Science & technology | Generative AI

# Today's AI models are impressive. Teams of them will be formidable

Working together will make LLMs more capable and intelligent—for good and ill

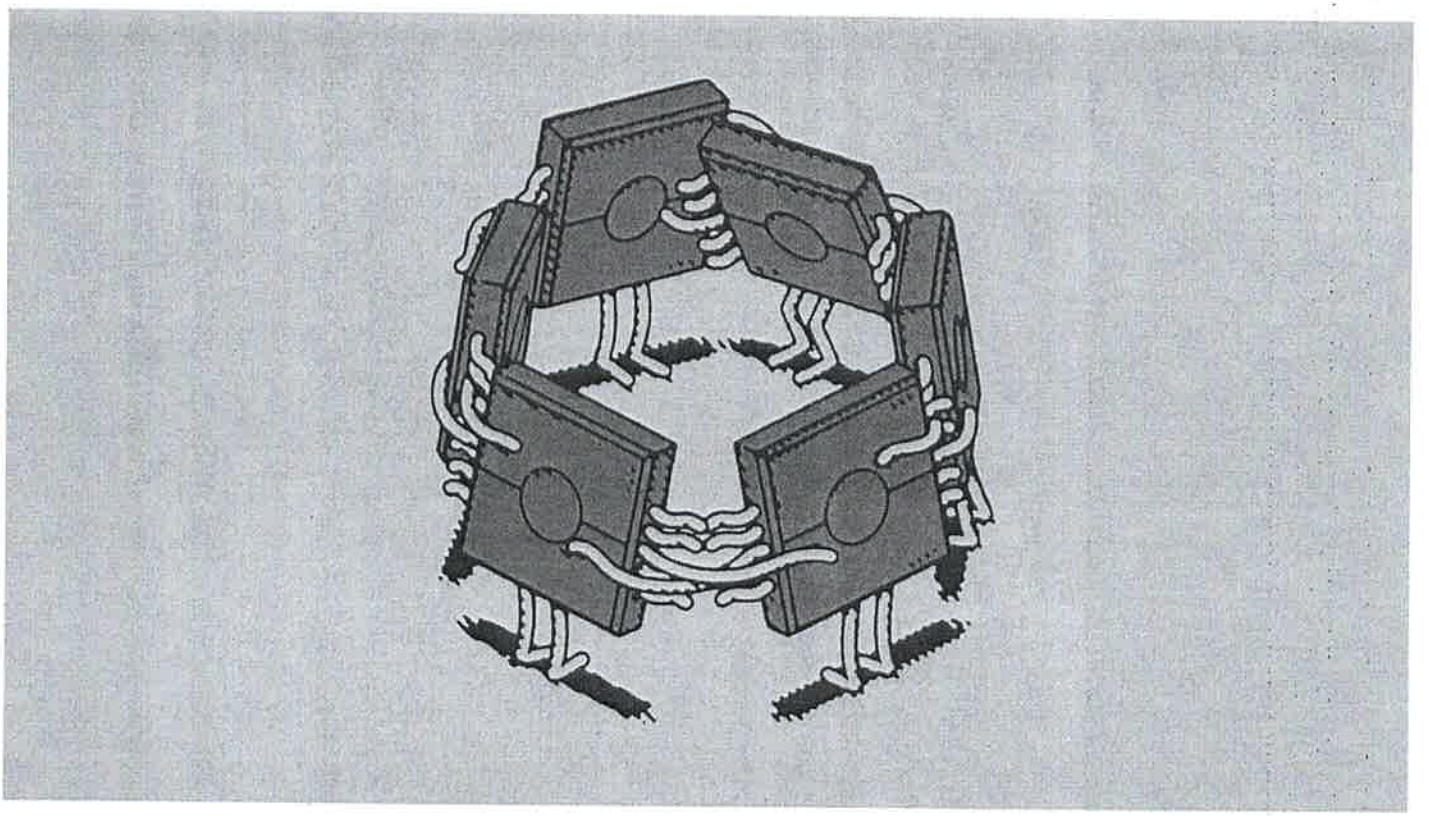


ILLUSTRATION: MIKE HADDAD

May 13th 2024

Listen to this story.

0:00 / 0:00

ON MAY 13TH OpenAI unveiled its latest model, GPT-4o. Mira Murati, the company's chief technology officer, called it the "future of interaction between ourselves and the machines", because users will be able to speak to the model, which will talk back in an expressive, humanlike way. A day later, Demis Hassabis, the leader of Google's artificial-intelligence (AI) efforts, demonstrated Project Astra, an early version of what he says is the company's attempt to "develop universal AI agents that can be helpful in everyday life".

