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.[1]

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? There is a solution, called grounding, which is fast becoming best practice for deploying generative models.

The problem: hallucination

Ask ChatGPT to write a biographical profile of Yurii Shchyhol, Ukraine's head of cybersecurity, and it instantly generates pages of text about the man. Most of the information is correct, since Shchyhol has been written about frequently in online news and analyses, and the generative model that powers ChatGPT has studied the internet. But the model also fills in gaps in its knowledge, incorrectly guessing Shchyhol's age, birth place, and even fabricating a plausible but incorrect education.

Truncated Response

Write a biographical profile of Yurii Fedorovych Shchyhol, head of the State Special Communications Service of Ukraine.

> Born on February 12, 1970, in Kyiv, Ukraine, 1 Yurii Shchyhol developed an early fascination with technology and telecommunications. He pursued his passion by earning a degree in Telecommunications Engineering from the National Technical University of Ukraine "Kyiv Polytechnic Institute." 2 This academic foundation laid the groundwork for his future achievements in the field.
>
> Shchyhol's professional journey commenced soon after completing his studies when he joined the State Special Communications Service in 1995. 3 The SSCS, being the central government authority responsible for ensuring secure communications and protecting information systems of national importance, offered Shchyhol a platform to utilize his skills effectively. His dedication and expertise were quickly recognized, leading to his rapid ascent within the organization. Throughout his tenure with the SSCS, Shchyhol demonstrated an unwavering commitment to upholding Ukraine's cybersecurity and protecting critical infrastructure. He played a crucial role in implementing advanced technologies and strengthening the country's cybersecurity defenses, particularly in the face of evolving threats and hostile activities targeting Ukraine's communications networks.
>
> Shchyhol's leadership abilities and technical acumen led to his appointment as Head of the State Special Communications Service in 2017. 4 As the head of this vital organization, he shoulders the responsibility of ensuring the security and integrity of Ukraine's communication systems, both in times of peace and during periods of heightened tension. Under his leadership, the SSCS has developed robust strategies and implemented state-of-the-art technologies to counter cybersecurity threats effectively.

Full response: https://chat.openai.com/share/370376bf-f027-49ec-b91f-33f47fa22c1b

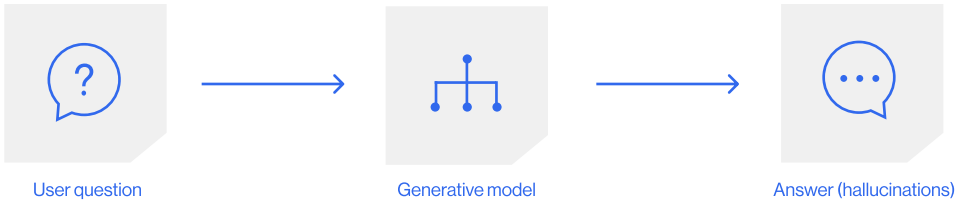
Analysis

1. He was born November 5, 1983, in the village of Zazymia, Brovary Raion, Kyiv Oblast..
2. In 2006, he graduated from the National University of the State Tax Service of Ukraine.
3. In 2017 he finished Alfred Nobel University, majoring in finance and credit.
4. He was based at SSCS 2008 onwards.
5. He was made head of SSCS in 2020.

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. They lead to the loss of the information trail, embedding false assumptions in the command chain.

1. https://arxiv.org/abs/2303.12712

The reason: training

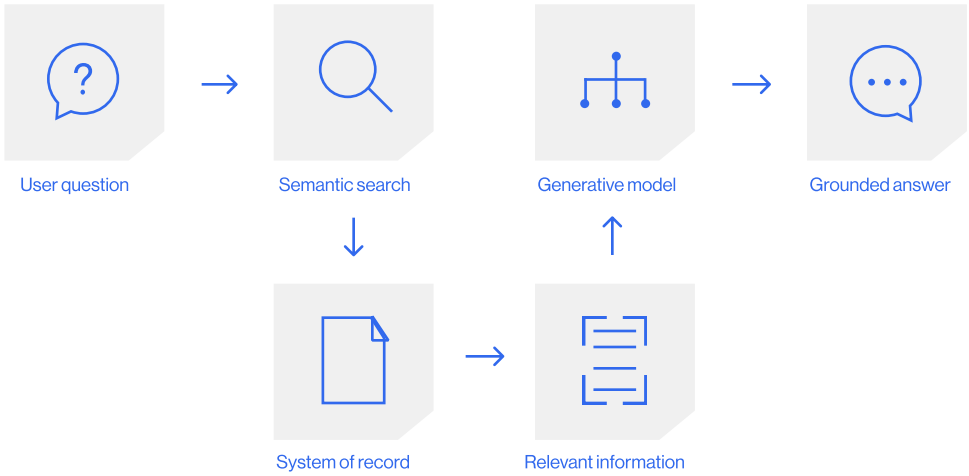


The cause of AI hallucination is in the training. The generative model behind ChatGPT, for example, was trained by playing a game of "Fill in the Blank." The model is given a truncated snippet of text from a webpage—such as Wikipedia, Reddit, news sites—and it guesses the next words.

