

OpenAI o3 and o4-mini System Card

OpenAI

April 16, 2025

1 Introduction

OpenAI o3 and OpenAI o4-mini combine state-of-the-art reasoning with full tool capabilities — web browsing, Python, image and file analysis, image generation, canvas, automations, file search, and memory. These models excel at solving complex math, coding, and scientific challenges while demonstrating strong visual perception and analysis. The models use tools in their chains of thought to augment their capabilities; for example, cropping or transforming images, searching the web, or using Python to analyze data during their thought process.

The OpenAI o-series models are trained with large-scale reinforcement learning on chains of thought. These advanced reasoning capabilities provide new avenues for improving the safety and robustness of our models. In particular, our models can reason about our safety policies in context when responding to potentially unsafe prompts, through deliberative alignment [1]¹.

This is the first launch and system card to be released under Version 2 of our [Preparedness Framework](#). OpenAI’s Safety Advisory Group (SAG) reviewed the results of our Preparedness evaluations and determined that OpenAI o3 and o4-mini do not reach the High threshold in any of our three Tracked Categories: Biological and Chemical Capability, Cybersecurity, and AI Self-improvement. We describe these evaluations below, and provide an update on our work to mitigate risks in these areas.

2 Model Data and Training

OpenAI reasoning models are trained to reason through reinforcement learning. Models in the o-series family are trained to think before they answer: they can produce a long internal chain of thought before responding to the user. Through training, these models learn to refine their thinking process, try different strategies, and recognize their mistakes. Reasoning allows these models to follow specific guidelines and model policies we’ve set, helping them act in line with our safety expectations. This means they provide more helpful answers and better resist attempts to bypass safety rules.

Like OpenAI’s other o-series models, OpenAI o3 and o4-mini were trained on diverse datasets, including information that is publicly available on the internet, information that we partner with third parties to access, and information that our users or human trainers and researchers provide

¹[Deliberative alignment](#) is a training approach that teaches LLMs to explicitly reason through safety specifications before producing an answer.