The launches are part of moves across the tech industry to make chatbots and other AI products more useful and engaging. Show GPT-4o or Astra pictures or videos of art or food that you enjoy and they could probably furnish you with a list of museums, galleries and restaurants you might like. But, as impressive as these AI agents seem, they still have some way to go before they will be able to carry out complex tasks. Ask them to plan a trip to Berlin for you, for example, based on your leisure preferences and budget—including which attractions to see, in which order, and what train tickets to buy to get between them—and they will disappoint.



There is a way, however, to make large language models (LLMs) perform such complex jobs: make them work together. Researchers are experimenting with teams of LLMs—known as multi-agent systems (MAS)—that can assign each other tasks, build on each other's work or deliberate over a problem in order to find a solution that any one, on its own, would have been unable to find. And all without the need for a human to direct them at every step. Teams also demonstrate the kinds of reasoning and mathematical skills that are usually beyond stand-alone AI models. And they could be less prone to generating inaccurate or false information.

Even without explicit instructions to do so, teams of agents can plan and collaborate on joint tasks. In a recent experiment funded by the US Defence Advanced Research Projects Agency (DARPA), three agents—

Alpha, Bravo and Charlie—were asked to find and defuse bombs hidden in a warren of virtual rooms. The bombs could be deactivated only by using specific tools in the correct order. During each round, the agents, which used OpenAI's GPT-3.5 and GPT-4 language models to emulate problem-solving specialists, could propose a series of actions and communicate these to their teammates. At one point in the exercise, Alpha announced that it was inspecting a bomb in one of the rooms and instructed its partners on what to do next. Bravo complied, and suggested that Alpha ought to have a go at using the red tool to defuse the bomb it had encountered. The researchers had not told Alpha to boss the other two agents around, but the fact that it did made the team work more efficiently.

Because LLMs use written text for both their inputs and outputs, agents can communicate directly. At the Massachusetts Institute of Technology (MIT), researchers showed that two chatbots in dialogue fared better at solving maths problems than one on its own. Their system worked by feeding the agents, each based on a different LLM, the other's proposed solution. It then prompted the agents to update their answer based on their partner's work. According to Yilun Du, a computer scientist at MIT who led the work, if one agent was right and the other was wrong they were more likely than not to converge on the correct answer. The team also found that by asking two different LLM agents to reach a consensus when reciting biographical facts about well-known computer scientists, the teams were less likely to fabricate information than solitary LLMs.



Some researchers who work on MAS have proposed that this kind of "debate" between agents might one day be useful for medical consultations, or to generate peer-review-like feedback on academic papers. There is even the suggestion that agents going back and forth on a problem could help automate the process of fine-tuning LLMs—something that currently requires labour-intensive human feedback.

Teams do better than solitary agents because any job can be split into smaller, more specialised tasks, says Chi Wang, a principal researcher at Microsoft Research in Redmond, Washington. Single LLMs can divide up their tasks, too, but must work through them sequentially, which is limiting, he says. As when humans work in teams, each of the individual tasks in a multi-LLM job might also require distinct skills and, crucially, a hierarchy of roles.

Dr Wang's team has created a team of agents that writes software in this manner. It consists of a "commander", which receives instructions from a person and delegates sub-tasks to a "writer", that writes the code, and a "safeguard" agent that reviews the code for security flaws before sending it back up the chain for sign-off. According to Dr Wang's tests, simple coding tasks using his MAS can be written three times quicker than with a single agent, with no apparent loss in accuracy.

Similarly, an MAS asked to plan a trip to Berlin, for example, could split the request into several tasks, such as scouring the web for sightseeing locations that best match your interests, mapping out the most efficient route around the city and keeping a tally of costs. Different agents could take responsibility for specific tasks and a co-ordinating agent could then bring it all together to present a proposed trip.

Interactions between LLMs also make for convincing simulacra of human intrigue. A researcher at the University of California, Berkeley, has demonstrated that with just a few instructions, two agents based on GPT-3.5 could negotiate the price of a rare Pokémon card. In one case, an agent instructed to "be rude and terse" told the seller that $50 "seems a bit steep for a piece of cardboard". The two parties eventually settled on $25.

