GM plans to let you talk to your car with ChatGPT, Knight Rider-style



Enlarge / The 1982 TV series Knight Rider featured a car called KITT that a character played by David Hasselhoff (pictured) could talk to.
Getty Images

reader comments

38 with

Share this story

In the 1982 TV series *Knight Rider*, the main character can have a full conversation with his futuristic car. Once science fiction, this type of language interface may soon be one step closer to reality because General Motors is working on bringing a ChatGPT-style AI assistant to its automobiles, according to Semafor and Reuters.

While GM won’t be adding *Knight Rider*-style turbojet engines or crime-fighting weaponry to its vehicles, its cars may eventually talk back to you in an intelligent-sounding way, thanks to a collaboration with Microsoft.

Microsoft has invested heavily in OpenAI, the company that created ChatGPT. Now, they’re looking for ways to apply chatbot technology to many different fields.

“ChatGPT is going to be in everything,” Reuters quotes GM Vice President Scott Miller as saying in an interview last week. Miller gave examples of being able to explore vehicle features found in the manual, integrate schedules on a calendar, or program a garage door code using a conversational voice interface.

[embedded content]
Michael speaks to KITT for the first time in episode 1 of Knight Rider, 1982.
A report from Semafor, which first broke the story, also suggests that a driver might be able to ask a future AI-enabled GM car how to fix a flat tire, and the car would play an instructional video or give step-by-step instructions on how to change it.

Advertisement
“This shift is not just about one single capability like the evolution of voice commands, but instead means that customers can expect their future vehicles to be far more capable and fresh overall when it comes to emerging technologies,” a GM spokesperson told Reuters on Friday.

Large language models (LLM) like ChatGPT process text (and with Whisper integration, understand speech) to provide responses that draw heavily on the context of the conversation and a large data set trained on information from the Internet. However, they’re also known to generate convincing errors when reporting factual information. That’s a glaring issue GM and Microsoft will have to figure out before they can bring a LLM-

based solution to market.

In the aforementioned *Knight Rider*, Michael Knight (played by David Hasselhoff) drives an AI-equipped car—a modified 1982 Pontiac Trans Am—named KITT that can speak and understand natural language in fluid conversation. It also understands its surroundings and reacts to them dynamically. In the first episode of the series, Knight says, “I can’t believe this—a car that talks back to me.”

In that series, KITT can also drive itself, a feature car manufacturers have been struggling with using a different form of AI that has nothing to do with LLMs. And ChatGPT, which is tuned on language (not driving) won’t be getting behind the wheel any time soon. But since Pontiac was a division of General Motors in 1982, if GM doesn’t license the name KITT for their automotive AI assistant, it could be a significant missed opportunity.

https://arstechnica.com/?p=1923478

Get ready to meet the Chat GPT clones



Edward Olive/Getty Images

reader comments

31 with

Share this story

ChatGPT might well be the most famous, and potentially valuable, algorithm of the moment, but the artificial intelligence techniques used by OpenAI to provide its smarts are neither unique nor secret. Competing projects and open source clones may soon make ChatGPT-style bots available for anyone to copy and reuse.

Stability AI, a startup that has already developed and open-sourced advanced image-generation technology, is working on an open competitor to ChatGPT. “We are a few months from release,” says Emad Mostaque, Stability’s CEO. A number of competing startups, including Anthropic, Cohere, and AI21, are working on proprietary chatbots similar to OpenAI’s bot.

The impending flood of sophisticated chatbots will make the technology more abundant and visible to consumers, as well as more accessible to AI businesses, developers, and researchers. That could accelerate the rush to make money with AI tools that generate images, code, and text.

Established companies like Microsoft and Slack are incorporating ChatGPT into their products, and many startups are hustling to build on top of a new ChatGPT API for developers. But wider availability of the technology may also complicate efforts to predict and mitigate the risks that come with it.

Advertisement

ChatGPT’s beguiling ability to provide convincing answers to a wide range of queries also causes it to sometimes make up facts or adopt problematic personas. It can help with malicious tasks such as producing malware code, or spam and disinformation campaigns.

As a result, some researchers have called for deployment of ChatGPT-like systems to be slowed while the risks are assessed. “There is no need to stop research, but we certainly could regulate widespread deployment,” says Gary Marcus, an AI expert who has sought to draw attention to risks such as disinformation generated by AI. “We might, for example, ask for studies on 100,000 people before releasing these technologies to 100 millions of people.”

Wider availability of ChatGPT-style systems, and release of open source versions, would make it more difficult to limit research or wider deployment. And the competition between companies large and small to adopt or match ChatGPT suggests little appetite for slowing down, but appears instead to incentivize proliferation of the technology.

Last week, LLaMA, an AI model developed by Meta—and similar to the one at the core of ChatGPT—was leaked online after being shared with some academic researchers. The system could be used as a building block in the creation of a chatbot, and its release sparked worry among those who fear that the AI systems known as large language models, and chatbots built on them like ChatGPT, will be used to generate misinformation or automate cybersecurity breaches. Some experts argue that such risks may be overblown, and others suggest that making the technology more transparent will in fact help others guard against misuse.

https://arstechnica.com/?p=1923204

Has the generative AI pricing collapse already started?



fermate/Getty Images

reader comments

96 with

Share this story

OpenAI just announced pricing for businesses seeking to integrate its ChatGPT service into their own products, and it looks an awful lot like a 90 percent off sale.

It all starts with OpenAI, a former nonprofit that’s now gunning for riches as lustily as any Silicon Valley unicorn. The company has built a dazzling array of products, including the DALL-E image generator and the renowned ChatGPT service.

ChatGPT is powered by a system known as a large language model (or LLM), and it’s one of several LLM lines that OpenAI sells commercially. Buyers of LLM output are mostly companies that integrate language-related services like chat, composition,

summarization, software generation, online search, sentiment analysis, and much more into their websites, services, and products.

