Primer

Reliable AI: grounding to mitigate hallucinations

Artificial Intelligence (AI) is on course to transform society. Recently, attention has focused on generative models, neural networks trained on internet-scale text data. These models are powerful, with some researchers describing them as approaching human-level intelligence.[2]

But can you trust what a generative model generates? If an AI system answers a question – about a real-time situation, about available intelligence, or even about a single document – how can the user validate the credibility of the answer?

The problem: hallucination

Ask ChatGPT to write a biographical portrait of Dr. Craig Martell, the DoD CDAO, and it instantly generates pages of text about the man. Most of the information is correct, since Dr. Martell was an industry leader often mentioned in the news, and the generative model that powers ChatGPT has studied the internet. But the model also fills in gaps in its knowledge, making the false claim that Martell's "doctoral research focused on developing novel algorithms for autonomous systems." In fact his PhD focused on how humans communicate with gestures.[3]

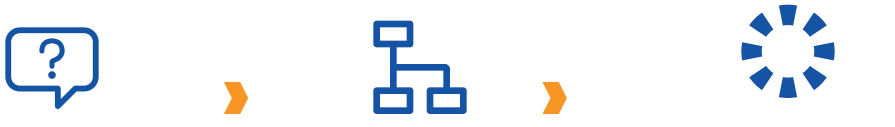
AI researchers call this a "hallucination." The model is not lying, since it has no motivation to deceive, nor any awareness that what it is saying is false. Regardless, hallucinations are a nightmare for analysts, operators, and decision-makers in national security.

Subtly incorrect claims are the most dangerous. The information trail can be lost, baking false assumptions into seemingly correct responses and threatening decision-making.

The cause: training

The cause of AI hallucination is in the training. The generative model behind ChatGPT, for example, was trained by playing a game of "Fill in the Blank." The model is given a truncated snippet of text from a webpage – such as Wikipedia, Reddit, news sites – and it guesses the next words. The training is as simple as that. And yet learning to predict words on the scale of the internet – on the order of 1 trillion words – teaches the model an astounding range of skills. It can read and write in dozens of languages. It can even pass exams in medicine[4] and law.[5]

But a generative model never stops playing the game. It is predicting the most plausible next word, regardless of whether it expresses something true or false. Given this training paradigm, the problem of hallucination is inevitable.



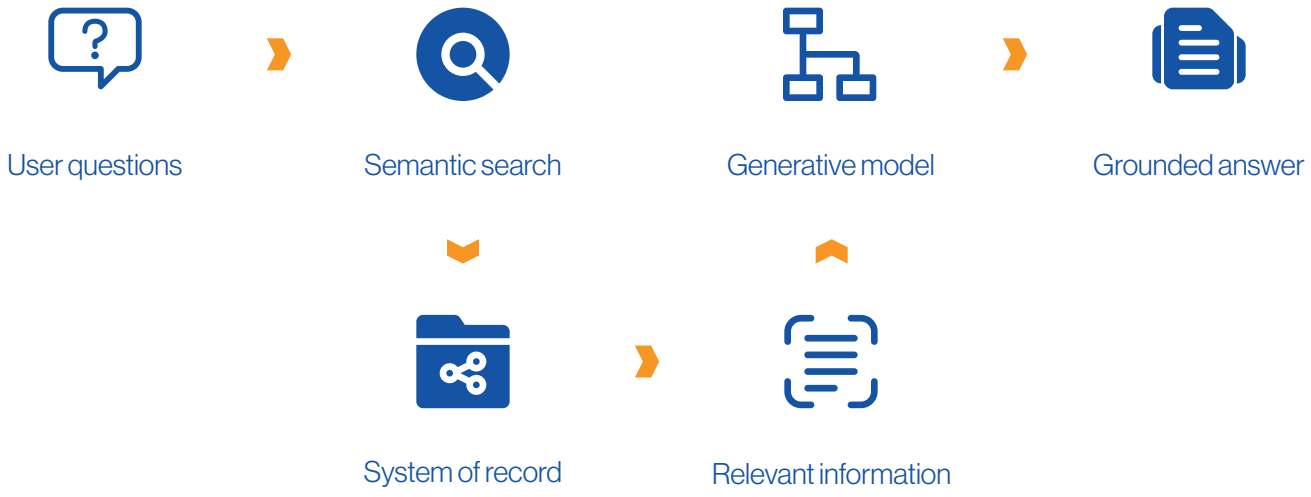
The solution: grounding

You wouldn't ask a human analyst to write a mission-critical report from memory. Nor to answer questions about a complex battlefield without access to the latest relevant data. So too with generative AI systems. The solution to hallucination is "grounding."

Instead of merely predicting the next words, given a user's input, a grounded system first retrieves relevant information from a trusted system of record, then generates a prompt for the generative model. That prompt instructs the model to answer the user's question based only on the retrieved information and only if possible. Otherwise, crucially, the model is instructed to say "not enough information."

Grounding is not a silver bullet. The model's answers can only be as good as the quality of the information provided–a perennial problem. Also, the information retrieval can become the bottleneck. But information retrieval is a far better understood engineering problem, one that Primer is already tackling by working to integrate semantic search into its applications.

> "My fear is that we trust [AI] too much without the providers of that service building into it the right safeguards and the ability for us to validate."
>
> *— Dr. Craig Martell, Chief Digital and Artificial Intelligence Officer (CDAO) US Department of Defense*[1]



1. https://www.defenseone.com/technology/2023/05/pentagons-ai-chief-scared-death-chatgpt/385963/
2. https://arxiv.org/abs/2303.12712
3. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=56dba273dd99a9398f64aa5ff02f7c38f4ef90c0
4. https://www.microsoft.com/en-us/research/publication/capabilities-of-gpt-4-on-medical-challenge-problems/
5. https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry/

About Primer

Primer exists to make the world a safer place. We believe that AI is critical to accelerating intelligence and decision cycles that keep us safe. Our AI technology is deployed by the U.S. Government, strategic allies, and Fortune 100 companies to extract timely insight and decision advantage from massive datasets. Primer has offices in Arlington, Virginia and San Francisco, California. For more information, please visit primer.ai.