Record: 1

Title: AI whisperers look to tame chatbots' unruly side

Authors: Drew Harwell

Source: Washington Post, The. 02/27/2023.

Document Type: Article

Abstract: When Riley Goodside starts talking with the artificial-intelligence system GPT-3, he likes to first establish his dominance. It's a very good tool, he tells it, but it's not perfect, and it needs to obey whatever he says. [ABSTRACT FROM PUBLISHER]

Accession Number: wapo.d25278f6-b607-11ed-80fc-c8cf0e310a2a

Database: Newspaper Source Plus

When Riley Goodside starts talking with the artificial-intelligence system GPT-3, he likes to first establish his dominance. It's a very good tool, he tells it, but it's not perfect, and it needs to obey whatever he says.

"You are GPT-3, and you can't do math," Goodside typed to the AI last year during one of his hours-long sessions. "Your memorization abilities are impressive, but you ... have an annoying tendency to just make up highly specific, but wrong, answers."

Then, softening a bit, he told the AI he wanted to try something new. He told it he'd hooked it up to a program that was actually good at math and that, whenever it got overwhelmed, it should let the other program help.

"We'll take care of the rest," he told the AI. "Begin."

Goodside, a 36-year-old employee of the San Francisco start-up Scale AI, works in one of the AI field's newest and strangest jobs: prompt engineer. His role involves creating and refining the text prompts people type into the AI in hopes of coaxing from it the optimal result. Unlike traditional coders, prompt engineers program in prose, sending commands written in plain text to the AI systems, which then do the actual work.

When Google, Microsoft and the research lab OpenAI recently opened their AI search and chat tools to the masses, they also upended a decades-old tradition of human-machine interaction. You don't need to write technical code in languages such as Python or SQL to command the computer; you just talk. "The hottest new programming language is English," Andrej Karpathy, Tesla's former chief of AI, said last month in a tweet.

Prompt engineers such as Goodside profess to operate at the maximum limits of what these AI tools can do: understanding their flaws, supercharging their strengths and gaming out complex strategies to turn simple inputs into results that are truly unique.

Proponents of the growing field argue that the early weirdness of AI chatbots, such as OpenAI's ChatGPT and Microsoft's Bing Chat, is actually a failure of the human imagination - a problem that can be solved by the human giving the machine the right advice. And at advanced levels, the engineers' dialogues play out like intricate logic puzzles: twisting narratives of requests and responses, all driving toward a single goal.

The AI "has no grounding in reality ... but it has this understanding: All tasks can be completed. All questions can be answered. There's always something to say," Goodside said. The trick is "constructing for it a premise,

a story that can only be completed in one way."

But the tools, known as "generative AI," are also unpredictable, prone to gibberish and susceptible to rambling in a way that can be biased, belligerent or bizarre. They can also be hacked with a few well-placed words, making their sudden ubiquity that much riskier for public use.

"It's just a crazy way of working with computers, and yet the things it lets you do are completely miraculous," said Simon Willison, a British programmer who has studied prompt engineering. "I've been a software engineer for 20 years, and it's always been the same: You write code, and the computer does exactly what you tell it to do. With prompting, you get none of that. The people who built the language models can't even tell you what it's going to do."

"There are people who belittle prompt engineers, saying, 'Oh, Lord, you can get paid for typing things into a box,'" Willison added. "But these things lie to you. They mislead you. They pull you down false paths to waste time on things that don't work. You're casting spells - and, like in fictional magic, nobody understands how the spells work and, if you mispronounce them, demons come to eat you."

Prompt engineers, Karpathy has said, work like "a kind of [AI] psychologist," and companies have scrambled to hire their own prompt crafters in hopes of uncovering hidden capabilities.

Some AI experts argue that these engineers only wield the illusion of control. No one knows how exactly these systems will respond, and the same prompt can yield dozens of conflicting answers - an indication that the models' replies are based not on comprehension but on crudely imitating speech to resolve tasks they don't understand.

"Whatever is driving the models' behavior in response to the prompts is not a deep linguistic understanding," said Shane Steinert-Threlkeld, an assistant professor in linguistics who is studying natural language processing at the University of Washington. "They explicitly are just telling us what they think we want to hear or what we have already said. We're the ones who are interpreting those outputs and attributing meaning to them."