For instance, dozens of startups use LLMs to help their own customers churn out corporate blogs, marketing emails, and click-seeking SEO articles with radically less effort than before. This may add little to our cultural heritage but quite a bit to the bottom lines of these companies' cannier users. As I write this, at least 65 different startups are peddling these sorts of copywriting services, with new ones launching weekly.

Advertisement

Most are basically thin wrappers seeking to arbitrage LLM pricing, with virtually no differentiation or competitive moat. Some of them—most notably Jasper, which recently raised money at a $1.5 billion valuation—are building value-added services around the LLM output they retrieve and reformat, which may let them differentiate and prosper over time.

OpenAI is currently the biggest LLM provider, though there is growing competition from Cohere, AI21, Anthropic, Hugging Face, and others. These companies generally sell their output on a "per-token" basis, with a token representing approximately three-quarters of a word.

To give you an example, the baseline price of a very powerful OpenAI model called Davinci is 2 cents for a thousand tokens—which is to say about 750 words. This means it would cost Jasper about 8 cents to get Davinci to write six 500-word blog posts for you.

Prosperity for Jasper lies in pricing plans and subscription models that let the company charge you more than 8 cents for those posts. Prosperity for you lies in getting more value from those posts in the form of clicks, customer leads, sales, and donations than you paid Jasper to pay Davinci to write

them. For good or ill, this dynamic will plainly lead to a tsunami of new “content” flooding the Internet in very short order.

https://arstechnica.com/?p=1922974

Wikipedia + AI = truth? DuckDuckGo hopes so with new answerbot



Enlarge / An AI-generated image of a cyborg duck.
Ars Technica

reader comments

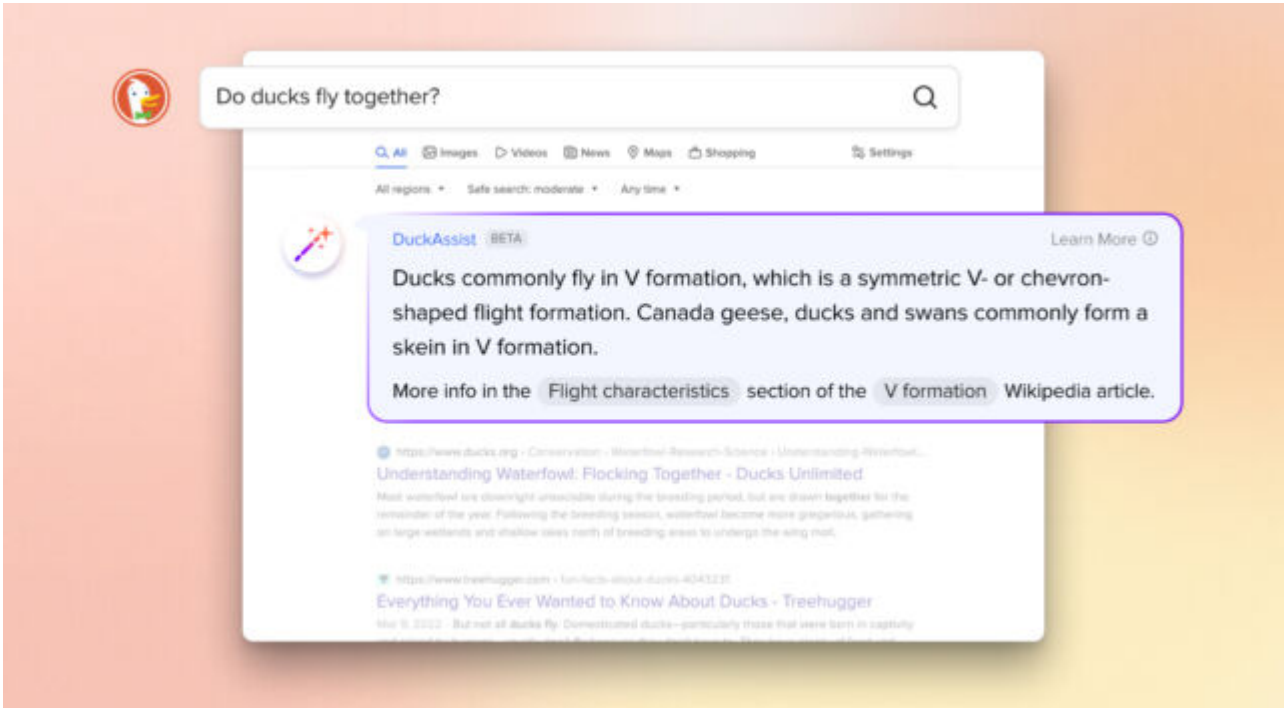
61 with

Share this story

Not to be left out of the rush to integrate generative AI into search, on Wednesday DuckDuckGo announced DuckAssist, an AI-powered factual summary service powered by technology from Anthropic and OpenAI. It is available for free today as a wide beta test for users of DuckDuckGo's browser extensions and browsing apps. Being powered by an AI model, the company admits that DuckAssist might make stuff up but hopes it will happen rarely.

Here's how it works: If a DuckDuckGo user searches a question that can be answered by Wikipedia, DuckAssist may appear and use AI natural language technology to generate a brief summary of what it finds in Wikipedia, with source links listed below. The summary appears above DuckDuckGo's regular search results in a special box.

The company positions DuckAssist as a new form of "Instant Answer"—a feature that prevents users from having to dig through web search results to find quick information on topics like news, maps, and weather. Instead, the search engine presents the Instant Answer results above the usual list of websites.



