Hype grows over “autonomous” AI agents that loop GPT-4 outputs

Enlarge / An AI-generated image of a “self-improving robot.” Midjourney

reader comments

87 with

Share this story

Since the launch of OpenAI’s GPT-4 API last month to beta testers, a loose group of developers have been experimenting with making agent-like (“agentic”) implementations of the AI model that attempt to carry out multistep tasks with as little human intervention as possible. These homebrew scripts can loop, iterate, and spin off new instances of an AI model as needed.

Two experimental open source projects, in particular, have captured much attention on social media, especially among those who hype AI projects relentlessly: Auto-GPT, created by Toran Bruce Richards, and BabyAGI, created by Yohei Nakajima.

What do they do? Well, right now, not very much. They need a lot of human input and hand-holding along the way, so they’re not yet as autonomous as promised. But they represent early steps toward more complex chaining AI models that could potentially be more capable than a single AI model working alone.

“Autonomously achieve whatever goal you set”

Richards bills his script as “an experimental open source application showcasing the capabilities of the GPT-4 language model.” The script “chains together LLM ‘thoughts’ to autonomously achieve whatever goal you set.”

Basically, Auto-GPT takes output from GPT-4 and feeds it back into itself with an improvised external memory so that it can further iterate on a task, correct mistakes, or suggest improvements. Ideally, such a script could serve as an AI assistant that could perform any digital task by itself.

To test these claims, we ran Auto-GPT (a Python script) locally on a Windows machine. When you start it, it asks for a name for your AI agent, a description of its role, and a list of five goals it attempts to fulfill. While setting it up, you need to provide an OpenAI API key and a Google search API key. When running, Auto-GPT asks for permission to perform every step it generates by default, although it also includes a fully automatic mode if you’re feeling adventurous.

Advertisement

If tasked to do something like “Purchase a vintage pair of Air Jordans,” Auto-GPT will develop a multistep plan and attempt to execute it. For example, it might search for shoe sellers, then look for a specific pair that meets your criteria. But that’s when it stops because it can’t actually buy anything—at the moment. If hooked into an appropriate purchasing API, that could be possible.

If you want to get a taste of what Auto-GPT does yourself, someone created a web-based version called AgentGPT that functions in a similar way.

Richards has been very open about his goal with Auto-GPT: to develop a form of AGI (artificial general intelligence). In

AI, “general intelligence” typically refers to the still-hypothetical ability of an AI system to perform a wide range of tasks and solve problems that are not specifically programmed or trained for.

Enlarge / A screenshot of AgentGPT, based on Auto-GPT, executing a task of attempting to buy a vintage pair of Air Jordan shoes.
Ars Technica

Like a reasonably intelligent human, a system with general intelligence should be able to adapt to new situations and learn from experience, rather than just following a set of pre-defined rules or patterns. This is in contrast to systems with narrow or specialized intelligence (sometimes called “narrow AI”), which are designed to perform specific tasks or operate within a limited range of contexts.

Meanwhile, BabyAGI (which gets its name from an aspirational goal of working toward artificial general intelligence) works in a similar way to Auto-GPT but with a different task-oriented flavor. You can try a version of it on the web at a

site not-so-modestly titled “God Mode.”

Advertisement

Nakajima, the creator of BabyAGI, tells us that he was inspired to create his script after witnessing the “HustleGPT” movement in March, which sought to use GPT-4 to build businesses automatically as a type of AI cofounder, so to speak. “It made me curious if I could build a fully AI founder,” Nakajima says.

Why Auto-GPT and BabyAGI fall short of AGI is due to the limitations of GPT-4 itself. While impressive as a transformer and analyzer of text, GPT-4 still feels restricted to a narrow range of interpretive intelligence, despite some claims that Microsoft has seen “sparks” of AGI-like behaviors in the model. In fact, the limited usefulness of tools like Auto-GPT at the moment may serve as the most potent evidence yet of the current limitations of large language models. Still, that does not mean those limitations will not eventually be overcome.

Also, the issue of confabulations—when LLMs just make things up—may prove a significant limitation to the usefulness of these agent-like assistants. For example, in a Twitter thread, someone used Auto-GPT to generate a report about companies that produce waterproof shoes by searching the web and looking at reviews of each company’s products. At any step along the way, GPT-4 could have potentially “hallucinated” reviews, products, or even entire companies that factored into its analysis.

When asked for useful application of BabyAGI, Nakajima couldn’t come up with substantive examples aside from “Do Anything Machine,” a project build by Garrett Scott that aspires to create a self-executing to-do list, which is currently in development. To be fair, the BabyAGI project is only about a week old. “It’s more of an introduction to a framework/approach, and what’s most exciting are what people are building on top of this idea,” he says.

https://arstechnica.com/?p=1929067