MIT Technology Review

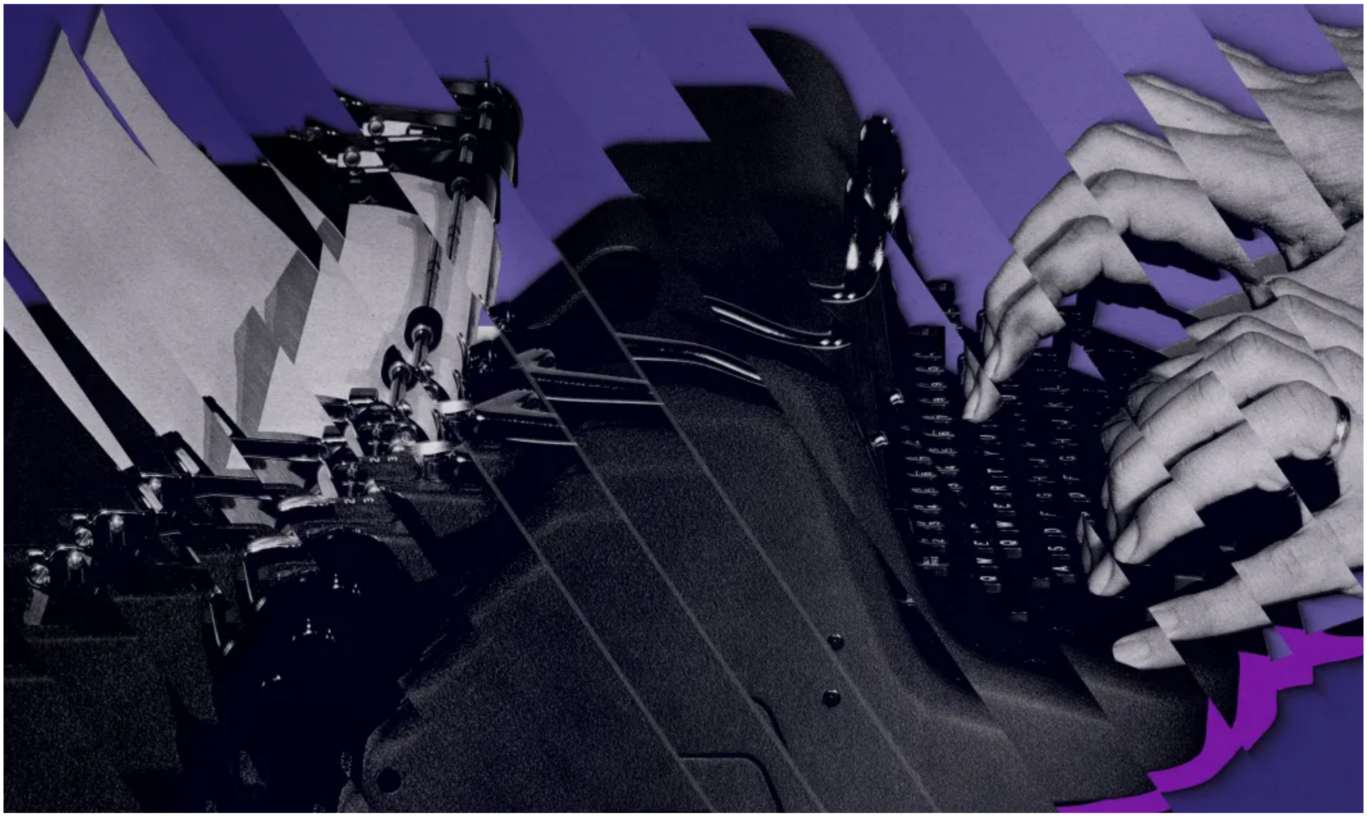
ARTIFICIAL INTELLIGENCE

Anthropic's chief scientist on 4 ways agents will be even better in 2025

The hottest topic in AI is only just getting started.

By Melissa Heikkilä & Will Douglas Heaven

January 11, 2025



STEPHANIE ARNETT/MIT TECHNOLOGY REVIEW | GETTY

Agents are the hottest thing in tech right now. Top firms from Google DeepMind to OpenAI to Anthropic are racing to augment large language models with the ability to carry out tasks by themselves. Known as agentic AI in industry jargon, such systems have fast become the new target of Silicon Valley buzz. Everyone from Nvidia to Salesforce is talking about how they are going to upend the industry.

"We believe that, in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies," Sam Altman claimed in a blog post last week.

In the broadest sense, an agent is a software system that goes off and does something, often with minimal to zero supervision. The more complex that thing is, the smarter the agent needs to be. For many, large language models are now smart enough to power agents that can do a whole range of useful tasks for us, such as filling out forms, looking up a recipe and adding the ingredients to an online grocery basket, or using a search engine to do last-minute research before a meeting and producing a quick bullet-point summary.

