SDS PODCAST EPISODE 788: FIVE-MINUTE FRIDAY: MULTI-AGENT SYSTEMS: HOW TEAMS OF LLMS EXCEL AT COMPLEX TASKS

(00:05):

This is Five-Minute Friday on Multi-Agent Systems.

(00:19):

Welcome back to The Super Data Science Podcast. I'm your host, Jon Krohn. As we're in the habit of doing most Fridays now, let's start off with a couple of reviews.

(00:28):

Our first today is a five-star review from Apple Podcasts that is by Nathaniel Brewer, who's a CTO in Charlotte, North Carolina, and Nate very generously titled his Apple Podcast review "Greatest AI podcast". Terrifically kind of you to say, Nate, thank you! The body includes some detail on why he thinks this show is the greatest AI podcast so he said: "As someone who is in the AI field, with limited people to talk to in the field, this podcast is food for thought. Jon has the greatest guests with really awesome topics. If you're looking to stay up to date with AI, this is the podcast." All right, thank you Nate.

(01:07):

And our second review comes from Seyi Adeeko, who is a healthcare analyst in Nigeria. Seyi says: "Your podcast was a top resource through my Masters' program in Applied AI and Data Science and continues to inspire and keep me abreast of the possibilities of data science." So happy to hear it, Seyi. I hope we can continue to inspire you.

(01:29):

Thanks for all the recent ratings and feedback on wherever you listen to your podcasts, Apple Podcasts, Spotify and all the other podcasting platforms out there, as well as for likes and comments we provide on our

YouTube videos. Apple Podcast reviews in particular are especially helpful to us because they allow you to leave written feedback if you want to and that allows me to keep a close eye on reviews and, if you leave one, I'll be sure to read it on air like I did the couple of reviews we had today.

(01:59):

All right, let's get into the meat of this episode, which is focused on groundbreaking multi-agent systems and how these multi-agent systems are transforming the way AI models collaborate to tackle complex challenges.

(02:11):

For a bit of timely, high-profile context, two weeks ago, OpenAI unveiled its latest model, GPT-4o. Mira Murati, the company's chief technology officer, hailed it as the "future of interaction between ourselves and the machines." What sets GPT-4o apart is its ability to engage in expressive, humanlike conversations with users in real-time. You can now speak to a state-of-the-art LLM that not only understands your words but is now also engineered to respond in a natural, intuitive way. This isn't so much an LLM innovation as a stitching together of an LLM with existing tech, but in terms of usability it's an awesome step forward… for some in the direction of the famous Spike Jonze A.I. film "Her".

(02:57):

Not to be outdone by OpenAI, Google DeepMind head Demis Hassabis showcased Project Astra, just a day after the GPT-4o release. This early version of Project Astra is what Hassabis describes as the company's endeavor to "develop universal AI agents that can be helpful in everyday life" and this marks another significant step forward in the AI revolution. You can check out the link in the show notes to see demos of the Project Astra agent being used via Google Pixel Phones or prototype glasses that

aim to build on the ultimately floppy release of Google Glass a decade ago. Anyway, regardless of the implement, Project Astra was shown in these demos to be able to analyze real time video in order to explain physics, literature and landmarks, even to be able to solve math problems on a whiteboard, was very cool to see.

(03:54):

These launches, so things like GPT-4o and Project Astra, are part of a larger trend across the tech industry to create chatbots and AI products that are more useful and engaging to users and more useful in a wider range of situations. So if you show GPT-4o or Astra pictures or videos of art or food that you enjoy, they can provide you with a list of museums, galleries, and restaurants say tailored to your preferences and answer questions about the food or the art or whatever.

(04:23):

As impressive as these AI agents are, however, they still have plenty of limitations when it comes to executing complex tasks. For example, if you ask if you ask today's LLMs to plan your trip to Berlin based on your leisure preferences and your budget, including asking it to provide you attractions to see, in what order to see them in, ask it to actually buy you tickets to these attractions, appropriate train tickets to get you there, you are likely to be disappointed.

