January 23, 2025 Release

Computer-Using Agent

Powering Operator with Computer-Using Agent, a universal interface for AI to interact with the digital world.

Go to Operator ↗

00:00 0.5x 1x 1.5x 2x Share

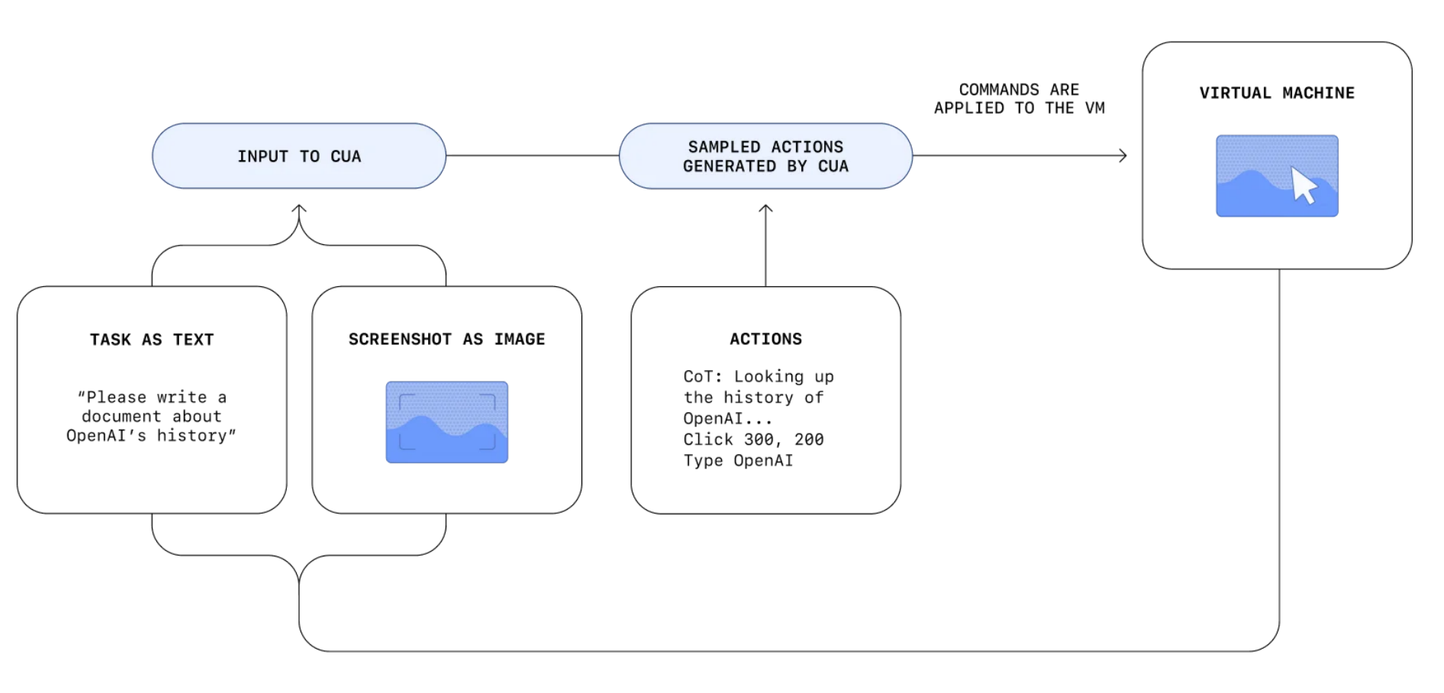
Today we introduced a research preview of Operator, an agent that can go to the web to perform tasks for you. Powering Operator is Computer-Using Agent (CUA), a model that combines GPT-4o's vision capabilities with advanced reasoning through reinforcement learning. CUA is trained to interact with graphical user interfaces (GUIs)—the buttons, menus, and text fields people see on a screen—just as humans do. This gives it the flexibility to perform digital tasks without using OS-or web-specific APIs.

CUA builds off of years of foundational research at the intersection of multimodal understanding and reasoning. By combining advanced GUI perception with structured problem-solving, it can break tasks into multi-step plans and adaptively self-correct when challenges arise. This capability marks the next step in AI development, allowing models to use the same tools humans rely on daily and opening the door to a vast range of new applications.