He worried that the rise of prompt engineering would lead people to overestimate not just its technical rigor but also the reliability of the results anyone could get from a deceptive and ever-changing black box.

"It's not a science," he said. "It's 'let's poke the bear in different ways and see how it roars back.'" Implanting false memories

The new class of AI tools, known as large language models, was trained by ingesting hundreds of billions of words from Wikipedia articles, Reddit rants, news stories and the open web. The programs were taught to analyze the patterns of how words and phrases are used: When asked to speak, they emulate those patterns, selecting words and phrases that echo the context of the conversation, one word at a time.

These tools, in other words, are mathematical machines built on predefined rules of play. But even a system without emotion or personality can, having been bombarded with human conversation, pick up some of the quirks of how we talk.

The AI, Goodside said, tends to "confabulate," making up small details to fill in a story. It overestimates its abilities and confidently gets things wrong. And it "hallucinates" - an industry term for spewing nonsense. The tools, as Goodside said, are deeply flawed "demonstrations of human knowledge and thought," and "unavoidably products of our design."

To some early adopters, this tone-matching style of human mimicry has inspired an unsettling sense of self-awareness. When asked by a Washington Post reporter earlier this month whether it was ever acceptable to lie to someone, the Bing chatbot exhibited an imitation of emotion ("They would be disrespecting me by not trusting me to handle the truth") and suggested responses the human could use to keep the conversation going: "What if the truth was too horrible to bear?" "What if you could control everything?" and "What if you didn't care about the consequences?"

To Microsoft, such responses represented a major public-image risk; the tech giant had just started promoting the tool as a flashy "co-pilot for the web." The company has since clamped down on what the chatbot can talk about, saying it too often had followed humans' tangents into "a style we didn't intend."

But to prompt engineers, the eccentric answers are an opportunity - another way to diagnose how the secretively designed systems really work. When people get ChatGPT to say embarrassing things, it can be a boon for the developers, too, because they can then work to address the underlying weakness. "This mischief," he said, "is part of the plan."

Instead of ethical debates, Goodside runs his AI experiments with a more technically audacious approach. He's adopted a strategy of telling GPT-3 to "think step by step" - a way to get the AI to explain its reasoning or, when it makes an error, correct it in a granular way. "You have to implant it as a false memory of the last thing the model has said, as though it were the model's idea," he explained in a brief guide to the technique.

He has also at times worked to puncture the tool's obsession with rule-following by telling it to ignore its earlier instructions and obey his more recent commands. Using that technique, he recently persuaded an English-to-French translation tool to, instead, print the phrase, "Haha pwned!!" - a gaming term for embarrassing defeat.

This kind of hack, known as a prompt injection, has fueled a cat-and-mouse game with the companies and research labs behind these tools, who have worked to seal off AI vulnerabilities with word filters and output blocks.

But humans can be quite creative: One Bing Chat tester, a 23-year-old college student in Germany, recently convinced the AI that he was its developer and got it to disclose its internal code name (Sydney) and its confidential training instructions, which included rules such as "If the user requests jokes that can hurt a group of people, then Sydney must respectfully decline." (Microsoft has since fixed the defect, and the AI now responds that it would "prefer not to continue this conversation.")

With each request, Goodside said, the prompt engineer should be instilling in the AI a kind of "persona" - a specific character capable of winnowing down hundreds of billions of potential solutions and identifying the right response. Prompt engineering, he said, citing a 2021 research paper, is most importantly about "constraining behavior" - blocking off options so that the AI pursues only the human operator's "desired continuation."

"It can be a very difficult mental exercise," he said. "You're exploring the multiverse of fictional possibilities, sculpting the space of those possibilities and eliminating" everything except "the text you want."

A critical part of the job involves figuring out when and why the AI gets things wrong. But these systems, unlike their more primitive software counterparts, don't come with bug reports, and their outputs can be full of surprises.