(04:55):

But this is where multi-agent systems come into play. By enabling LLMs to work together, researchers are unlocking new possibilities for AI to perform intricate jobs. Recent experiments for example, have shown that teams of LLMs in a multi-agent system can assign each other tasks, build upon one another's work, and even engage in deliberation to find solutions that would

be out of reach for any single LLM. And all this happens without the need for constant human direction in the loop.

(05:27):

In one remarkable example from DARPA, a team of three agents named Alpha, Bravo, and Charlie worked together to find and defuse virtual bombs. Alpha took the lead, instructing its partners Bravo and Charlie on what to do next, resulting in a more efficient problem-solving process. Critically, this emergent behavior between Alpha, Bravo and Charlie wasn't explicitly programmed, but rather was a result of the agents' collaboration. Cool, right?

(05:56):

In another example, researchers at MIT have also demonstrated that two chatbots in dialogue perform better at solving math problems than a single agent. By feeding each other's proposed solutions and updating their answers based on their partner's work, the agents were more likely to converge on the correct answer. In other potential real-world examples, this kind of "debate" between agents could potentially be applied to say things like medical consultations or peer-review feedback on academic papers. It could even be used to fine tune LLMs themselves, therefore limiting the need for humans to be labeling data with processes like RLHF.

(06:38):

Anyway, the power of multi-agent systems lies in the ability of these systems to split jobs into smaller, specialized tasks, with each agent possessing distinct skills and roles. At Microsoft Research, for example, a team of humans created a software-writing MAS consisting of a "commander" that delegates sub-tasks, a "writer" that writes the code, and a "safeguard" agent that reviews the code for security flaws. This approach of having these different specialized characters, a commander, a writer, a

safeguard, this resulted in code being written three times faster than with a single agent, without sacrificing accuracy. Cool, so some really great examples there.

(07:21):

However, as with any AI advances, there are potential downsides to multi-agent systems as well. LLMs can sometimes generate illogical solutions, and in a multi-agent system, these so-called "hallucinations" can cascade through the entire team. Agents have also been known to occasionally get stuck in loops, for example by repeatedly bidding each other farewell without breaking free from that loop.

(07:48):

Despite these challenges, the commercial interest in AI teams is growing. Microsoft's CEO Satya Nadella has emphasized the importance of AI agents' ability to converse and coordinate, and the company has released AutoGen, an open-source framework for building LLM-based multi-agent systems. Other frameworks, like Camel, offer no-code functionality, allowing users without any coding ability to input tasks in plain English and watch the agents get right to work. I've got links to both of these projects, Microsoft's AutoGen and Camel, for you to check out in the show notes.

(08:23):

As MAS technology advances, like any AI advances, yeah, yeah, there are of course potential risks. Malicious actors could exploit these systems by conditioning agents with "dark personality traits," enabling them to bypass safety mechanisms and carry out harmful tasks. The same techniques used for multi-agent collaboration could also be used to attack commercial LLMs through "jailbreaking" or do all kinds of other nefarious things. Hopefully, however, the positive applications of MAS end up greatly overwhelming the negative ones, including by us doing research, by people like you potentially

doing research on systems for defending against multi-agent system misuse.

(09:04):

Yeah, so assuming we can contain the negative effects, this is very exciting indeed. And there are truly limitless applications for MAS out there. What in your industry could be automated or improved by MAS where say single-LLM approaches aren't sufficiently effective? Check out the links in the show notes to get started on experimenting with MAS today.

(09:31):

All right, that's it for today's episode. If you enjoyed it or you know someone who might, consider sharing this episode with them, leave a review of the show on your favorite podcasting platform, tag me in a LinkedIn or Twitter post with your thoughts, I will respond to those, or if you aren't already, be sure to subscribe to the show. Most importantly, however, we hope you'll just keep on listening. Until next time, keep on rockin' it out there and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.