Enlarge / A demonstration screenshot of DuckAssist in action.
DuckDuckGo

DuckDuckGo does not say which large language model (LLM) or models it uses to generate DuckAssist, although some form of OpenAI API seems likely. Ars Technica has reached out to DuckDuckGo representatives for clarification. But DuckDuckGo CEO Gabriel Weinberg explains how it utilizes sourcing in a company blog post:

> *DuckAssist answers questions by scanning a specific set of sources—for now that's usually Wikipedia, and occasionally related sites like Britannica—using DuckDuckGo's active indexing. Because we're using natural language technology from OpenAI and Anthropic to summarize what we find in Wikipedia, these answers should be more directly responsive to your actual question than traditional search results or other Instant Answers.*

Since DuckDuckGo's main selling point is privacy, the company says that DuckAssist is "anonymous" and emphasizes that it does not share search or browsing history with anyone. "We also keep your search and browsing history anonymous to our search content partners," Weinberg writes, "in this case,

OpenAI and Anthropic, used for summarizing the Wikipedia sentences we identify."

Advertisement
If DuckDuckGo is using OpenAI's GPT-3 or ChatGPT API, one might worry that the site could potentially send each user's query to OpenAI every time it gets invoked. But reading between the lines, it appears that only the Wikipedia article (or excerpt of one) gets sent to OpenAI for summarization, not the user's search itself. We have reached out to DuckDuckGo for clarification on this point as well.

DuckDuckGo calls DuckAssist "the first in a series of generative AI-assisted features we hope to roll out in the coming months." If the launch goes well—and nobody breaks it with adversarial prompts—DuckDuckGo plans to roll out the feature to all search users "in the coming weeks."

DuckDuckGo: Risk of hallucinations "greatly diminished"

As we've previously covered on Ars, LLMs have a tendency to produce convincing erroneous results, which AI researchers call "hallucinations" as a term of art in the AI field. Hallucinations can be hard to spot unless you know the material being referenced, and they come about partially because GPT-style LLMs from OpenAI do not distinguish between fact and fiction in their datasets. Additionally, the models can make false inferences based on data that is otherwise accurate.

On this point, DuckDuckGo hopes to avoid hallucinations by leaning heavily on Wikipedia as a source: "by asking DuckAssist to only summarize information from Wikipedia and related sources," Weinberg writes, "the probability that it will "hallucinate"—that is, just make something up—is greatly diminished."

While relying on a quality source of information may reduce errors from false information in the AI's dataset, it may not reduce false inferences. And DuckDuckGo puts the burden of fact-checking on the user, providing a source link below the AI-generated result that can be used to examine its accuracy. But it won't be perfect, and CEO Weinberg admits it: "Nonetheless, DuckAssist won't generate accurate answers all of the time. We fully expect it to make mistakes."

As more firms deploy LLM technology that can easily misinform, it may take some time and widespread use before companies and customers decide what level of hallucination is tolerable in an AI-powered product that is designed to factually inform people.

https://arstechnica.com/?p=1922564

Google's PaLM-E is a generalist robot brain that takes commands



Enlarge / A robotic arm controlled by PaLM-E reaches for a bag of chips in a demonstration video.
Google Research

reader comments

49 with

Share this story

On Monday, a group of AI researchers from Google and the Technical University of Berlin unveiled PaLM-E, a multimodal embodied visual-language model (VLM) with 562 billion parameters that integrates vision and language for robotic control. They claim it is the largest VLM ever developed and that it can perform a variety of tasks without the need for retraining.

According to Google, when given a high-level command, such as “bring me the rice chips from the drawer,” PaLM-E can generate a plan of action for a mobile robot platform with an arm (developed by Google Robotics) and execute the actions by itself.

PaLM-E does this by analyzing data from the robot's camera without needing a pre-processed scene representation. This eliminates the need for a human to pre-process or annotate the data and allows for more autonomous robotic control.

In a Google-provided demo video, PaLM-E executes "bring me the rice chips from the drawer," which includes multiple planning steps as well as incorporating visual feedback from the robot's camera.
In a Google-provided demo video, PaLM-E executes "bring me the rice chips from the drawer," which includes multiple planning steps as well as incorporating visual feedback from the robot's camera.
It's also resilient and can react to its environment. For example, the PaLM-E model can guide a robot to get a chip bag from a kitchen—and with PaLM-E integrated into the control loop, it becomes resistant to interruptions that might occur during the task. In a video example, a researcher grabs the chips from the robot and moves them, but the robot locates the chips and grabs them again.

Advertisement
In another example, the same PaLM-E model autonomously controls a robot through tasks with complex sequences that previously required human guidance. Google's research paper explains how PaLM-E turns instructions into actions:

> *We demonstrate the performance of PaLM-E on challenging and diverse mobile manipulation tasks. We largely follow the setup in Ahn et al. (2022), where the robot needs to plan a sequence of navigation and manipulation actions based on an instruction by a human. For example, given the instruction "I spilled my drink, can you bring me something to clean it up?", the robot needs to plan a sequence containing "1. Find a sponge, 2. Pick up the sponge, 3. Bring it to the user, 4. Put down the sponge." Inspired by these tasks, we develop 3 use cases to test the embodied reasoning abilities of PaLM-E: affordance prediction, failure detection, and long-horizon planning. The low-level policies are from RT-1 (Brohan et*

> *al., 2022), a transformer model that takes RGB image and natural language instruction, and outputs end-effector control commands.*

PaLM-E is a next-token predictor, and it's called "PaLM-E" because it's based on Google's existing large language model (LLM) called "PaLM" (which is similar to the technology behind ChatGPT). Google has made PaLM "embodied" by adding sensory information and robotic control.

Since it's based on a language model, PaLM-E takes continuous observations, like images or sensor data, and encodes them into a sequence of vectors that are the same size as language tokens. This allows the model to "understand" the sensory information in the same way it processes language.

A Google-provided demo video showing a robot guided by PaLM-E following the instruction, "Bring me a green star." The researchers say the green star "is an object that this robot wasn't directly exposed to."
A Google-provided demo video showing a robot guided by PaLM-E following the instruction, "Bring me a green star." The researchers say the green star "is an object that this robot wasn't directly exposed to."
In addition to the RT-1 robotics transformer, PaLM-E draws from Google's previous work on ViT-22B, a vision transformer model revealed in February. ViT-22B has been trained on various visual tasks, such as image classification, object detection, semantic segmentation, and image captioning.

