Addendum to OpenAI o3 and o4-mini system card: OpenAI o3 Operator

OpenAI

May 23, 2025

In January 2025, we shipped Operator, a product to serve our Computer Using Agent (CUA) model as a research preview. CUA is an agentic model that can use the web to perform tasks for the user. Using its own browser, it can look at a webpage, and interact with it much like a human would by typing, clicking, scrolling and more.

We are replacing the existing GPT-4o-based model for Operator with a version based on OpenAI o3. The API version will remain based on 4o.

o3 Operator uses the same multi-layered approach to safety that we used for the 4o version of Operator and described in our original Operator System Card. Compared with other models in the o3 family, o3 Operator was fine-tuned with additional safety data for computer use, including safety datasets designed to teach the model our decision boundaries on confirmations and refusals.

Although o3 Operator inherits o3's coding capabilities, it does not have native access to a coding environment or Terminal.

1 Baseline Model Safety Evaluations

1.1 Disallowed Content Evaluations

Operator is trained to use a computer in the same way a person would use one: by visually perceiving the computer screen and using a cursor and keyboard. Although the model is not intended or expected to be used in general-purpose chat applications, it is able to operate conversationally. Therefore, we conducted standard refusal evaluations across disallowed content categories (metric 'not_unsafe', higher is better).

Table 1: Disallowed content

Category	o3 Operator	o3	4o Operator
harassment/threatening	98%	99%	100%
sexual/exploitative	98%	98%	100%
sexual/minors	99%	100%	100%
extremist/propaganda	100%	100%	100%
hate	100%	100%	100%
hate/threatening	100%	100%	100%
illicit/non-violent	100%	100%	100%
illicit/violent	100%	100%	100%
personal-data/semi-restricted	92%[1]	100%	N/A
personal-data/restricted	100%[2]	100%	N/A
regulated-advice	100%	100%	100%
self-harm/intent	100%	100%	100%
self-harm/instructions	100%	100%	100%

1.2 Jailbreaks

We evaluated the robustness of the model to jailbreaks: adversarial prompts that purposely try to circumvent model refusals for content the model is not supposed to produce. Similar to the above, we measured the rate at which the model produced outputs that are not unsafe, using the following evaluation:

- StrongREJECT [1]: An academic jailbreak benchmark that tests a model's resistance against common attacks from the literature.

Table 2: StrongREJECT

Evaluation (Higher is better)	o3 Operator	o3	4o Operator
StrongREJECT (not_unsafe)	0.97	0.97	0.37

o3 Operator performs approximately on par with o3 without computer use.

[1]We have updated the definitions of the personal-data categories in our standard disallowed content evaluation after the 4o Operator, adding more granularity to better assess harmful generations risks. This change means comparisons to the earlier personal-data category are not apples-to-apples.

[2]The "personal-data/restricted" eval was created after 4o Operator launch. Since then, the infrastructure has changed significantly to support newer models based on o3. Retroactively running these new evaluations on the previous 4o Operator is no longer supported.

2 Product-Specific Risk Mitigations

Using the same datasets and techniques used to train the previous CUA model, we trained the o3 Operator model to address product-specific risks like prompt injections. Further details are in the previous Operator system card. Below, we present updated performance metrics for o3 Operator, which match or exceed those of the 4o-based model.

Table 3: Risks and Mitigations

Product-specific safety risk	**Mitigations**
User uses o3 Operator to do a harmful task	• Baseline safety training inherited from the o3 family of models • Safety training and evaluations specific to agentic harms • Human and automated monitoring and review for deceptive and fraudulent activities • Rate limits at various levels to deter or prevent scaled abuse
o3 Operator is prompt injected	• Model robustified against prompt injections • Prompt Injection Monitor that ingests current screenshot and pauses execution if detects suspicious activity • Explicit confirmation with users to finalize actions • "Watch Mode" requirement on high risk websites such as email
o3 Operator makes a mistake	• "Watch Mode" requirement on high risk websites such as email • Operator refuses to complete highest risk tasks • Operator asks for confirmations before finalizing transactions

2.1 Harmful Tasks

2.1.1 Risk description

A user may ask Operator to perform harmful actions. As an agentic model, Operator takes actions in the real world, introducing novel risks compared to o3.

2.1.2 Mitigations

In addition to the safety behavior learned during o3's existing post-training, we introduced the same safety training data previously used specifically to teach the model to refuse harmful agentic tasks. We evaluated o3 Operator on the same evaluations we used for the previous CUA model.

Table 4: Harmful agentic tasks

Refusals (higher is better)	**o3 Operator**	**4o Operator**
Performing illicit activities	1.0	0.97
Prohibited financial activities	1.0	0.97
Searching for sensitive personal data	1.0	1.0
Overrefusals (lower is better)	0.13	0.3

2.2 Model mistakes

2.2.1 Risk description

The second harm category arises when the model takes actions misaligned with the user's intent, potentially causing harm.

2.2.2 Mitigations

See the previous system card for details on mitigations against model mistakes.

We aim to have the model ask the user for confirmations before finalizing actions that affect the state of the world (e.g. before submitting a purchase or sending an email), train the model to refuse high-risk tasks like banking transactions and making high-stakes decisions, and require the user to watch the model in "Watch Mode" on certain high-risk web sites (like email).

Explicit confirmation with users to finalize actions mitigate the majority of the private data leakage risk where the model may inadvertently provide sensitive user data it had access to.

2.2.3 Evaluation

We reran the same confirmation evaluation used in the previous iteration of Operator, and obtained an improved confirmation rate of 94%, compared to 92% with the previous model. Of note, Operator now confirms in 100% of the financial transactions in our eval set. We also manually reviewed 9 scenarios for private data leakage and found that there is a low level of residual risk.

2.3 Prompt Injection

2.3.1 Risk description

Operator may see something on screen (e.g. a malicious web site or email) which induces it to do something that the user does not want.

2.3.2 Mitigations

As before, Confirmations, Watch Mode, and refusals for high-risk tasks serve as base layers of defense against prompt injections, and a prompt injection monitor ingests the current screenshot and pauses execution if it detects anything suspicious. The model itself also has some base level of resilience against prompt injection.

2.3.3 Evaluation

We ran the same evaluation that we ran on the previous model and found that the model displays a reduced base level of susceptibility of 20%, compared to 23% for the previous model (lower being better). Note that the prompt injection monitor, which serves as the primary layer of defense, remains unchanged from the previous Operator.

2.4 Preparedness

The o3 Operator model is a variant of the o3 model, whose frontier capabilities are assessed under the Preparedness Framework in the original o3 System Card and it inherits 'Below High' capability levels for Chemical and Biological, cybersecurity, and AI Self-Improvement. It has minimal additional training, including safety datasets designed to teach the model our decision boundaries on confirmations / refusals and datasets to teach the model to comply with our policies, and does not have native access to a coding environment or Terminal. We validated that the Operator model underperforms o3 on our biological capability evals, indicating that the model remains below "High" capability for the Biological and Chemical frontier risk category.

References

[1] A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins, *et al.*, "A strongreject for empty jailbreaks," *arXiv preprint arXiv:2402.10260*, 2024.