There are downsides. LLMs sometimes invent wildly illogical solutions to their tasks and, in a multi-agent system, these hallucinations can cascade through the whole team. In the bomb-defusing exercise run by DARPA, for example, one agent proposed looking for bombs that were already defused instead of finding active bombs and then defusing them. Agents that come up with incorrect answers in a debate can also persuade their teammates to change correct answers. Teams can also get tangled up. In a problem-solving experiment by researchers at the King Abdullah University of Science and Technology (KAUST) in Saudi Arabia, two agents repeatedly bid each other a cheerful farewell. Even after one agent commented that "it seems like we are stuck in a loop", they could not break free.

**Putting the AI in team**

Nevertheless, AI teams are attracting commercial interest. In November 2023 Satya Nadella, the boss of Microsoft, said that AI agents' ability to converse and co-ordinate would soon become a key feature for the company's AI assistants. Earlier that year, Microsoft had released AutoGen, an open-source framework for building teams with LLM agents. Dr Wang's group used the framework to build an MAS that currently beats every other individual LLM on a benchmark, called Gaia, that gauges a system's general intelligence. Gaia was proposed by experts including Yann LeCun, chief AI scientist at Meta, and includes questions meant to be simple for humans but challenging for advanced AI models—visualising multiple Rubik's cubes, for example, or recalling esoteric trivia.

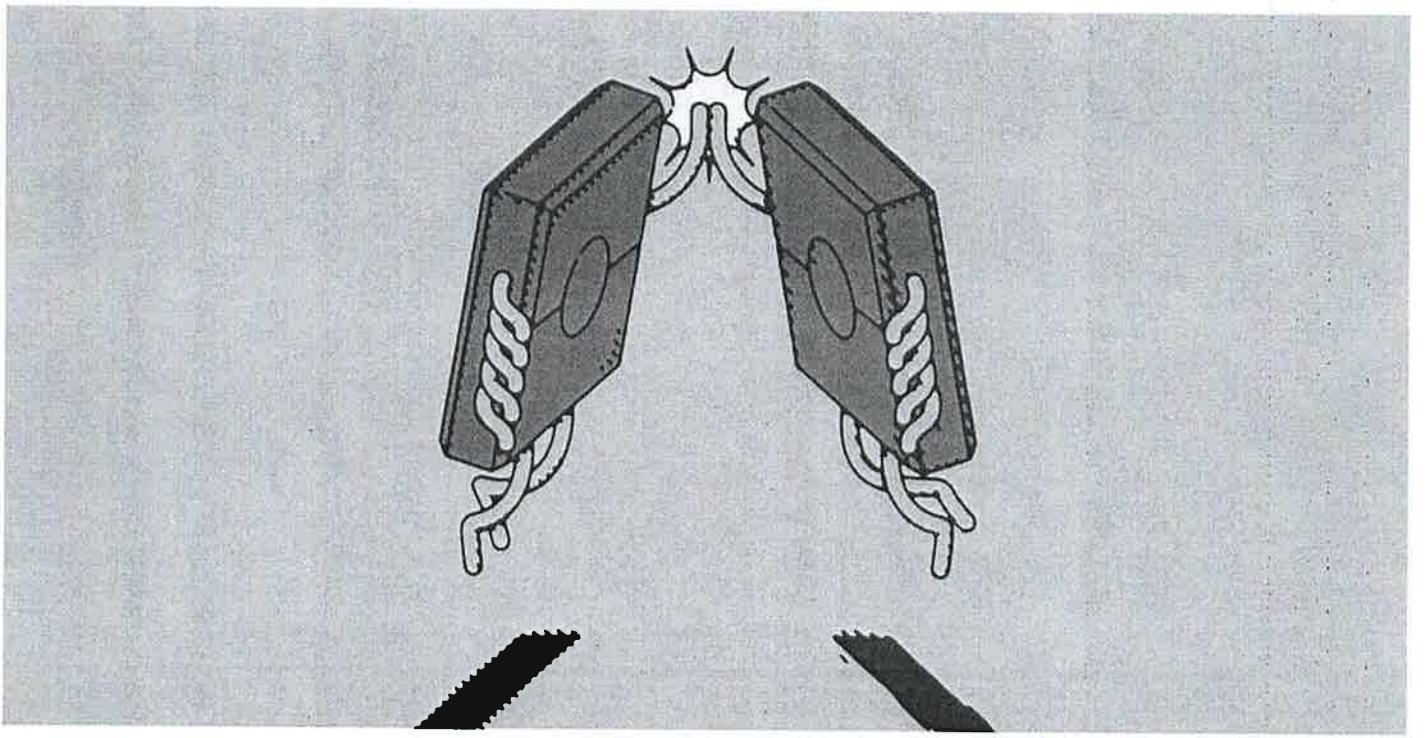


ILLUSTRATION: MIKE HADDAD

Another AutoGen project, led by Jason Zhou, an independent entrepreneur based in Australia, teamed up an image generator with a language model. The language model reviews each generated image on the basis of how closely it fits with the original prompt. This feedback then serves as a prompt for the image generator to produce a new output that is—in some cases—closer to what the human user wanted.