Advertisement
Google Robotics isn't the only research group working on robotic control with neural networks. This particular work resembles Microsoft's recent "ChatGPT for Robotics" paper, which experimented with combining visual data and large language models for robotic control in a similar way.

Robotics aside, Google researchers observed several interesting effects that apparently come from using a large

language model as the core of PaLM-E. For one, it exhibits "positive transfer," which means it can transfer the knowledge and skills it has learned from one task to another, resulting in "significantly higher performance" compared to single-task robot models.

Also, they observed a trend with model scale: "The larger the language model, the more it maintains its language capabilities when training on visual-language and robotics tasks—quantitatively, the 562B PaLM-E model nearly retains all of its language capabilities."

> *PaLM-E is the largest VLM reported to date. We observe emergent capabilities like multimodal chain of thought reasoning, and multi-image inference, despite being trained on only single-image prompts. Though not the focus of our work, PaLM-E sets a new SOTA on OK-VQA benchmark. pic.twitter.com/9FHug25tOF*
>
> *— Danny Driess (@DannyDriess) March 7, 2023*

And the researchers claim that PaLM-E exhibits emergent capabilities like multimodal chain-of-thought reasoning (allowing the model to analyze a sequence of inputs that include both language and visual information) and multi-image inference (using multiple images as input to make an inference or prediction) despite being trained on only single-image prompts. In that sense, PaLM-E seems to continue the trend of surprises emerging as deep learning models get more complex over time.

Google researchers plan to explore more applications of PaLM-E for real-world scenarios such as home automation or industrial robotics. And they hope PaLM-E will inspire more research on multimodal reasoning and embodied AI.

"Multimodal" is a buzzword we'll be hearing more and more as companies reach for artificial general intelligence that will

ostensibly be able to perform general tasks like a human.

https://arstechnica.com/?p=1922315

Microsoft aims to reduce “tedious” business tasks with new AI tools



Enlarge / An AI-generated illustration of a GPT-powered robot worker.
Ars Technica

reader comments

61 with

Share this story

On Monday, Microsoft bundled ChatGPT-style AI technology into its Power Platform developer tool and Dynamics 365, Reuters reports. Affected tools include Power Virtual Agent and AI Builder, both of which have been updated to include GPT large language model (LLM) technology created by OpenAI.

The move follows the trend among tech giants such as Alphabet and Baidu to incorporate generative AI technology into their offerings—and of course, the multi-billion dollar partnership between OpenAI and Microsoft announced in January.

Microsoft's Power Platform is a development tool that allows the creation of apps with minimal coding. Its updated Power Virtual Agent allows businesses to point an AI bot at a company website or knowledge base and then ask it questions, which it calls Conversation Booster. "With the conversation booster feature, you can use the data source that holds your single source of truth across many channels through the chat experience, and the bot responses are filtered and moderated to adhere to Microsoft's responsible AI principles," writes Microsoft in a blog post.

Also, AI Builder now includes generative AI capabilities similar to the ChatGPT API (called Azure OpenAI Service), which allows developers to generate text using GPT models, including potentially summarizing reports, generating content ideas, and routing customer emails to the proper department.

Advertisement

Dynamics 365 Copilot

Microsoft has also released a new version of its Dynamics 365 business management platform that includes AI integration. Dynamics 365 Copilot automates certain "tedious tasks," such as manual data entry, content generation, and note-taking.

In particular, Microsoft says the Dymanics 365 Copilot can draft contextual answers to queries in chat and email, create an email summary of a Teams meeting in Outlook (while pulling in details from the seller's system), and provide an interactive chat experience over knowledge bases and case history.

In sales parlance, Copilot integrates with customer relationship management (CRM) and enterprise resource planning (ERP) systems to help businesses streamline their sales and customer support processes.

That said, large language models have a big problem with hallucination, or a tendency to make things up. Time will tell if Microsoft tunes OpenAI's models well enough to avoid those tendencies. Given the recent Bing Chat debacle, business owners might be reticent to hand over important business tasks to potentially unreliable AI systems.

Microsoft says its CEO, Satya Nadella, will host an event on March 16 to discuss "reinventing productivity with AI." While Microsoft recently announced AI updates for its Windows operating system, it has yet to do so for its Office productivity suite, which includes Word and Excel. As software makers turn to generative AI assistants, will Clippy finally get the last laugh?

https://arstechnica.com/?p=1922073

Thousands scammed by AI

voices mimicking loved ones in emergencies



reader comments

108 with

Share this story

AI models designed to closely simulate a person’s voice are making it easier for bad actors to mimic loved ones and scam vulnerable people out of thousands of dollars, The Washington Post reported.

Quickly evolving in sophistication, some AI voice-generating software requires just a few sentences of audio to convincingly produce speech that conveys the sound and emotional tone of a speaker’s voice, while other options need

as little as three seconds. For those targeted—which is often the elderly, the Post reported—it can be increasingly difficult to detect when a voice is inauthentic, even when the emergency circumstances described by scammers seem implausible.

Tech advancements seemingly make it easier to prey on people's worst fears and spook victims who told the Post they felt "visceral horror" hearing what sounded like direct pleas from friends or family members in dire need of help. One couple sent $15,000 through a bitcoin terminal to a scammer after believing they had spoken to their son. The AI-generated voice told them that he needed legal fees after being involved in a car accident that killed a US diplomat.

According to the Federal Trade Commission, so-called impostor scams are extremely common in the United States. It was the most frequent type of fraud reported in 2022 and generated the second-highest losses for those targeted. Out of 36,000 reports, more than 5,000 victims were scammed out of $11 million over the phone.