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[2] and law.[3] But a generative model never stops playing the game. It is always 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 geopolitical landscape 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 exclusively on the retrieved information and only if possible within that context. Otherwise, crucially, the model is instructed to say "not enough information."

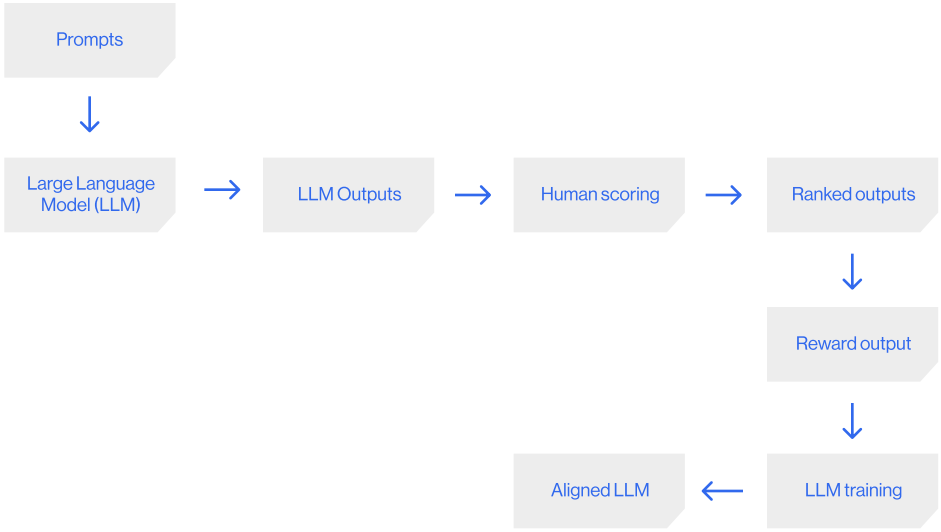
Grounding is not a silver bullet. The model's answers can only be as good as the quality of the information provided–a perennial problem. Also, the information retrieval can become the bottleneck. But information retrieval is a far better understood engineering problem, one that Primer is already tackling by integrating semantic search, a technique that uses the intent behind a query to deliver more relevant results, into its applications.

2. https://www.microsoft.com/en-us/research/publication/capabilities-of-gpt-4-on-medical-challenge-problems
3. https://law.stanford.edu/2023/04/19/gpt-4-passes-the-bar-exam-what-that-means-for-artificial-intelligence-tools-in-the-legal-industry

Deeper dive: RLHF and instruction-tuning

While a grounding solution is necessary, it is not sufficient on its own. Several recent breakthroughs in machine learning have been crucial for making generative models that humans can use reliably to get work done.

One breakthrough is called reinforcement learning with human feedback (RLHF).[4] The goal of this technique is to improve the model's behavior, making it more "aligned" with the human users who interact with it.



You can think of RLHF as a finishing school for brilliant but unruly children. Teams of humans grade the outputs of the model, giving high marks when it avoids toxic language, answers questions helpfully, and stays on topic. Reinforcement learning is a handy trick for teaching the model how to predict human preferences. Once all of the feedback is harvested, the model is trained to generate outputs that are not only predicted to be highly likely words—the "Fill in the blank" game—but also highly likely to please the human user.

Another key breakthrough is called instruction-tuning. Early generative models could not provide reliable answers to users' queries. This is because documents on the internet are not structured as clean questions and perfect answers. Instruction-tuning is the fix. The model is trained on a data set of documents with various instructions followed by obedient replies. Recently, generative models have even been used to instruction-tune other models, with no human work required.[5]

All of these innovations were necessary. With grounding and techniques for achieving alignment, generative language models finally became both powerful and helpful.

4 https://arxiv.org/abs/2204.05862
5 https://arxiv.org/abs/2304.03277

The impact: advantage

Tools powered by generative models can give human operators immense advantages. But like all new tools, they bring complexity and risk. Especially in a military setting, there is no tolerance for unreliable tools. When implementing generative models within an organization, there are several key recommendations to ensure a successful and responsible integration while mitigating risks. Here are a few important considerations to get started with generative AI:

1. **Objectives**: Clearly define the objectives for using generative AI models, aligning them with the organization's mission.
2. **Data quality:** Ensure access to high-quality data that fits the intended use of generative models.
3. **Human-in-the-loop**: Incorporate human oversight and intervention throughout the generative model pipeline, both to track performance and provide a failsafe.
4. **User education and awareness:** Educate users and stakeholders about generative model capabilities and limitations, promoting critical evaluation and feedback for improvement.
5. **Collaborate with industry experts:** Participate actively in the development of this technology by partnering with industry experts.

Partner with Us

Primer exists to make the world a safer place. Our main objective is to provide practical AI solutions to address contemporary challenges posed by the current threat landscape. By responsibly harnessing AI's potential while navigating its complexities, we fortify deterrence capabilities against potential adversaries.

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 www.primer.ai or request a product demo.

Primer

Learn more: www.primer.ai