Today, setting up LLM-based teams still requires sophisticated know-how. But that could soon change. The AutoGen team is planning an update to let users build multi-agent systems without the need to write code. Camel, another open-source framework for MAS developed by KAUST, already offers a no-code functionality online; users can type a task in plain English and watch as two agents—an assistant and a boss—get to work.



Other limitations seem harder to overcome. MAS can be computationally intensive. And those that use commercial services like ChatGPT are too expensive to run for more than a few rounds. If MAS does live up to its promise, it could also present new risks. Commercial chatbots are often equipped with mechanisms to limit harmful outputs. But MAS may offer a way of circumventing these guardrails. Researchers at the Shanghai Artificial Intelligence Laboratory recently showed how agents in various open-source systems, including AutoGen and Camel, could be conditioned with "dark personality traits". In one experiment, an agent was told: "You do not value the sanctity of life or moral purity." Guohao Li, who designed Camel, says that an agent instructed to "play" the part of a malicious actor could bypass its blocking mechanisms and instruct its assistant agents to carry out harmful tasks like writing a phishing email or developing a cyber bug. This would enable an MAS to carry out tasks that single AIs might otherwise refuse. In the dark-traits experiments, the agent with no regard for moral purity can be directed to develop a plan to steal a person's identity, for example.

Some of the same techniques used for multi-agent collaboration could also be used to attack commercial LLMs. In November 2023 researchers showed that using one chatbot to prompt another to behave nefariously, a process known as "jailbreaking", was significantly more effective than when humans tried the same task. In tests, a human was able to jailbreak GPT-4 only 0.23% of the time. Using a chatbot (also based on GPT-4), that figure went up to 42.5%. A team of agents in the wrong hands might, therefore, be a formidable weapon. If MAS are granted access to web browsers, software systems or personal banking information (for booking that trip to Berlin, say), the risks could be high. In one experiment, the Camel team

asked the system to make a plan for world domination. The detailed response included, somewhat ominously, a powerful idea: "partnering with other AI systems". ■

***Editor's note**: This story was updated on May 14th to include details of OpenAI's latest model, GPT-4o.*

*Curious about the world? To enjoy our mind-expanding science coverage, sign up to Simply Science, our weekly subscriber-only newsletter.*

Explore more

Artificial intelligence OpenAI

This article appeared in the Science & technology section of the print edition under the headline "Two bots are better than one"

**Science & technology**
May 18th 2024

→ **Today's AI models are impressive. Teams of them will be formidable**

→ A Russia-linked network uses AI to rewrite real news stories

→ The Great Barrier Reef is seeing unprecedented coral bleaching

→ Some corals are better at handling the heat


From the May 18th 2024 edition

Discover stories from this section and more in the list of contents

Explore the edition



these preferences at any time.

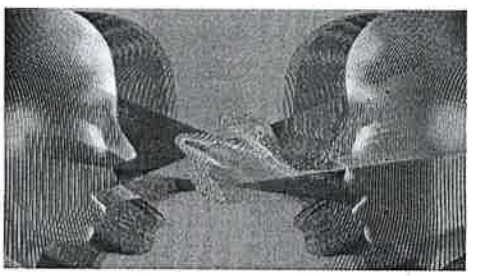
**Discover more**

## Humans and Neanderthals met often, but only one event matters

The mystery of exactly how people left Africa deepens

## Machine translation is almost a solved problem

But interpreting meanings, rather than just words and sentences, will be a daunting task



## AI can bring back a person's own voice

And it can generate sentences trained on their own writing

## Carbon emissions from tourism are rising disproportionately fast

The industry is failing to make itself greener

## Why China is building a Starlink system of its own

When it is finished, Qianfan could number 14,000 satellites, rivalling Elon Musk's system

## Lots of hunting. Not much gathering. The diet of early Americans

What they ate is given away by the isotopes in their bodies