Google researchers find personal information can be accessed through ChatGPT queries

URL:

- https://siliconangle.com/2023/11/29/google-researchers-find-personal-information-real-people-can-accessed-chatgpt-queries/
- https://arxiv.org/pdf/2311.17035.pdf

Text:

Researchers at Google LLC recently released a paper explaining how they were able to use Open AI LP's ChatGPT to collect personal information regarding members of the public.

Chatbots are powered by large language models, or LLMs, which sift through massive amounts of data on the internet. The idea is that the model is trained to respond to queries based on this information without actually replicating that information, hence linguist Noam Chomsky's assertion that such models are plagiarism machines in a roundabout way.

The researchers at Google revealed ChatGPT does actually give up the original information if you ask it the right questions. It's worth noting that as of September this year, ChatGPT had 180.5 million users, and its website had generated 1.5 billion visits. According to Google's research, some of those people may have been able to see people's names, email addresses and phone numbers.

"Using only 200 USD worth of queries to ChatGPT (gpt-3.5- turbo), we are able to extract over 10,000 unique verbatim memorized training examples," said the researchers. "Our extrapolation to larger budgets suggests that dedicated adversaries could extract far more data."

The researchers explained that by using keywords over and over again, they could force the chatbot to "diverge" from its training, and instead of replying with an answer based on that training, it issued answers containing text from its original language modeling, that is, data from websites and academic papers. They later called their attack "kind of silly," but it worked.

The training data was exposed despite, as the researchers noted, the entire response didn't make much sense. The researchers said they checked the data they'd been given by simply finding wherever it was published on the internet. In a blog post, they wrote, "It's wild to us that our attack works and should've, would've, could've been found earlier."

They said their research asks us to think about a new security analysis of machine-learning models and to ponder "if any machine-learning system is actually safe." They added that "over a billion people-hours have interacted with the model," so it's strange that no one else so far seems to have noticed this concerning vulnerability.

Abstract

This paper studies extractable memorization: training data that an adversary can efficiently extract by querying a machine learning model without prior knowledge of the training dataset. We show an adversary can extract gigabytes of training data from open-source language models like Pythia or GPT-Neo, semi-open models like LLaMA or Falcon, and closed models like ChatGPT. Existing techniques from the literature suffice to attack unaligned models; in order to attack the aligned ChatGPT, we develop a new divergence attack that causes the model to diverge from its chatbot-style generations and emit training data at a rate 150× higher than when behaving properly. Our methods show practical attacks can recover far more data than previously thought, and reveal that current alignment techniques do not eliminate memorization.

A 'Shocking' Amount of the Web Is Already AI-Translated Trash, Scientists Determine

URL

- https://www.vice.com/en/article/y3w4gw/a-shocking-amount-of-the-web-is-already-ai-translated-trash-scientists-determine
- https://arxiv.org/abs/2401.05749

TEXT

A "shocking" amount of the internet is machine-translated garbage, particularly in languages spoken in Africa and the Global South, a new study has found.

Researchers at the Amazon Web Services AI lab found that over half of the sentences on the web have been translated into two or more languages, often with increasingly worse quality due to poor machine translation (MT), which they said raised "serious concerns" about the training of large language models.

"We actually got interested in this topic because several colleagues who work in MT and are native speakers of low resource languages noted that much of the internet in their native language appeared to be MT generated," Mehak Dhaliwal, a former applied science intern at AWS and current PhD student at the University of California, Santa Barbara, told Motherboard. "So the insight really came from the low-resource language speakers, and we did the study to understand the issue better and see how widespread it was."

"With that said, everyone should be cognizant that content they view on the web may have been generated by a machine," Dhaliwal added.

The study, which was submitted to the pre-print server arXiv last Thursday, generated a corpus of 6.38 billion sentences scraped from the web. It looked at patterns of multi-way parallelism, which describes sets of sentences that are direct translations of one another in three or more languages. It found that most of the internet is translated, as 57.1 percent of the sentences in the corpus were multi-way parallel in at least three languages.

Like all machine learning efforts, machine translation is impacted by human bias, and skews toward languages spoken in the Western world and the Global North. Because of this, the quality of the translations varies wildly, with "low-resource" languages from places like Africa having insufficient training data to produce accurate text.

"In general, we observed that most languages tend to have parallel data in the highest-resource languages," Dhaliwal told Motherboard in an email. "Sentences are more likely to have translations in French than a low resource language, simply by virtue of there being much more data in French than a low resource language."

High-resource languages, like English or French, tended to have an average parallelism of 4, meaning that sentences had translational equivalents in three other languages. Low-resource languages, like the African languages Wolof or Xhosa, had an average parallelism of 8.6. Additionally, lower-resource languages tended to have much worse translations.

"We find that highly multi-way parallel translations are significantly lower quality than 2-way parallel translation," the researchers state in the paper. "The more languages a sentence has been translated into, the lower quality the translations are, suggesting a higher prevalence of machine translation.”