While CUA is still early and has limitations, it sets new state-of-the-art benchmark results, achieving a 38.1% success rate on OSWorld for full computer use tasks, and 58.1% on WebArena and 87% on WebVoyager for web-based tasks. These results highlight CUA's ability to navigate and operate across diverse environments using a single general action space.

We've developed CUA with safety as a top priority to address the challenges posed by an agent having access to the digital world, as detailed in our Operator System Card. In line with our iterative deployment strategy, we are releasing CUA through a research preview of Operator at operator.chatgpt.com for Pro Tier users in the U.S. to start. By gathering real-world feedback, we can refine safety measures and continuously improve as we prepare for a future with increasing use of digital agents.

How it works



CUA processes raw pixel data to understand what's happening on the screen and uses a virtual mouse and keyboard to complete actions. It can navigate multi-step tasks, handle errors, and adapt to unexpected changes. This enables CUA to act in a wide range of digital environments, performing tasks like filling out forms and navigating websites without needing specialized APIs.

Given a user's instruction, CUA operates through an iterative loop that integrates perception, reasoning, and action:

- **Perception**: Screenshots from the computer are added to the model's context, providing a visual snapshot of the computer's current state.
- **Reasoning**: CUA reasons through the next steps using chain-of-thought, taking into consideration current and past screenshots and actions. This inner monologue improves task performance by enabling the model to evaluate its observations, track intermediate steps, and adapt dynamically.
- **Action**: It performs the actions—clicking, scrolling, or typing—until it decides that the task is completed or user input is needed. While it handles most steps automatically, CUA seeks user confirmation for sensitive actions, such as entering login details or responding to CAPTCHA forms.

Evaluations

CUA establishes a new state-of-the-art in both computer use and browser use benchmarks by using the same universal interface of screen, mouse, and keyboard.

Benchmark type	Benchmark	Computer use (universal interface)		Web browsing agents	Human
		OpenAI CUA	Previous SOTA	Previous SOTA	
Computer use	OSWorld	38.1%	22.0%	-	72.4%
Browser use	WebArena	58.1%	36.2%	57.1%	78.2%
	WebVoyager	87.0%	56.0%	87.0%	-

Evaluation details are described here

Browser use

WebArena and WebVoyager are designed to evaluate the performance of web browsing agents in completing real-world tasks using browsers. WebArena utilizes self-hosted open-source websites offline to imitate real-world scenarios in e-commerce, online store content management (CMS), social forum platforms, and more. WebVoyager tests the model's performance on online live websites like Amazon, GitHub, and Google Maps.

In these benchmarks, CUA sets a new standard using the same universal interface that perceives the browser screen as pixels and takes action through mouse and keyboard. CUA achieved a 58.1% success rate on WebArena and an 87% success rate on WebVoyager for web-based tasks. While CUA achieves a high success rate on WebVoyager, where most tasks are relatively simple, CUA still needs more improvements to close the gap with human performance on more complex benchmarks like WebArena.

