

# [***Amid OpenAI Chaos, Rival Inflection AI Releases Model On GPT-4 s Heels***](https://advance.lexis.com/api/document?collection=news&id=urn:contentItem:69PH-D0N1-JBCM-F01T-00000-00&context=1516831)

Forbes.com

November 22, 2023 Wednesday

Copyright 2023 Forbes LLC All Rights Reserved

**Length:** 1147 words

**Byline:** Alex Konrad, Forbes Staff

**Highlight:** Called Inflection-2, the model performed better than Google s PaLM Large 2 model previously announced in May on a number of standard benchmarks, Inflection said, while...

**Body**

<figure>

<figcaption>

Inflection CEO Mustafa Suleyman has already started work on his next model, aimed to exceed GPT-4's capabilities.

Inflection AI

</figcaption></figure>

Inflection AI, the startup behind the conversational chatbot Pi, has unveiled a new AI model that the company claims can outperform two popular alternatives developed by Google and Meta and is hot on the heels of OpenAI s larger, flagship model GPT-4.

Called Inflection-2, the model performed better than Google s PaLM Large 2 model previously announced in May on a number of standard benchmarks, Inflection said, while beating the open-source LLaMA 2 model largely developed by Meta on different measures. Overall, Inflection s model is the top-performing of its size, the startup said. It only trails GPT-4, the flagship released model from OpenAI, thought to be significantly larger.

Our belief is that we re just at the very beginning of this scaling curve, and the new capabilities that are going to arise are truly mind blowing, CEO Mustafa Suleyman said in an interview.

The newly-released model will soon be integrated into Pi, the chatbot that Inflectionreleasedin May. First it needs a bit more extra work known as alignment to teach it Pi s tone and answering style, and to help Pi function better while absorbing up-to-date information without additional hallucinations, Suleyman said.

Whether you want to have a sensitive conversation about race, or gender, or ***politics***, or OpenAI or whatever contentious issue of the day, Pi very subtly and carefully engages with you in a factual way, with real-time access to information, he said. Pi will be updated with the new model very soon, it won t be long, Suleyman added, while declining to provide a date.

He would not provide updated user numbers on Pi but said it s massively popular, with huge retention. Two weeks ago, OpenAIrevealedits free ChatGPT service had reached 100 million weekly users, just before the board s shock temporaryfiringof CEO Sam Altman on Friday.

It s hard not to view Inflection s large language model release, billed as the second most capable LLM in the world today, within the context of the turmoil at the industry s best-known model maker. Earlier on Wednesday morning, OpenAIannouncedthat Altman was returning as CEO of OpenAI, following a weekend of alliesnegotiatingfor Altman s return and OpenAI employees signing a petitioncallingfor their board to resign en masse.

At Inflection, Suleyman insisted that his startup, whichraiseda $1.3 billion funding round earlier this year, had not moved up the release date of Inflection s new model, despite earlier public remarks that seemed to suggest an end-of-year release. This release was actually a week overdue, he claimed, with training finished last week.

To train Inflection-2, the startup used 5,000 Nvidia H100 graphics processing units, or GPUs, up from the several thousand older A100s that trained its predecessor. The new model, which Suleyman said was faster and cheaper to train, still handled a massive amount of operations (10^25 FLOPs, or floating point operations). Inflection partners closely with Microsoft, Nvidia and CoreWeave to manage its compute cluster.

Inflection tested its new model against a popular benchmark of high school- and professional-level tasks known as MMLU, where models are asked questions on 57 subjects from world knowledge to problem-solving and ethics.

<figure>

<figcaption>

Inflection-2 performance against other models.

Inflection AI

</figcaption></figure>

Suleyman said Inflection-2 outperformed the largest, 70 billion parameter version of LLaMA 2, Elon Musk s xAI startup s Grok-1, Google s PaLM 2 Large and startup Anthropic s Claude 2, trailing only GPT-4. The new model beat the LLaMA 2 and PaLM 2 models on all but two of seven scientific answering benchmarks, Inflection reported; it also performed best on two of three questions-and-answers task benchmarks, losing to PaLM 2 Large in one. Across four math and code benchmarks, it finished respectably despite those areas not being explicit focuses, the company added; it trailed GPT-4 considerably on the two for which OpenAI has shared results, however.

While such benchmarks might not seem significant to anyone except an AI researcher or developer, small improvements can make the difference between a hacky prototype and a production grade, reliable and high quality model, Suleyman said. Overall, Suleyman claimed Inflection-2 could be considered best-in-class for its size very, very close to GPT-4. Starting today, Inflection is shifting its training focus to its next model, which he predicted would reach 10x the scale of its predecessor within six months, with another 10x bigger than that in another six months. (One hundred times bigger than now, for those keeping score.)

Speaking hours before OpenAI s messy board battle reached provisional resolution, Suleyman urged the public to approach those involved with empathy and forgiveness. There are a lot of very well-intentioned people there, he noted, particularly shouting out cofounder Ilya Sutskever, originally one of the board directors who fired Altman, and who later announced he regretted that decision on Sunday night. (It wasn t immediately clear whether Sutskever would remain at OpenAI, but he liked posts by the company and Altman announcing the return on X.)

Sutskever worked for Suleyman back in 2011 as a contractor and consultant at DeepMind, the AI startup he sold to Google three years later. I have great respect for him, not just technically, but he s very principled and very sincere. I assume genuinely good intentions from him and the rest of the team, Suleyman said.

Inflection and Suleyman have previously beenoutspokenabout AI safety,signingPresident Biden s voluntary order in July, among other moves. He s alsovowedto stay away from having Pi answer election-related questions or participate in creating materials for political campaigns. Suleyman recently published a book that in part details risks with AI, called The Coming Wave.

Obviously there've been some very, very major mistakes, he said of OpenAI s board and the actions of its unusual non-profit board, which had been maligned by some in the tech industry in recent days. I hope people will try to be forgiving, and see this as a minor blip on the overall journey to trying to build better companies that do the right thing and try to make the world a better place with new governance structures.

That said, Suleyman is also a capitalist: asked if OpenAI stumbling ChatGPT was down for at least parts of Monday and Tuesday, as customers watched proceedings with concern gave rivals more opportunity, Inflection s CEO didn t disagree.

Fundamentally I m building a business, Suleyman said. And it s extremely competitive. This is the most competitive and creative time in Silicon Valley in years.

**Load-Date:** May 10, 2024

**End of Document**