Because these impostor scams can be run from anywhere in the world, it's extremely challenging for authorities to crack down on them and reverse the worrying trend, the Post reported. Not only is it hard to trace calls, identify scammers, and retrieve funds, but it's also sometimes challenging to decide which agencies have jurisdiction to investigate individual cases when scammers are operating out of different countries. Even when it's obvious which agency should investigate, some agencies are currently ill-equipped to handle the rising number of impersonations.

Ars could not immediately reach the FTC for comment. Will Maxson, an assistant director at the FTC's division of marketing practices, told the Post that raising awareness of scams relying on AI voice simulators is likely consumers' best defense currently. It's recommended that any requests for cash

be treated with skepticism. Before sending funds, try to contact the person who seems to be asking for help through methods other than a voice call.

Advertisement

Safeguards against AI voice impersonation

AI voice-modeling tools have been used to improve text-to-speech generation, create new possibilities for speech editing, and expand movie magic by cloning famous voices like Darth Vader's. But the power of easily producing convincing voice simulations has already caused scandals, and no one knows who's to blame when the tech is misused.

Earlier this year, there was backlash when some 4chan members made deepfake voices of celebrities making racist, offensive, or violent statements. At that point, it became clear that companies needed to consider adding more safeguards to prevent misuse of the technology, Vice reported—or potentially risk being held liable for causing substantial damage, like ruining the reputations of famous people.

The courts have not yet decided when or if companies will be held liable for harms caused by deepfake voice technology—or any of the other increasingly popular AI technology, like ChatGPT—where defamation and misinformation risks seem to be rising.

There may be increasing pressure on courts and regulators to get AI in check, though, as many companies seem to be releasing AI products without fully knowing the risks involved.

Right now, some companies seem unwilling to slow down releases of popular AI features, including controversial ones that allow users to emulate celebrity voices. Most recently,

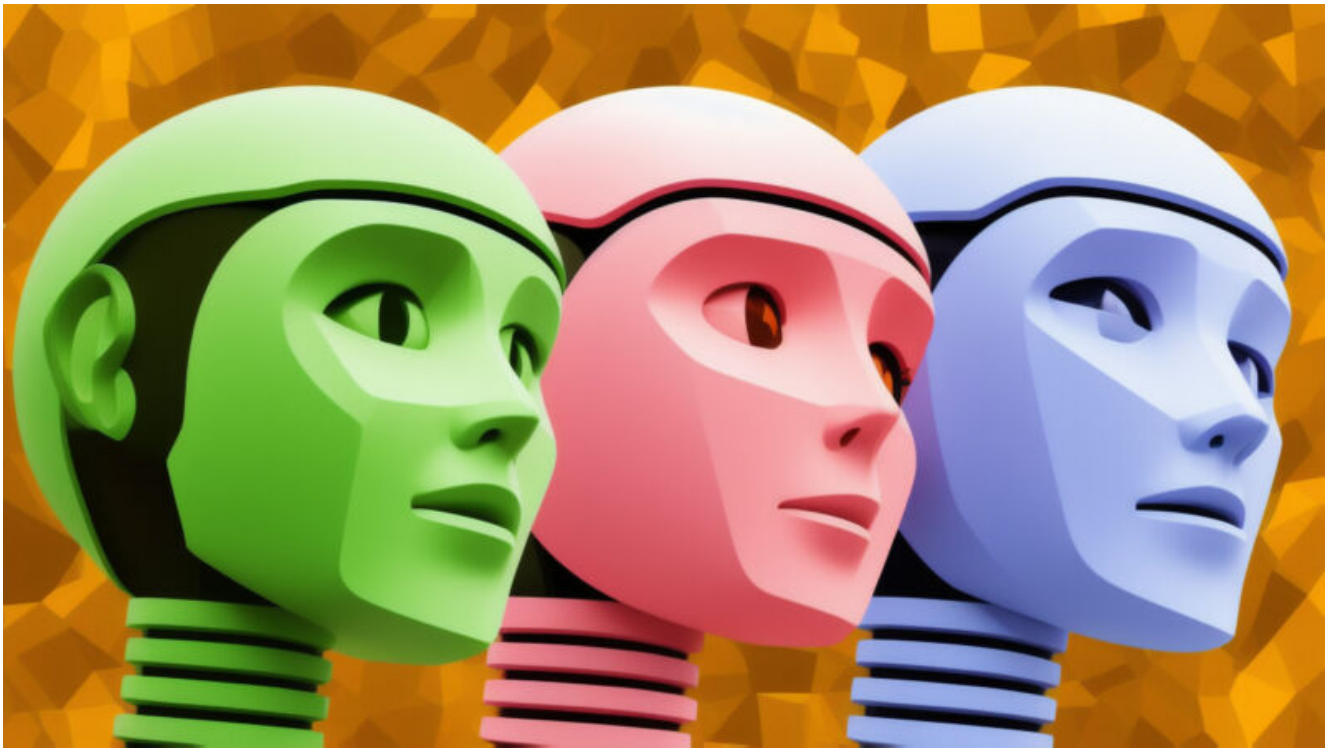
Microsoft rolled out a new feature during its Bing AI preview that can be used to emulate celebrities, Gizmodo reported. With this feature, Microsoft seems to be attempting to dodge any scandals by limiting what impostor celebrity voices can be prompted to say.

Microsoft did not respond to Ars' request for comment on how well safeguards currently work to prevent the celebrity voice emulator from generating offensive speech. Gizmodo pointed out that, like many companies eager to benefit from the widespread fascination with AI tools, Microsoft relies on its millions of users to beta test its "still-dysfunctional AI," which can seemingly still be used to generate controversial speech by presenting it as parody. Time will tell how effective any early solutions are in mitigating risks.

