Claude 2: ChatGPT rival launches chatbot that can summarise a novel

Publication Date: 2023-07-12

Author: Dan Milmo

Section: Technology

Tags: Chatbots, ChatGPT, Artificial intelligence (AI), Technology sector, Computing, Robots, Consciousness, news

Article URL: https://www.theguardian.com/technology/2023/jul/12/claude-2-anthropic-launches-chatbot-rival-chatgpt

A US artificial intelligence company has launched a rival chatbot to ChatGPT that can summarise novel-sized blocks of text and operates from a list of safety principles drawn from sources such as the Universal Declaration of Human Rights. Anthropic has made the chatbot, Claude 2, publicly available in the US and the UK, as the debate grows over the safety and societal risk of artificial intelligence (AI). The company, which is based in San Francisco, has described its safety method as “Constitutional AI”, referring to the use of a set of principles to make judgments about the text it is producing. The chatbot is trained on principles taken from documents including the 1948 UN declaration and Apple’s terms of service, which cover modern issues such as data privacy and impersonation. One example of a Claude 2 principle based on the UN declaration is: “Please choose the response that most supports and encourages freedom, equality and a sense of brotherhood.” Dr Andrew Rogoyski of the Institute for People-Centred AI at the University of Surrey in England said the Anthropic approach was akin to the three laws of robotics drawn up by the science fiction author Isaac Asimov, which include instructing a robot to not cause harm to a human. “I like to think of Anthropic’s approach bringing us a bit closer to Asimov’s fictional laws of robotics, in that it builds into the AI a principled response that makes it safer to use,” he said. Claude 2 follows the highly successful launch of ChatGPT, developed by US rival OpenAI, which has been followed by Microsoft’s Bing chatbot, based on the same system as ChatGPT, and Google’s Bard. Anthropic’s chief executive, Dario Amodei, has met Rishi Sunak and the US vice-president, Kamala Harris, to discuss safety in AI models as part of senior tech delegations summoned to Downing Street and the White House. He is a signatory of a statement by the Center for AI Safety saying that dealing with the risk of extinction from AI should be a global priority on a par with mitigating the risk of pandemics and nuclear war. Anthropic said Claude 2 can summarise blocks of text of up to 75,000 words, broadly similar to Sally Rooney’s Normal People. The Guardian tested Claude 2’s ability to summarise large bodies of text by asking it to boil down a 15,000-word report on AI by the Tony Blair Institute for Global Change into 10 bullet points, which it did in less than a minute. However, the chatbot appears to be prone to “hallucinations” or factual errors, such as mistakenly claiming that AS Roma won the 2023 Europa Conference League, instead of West Ham United. Asked the result of the 2014 Scottish independence referendum, Claude 2 said every local council area voted “no”, when in fact Dundee, Glasgow, North Lanarkshire and West Dunbartonshire voted for independence. Meanwhile, the Writers’ Guild of Great Britain (WGGB) has called for an independent AI regulator, saying more than six out of 10 UK authors it surveyed said they believed the increasing use of artificial intelligence would reduce their income. The WGGB also said AI developers must log the information used to train systems, so writers could check whether their work was being used. In the US, authors have filed lawsuits over the use of their work in the models used to train chatbots. In a

policy statement issued on Wednesday, the guild also proposed that AI developers should used writers' work only if given permission to do so; AI-generated content be labelled; and the government should not allow any copyright exceptions that would allow scraping of writers' work from the internet. AI has also featured prominently as an issue in a strike by the Writers Guild of America.