In October, Anthropic showed off one of the most advanced agents yet: an extension of its Claude large language model called computer use. As the name suggests, it lets you direct Claude to use a computer much as a person would, by moving a cursor, clicking buttons, and typing text. Instead of simply having a conversation with Claude, you can now ask it to carry out on-screen tasks for you.

Related Story

What are AI agents?

The next big thing is AI tools that can do more complex tasks. Here's how they will work.

Anthropic notes that the feature is still cumbersome and error-prone. But it is already available to a handful of testers, including third-party developers at companies such as DoorDash, Canva, and Asana.

Computer use is a glimpse of what's to come for agents. To learn what's coming next, *MIT Technology Review* talked to Anthropic's cofounder and chief scientist Jared Kaplan. Here are four ways that agents are going to get even better in 2025.

(Kaplan's answers have been lightly edited for length and clarity.)

1/ Agents will get better at using tools

"I think there are two axes for thinking about what AI is capable of. One is a question of how complex the task is that a system can do. And as AI systems get smarter, they're getting better in that direction. But another direction that's very relevant is what kinds of environments or tools the AI can use.

"So, like, if you go back almost 10 years now to [DeepMind's Go-playing model] AlphaGo, we had AI systems that were superhuman in terms of how well they could play board games. But if all you can work with is a board game, then that's a very restrictive environment. It's not actually useful, even if it's very smart. With text models, and then multimodal models, and now computer use—and perhaps in the future with robotics—you're moving toward bringing AI into different situations and tasks, and making it useful.

"We were excited about computer use basically for that reason. Until recently, with large language models, it's been necessary to give them a very specific prompt, give them very specific tools, and then they're restricted to a specific kind of environment. What I see is that computer use will probably improve quickly in terms of how well models can do different tasks and more complex tasks. And also to realize when they've made mistakes, or realize when there's a high-stakes question and it needs to ask the user for feedback."

2/ Agents will understand context

"Claude needs to learn enough about your particular situation and the constraints that you operate under to be useful. Things like what particular role you're in, what styles of writing or what needs you and your organization have.

"I think that we'll see improvements there where Claude will be able to search through things like your documents, your Slack, etc., and really learn what's useful for you. That's underemphasized a bit with agents. It's necessary for systems to be not only useful but also safe, doing what you expected.

"Another thing is that a lot of tasks won't require Claude to do much reasoning. You don't need to sit and think for hours before opening Google Docs or something. And so I think that a lot of what we'll see is not just more reasoning but the application of reasoning when it's really useful and important, but also not wasting time when it's not necessary."

ANTHROPIC

3/ Agents will make coding assistants better

"We wanted to get a very initial beta of computer use out to developers to get feedback while the system was relatively primitive. But as these systems get better, they might be more widely used and really collaborate with you on different activities.

"I think DoorDash, the Browser Company, and Canva are all experimenting with, like, different kinds of browser interactions and designing them with the help of AI.

"My expectation is that we'll also see further improvements to coding assistants. That's something that's been very exciting for developers. There's just a ton of interest in using Claude 3.5 for coding, where it's not just autocomplete like it was a couple of years ago. It's really understanding what's wrong with code, debugging it—running the code, seeing what happens, and fixing it."

4/ Agents will need to be made safe

"We founded Anthropic because we expected AI to progress very quickly and [thought] that, inevitably, safety concerns were going to be relevant. And I think that's just going to become more and more visceral this year, because I think these agents are going to become more and more integrated into the work we do. We need to be ready for the challenges, like prompt injection.

[Prompt injection is an attack in which a malicious prompt is passed to a large language model in ways that its developers did not foresee or intend. One way to do this is to add the prompt to websites that models might visit.]

"Prompt injection is probably one of the No.1 things we're thinking about in terms of, like, broader usage of agents. I think it's especially important for computer use, and it's something we're working on very actively, because if computer use is deployed at large scale, then there could be, like, pernicious websites or something that try to convince Claude to do something that it shouldn't do.

"And with more advanced models, there's just more risk. We have a robust scaling policy where, as AI systems become sufficiently capable, we feel like we need to be able to really prevent them from being misused. For example, if they could help terrorists—that kind of thing.

"So I'm really excited about how AI will be useful—it's actually also accelerating us a lot internally at Anthropic, with people using Claude in all kinds of ways, especially with coding. But, yeah, there'll be a lot of challenges as well. It'll be an interesting year."

by Melissa Heikkilä & Will Douglas Heaven

DEEP DIVE