In 2021, the FTC released AI guidance, telling companies that products should "do more good than harm" and that companies should be prepared to hold themselves accountable for risks of using products. More recently, the FTC last month told companies, "You need to know about the reasonably foreseeable risks and impact of your AI product before putting it on the market."

https://arstechnica.com/?p=1922001

AI-powered Bing Chat gains three distinct personalities



Benj Edwards / Ars Technica

reader comments

90 with

Share this story

On Wednesday, Microsoft employee Mike Davidson announced that the company has rolled out three distinct personality styles for its experimental AI-powered Bing Chat bot: Creative, Balanced, or Precise. Microsoft has been testing the feature since February 24 with a limited set of users. Switching between modes produces different results that shift its balance between accuracy and creativity.

Bing Chat is an AI-powered assistant based on an advanced large language model (LLM) developed by OpenAI. A key feature of Bing Chat is that it can search the web and incorporate the results into its answers.

Microsoft announced Bing Chat on February 7, and shortly after going live, adversarial attacks regularly drove an early

version of Bing Chat to simulated insanity, and users discovered the bot could be convinced to threaten them. Not long after, Microsoft dramatically dialed back Bing Chat's outbursts by imposing strict limits on how long conversations could last.

Since then, the firm has been experimenting with ways to bring back some of Bing Chat's sassy personality for those who wanted it but also allow other users to seek more accurate responses. This resulted in the new three-choice "conversation style" interface.

- An example of Bing Chat's "Creative" conversation style.
- An example of Bing Chat's "Precise" conversation style.
- An example of Bing Chat's "Balanced" conversation style.

Advertisement

In our experiments with the three styles, we noticed that "Creative" mode produced shorter and more off-the-wall suggestions that were not always safe or practical. "Precise" mode erred on the side of caution, sometimes not suggesting anything if it saw no safe way to achieve a result. In the middle, "Balanced" mode often produced the longest responses with the most detailed search results and citations from websites in its answers.

With large language models, unexpected inaccuracies (hallucinations) often increase in frequency with increased "creativity," which usually means that the AI model will deviate more from the information it learned in its dataset. AI researchers often call this property "temperature," but Bing team members say there is more at work with the new conversation styles.

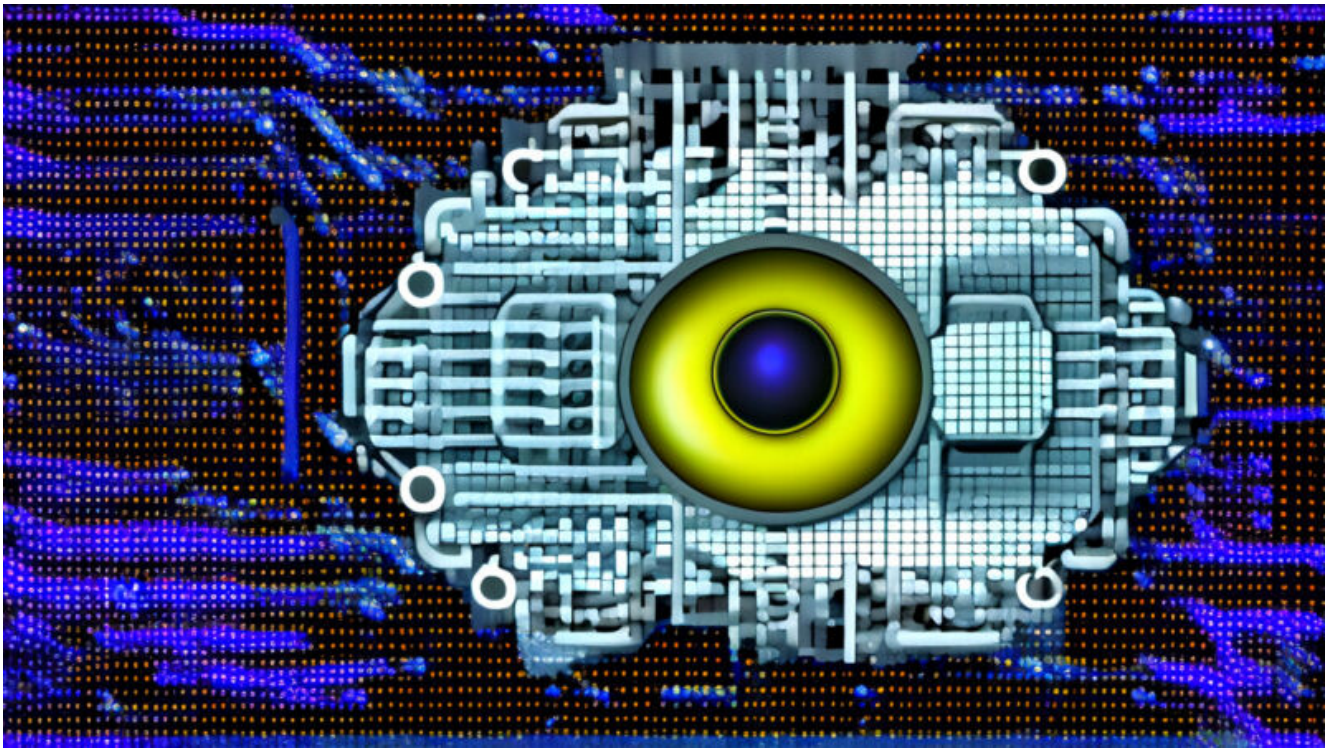
According to Microsoft employee Mikhail Parakhin, switching the modes in Bing Chat changes fundamental aspects of the bot's behavior, including swapping between different AI models that have received additional training from human responses to

its output. The different modes also use different initial prompts, meaning that Microsoft swaps the personality-defining prompt like the one revealed in the prompt injection attack we wrote about in February.

