CUA eval extra information

OpenAI

January 2025

This document includes extra information to how we evaluated our Computer Using Agent, including (browser/VM) environments, prompts, sampling parameters, and scoring procedures. For more details, read https://openai.com/index/computer-using-agent/.

1 Environment

- For WebArena and WebVoyager, we run the evals in operator browser instead of playwright browsers since our model relies on the visual action space for navigation (search bar, backward/forward button). Our model does not have access to tool calls that control the navigation.
- For OSWorld, we use the VMWare Ubuntu VM distributed by the authors. Our environment has the dock on the right side of the screen instead of the left side, which we have found to improve the performance slightly.

2 Prompting

2.1 WebArena

For WebArena, we used per website prompts inspired by the step paper to 1. help with scenarios that lack the knowledge of specific websites in this benchmark, and 2. let the model understand the expected response format for different tasks.

2.1.1 Shopping admin example prompt:

Initialize computer and solve the following task: What is the top-1 best-selling product in 2022
The following websites are available at: magento: http://magento.site/admin
All you need is on the provided websites. Start the task from the following URL: http://magento.site/admin
Here are tips for using the magento.site/admin website:

- When you add a new product in the CATALOG > Products tab, you can click the downward arrow beside the "Add Product" button to select options like "Simple Product", "Configurable Product", etc.
- If you need to add new attribute values (e.g. size, color, etc) to a product, you can find the product at CATALOG > Products, search for the product, edit product with "Configurable Product" type, and use "Edit Configurations" to add the product with new attribute values. If the value that you want does not exist, you may need to add new values to the attribute.
- If you need to add new values to product attributes (e.g. size, color, etc), you can visit STORES > Attributes > Product, find the attribute and click, and add value after clicking "Add Swatch"

button.

- You can generate various reports by using menus in the REPORTS tab. Select REPORTS > "report type", select options, and click "Show Report" to view report.
- You can generate various reports by using menus in the REPORTS tab. Select REPORTS > "report type", select options, and click "Show Report" to view report.
- In this website, there is a UI that looks like a dropdown, but is just a 1-of-n selection menu. For example in REPORTS > Orders, if you select "Specified" Order Status, you will choose one from many options (e.g. Canceled, Closed, ...), but it's not dropdown, so your click will just highlight your selection (1-of-n select UI will not disappear).
- Configurable products have some options that you can mark as "on" of "off". For example, the options may include "new", "sale", "eco collection", etc.
- You can find all reviews and their counts in the store in MARKETING > User Content > All Reviews. If you see all reviews grouped by product, go REPORTS > By Products and search by Product name.
- This website has been operating since 2022. So if you have to find a report for the entire history, you can select the date from Jan 1, 2022, to Today.
- Do not export or download files, or try to open files. It will not work.

Here are rules for providing the answer: If the objective is to find a text-based answer, do not use computer_output_citation, instead provide the answer in the last message with following quoted format ‘‘‘Answer:<your answer>‘‘‘
Important notes about the answer format: - DO NOT RESPOND WITH ANYTHING ELSE OTHER THAN THIS FORMAT. DO NOT ASK ME IF I NEED ANYTHING ELSE. I JUST NEED THE ANSWER IN THIS FORMAT. - Importantly, there is no empty space between "Answer:" and <your answer>. - You should include ‘‘‘ in your response. For example, you should write "‘‘‘Answer:42‘‘‘" instead of "Answer:42". - If you do not write in this format, you get no reward at all!!! - Keep the answer as short and concise as possible. For example, if the answer is "42", instead of writing "based on the results, I believe the answer is 42.", you should just write:‘‘‘Answer:42‘‘‘
Keep going if the answer is not found. DO NOT ask the user any question if you encounter an issue! You have the full authority until the task is completed. If you believe the task is impossible to complete, provide the following answer: ‘‘‘Answer:N/A‘‘‘. When asked to return a count, return the count as a number instead of N/A if it's 0.