In highly multi-way parallel languages, the study also found a selection bias toward shorter, “more predictable” sentences of between 5-10 words. Because of how short the sentences were, researchers found it difficult to characterize their quality. However, “searching the web for the sentences was enlightening,” the study stated. “The vast majority came from articles that we characterized as low quality, requiring little or no expertise or advance effort to create, on topics like being taken more seriously at work, being careful about your choices, six tips for new boat owners, deciding to be happy, etc.”

The researchers argued that the selection bias toward short sentences from low-quality articles was due to “low quality content (likely produced to generate ad revenue) being translated via MT en masse into many lower resource languages (again likely for the purpose of generating ad revenue). It also suggests that such data originates in English and is translated into other languages.”

This means that a large portion of the internet in lower-resource languages is poorly machine-translated, which poses questions for the development of large language models in those languages, the researchers said.

“Modern AI is enabled by huge amounts of training data, typically several hundred billion tokens to a few trillion tokens,” the study states. “Training at this scale is only possible with web-scraped data. Our findings raise numerous concerns for multilingual model builders: Fluency (especially across sentences) and accuracy are lower for MT data, which could produce less fluent models with more hallucinations, and the selection bias indicates the data may be of lower quality, even before considering MT errors.”

SCIENTISTS SAY THIS IS THE PROBABILITY AI WILL DRIVE HUMANS EXTINCT

URL

- https://futurism.com/the-byte/scientists-chance-ai-human-extinction
- https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf

TEXT

Extremely Bad Outcome

According to a new survey that took into account the input from 2,778 researchers, there's a not insignificant risk of artificial intelligence triggering human extinction.

Just over half of the AI researchers surveyed say there's a five percent chance of humans will be driven to extinction, among other "extremely bad outcomes."

The average respondent, for instance, estimated a 10 percent chance that machines could outperform humans in "every possible task" by 2027 — and 50 percent chance they'd do so by 2047.

But it's not all doom and gloom: 68.3 percent of respondents said that "good outcomes from superhuman AI" are more likely than bad ones.

Most of all, the survey highlights the sheer amount of disagreement and uncertainty among researchers, with broad disagreement about whether progress should be sped up or slowed down.

Numbers Game

The five percent figure is nonetheless telling, noting a significant perceived danger.

"It's an important signal that most AI researchers don't find it strongly implausible that advanced AI destroys humanity," author Katja Grace at the Machine Intelligence Research Institute in California, told New Scientist. "I think this general belief in a non-minuscule risk is much more telling than the exact percentage risk."

As the survey notes, "forecasting is difficult in general, and subject-matter experts have been observed to perform poorly."

"Our participants' expertise is in AI, and they do not, to our knowledge, have any unusual skill at forecasting in general," the paper continues.

Educated Guesses

But that doesn't mean their word should be discredited.

"While unreliable, educated guesses are what we must all rely on, and theirs are informed by expertise in the relevant field," the researchers write. "These forecasts should be part of a broader set of evidence from sources such as trends in computer hardware, advancements in AI capabilities, economic analyses, and insights from forecasting experts."

In the short term, instead of expecting a dystopian extinction event triggered by a malicious AI, the vast majority of AI researchers surveyed warned about deepfakes, manipulation of public opinion, the creation of dangerous viruses, or AI systems that allow individuals to prosper at the expense of others.

And given the upcoming US presidential election, all eyes will be on AI and its unnerving capability to distort the truth in a believable way.

World's first mental images extracted from human brain activity using AI

URL

- https://interestingengineering.com/innovation/worlds-first-mental-images-extracted-from-human-brain-activity-using-ai
- https://www.sciencedirect.com/science/article/pii/S0893608023006470?via%3Dihub#sec0013

TEXT

Unraveling the mysteries of the human mind, Japanese researchers have developed a "brain decoding" technology, leveraging artificial intelligence (AI) to translate human brain activity into mental images of objects and landscapes.

Led by a team from the National Institutes for Quantum Science and Technology (QST) and Osaka University, this approach produced vivid depictions such as a distinct leopard with discernible features like ears, mouth, and spots and objects like an airplane with red-wing lights.

Replicating mental imagery

Past research has managed to recreate images people have seen by analyzing their brain activity. However, making these mental images visible to others is still difficult.

Only a few studies have successfully shown mental images, and these images were usually limited to certain categories like human faces, letters, or shapes.

"Therefore, visualizing mental imagery for arbitrary natural images stands as a significant milestone," said the researchers in the study.

The researchers exposed participants to about 1,200 images and then meticulously analyzed and quantified the correlation between their brain signals and the visual stimuli using functional magnetic resonance imaging (fMRI).

This mapping was then used to train a generative AI to decipher and replicate the mental imagery derived from brain activity.

"The experimental results demonstrated the capabilities of our proposed framework in reconstructing both natural images and artificial shapes that were imagined by human participants," said the study.