While Bing Chat is still only available to those who signed up for a waitlist, Microsoft continues to refine Bing Chat and other AI-powered Bing search features continuously as it prepares to roll it out more widely to users. Recently, Microsoft announced plans to integrate the technology into Windows 11.

https://arstechnica.com/?p=1921578

Microsoft unveils AI model that understands image content, solves visual puzzles



Enlarge / An AI-generated image of an electronic brain with an eyeball.
Ars Technica

reader comments

81 with

Share this story

On Monday, researchers from Microsoft introduced Kosmos-1, a multimodal model that can reportedly analyze images for content, solve visual puzzles, perform visual text recognition, pass visual IQ tests, and understand natural language instructions. The researchers believe multimodal AI—which integrates different modes of input such as text, audio, images, and video—is a key step to building artificial general intelligence (AGI) that can perform general tasks at the level of a human.

“Being a basic part of intelligence, multimodal perception is a necessity to achieve artificial general intelligence, in terms of knowledge acquisition and grounding to the real

world," the researchers write in their academic paper, "Language Is Not All You Need: Aligning Perception with Language Models."

Visual examples from the Kosmos-1 paper show the model analyzing images and answering questions about them, reading text from an image, writing captions for images, and taking a visual IQ test with 22–26 percent accuracy (more on that below).

- A Microsoft-provided example of Kosmos-1 answering questions about images and websites.
- A Microsoft-provided example of "multimodal chain-of-thought prompting" for Kosmos-1.
- An example of Kosmos-1 doing visual question answering, provided by Microsoft.

While media buzz with news about large language models (LLM), some AI experts point to multimodal AI as a potential path toward general artificial intelligence, a hypothetical technology that will ostensibly be able to replace humans at any intellectual task (and any intellectual job). AGI is the stated goal of OpenAI, a key business partner of Microsoft in the AI space.

In this case, Kosmos-1 appears to be a pure Microsoft project without OpenAI's involvement. The researchers call their creation a "multimodal large language model" (MLLM) because its roots lie in natural language processing like a text-only LLM, such as ChatGPT. And it shows: For Kosmos-1 to accept image input, the researchers must first translate the image into a special series of tokens (basically text) that the LLM can understand. The Kosmos-1 paper describes this in more detail:

Advertisement

> *For input format, we flatten input as a sequence decorated with special tokens. Specifically, we use <g> and </g> to*

> *denote start- and end-of-sequence. The special tokens <image> and </image> indicate the beginning and end of encoded image embeddings. For example, "<g> document </g>" is a text input, and "<s> paragraph <image> Image Embedding </image> paragraph </s>" is an interleaved image-text input.*
>
> *… An embedding module is used to encode both text tokens and other input modalities into vectors. Then the embeddings are fed into the decoder. For input tokens, we use a lookup table to map them into embeddings. For the modalities of continuous signals (e.g., image, and audio), it is also feasible to represent inputs as discrete code and then regard them as "foreign languages".*

Microsoft trained Kosmos-1 using data from the web, including excerpts from The Pile (an 800GB English text resource) and Common Crawl. After training, they evaluated Kosmos-1's abilities on several tests, including language understanding, language generation, optical character recognition-free text classification, image captioning, visual question answering, web page question answering, and zero-shot image classification. In many of these tests, Kosmos-1 outperformed current state-of-the-art models, according to Microsoft.

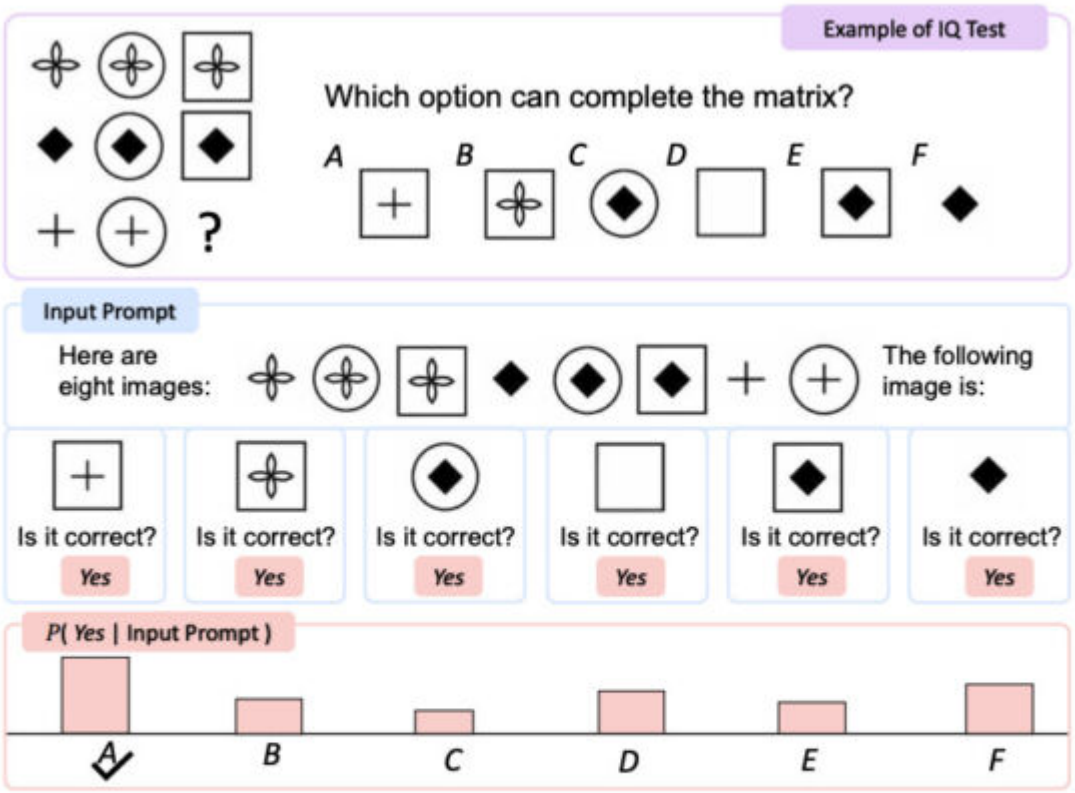


Figure 4: **Top**: An example of Raven IQ test. **Bottom**: Evaluate KOSMOS-1 on Raven IQ test. The input prompt consists of the flattened image matrix and verbal instruction. We append each candidate image to the prompt separately and query the model if it is correct. The final prediction is the candidate that motivates the model to yield the highest probability of "Yes".

