

# OpenAI launches Operator, an AI agent that can operate your computer

While it's working, Operator shows a miniature browser window of its actions.

However, the technology behind Operator is still relatively new and far from perfect. The model reportedly performs best at repetitive web tasks like creating shopping lists or playlists. It struggles more with unfamiliar interfaces like tables and calendars, and does poorly with complex text editing (with a 40 percent success rate), according to OpenAI's internal testing data.

OpenAI reported the system achieved an 87 percent success rate on the WebVoyager benchmark, which tests live sites like Amazon and Google Maps. On WebArena, which uses offline test sites for training autonomous agents, Operator's success rate dropped to 58.1 percent. For computer operating system tasks, CUA set an apparent record of 38.1 percent success on the OSWorld benchmark, surpassing previous models but still falling short of human performance at 72.4 percent.

With this imperfect research preview, OpenAI hopes to gather user feedback and refine the system's capabilities. The company acknowledges CUA won't perform reliably in all scenarios but plans to improve its reliability across a wider range of tasks through user testing.

# Safety and privacy concerns

For any AI model that can see how you operate your computer and even control some aspects of it, privacy and safety are very important. OpenAI says it built multiple safety controls into Operator, requiring user confirmation before completing sensitive actions like sending emails or making purchases. Operator also has limits on what it can browse, set by OpenAI. It cannot access certain website categories, including gambling and adult content.

Traditionally, AI models based on large language model-style Transformer technology like Operator have been relatively easy to fool with jailbreaks and prompt injections.

To catch attempts at subverting Operator, which might hypothetically be embedded in websites that the AI model browses, OpenAI says it has implemented real-time moderation and detection systems. OpenAI reports the system recognized all but one case of prompt injection attempts during an early internal red-teaming session.

<https://arstechnica.com/ai/2025/01/openai-launches-operator-an-ai-agent-that-can-operate-your-computer/>