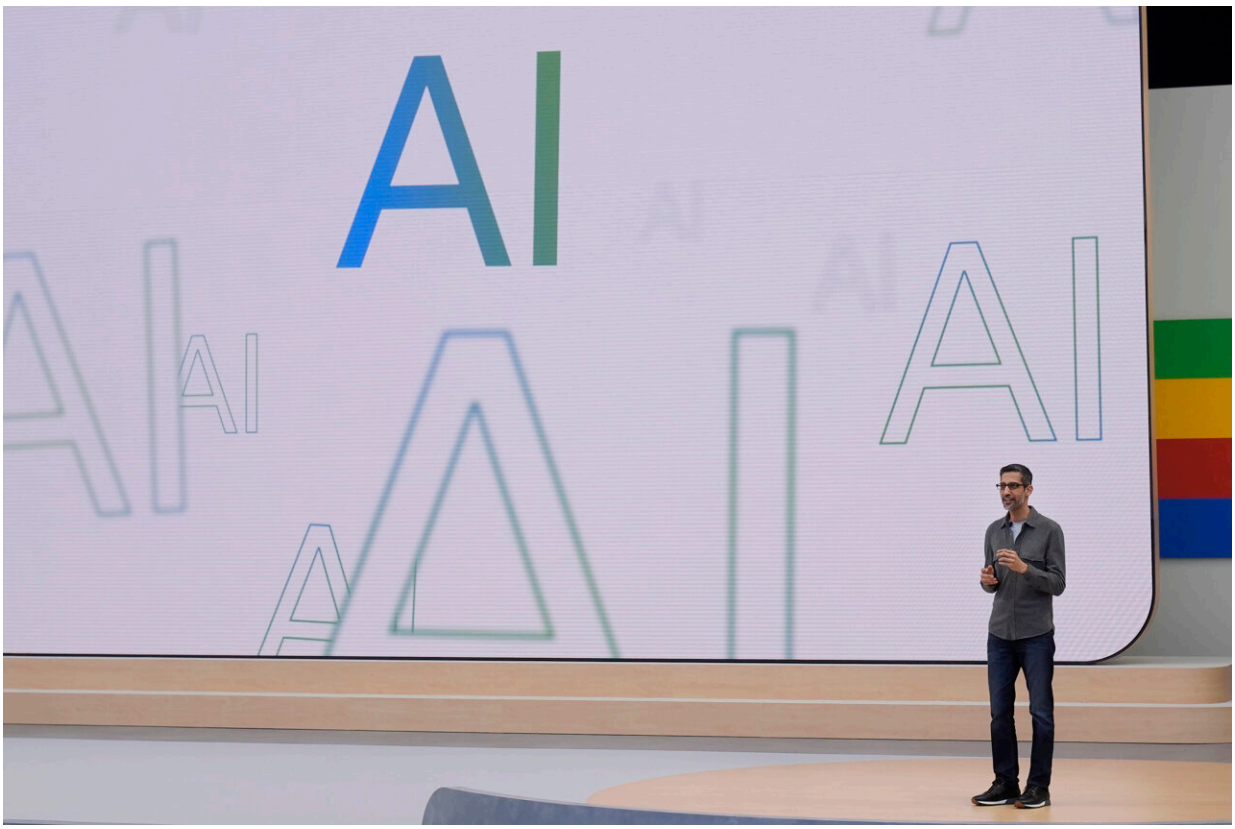
Cats on the moon? Google's AI tool is producing misleading responses that have experts worried

May 25 2024, by MATT O'BRIEN and ALI SWENSON



Alphabet CEO Sundar Pichai speaks at a Google I/O event in Mountain View, Calif., May 14, 2024. Bloopers — some funny, others disturbing — have been shared on social media since Google unleashed a makeover of its search page that frequently puts AI-generated summaries on top of search results. Credit: AP Photo/Jeff Chiu, File

Ask Google if cats have been on the moon and it used to spit out a ranked list of websites so you could discover the answer for yourself.