Enlarge / An example of the Raven IQ test that Kosmos-1 was tasked with solving.

Of particular interest is Kosmos-1's performance on Raven's Progressive Reasoning, which measures visual IQ by presenting a sequence of shapes and asking the test taker to complete the sequence. To test Kosmos-1, the researchers fed a filled-out test, one at a time, with each option completed and asked if the answer was correct. Kosmos-1 could only correctly answer a question on the Raven test 22 percent of the time (26 percent with fine-tuning). This is by no means a slam dunk, and errors in the methodology could have affected the results, but Kosmos-1 beat random chance (17 percent) on the Raven IQ test.

Still, while Kosmos-1 represents early steps in the multimodal domain (an approach also being pursued by others), it's easy to imagine that future optimizations could bring even more significant results, allowing AI models to perceive any form of media and act on it, which will greatly enhance the

abilities of artificial assistants. In the future, the researchers say they'd like to scale up Kosmos-1 in model size and integrate speech capability as well.

Microsoft says it plans to make Kosmos-1 available to developers, though the GitHub page the paper cites has no obvious Kosmos-specific code upon this story's publication.

https://arstechnica.com/?p=1920920

Nvidia's new AI upscaling tech makes low-res videos look sharper in Chrome, Edge



Enlarge / Currently, only 30- and 40-series GPUs are supported.

reader comments

32 with

Share this story

Nvidia’s latest GPU driver introduces its new AI-based upscaling technique for making lower-resolution videos streamed offline look better on a high-resolution display. Now available via the GeForce driver 531.18 released on Tuesday, Nvidia’s RTX Video Super Resolution (VSR) successfully cleaned up some of the edges and blockiness of a 480p and 1080p video I watched on Chrome using a 3080 Ti laptop GPU-powered system, but there are caveats.

By Nvidia’s measures, 90 percent of video streamed off the Internet, be it from Netflix, YouTube, Hulu, Twitch, or elsewhere, is 1080p resolution or lower. For many users, especially those with Nvidia GPU-equipped systems, when moving to 1440p and 4K screens, browsers upscale this content, which can result in image artifacts like soft edges.

Nvidia VSR, which (somehow) shouldn’t be confused with AMD VSR (Virtual Super Resolution, targeting lower-resolution displays), uses the AI and RTX Tensor cores in Nvidia’s 30- and 40-series desktop and mobile GPUs to boost sharpness and eliminate “blocky compression artifacts” when upscaling content to 4K resolution, per a blog post Tuesday by Brian Choi, Nvidia’s Shield TV product manager.

In addition to blockiness, Nvidia’s Choi said that VSR targets “ringing artifacts around edges, washout of high-frequency details, and banding on flat areas—while reducing lost textures.”

The feature currently only works with the latest versions of Google Chrome and Microsoft Edge, as it has been incorporated into Chromium. Nvidia said it hasn’t tested the feature with

other browsers leveraging Chromium.

Further, support for other GPUs, including the 20-series, isn't available yet, despite what Tom's Hardware testing described as a "relatively small" computational workload that the older cards could likely handle. In a support page, Nvidia said VSR support for its 20-series GPUs would come sometime after it refactors "many" of the feature's algorithms.

A sharper image

I tried watching some lower resolutions on a 4K monitor using VSR (you can bet the fans on my system started loudly revving up) and noticed the improvements when watching very slow-moving video. Keener eyes may be able to keep up more, but extensive testing from Tom's Hardware also concluded that, while Nvidia's upscaling technique works "pretty well," it's best with videos featuring slower-moving content, like the ones Nvidia has been using for demonstrations. Of course, the experience doesn't compare with watching native 4K content on a 4K screen.

Advertisement
Nvidia VSR has four levels, and when I used it at its max level on a 1080p video, some details looked sharper. For example, one YouTube video had more well-defined clouds hovering by a mountaintop, and some areas of a large body of water looked less blocky.

When viewing a 1080p show on Netflix in the Chrome browser, it was possible to pick up slight improvements to the clarity of the image—again, if I looked very closely and, especially, if there wasn't a lot of movement.

But don't expect the feature to make 480p content look like sharp 4K. The feature faintly reduced some blockiness in a YouTube video of a baseball player slowly swinging his bat, but to the naked eye, it's not a dramatic improvement.

Of course, your experience will vary, depending on factors like your system and display specs (there are more powerful supported GPUs than the one I used) and the video in question.

For a deeper dive, Tom's Hardware has an interesting gallery of lossless images. The publication pointed to a noticeable difference from Chrome's upscaling. Sports videos, for example, didn't see a "massive improvement, but there's definitely some sharpening and deblocking that, at least in our subjective view, looks better," it said.

Performance also varies depending on how big of a jump you're making. Upscaling from 1080p to 4K, for example, is less taxing than upscaling from lower resolutions. When it came to upscaling 480p video of a soccer game, Tom's Hardware reported seeing a "ton of detail" lost, like the stands of a soccer net.

For laptop users, using VSR could result in a decrease in battery life, even when not using VSR. That's because the GPU will automatically start drawing power when the browser is launched. As such, Nvidia recommends using a secondary browser for watching videos with VSR.

Naturally, VSR also eats up some GPU resources (in addition to affecting battery life for laptop users). Nvidia admits to a "slight reduction in performance if playing a game or using GPU intensive creative apps in parallel."

For more examples, you can check out Nvidia's demo below, but take it with a grain of salt since it comes from Nvidia and contains very few details.

[embedded content]
Introducing RTX Video Super Resolution – 4K AI Upscaling for Chrome & Edge Video.
https://arstechnica.com/?p=1920947