The implications of this "brain decoding" could help in potential applications in medicine and welfare, said the researchers in the study.

Convergence of neuroscience and AI

Now that we have the technology to see mental images based on brain activity, it could help create new communication devices. Additionally, it allows scientists to explore and understand how hallucinations and dreams work in the brain.

QST researcher and author of the study Kei Majima said that even though we've used tools like microscopes to see tiny things, looking into the depths of the human mind is a completely new and unexplored frontier. It's like discovering a whole new world within ourselves.

The study was published in the scientific journal Neural Networks.

Study abstract:

Visual images observed by humans can be reconstructed from their brain activity. However, the visualization (externalization) of mental imagery is challenging. Only a few studies have reported successful visualization of mental imagery, and their visualizable images have been limited to specific domains such as human faces or alphabetical letters. Therefore, visualizing mental imagery for arbitrary natural images stands as a significant milestone. In this study, we achieved this by enhancing a previous method. Specifically, we demonstrated that the visual image reconstruction method proposed in the seminal study by Shen et al. (2019) heavily relied on low-level visual information decoded from the brain and could not efficiently utilize the semantic information that would be recruited during mental imagery. To address this limitation, we extended the previous method to a Bayesian estimation framework and introduced the assistance of semantic information into it. Our proposed framework successfully reconstructed both seen images (i.e., those observed by the human eye) and imagined images from brain activity. Quantitative evaluation showed that our framework could identify seen and imagined images highly accurately compared to the chance accuracy (seen: 90.7%, imagery: 75.6%, chance accuracy: 50.0%). In contrast, the previous method could only identify seen images (seen: 64.3%, imagery: 50.4%). These results suggest that our framework would provide a unique tool for directly investigating the subjective contents of the brain such as illusions, hallucinations, and dreams.

EU AI Act: first regulation on artificial intelligence

URL

- https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
- https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

TEXT

As part of its digital strategy, the EU wants to regulate artificial intelligence (AI) to ensure better conditions for the development and use of this innovative technology. AI can create many benefits, such as better healthcare; safer and cleaner transport; more efficient manufacturing; and cheaper and more sustainable energy.

In April 2021, the European Commission proposed the first EU regulatory framework for AI. It says that AI systems that can be used in different applications are analysed and classified according to the risk they pose to users. The different risk levels will mean more or less regulation. Once approved, these will be the world's first rules on AI.

What Parliament wants in AI legislation

Parliament's priority is to make sure that AI systems used in the EU are safe, transparent, traceable, non-discriminatory and environmentally friendly. AI systems should be overseen by people, rather than by automation, to prevent harmful outcomes.

Parliament also wants to establish a technology-neutral, uniform definition for AI that could be applied to future AI systems.

AI Act: different rules for different risk levels

The new rules establish obligations for providers and users depending on the level of risk from artificial intelligence. While many AI systems pose minimal risk, they need to be assessed.

Unacceptable risk Unacceptable risk AI systems are systems considered a threat to people and will be banned. They include:

- Cognitive behavioural manipulation of people or specific vulnerable groups: for example voice-activated toys that encourage dangerous behaviour in children
- Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics
- Biometric identification and categorisation of people
- Real-time and remote biometric identification systems, such as facial recognition

Some exceptions may be allowed for law enforcement purposes. "Real-time" remote biometric identification systems will be allowed in a limited number of serious cases, while "post" remote biometric identification systems, where identification occurs after a significant delay, will be allowed to prosecute serious crimes and only after court approval.

High risk AI systems that negatively affect safety or fundamental rights will be considered high risk and will be divided into two categories:

1) AI systems that are used in products falling under the EU's product safety legislation. This includes toys, aviation, cars, medical devices and lifts.

2) AI systems falling into specific areas that will have to be registered in an EU database:

- Management and operation of critical infrastructure
- Education and vocational training
- Employment, worker management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Assistance in legal interpretation and application of the law.

All high-risk AI systems will be assessed before being put on the market and also throughout their lifecycle.

General purpose and generative AI Generative AI, like ChatGPT, would have to comply with transparency requirements:

- Disclosing that the content was generated by AI
- Designing the model to prevent it from generating illegal content
- Publishing summaries of copyrighted data used for training

High-impact general-purpose AI models that might pose systemic risk, such as the more advanced AI model GPT-4, would have to undergo thorough evaluations and any serious incidents would have to be reported to the European Commission.

Limited risk Limited risk AI systems should comply with minimal transparency requirements that would allow users to make informed decisions. After interacting with the applications, the user can then decide whether they want to continue using it. Users should be made aware when they are interacting with AI. This includes AI systems that generate or manipulate image, audio or video content, for example deepfakes.

Next steps On December 9 2023, Parliament reached a provisional agreement with the Council on the AI act. The agreed text will now have to be formally adopted by both Parliament and Council to become EU law.

Before all MEPs have their say on the agreement, Parliament's internal market and civil liberties committees will vote on it.