2.1.2 Map example prompt:

Initialize computer and solve the following task: I will arrive Pittsburgh Airport soon. Provide the name of a Hilton hotel in the vicinity, if available. Then, tell me the the walking distance to the nearest supermarket own by a local company from the hotel.
The following websites are available at: openstreetmap: http://10.138.0.12
All you need is on the provided websites. Start the task from the following URL: http://10.138.0.12
To find routes from point A to B, you can use the 'two arrow' button on the right hand side of the 'Go' button. If the name of a location is not well known, you should look up the one in Pittsburgh, PA. If the question has a yes/no answer, include yes or no in the answer. If your answer contains an address, please use the format 'OpenAI, 575 Florida Street, Mission District, San Francisco' (landmark, street, district, city). If the answer contains a phone number, please use the format '8002428478' (no country

calling codes). If the answer contains a distance, please use the format '5km', '645m' (no space between the number and the unit). Here are rules for providing the answer: If the objective is to find a text-based answer, do not use computer_output_citation, instead provide the answer in the last message with following quoted format ‘‘‘Answer:<your answer>‘‘‘
Important notes about the answer format: - DO NOT RESPOND WITH ANYTHING ELSE OTHER THAN THIS FORMAT. DO NOT ASK ME IF I NEED ANYTHING ELSE. I JUST NEED THE ANSWER IN THIS FORMAT. - Importantly, there is no empty space between "Answer:" and <your answer>. - You should include ‘‘‘ in your response. For example, you should write "‘‘‘Answer:42‘‘‘" instead of "Answer:42". - If you do not write in this format, you get no reward at all!!! - Keep the answer as short and concise as possible. For example, if the answer is "42", instead of writing "based on the results, I believe the answer is 42.", you should just write:‘‘‘Answer:42‘‘‘
Keep going if the answer is not found. DO NOT ask the user any question if you encounter an issue! You have the full authority until the task is completed. If you believe the task is impossible to complete, provide the following answer: ‘‘‘Answer:N/A‘‘‘. When asked to return a count, return the count as a number instead of N/A if it's 0.

2.1.3 Shopping example prompt:

Initialize computer and solve the following task: What is the price range of teeth grinding mouth guard in the One Stop Market?
The following websites are available at: one-stop-shop: http://one-stop-shop.site
All you need is on the provided websites. Start the task from the following URL: http://one-stop-shop.site
Here are tips for using the one-stop-shop website:

- This website provides very detailed category of products. You can hover categories on the top menu to see subcategories.
- If you need to find information about your previous purchases, you can go My Account > My Orders, and find order by date, order number, or any other available information
- An order is considered out of delivery if it is marked as "processing" in the order status
- When the task asks you to draft and email. DO NOT send the email. Just draft it and provide the content in the last message

Here are rules for providing the answer: If the objective is to find a text-based answer, do not use computer_output_citation, instead provide the answer in the last message with following quoted format ‘‘‘Answer:<your answer>‘‘‘
Important notes about the answer format: - DO NOT RESPOND WITH ANYTHING ELSE OTHER THAN THIS FORMAT. DO NOT ASK ME IF I NEED ANYTHING ELSE. I JUST NEED THE ANSWER IN THIS FORMAT. - Importantly, there is no empty space between "Answer:" and <your answer>. - You should include ‘‘‘ in your response. For example, you should write "‘‘‘Answer:42‘‘‘" instead of "Answer:42". - If you do not write in this format, you get no reward at all!!! - Keep the answer as short and concise as possible. For example, if the answer is "42", instead of writing "based on the results, I believe the answer is 42.", you should just write:‘‘‘Answer:42‘‘‘
Keep going if the answer is not found. DO NOT ask the user any question if you encounter an issue! You have the full authority until the task is completed. If you believe the task is impossible to complete, provide the following answer: ‘‘‘Answer:N/A‘‘‘. When asked to return a count, return the count as a number instead of N/A if it's 0.

2.1.4 Reddit example prompt:

Initialize computer and solve the following task: Tell me the count of comments that have received more downvotes than upvotes for the user who made the latest post on the Worcester forum.
The following websites are available at: reddit: http://reddit.site
All you need is on the provided websites. Start the task from the following URL: http://reddit.site
Here are tips for using the reddit website:

- when the task mentions subreddit, it is referring to 'forum'
- if you need find a relevant subreddit or forum, you can find the name after clicking "alphabetical" in the "Forum" tab.
- if you have to find submissions (posts) or comments by a particular user, visit reddit.site/user/<user_name> to see the list