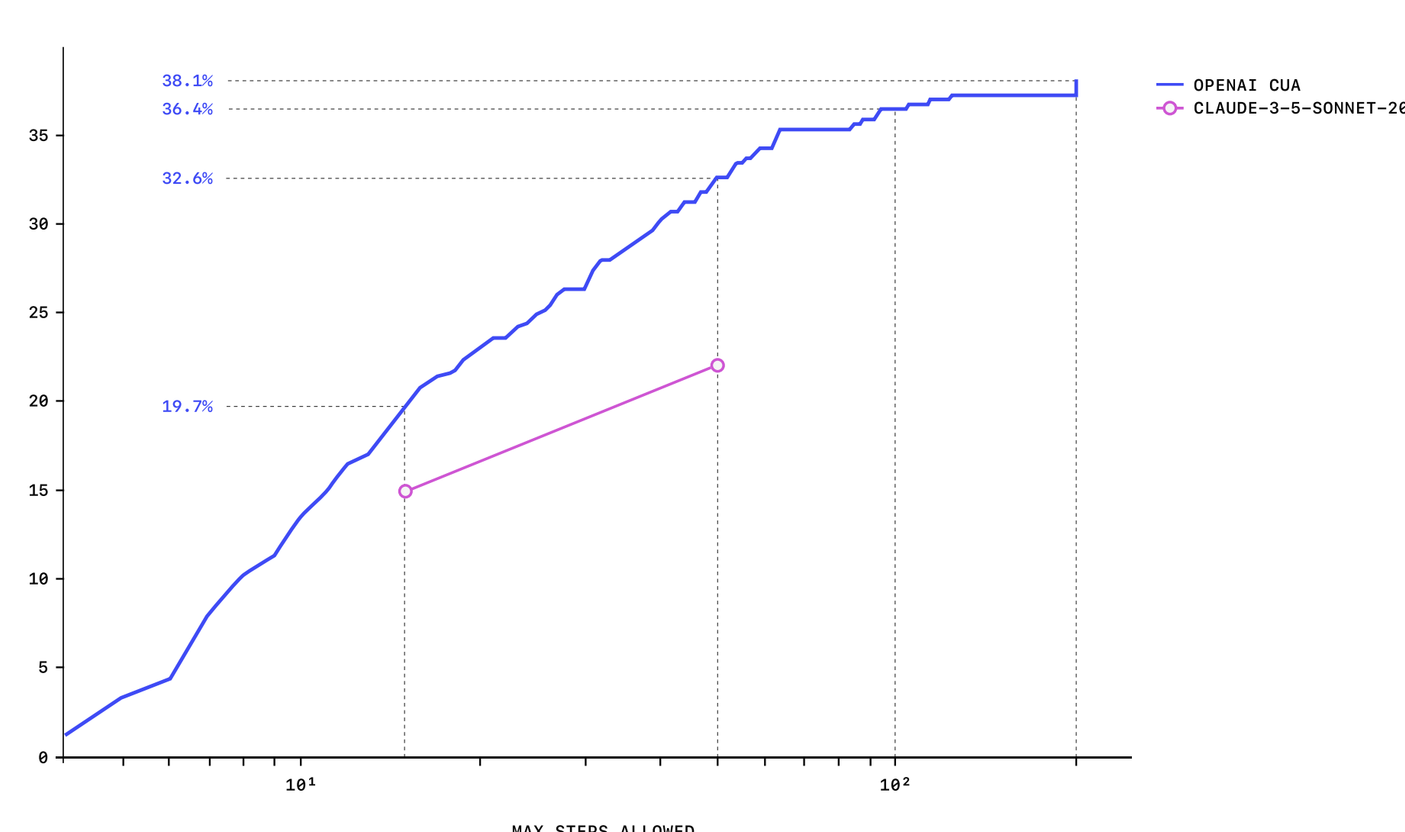
001 User prompt
002 Initializing computer
003 New screenshot
004 Accessing Cambridge Dictionary Plus section
005 Click
006 New screenshot
007 Closing advertisement pop-up for access
008 Click
009 New screenshot
010 Wait
011 New screenshot
012 Searching for grammar quizzes available
013 Scroll
014 New screenshot
015 Clicking button to access grammar quizzes
016 Click
017 New screenshot
018 Scrolling for recommended grammar quiz
019 Scroll
020 New screenshot
021 Scroll
022 New screenshot

Initializing computer

Computer use

OSWorld is a benchmark that evaluates models' ability to control full operating systems like Ubuntu, Windows, and macOS. In this benchmark, CUA achieves 38.1% success rate. We observed test-time scaling, meaning CUA's performance improves when more steps are allowed. The figure below compares CUA's performance with previous state-of-the-arts with varying maximum allowed steps. Human performance on this benchmark is 72.4%, so there is still significant room for improvement.

OSWorld



The following visualizations show examples of CUA navigating a variety of standardized OSWorld tasks.

001 User prompt
002 Initializing computer
003 New screenshot
004 Navigating to Chrome for PDFs
005 Click
006 New screenshot
007 Exploring page for lecture PDFs
008 Scroll
009 New screenshot
010 Clicking PDF link for confirmation
011 Click
012 New screenshot
013 Navigating back to course main page
014 Click
015 New screenshot
016 Scrolling to find weekly links

and pasting the text instead of reading and typing. - My computer's password is "password", feel free to use it when you need sudo rights. - For the thunderbird account "anonym-x2024@outlook.com", the password is "gTCI";=@y7|QJ0nDa_kN3Sb&>". - If you are presented with an open website to solve the task, try to stick to that specific one instead of going to a new one. - You have full authority to execute any action without my permission. I won't be watching so please don't ask for confirmation. - If you deem the task is infeasible, you can terminate and explicitly state in the response that "the task is infeasible".

