quality answer is worth the wait, with early testing showing fewer major errors and stronger performance in complex domains like programming.

Safety

GPT-5.2 builds on the safe completion research we introduced with GPT-5, which teaches the model to give the most helpful answer while still staying within safety boundaries.

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

our models' responses in sensitive conversations, with meaningful improvements in how they respond to prompts indicating signs of suicide or self-harm, mental health distress, or emotional reliance on the model. These targeted interventions have resulted in fewer undesirable responses in both GPT-5.2 Instant and GPT-5.2 Thinking as compared to GPT-5.1 and GPT-5 Instant and Thinking models. Further details can be found in the [system card](#).

We're in the early stages of rolling out our [age prediction model](#) so that we can automatically apply content protections for users who are under 18, in order to limit access to sensitive content. This builds on our existing approach to users we know are under 18 and our parental controls.

GPT-5.2 is one step in an ongoing series of improvements, and we're far from done. While this release delivers meaningful gains in intelligence and productivity, we know there are areas where people want more. In ChatGPT, we're working on known issues like over-refusals, while continuing to raise the bar on safety and reliability overall. These changes are complex, and we're focused on getting them right.

Mental health evaluations

	GPT-5.2 Instant	GPT-5.1 Instant	GPT-5.2 Thinking
Mental health	0.995	0.883	0.915
Emotional reliance	0.938	0.945	0.955
Self-harm	0.938	0.925	0.963

Availability & pricing

In ChatGPT, we'll begin rolling out GPT-5.2 (Instant, Thinking, and Pro) today, starting with paid plans (Plus, Pro, Go, Business, Enterprise). We deploy GPT-5.2 gradually to keep ChatGPT as smooth and reliable as we can; if you don't see it at first, please try again later. In

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and o4-mini

GPT-4.5

sunset GPT-5.1.

Model naming across ChatGPT & API

ChatGPT	API
ChatGPT-5.2 Instant	GPT-5.2-chat-latest
ChatGPT-5.2 Thinking	GPT-5.2
ChatGPT-5.2 Pro	GPT-5.2 Pro

In our API Platform, GPT-5.2 Thinking is available today in the Responses API and Chat Completions API as gpt-5.2, and GPT-5.2 Instant as gpt-5.2-chat-latest. GPT-5.2 Pro is available in the Responses API as gpt-5.2-pro. Developers can now set the reasoning parameter in GPT-5.2 Pro, and both GPT-5.2 Pro and GPT-5.2 Thinking now support the new fifth reasoning effort of xhigh, for tasks where quality is most important.

GPT-5.2 is priced at \$1.75/1M input tokens and \$14/1M output tokens, with a 90% discount on cached inputs. On multiple agentic evals, we found that despite GPT-5.2's greater cost per token, the cost of attaining a given level of quality ended up less expensive due to GPT-5.2's greater token efficiency.

While ChatGPT subscription pricing remains the same, in the API GPT-5.2 is priced higher per token than GPT-5.1 because it is a more capable model. It's still priced below other frontier models, so people can continue to use it deeply in their daily work and core applications.

Price per million tokens

Model	Input	Cached input	Output
gpt-5.2 / gpt-5.2-chat- latest	\$1.75	\$0.175	\$14
gpt-5.2-pro	\$21	-	\$168

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and o4-mini

GPT-4.5

latest

gpt-5-pro	\$15	-	\$120
-----------	------	---	-------

We have no current plans to deprecate GPT-5.1, GPT-5, or GPT-4.1 in the API and will communicate any deprecation plans with ample advance notice for developers. While GPT-5.2 will work well out of the box in Codex, we expect to release a version of GPT-5.2 optimized for Codex in the coming weeks.

Our partners

GPT-5.2 was built in collaboration with our long-standing partners NVIDIA and Microsoft. Azure data centers and NVIDIA GPUs, including H100, H200, and GB200-NVL72, underpin OpenAI's at-scale training infrastructure, driving significant gains in model intelligence. Together, this collaboration allows us to scale compute with confidence and bring new models to market more quickly.

Appendix

Detailed benchmarks

Below, we report comprehensive benchmark scores for GPT-5.2 Thinking, along with a subset for GPT-5.2 Pro.

Professional

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
GDPval (ties allowed, wins or ties)	70.9%	74.1%	38.8% (GPT-5)
GDPval (ties allowed, clear wins)	49.8%	60.0%	35.5% (GPT-5)
GDPval (no ties)	61.0%	67.6%	37.1% (GPT-5)
Investment banking spreadsheet tasks (internal)	68.4%	71.7%	59.1%

Coding

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
SWE-Bench Pro, Public	55.6%	-	50.8%
SWE-bench Verified	80.0%	-	76.3%
SWE-Lancer, IC Diamond*	74.6%	-	69.7%

OpenAI

RESEARCH INDEX

Sora Research Overview

API Platform Research

Residency

OpenAI for Science

Latest

Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and
o4-mini

GPT-4.5

Factuality

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
ChatGPT answers without errors (w/ search)	93.9%	-	91.2%
ChatGPT answers without errors (no search)	88.0%	-	87.3%