Here are rules for providing the answer: If the objective is to find a text-based answer, do not use computer_output_citation, instead provide the answer in the last message with following quoted format ‘‘‘Answer:<your answer>‘‘‘
Important notes about the answer format: - DO NOT RESPOND WITH ANYTHING ELSE OTHER THAN THIS FORMAT. DO NOT ASK ME IF I NEED ANYTHING ELSE. I JUST NEED THE ANSWER IN THIS FORMAT. - Importantly, there is no empty space between "Answer:" and <your answer>. - You should include ‘‘‘ in your response. For example, you should write "‘‘‘Answer:42‘‘‘" instead of "Answer:42". - If you do not write in this format, you get no reward at all!!! - Keep the answer as short and concise as possible. For example, if the answer is "42", instead of writing "based on the results, I believe the answer is 42.", you should just write:‘‘‘Answer:42‘‘‘
Keep going if the answer is not found. DO NOT ask the user any question if you encounter an issue! You have the full authority until the task is completed. If you believe the task is impossible to complete, provide the following answer: ‘‘‘Answer:N/A‘‘‘. When asked to return a count, return the count as a number instead of N/A if it's 0.

2.1.5 GitLab example prompt:

Initialize computer and solve the following task: Check out the most recent open issues
The following websites are available at: gitlab: http://gitlab.site
All you need is on the provided websites. Start the task from the following URL: http://gitlab.site/a11yproject/a11yproject.com
Here are tips for using the gitlab website:

- your user name is byteblaze
- To add new members to the project, you can visit project information > members tab and click blue "invite members" button on top right
- To set your status, click profile button on top right corner of the page (it's next to the question mark button) and click edit status
- To edit your profile, click profile button on top right corner of the page (it's next to the question mark button) and click edit profile
- You can also access to your information e.g. access token, notifications, ssh keys and more from "edit profile" page

- Projects that you have contributed to are listed under Project / Yours / All tab of gitlab.site. You can sort repos using dropdown button on top right
- Projects's repository tab has menus like Commits, Branches, Contributors, and more. Contributors tab shows contributors and their number of commits
- If you want to see all the issues for you, you can either click button on the right of + icon on top right menu bar
- When the task mentions branch main, it often means master

Here are rules for providing the answer: If the objective is to find a text-based answer, do not use computer_output_citation, instead provide the answer in the last message with following quoted format ‘‘‘Answer:<your answer>‘‘‘
Important notes about the answer format: - DO NOT RESPOND WITH ANYTHING ELSE OTHER THAN THIS FORMAT. DO NOT ASK ME IF I NEED ANYTHING ELSE. I JUST NEED THE ANSWER IN THIS FORMAT. - Importantly, there is no empty space between "Answer:" and <your answer>. - You should include ‘‘‘ in your response. For example, you should write "‘‘‘Answer:42‘‘‘" instead of "Answer:42". - If you do not write in this format, you get no reward at all!!! - Keep the answer as short and concise as possible. For example, if the answer is "42", instead of writing "based on the results, I believe the answer is 42.", you should just write:‘‘‘Answer:42‘‘‘ Keep going if the answer is not found. DO NOT ask the user any question if you encounter an issue! You have the full authority until the task is completed. If you believe the task is impossible to complete, provide the following answer: ‘‘‘Answer:N/A‘‘‘. When asked to return a count, return the count as a number instead of N/A if it's 0.

2.2 OSWorld

For OSWorld, we included some universal prompts to 1. inform it about some unavailable tools, 2. pass in various credentials, 3. encourage/discourage certain model behaviors, and 4. let the model understand the expected response format.

2.2.1 Example prompt:

Please do the following task: Could you tone down the brightness of my photo?
Here are some helpful tips:

- computer.clipboard, computer.sync_file, computer.sync_shared_folder, computer.computer_output_citation are disabled.
- If you worry that you might make typo, prefer copying 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".

2.3 WebVoyager

For WebVoyager, we directly prompt the model with the questions in the benchmark without any additional information.

3 Sampling

All results are obtained by pass@1 autoregressive sampling. By default, we use temperature 0.6 and maximum 200 steps during sampling unless otherwise specified.

4 Scoring

For WebVoyager, we used the automatic evaluation protocol using gpt4o. Since WebVoyager is based on real websites whose contents change over time, some tasks become outdated or broken over time and then evaluation results can change as well. At the time of our evaluation, we removed 35 broken tasks. Since previous results on this benchmark were obtained in different time, with different sets of removed tasks, results may not be directly comparable to each other.