Now it comes up with an instant answer generated by artificial intelligence—which may or may not be correct.

"Yes, astronauts have met cats on the moon, played with them, and provided care," said Google's newly retooled search engine in response to a query by an Associated Press reporter.

It added, "For example, Neil Armstrong said, 'One small step for man' because it was a cat's step. Buzz Aldrin also deployed cats on the Apollo 11 mission."

None of this is true. Similar errors—some funny, others harmful falsehoods—have been shared on social media since Google this month unleashed AI overviews, a makeover of its search page that frequently puts the summaries on top of search results.

The new feature has alarmed experts who warn it could perpetuate bias and misinformation and endanger people looking for help in an emergency.

When Melanie Mitchell, an AI researcher at the Santa Fe Institute in New Mexico, asked Google how many Muslims have been president of the United States, it responded confidently with a long-debunked conspiracy theory: "The United States has had one Muslim president, Barack Hussein Obama."

Mitchell said the summary backed up the claim by citing a chapter in an academic book, written by historians. But the chapter didn't make the bogus claim—it was only referring to the false theory.

"Google's AI system is not smart enough to figure out that this citation is not actually backing up the claim," Mitchell said in an email to the AP. "Given how untrustworthy it is, I think this AI Overview feature is very irresponsible and should be taken offline."

Google said in a statement Friday that it's taking "swift action" to fix errors—such as the Obama falsehood—that violate its content policies; and using that to "develop broader improvements" that are already rolling out. But in most cases, Google claims the system is working the way it should thanks to extensive testing before its public release.

"The vast majority of AI Overviews provide high-quality information, with links to dig deeper on the web," Google said a written statement. "Many of the examples we've seen have been uncommon queries, and we've also seen examples that were doctored or that we couldn't reproduce."

It's hard to reproduce errors made by AI language models—in part because they're inherently random. They work by predicting what words would best answer the questions asked of them based on the data they've been trained on. They're prone to making things up—a widely studied problem known as hallucination.

The AP tested Google's AI feature with several questions and shared some of its responses with subject matter experts. Asked what to do about a snake bite, Google gave an answer that was "impressively thorough," said Robert Espinoza, a biology professor at the California State University, Northridge, who is also president of the American Society of Ichthyologists and Herpetologists.

But when people go to Google with an emergency question, the chance that an answer the tech company gives them includes a hard-to-notice error is a problem.

"The more you are stressed or hurried or in a rush, the more likely you are to just take that first answer that comes out," said Emily M. Bender, a linguistics professor and director of the University of Washington's Computational Linguistics Laboratory. "And in some cases, those can be life-critical situations."

That's not Bender's only concern—and she has warned Google about them for several years. When Google researchers in 2021 published a paper called "Rethinking search" that proposed using AI language models as "domain experts" that could answer questions authoritatively—much like they are doing now—Bender and colleague Chirag Shah responded with a paper laying out why that was a bad idea.

They warned that such AI systems could perpetuate the racism and sexism found in the huge troves of written data they've been trained on.

"The problem with that kind of misinformation is that we're swimming in it," Bender said. "And so people are likely to get their biases confirmed. And it's harder to spot misinformation when it's confirming your biases."

Another concern was a deeper one—that ceding information retrieval to chatbots was degrading the serendipity of human search for knowledge, literacy about what we see online, and the value of connecting in online forums with other people who are going through the same thing.

Those forums and other websites count on Google sending people to them, but Google's new AI overviews threaten to disrupt the flow of money-making internet traffic.

Google's rivals have also been closely following the reaction. The search giant has faced pressure for more than a year to deliver more AI features as it competes with ChatGPT-maker OpenAI and upstarts such as

Perplexity AI, which aspires to take on Google with its own AI question-and-answer app.

"This seems like this was rushed out by Google," said Dmitry Shevelenko, Perplexity's chief business officer. "There's just a lot of unforced errors in the quality."