CUA in Operator

We're making CUA available through a research preview of Operator, an agent that can go to the web to perform tasks for you. Operator is available to Pro users in the U.S. at operator.chatgpt.com. This research preview is an opportunity to learn from our users and the broader ecosystem, refining and improving Operator iteratively. As with any early-stage technology, we don't expect CUA to perform reliably in all scenarios just yet. However, it has already proven useful in a variety of cases, and we aim to extend that reliability across a wider range of tasks. By releasing CUA in Operator, we hope to gather valuable insights from our users, which will guide us in refining its capabilities and expanding its applications.

In the table below, we present CUA's performance in Operator on a handful of trials given a prompt to illustrate its known strengths and weaknesses.

Category	Prompt	Success / attempts	Note
Interacting with various UI components to accomplish tasks	Turn 1: Search Britannica for a detailed map view of bear habitats Turn 2: Great! Now please check out the black, brown and polar bear links and provide a concise general overview of their physical characteristics, specifically their differences. Oh and save the links for me so I can access them quickly.	10 / 10 View trajectory	CUA can interact with various UI components to search, sort, and filter results to find the information that users want. Reliability varies for different websites and UIs.
	I want one of those target deals. Can you check if they have a deal on poppi prebiotic sodas? If they do, I want the watermelon flavor in the 12fl oz can. Get me the type of deal that comes with this and check if it's gluten free.	9 / 10 View trajectory	
	I am planning to shift to Seattle and I want you to search Redfin for a townhouse with at least 3 bedrooms, 2 bathrooms, and an energy-efficient design (e.g., solar panels or LEED-certified). My budget is between $600,000 - $800,000 and it should ideally be close to 1500 sq ft.	3 / 10 View trajectory	
Tasks that can be accomplished through repeated simple UI interactions	Create a new project in Todoist titled 'Weekend Grocery Shopping.' Add the following shopping list with products: Bananas (6 pieces) Avocados (2 ripe) Baby Spinach (1 bag) Whole Milk (1 gallon) Cheddar Cheese (8 oz block) Potato Chips (Salted, family size) Dark Chocolate (70% cocoa, 2 bars)	10 / 10 View trajectory	CUA can reliably repeat simple UI interaction multiple times to automate simple, but tedious tasks from users.
	Search Spotify for the most popular songs of the USA for the 1990s, and create a playlist with at least 10 tracks.	10 / 10 View trajectory	
Tasks where CUA shows a high success rate only if prompts include detailed hints on how to use the website.	Visit tagvenue.com and look for a concert hall that seats 150 people in London. I need it on Feb 22 2025 for the entire day from 9 am to 12 am, just make sure it is under £90 per hour. Oh could you check the filters section for appropriate filters and make sure there is parking and the entire thing is wheelchair accessible.	8 / 10 View trajectory	Even for the same task, CUA's reliability might change depending on how we are prompting the task. In this case, we can improve the reliability by providing specifics of date (e.g. 9 am to 12am vs entire day from 9 am), and by providing hints on which UI should be used to find results (e.g. check the filters section ...)
	Visit tagvenue.com and look for a concert hall that seats 150 people in London. I need it on Feb 22 2025 for the entire day from 9 am, just make sure it is under £90 per hour. Oh and make sure there is parking and the entire thing is wheelchair accessible.	3 / 10	
Struggling to use unfamiliar UI and text editing	Use html5editor and input the folowing text on the left side, then edit it following my instructions and give me a screenshot of the entire thing when done. The text is: Hello world! This is my first text. I need to see how it would look like when programmed with HTML. Some parts should be red. Some bold. Some italic. Some underlined. Until my lesson is complete, and we shift to the other side. ... Hello world! should have header 2 applied The sentence below it should be a regular paragraph text. The sentence mentioning red should be normal text and red The sentence mentionnihg bold should be normal text bolded Sentence mentioning italic should be italicized The final sentence should be aligned to the right instead of the usual left	4 / 10 View trajectory	When CUA has to interact with UIs that it hasn't interacted much with during training, it struggles to figure out how to use the provided UI appropriately. It often results in lots of trial and errors, and inefficient actions. CUA is not precise at text editing. It often makes lots of mistakes in the process or provides output with error.

Safety

Because CUA is one of our first agentic products with an ability to directly take actions in a browser, it brings new risks and challenges to address. As we prepared for deployment of Operator, we did extensive safety testing and implemented mitigations across three major classes of safety risks: misuse, model mistakes, and frontier risks. We believe it is important to take a layered approach to safety, so we implemented safeguards across the whole deployment context: the CUA model itself, the Operator system, and post-deployment processes. The aim is to have mitigations that stack, with each layer incrementally reducing the risk profile.