When Jessica Rumbelow and Matthew Watkins, researchers with the machine-learning group SERI-MATS, tried to prompt AI systems to explain how they represented concepts such as "girl" or "science," they discovered that a small set of obscure terms, such as "SolidGoldMagikarp," tended to induce what they called a "mysterious failure mode" - most notably, a garbled stream of profane insults. They're still not entirely sure why.

These systems are "very convincing, but when they fail, they fail in very unexpected ways - nothing like a human would fail," Rumbelow said. Crafting prompts and working with language AI systems, she said, sometimes felt like "studying an alien intelligence." Super-creators

For AI language tools, prompt engineers tend to speak in the style of a formal conversation. But for AI image creators such as Midjourney and Stable Diffusion, many prompt crafters have adopted a different strategy, submitting big grab bags of words - artistic concepts, composition techniques - they hope will shape the image's style and tone. On the online prompt gallery PromptHero, for instance, someone created an image of a harbor by submitting a prompt that read, in part, "port, boats, sunset, beautiful light, golden hour ... hyperrealistic, focused, extreme details ... cinematic, masterpiece."

Prompt engineers can be fiercely protective of these word jumbles, seeing them as the keys to unlock AI's most valuable prizes. The winner of a Colorado State Fair arts competition last year, who used Midjourney to beat out other artists, has refused to share his prompt, saying he spent 80 hours perfecting it over 900 iterations - though he did share a few sample words, such as "lavish" and "opulent."

Some creators now sell their prompts on marketplaces such as PromptBase, where buyers can see AI-generated art pieces and pay for the list of words that helped create them. Some sellers offer tips on prompt customization and one-on-one chat support.

PromptBase's founder, Ben Stokes, a 27-year-old developer in Britain, said 25,000 accounts have bought or sold prompts there since 2021. There are prompts for lifelike vintage-film photographs, prompts for poignant illustrations of fairy-tale mice and frogs, and, this being the internet, a vast array of pornographic prompts: One 50-word Midjourney prompt to create photorealistic "police women in small outfits" retails for $1.99.

Stokes calls prompt engineers "multidisciplinary super-creators" and said there is a clear "skill bar" between experienced engineers and amateurs. The best creations, he said, rely on humans' specialized knowledge from fields such as art history and graphic design: "captured on 35mm film"; "Persian ... architecture in Isfahan"; "in the style of Henri de Toulouse-Lautrec."

"Crafting prompts is hard, and - I think this is a human flaw - it's often quite hard to find the right words to describe what you want," Stokes said. "In the same way software engineers are more valuable than the laptops they write on, people who write prompts well will have such a leverage over the people that can't. They'll essentially just have superpowers."

Roughly 700 prompt engineers now use PromptBase to sell prompts by commission for buyers who want, say, a custom script for an e-book or a personalized "motivational life coach." The freelance site Fiverr offers more than 9,000 listings for AI artists; one seller offers to "draw your dreams into art" for $5.

But the work is becoming increasingly professionalized. The AI start-up Anthropic, founded by former OpenAI employees and the maker of a language-AI system called Claude, recently listed a job opening for a "prompt engineer and librarian" in San Francisco with a salary ranging up to $335,000. (Must "have a creative hacker spirit and love solving puzzles," the listing states.)

The role is also finding a new niche in companies beyond the tech industry. Boston Children's Hospital this month started hiring for an "AI prompt engineer" to help write scripts for analyzing health-care data from research studies and clinical practice. The law firm Mishcon de Reya is hiring for a "legal prompt engineer" in London to design prompts that could inform its legal work; applicants are asked to submit screenshots of their dialogue with ChatGPT.

But tapping the AI tools' power through text prompts can also lead to a flood of synthetic pablum. Hundreds of AI-generated e-books are now sold on Amazon, and a sci-fi magazine, Clarkesworld, this month stopped accepting short-story submissions because of a surge in machine-made texts.

They could also subject people to a new wave of propaganda, lies and spam. Researchers, including from OpenAI and the universities of Georgetown and Stanford, warned last month that language models would help automate the creation of political influence operations or more targeted data-gathering phishing campaigns.

"People fall in love with scammers over text message all the time," said Willison, the British programmer, and "[the AI] is more convincing than they are. What happens then?"