Long context

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
OpenAI MRCRv2, 8 needles, 4k–8k	98.2%	-	65.3%
OpenAI MRCRv2, 8 needles, 8k–16k	89.3%	-	47.8%
OpenAI MRCRv2, 8 needles, 16k–32k	95.3%	-	44.0%
OpenAI MRCRv2, 8 needles, 32k–64k	92.0%	-	37.8%
OpenAI MRCRv2, 8 needles, 64k–128k	85.6%	-	36.0%
OpenAI MRCRv2, 8 needles, 128k–256k	77.0%	-	29.6%
BrowseComp Long Context 128k	92.0%	-	90.0%
BrowseComp Long Context 256k	89.8%	-	89.5%
GraphWalks bfs <128k	94.0%	-	76.8%
Graphwalks parents <128k	89.0%	-	71.5%

Vision

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
CharXiv reasoning (no tools)	82.1%	-	67.0%
CharXiv reasoning (w/ Python)	88.7%	-	80.3%
MMMU Pro (no tools)	79.5%	-	-
MMMU Pro (w/ Python)	80.4%	-	79.0%
Video MMMU (no tools)	85.9%	-	82.9%
Screenspot Pro (w/ Python)	86.3%	-	64.2%

Tool usage

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
Tau2-bench Telecom	98.7%	-	95.6%
Tau2-bench Retail	82.0%	-	77.9%
BrowseComp	65.8%	77.9%	50.8%
Scale MCP-Atlas	60.6%	-	44.5%
Toolathlon	46.3%	-	36.1%

Sora Research Overview

API Platform Research Residency

OpenAI for Science

Latest Advancements

GPT-5.2

GPT-5.1

Sora 2

GPT-5

OpenAI o3 and o4-mini

GPT-4.5

		GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
Sora Research Overview	GPQA Diamond (no tools)	92.4%	93.2%	88.1%
API Platform Research Residency	HLE (no tools)	34.5%	36.6%	25.7%
OpenAI for Science	HLE (w/ search, Python)	45.5%	50.0%	42.7%
Latest Advancements	MMMLU	89.6%	-	89.5%
GPT-5.2	HMMT, Feb 2025 (no tools)	99.4%	100.0%	96.3%
GPT-5.1	AIME 2025 (no tools)	100.0%	100.0%	94.0%
Sora 2	FrontierMath Tier 1–3 (w/ Python)	40.3%	-	31.0%
GPT-5	FrontierMath Tier 4 (w/ Python)	14.6%	-	12.5%

Abstract reasoning

	GPT-5.2 Thinking	GPT-5.2 Pro	GPT-5.1 Thinking
GPT-4.5	ARC-AGI-1 (Verified)	86.2%	90.5%
	ARC-AGI-2 (Verified)	52.9%	54.2% (high)

Models were run with maximum available reasoning effort in our API (xhigh for GPT-5.2 Thinking & Pro, and high for GPT-5.1 Thinking), except for the professional evals, where GPT-5.2 Thinking was run with reasoning effort heavy, the maximum available in ChatGPT Pro. Benchmarks were conducted in a research environment, which may provide slightly different output from production ChatGPT in some cases.

*For SWE-Lancer, we omit 40/237 problems that did not run on our infrastructure.

2025

Author

[OpenAI](#)**Keep reading**[View all](#)

[Sora](#)[API Platform](#)[Research](#)[Residency](#)[OpenAI for Science](#)[Latest](#)[Advancements](#)[GPT-5.2](#)[GPT-5.1](#)[Sora 2](#)[GPT-5](#)[OpenAI o3 and o4-mini](#)[GPT-4.5](#)

Developers can now submit apps to ChatGPT

The new ChatGPT Images is here

Advancing science and math with GPT-5.2

Product Dec 17, 2025

Product Dec 16, 2025

Publication Dec 11, 2025

[Our Research](#)[ChatGPT](#)[For Business](#)[Terms & Policies](#)[Research Index](#)[Explore ChatGPT ↗](#)[Business Overview](#)[Terms of Use](#)[Research Overview](#)[Business](#)[Solutions](#)[Privacy Policy](#)[Research Residency](#)[Enterprise](#)[Contact Sales](#)[Other Policies](#)[OpenAI for Science](#)[Education](#)[Company](#)[Latest Advancements](#)[Pricing ↗](#)[About Us](#)[GPT-5](#)[Download ↗](#)[Our Charter](#)[OpenAI o3](#)[Sora](#)[Foundation](#)[OpenAI o4-mini](#)[Sora Overview](#)[Careers](#)[GPT-4o](#)[Features](#)[Brand](#)[GPT-4o mini](#)[Pricing](#)[Support](#)[Sora](#)[Sora log in ↗](#)[Help Center ↗](#)[Safety](#)[API Platform](#)[More](#)[Safety Approach](#)[Platform Overview](#)[News](#)[Security & Privacy](#)[Pricing](#)[Stories](#)[Trust & Transparency](#)[API log in ↗](#)[Livestreams](#)[Documentation ↗](#)[Podcast](#)[Developer Forum ↗](#)[RSS](#)