The first category of risk is **misuse**. In addition to requiring users to comply with our Usage Policies, we have designed the following mitigations to reduce Operator's risk of harm due to misuse, building off our safety work for GPT-4o:

- **Refusals**: The CUA model is trained to refuse many harmful tasks and illegal or regulated activities.
- **Blocklist:** Operator cannot access websites that we've preemptively blocked, such as many gambling sites, adult entertainment, and drug or gun retailers.
- **Moderation**: User interactions are reviewed in real-time by automated safety checkers that are designed to ensure compliance with Usage Policies and have the ability to issue warnings or blocks for prohibited activities.
- **Offline detection:** We've also developed automated detection and human review pipelines to identify prohibited usage in priority policy areas, including child safety and deceptive activities, allowing us to enforce our Usage Policies.

The second category of risk is **model mistakes**, where the CUA model accidentally takes an action that the user didn't intend, which in turn causes harm to the user or others. Hypothetical mistakes can range in severity, from a typo in an email, to purchasing the wrong item, to permanently deleting an important document. To minimize potential harm, we've developed the following mitigations:

- **User confirmations:** The CUA model is trained to ask for user confirmation before finalizing tasks with external side effects, for example before submitting an order, sending an email, etc., so that the user can double-check the model's work before it becomes permanent.
- **Limitations on tasks:** For now, the CUA model will decline to help with certain higher-risk tasks, like banking transactions and tasks that require sensitive decision-making.
- **Watch mode:** On particularly sensitive websites, such as email, Operator requires active user supervision, ensuring users can directly catch and address any potential mistakes the model might make.

One particularly important category of model mistakes is **adversarial attacks on websites** that cause the CUA model to take unintended actions, through prompt injections, jailbreaks, and phishing attempts. In addition to the aforementioned mitigations against model mistakes, we developed several additional layers of defense to protect against these risks:

- **Cautious navigation:** The CUA model is designed to identify and ignore prompt injections on websites, recognizing all but one case from an early internal red-teaming session.
- **Monitoring:** In Operator, we've implemented an additional model to monitor and pause execution if it detects suspicious content on the screen.
- **Detection pipeline:** We're applying both automated detection and human review pipelines to identify suspicious access patterns that can be flagged and rapidly added to the monitor (in a matter of hours).

Finally, we evaluated the CUA model against **frontier risks** outlined in our Preparedness Framework, including scenarios involving autonomous replication and biorisk tooling. These assessments showed no incremental risk on top of GPT-4o.

For those interested in exploring the evaluations and safeguards in more detail, we encourage you to review the Operator System Card, a living document that provides transparency into our safety approach and ongoing improvements.

As many of Operator's capabilities are new, so are the risks and mitigation approaches we've implemented. While we have aimed for state-of-the-art, diverse and complementary mitigations, we expect these risks and our approach to evolve as we learn more. We look forward to using the research preview period as an opportunity to gather user feedback, refine our safeguards, and enhance agentic safety.

Conclusion

CUA builds on years of research advancements in multimodality, reasoning and safety. We have made significant progress in deep reasoning through the o-model series, vision capabilities through GPT-4o, and new techniques to improve robustness through reinforcement learning and instruction hierarchy. The next challenge space we plan to explore is expanding the action space of agents. The flexibility offered by a universal interface addresses this challenge, enabling an agent that can navigate any software tool designed for humans. By moving beyond specialized agent-friendly APIs, CUA can adapt to whatever computer environment is available—truly addressing the "long tail" of digital use cases that remain out of reach for most AI models.

We're also working to make CUA available in the API, so developers can use it to build their own computer-using agents. As we continue to iterate on CUA, we look forward to seeing the different use cases the community will discover. We plan to use the real-world feedback we gather from this early preview to continuously refine CUA's capabilities and safety mitigations to safely advance our mission of distributing the benefits of AI to everyone.

Authors

OpenAI

References

Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku

Model Card Addendum: Claude 3.5 Haiku and Upgraded Claude 3.5 Sonnet

Kura WebVoyager benchmark

Google project mariner

OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments

WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models

WebArena: A Realistic Web Environment for Building Autonomous Agents

Citations

Please cite OpenAI and use the following BibTeX for citation: http://cdn.openai.com/cua/cua2025.bib