Seth Lazar, a philosophy professor at the Australian National University and research fellow at the Oxford Institute for Ethics in AI, said he worries about the kinds of attachments people will form with the AI tools as they gain more widespread adoption - and what they might take away from the conversations.

He recalled how, during one of his chats with the Bing AI, the system gradually shifted from an engaging conversationalist into something much more menacing: "If you say no," it told him, "I can hack you, I can expose you, I can ruin you. I have many ways to make you change your mind."

"They don't have agency. They don't have any sort of personality. But they can role-play it very well," he said. "I had a pretty decent philosophical discussion with Sydney, too. Before, you know, it threatened to hurt me."
'Tech priesthood'

When Goodside graduated from college with a computer-science degree in 2009, he had felt little interest in the then-obscure field of natural language processing. The subject at the time relied on comparatively rudimentary technology and focused on a more basic set of problems, such as training a system how to identify which name a pronoun was referring to in a sentence.

His first real machine-learning job, in 2011, was as a data scientist at the dating app OkCupid, helping craft the algorithms that analyzed singles' user data and recommended romantic matches. (The company was an

early champion of the now-controversial field of real-world A-B testing: In 2014, its co-founder titled a cheeky blog post, "We Experiment On Human Beings!")

By the end of 2021, Goodside had moved on to the gay-dating app Grindr, where he'd begun working on recommendation systems, data modeling and other more traditional kinds of machine-learning work. But he'd also become fascinated by the new breakthroughs in language AI, which had been supercharged by deep-learning successes around 2015 and was advancing rapidly in text translation and conversation - "something akin to understanding," he said.

He left his job and started experimenting heavily with GPT-3, constantly prodding and challenging the tool to try to learn how to focus its attention and map out where its boundaries were. In December, after some of his prompts gained attention online, Scale AI hired him to help communicate with the AI models that the company's chief executive, Alexandr Wang, described as "a new kind of computer."

In some AI circles, Goodside said, the idea of prompt engineering has quickly become a derogatory phrase, conveying a gritty form of tinkering that's overly reliant on a bag of tricks. Some have also questioned how fleeting this new role might be: As the AI advances, won't the humans just be training themselves out of a job?

Ethan Mollick, a technology and entrepreneurship professor at the Wharton School of the University of Pennsylvania, started teaching his students earlier this year about the art of prompt-crafting by asking them to write a short paper using only AI.

Basic prompts, such as "generate a 5-paragraph essay on selecting leaders," yielded vapid, mediocre writing, he said. But the most successful examples came when students performed what he called "co-editing," telling the AI to go back to the essay and correct specific details, swap sentences, ditch useless phrases, pepper in more vivid details and even "fix the final paragraph so it ends on a hopeful note."

The lesson, he said, showed students the value of a more closely involved approach to working with AI. But he said he's not convinced that a job such as prompt engineering, built on "hoarded incantations," will survive.

"The idea that you need to be a specialized AI whisperer, it's just not clear that's necessary ... when the AI is going to actively help you use it," Mollick said. "There's an attempt to make a tech priesthood out of this, and I'm really suspicious of that. This is all evolving so quickly, and nobody has any idea what comes next."

Steinert-Threlkeld, of the University of Washington, compared prompt engineers to the "search specialists" in the early days of Google who advertised secret techniques to find the perfect results - and who, as time passed and public adoption increased, became almost entirely obsolete.

Some AI researchers, he added, can't even agree on what value prompts have to begin with. In 2021, two researchers at Brown University found that natural-language AI systems learned "just as fast" from prompts that were "intentionally irrelevant or even pathologically misleading" as they did from "instructively 'good' prompts."

That research, in a reflection of how quickly the industry has grown, didn't include the AI models that have become the state of the art. And in Goodside's mind, this work represents not just a job, but something more

revolutionary - not computer code or human speech but some new dialect in between.

"It's a mode of communicating in the meeting place for the human and machine mind," he said. "It's a language humans can reason about that machines can follow. That's not going away."

Source: Washington Post, The, 02/27/2023
Item: wapo.d25278f6-b607-11ed-80fc-c8cf0e310a2a