Kinder, Friendlier AI Chatbot 'Claude 2' Unveiled by Anthropic

John P. Mello Jr.

July 12, 2023

The wraps were pulled off a new AI chatbot billed as "helpful, harmless and honest" on Tuesday by its developer, Anthropic.

The chatbot, Claude 2, boasts a familiar repertoire. It can create summaries, write code, translate text, and perform tasks that have become de rigueur for the software genre.

This latest version of the generative AI offering can be accessed via API and through a new web interface that the public can tap into in the United States and the United Kingdom. Previously, it was only available to businesses by request or through Slack as an app.

"Think of Claude as a friendly, enthusiastic colleague or personal assistant who can be instructed in natural language to help you with many tasks," Anthropic said in a statement.

"Anthropic is trying to lean into the personal assistant space," observed Will Duffield, a policy analyst at the Cato Institute, a Washington, D.C., think tank

"While Microsoft has a leg up bringing Bing to its productivity suite, Claude wants to be a more useful personal assistant than the rest," he told TechNewsWorld.

Improved Reasoning Scores

Claude 2 is improved over previous models in the areas of coding, math, and reasoning, according to Anthropic.

On the multiple-choice section of a bar exam, for example, Claude 2 scored 76.5%. Previous models scored 73.0%.

On the GRE reading and writing exams for college students applying for graduate school, Claude 2 scored above the 90th percentile. On quantitative reasoning, it did as well as median applicants.

In the coding area, Claude 2 scored 71.2% on the Codex HumanEval test, a Python coding test. That's a significant improvement over prior models, which achieved a score of 56.0%.

However, it did only slightly better than its predecessor on the GSM8K, which encompasses a large set of grade-school math problems, racking up a score of 88.0%, compared to 85.2% for Claude 1.3.

Knowledge Lag

Anthropic improved Claude in another area: input.

Claude 2's context window can handle up to 75,000 words. That means Claude can digest hundreds of pages of technical documentation or even a book. By comparison, ChatGPT's maximum input is 3,000 words.

Anthropic added that Claude can now also write longer documents — from memos to letters to stories up to a few thousand words.

Like ChatGPT, Claude isn't connected to the internet. It's trained on data that abruptly ends in December 2022. That gives it a slight edge over ChatGPT, whose data cuts off currently in September 2021 — but lags behind Bing and Bard.

"With Bing, you get up-to-date search results, which you also get with Bard," explained Greg Sterling, co-founder of Near Media, a news, commentary and analysis website.

However, that may have a limited impact on Claude 2. "Most people aren't going to see major differences unless they use all of these apps side by side," Sterling told TechNewsWorld. "The differences people may perceive will be primarily in the UIs."

Anthropic also touted safety improvements made in Claude 2. It explained that it has an internal "red team" that scores its models based on a large set of harmful prompts. The tests are automated, but the results are regularly checked manually. In its latest evaluation, Anthropic noted Claude 2 was two times better at giving harmless responses than Claude 1.3.

In addition, it has a set of principles called a constitution built into the system that can temper its responses without the need to use a human moderator.

Tamping Down Harm

Anthropic isn't alone in trying to put a damper on potential harm caused by its generative AI software. "Everyone is working on helpful AIs that are supposed to do no harm, and the goal is nearly universal," observed Rob Enderle, president and principal analyst at the Enderle Group, an advisory services firm in Bend, Ore.

"It is the execution that will likely vary between providers," he told TechNewsWorld.

He noted that industrial providers like Microsoft, Nvidia, and IBM have taken AI safety seriously from the time they entered the domain. "Some other startups appear more focused on launching something than something safe and trustworthy," he said.

"I always take issue with the use of language like harmless because useful tools can usually be misused in some way to do harm," added Duffield.

Attempts to minimize harm in a generative AI program could potentially impact its value. That doesn't seem to be the case with Claude 2, however. "It doesn't seem neutered to the point of uselessness," Duffield said.

Conquering Noise Barrier

Having an "honest" AI is key to trusting it, Enderle maintained. "Having a harmful, dishonest AI doesn't do us much good," he said. "But if we don't trust the technology, we shouldn't be using it."

“AIs operate at machine speeds, and we don’t,” he continued, “so they could do far more damage in a short period than we’d be able to deal with.”

“AI can make things up that are inaccurate but plausible-sounding,” Sterling added. “This is highly problematic if people rely on incorrect information.”

“AI also can spew biased or toxic information in some cases,” he said.

Even if Claude 2 can fulfill its promise to be a “helpful, harmless and honest” AI chatbot, it will have to fight to get noticed in what’s becoming a very noisy market.

“We are being overwhelmed by the number of announced things, making it harder to rise above the noise,” Enderle noted.

“ChatGPT, Bing, and Bard have the most mindshare, and most people will see little reason to use other applications,” added Sterling.

He noted that trying to differentiate Claude as the “friendly” AI probably won’t be enough to distinguish it from the other players in the market. “It’s an abstraction,” he said. “Claude will need to perform better or be more useful to gain adoption. People won’t see any distinction between it and its better-known rival ChatGPT.”

As if high noise levels weren’t enough, there’s ennui to deal with. “It’s harder to impress people with any kind of new chatbot than it was six months ago,” Duffield observed. “There’s a little bit of chatbot fatigue setting in.”