Generative AI FAQs

May 2023

Google has a long history of using Artificial Intelligence (AI) to improve our products for billions of people.

For example, in Google Maps, AI analyzes data to provide up-to-date information about traffic conditions and delays; in Gmail, it helps block nearly 10M spam messages every minute; through Translate and Lens, AI helps instantly translate between 21 languages.

Now, Google's newest AI technologies — like LaMDA, PaLM, Imagen and MusicLM — are creating entirely new ways to engage with information, from language and images to video and audio. We're working to bring these latest AI advancements to more people.

Several of our latest AI advancements center on Language Models (notice the "LM" references in LaMDA, PaLM and MusicLM), which are statistical tools that find patterns in human language. These patterns can be used for a range of tasks including predicting the next words to complete a sentence, or providing grammatical suggestions that preserve what you mean. Just as there are many different applications for AI, there are also different types of Language Models, including Large Language Models (or LLMs).

Understanding the differences between these terms and concepts can be a challenge and you probably have questions. Like what exactly *are* LLMs? Or how does generative AI fit into the broader AI ecosystem?

This set of Frequently Asked Questions provides some quick answers on key AI-related topics. We hope these answers can help inform deeper policy conversations about the changing role of AI in society.

1

Understanding Generative AI

What is Generative AI?

Generative AI is a type of machine learning (ML) model that can take what it has learned from the examples it has been provided to create new content, such as text, images, music, and code. These models learn through observation and pattern matching, also known as training. For example, a model may learn what a cat looks like by observing many different examples of cats and recognizing characteristics consistent with a cat. The same goes for sonnets, resumes, or packing lists for a camping trip.

To understand the concept "cat", a generative text-to-image model (like Imagen) would be trained on millions of cat photos. Over time, it learns to recognize that cats are animals that usually have whiskers, fur, pointy ears, etc.

This allows the model to take an input such as "cat wearing an ice cream hat," connect what it has learned about cats, ice cream, and hats, and generate a new corresponding image, even if it has never seen an image of a cat wearing an ice cream hat in its training data.

Generative AI models are neither information databases nor deterministic information retrieval systems. Because they are prediction engines, they can create different outputs in response to the same prompt.

For example, when you ask a generative AI model to produce an image of a cat, it does not look through its training data and return a cat photo. Instead, it will generate a new image of a cat each time.

What is a Large Language Model (LLM)?

Large Language Models, or LLMs, are generative AI models which can predict words that are likely to come next, based on the user's prompt and the text it has generated so far.

In some cases, LLMs can respond to the same prompt with different responses. This is due to the flexibility that LLMs are often given to pick among probable words that could follow, based on patterns identified from their training data. This flexibility allows them to generate more interesting and creative responses.

For example, if prompted to fill in the phrase "cat and [blank]", an LLM might predict that the next word is "mouse", or it might predict "dog."

Importantly, LLMs are not databases or information retrieval systems. When prompted for facts, they can generate articulate responses that may give the impression that they are retrieving information.

However, they do not inherently understand the words they are generating, the concepts they represent, or their accuracy, which is why they can sometimes produce answers that, while sounding plausible, contain factual errors.

What's the difference between machine learning, deep learning, and generative AI?

Much of the recent progress we've seen in AI is based on machine learning (ML), a subfield of computer science where computers learn to mathematically recognize patterns from example data, rather than being programmed with specific rules.

Deep learning is a specific ML technique based on neural networks. Neural networks use nodes or "artificial neurons," inspired by models of brain neurons, as fundamental processing units which receive and pass numeric inputs and outputs from other neurons. Deep learning connects multiple layers of these artificial neurons.

An example of deep learning would be a model that can detect cats in a photo. An example of a generative AI model, meanwhile, would be one that can generate photos of a cat when prompted.

Is generative AI the same as artificial general intelligence (AGI)?

No, artificial general intelligence (AGI) is a hypothetical type of artificial intelligence (AI) that would have the ability to learn, understand and perform any intellectual task that a human being can. Meanwhile, generative AI is technology already used in a variety of applications — such as image generators used in creative applications.

Predicting the arrival of true AGI is difficult. There's no universally accepted notion of AGI and no consensus exists whether AGI is possible within years, decades or more. Many factors contribute to the possibility of AGI, including advancements in computing power and learning capabilities of AI models. But capabilities resembling, yet still far from, human-like intelligence are emerging in some forms of AI, including GAI.

Is AI sentient?

No, AI models are not showing evidence of sentience. AI's capabilities are based on identifying patterns and relationships in data and, in so doing, AI can generate outputs that are generally informed by those patterns. This means, at times, an AI model might generate responses that seem to suggest it has opinions or emotions, like love, sadness or anger, since it has trained on information and data created by people that reflects the human experience and is predicting a likely response.

2

Training a Large Language Model

How does an LLM "learn"?

The technical process of "learning" for LLMs begins with training the model to identify relationships and patterns among words in a large dataset. Through this process, a generative AI model will learn "parameters," which represent the mathematical relationships in data. Once the model has learned these parameters, it can then use them to generate new outputs based on these parameters.

For example, the PaLM Research Model (2022) learned 540 billion parameters from training on text, which gave it strong capabilities in natural language inference, question and answering, and translation, among other skills. Research is showing that, with the proper training, models with fewer parameters (20B - 200B) can be similarly capable.

What is pre-training? What is fine-tuning?

LLMs are developed in multiple stages, including pre-training and fine-tuning. Pre-training is a way of training an ML model on a variety of data. This gives the model a head start when it is later trained on a smaller dataset of labeled data for a specific task.

Following pre-training, more data can be added to an existing LLM through a process called fine-tuning. Fine-tuning an LLM is the process of adapting a pre-trained LLM to improve its performance on a specific task. The model learns from additional example data to help hone its capabilities.

For example, fine-tuning a general purpose language model can teach it to summarize technical reports in general by using just a few examples of technical reports and accurate summaries.

What is grounding?

Grounding a model refers to the process of linking the abstract concepts in the model to real world entities. Developers use a variety of techniques for grounding generative AI models, including training with real-world data, simulating interactive environments, or even using equipment that can provide actual sensory input. Grounding an LLM can help equip it to better understand language and other abstract concepts in the context of the real world, which may be helpful for tasks such as natural language processing or improving the factuality of model responses.

For example, if a model is trained on soccer data accurate through June 2022, it would not be able to provide an accurate response to the question, “Who won the 2022 World Cup in December 2022?”, as it has no information on the tournament’s results. In this case, grounding the model with techniques for conducting factual checks with recent data, while not foolproof, aids LLMs in providing a better answer.

What data do LLMs need to train?

LLMs are trained on a variety of data, and they learn through observation and pattern matching. Naturally, the value and quality of individual documents within a given dataset may vary widely. The quality of some models’ predictions and outputs may benefit from having access to larger and/or more diverse pools of data.

The amount of data needed for training generally includes millions or billions of data points. In terms of scale, pre-training for text models usually involves hundreds of billions of words, while pre-training for image models may train on hundreds of millions of images or more. Fine-tuning, meanwhile, requires a smaller dataset. For example, fine-tuning for text LLMs might involve hundreds of thousands or millions of examples.

For example, LaMDA (short for Language Model for Dialogue Applications) is a language model that analyzes human speech and recognizes commands. It is trained on human dialogue and stories. The largest LaMDA model has 137 billion parameters and is trained with 1.56 trillion words.

In the training phase for LLMs, how do you manage risk responsibly?

Careful risk assessment is essential for LLM development and can involve multiple steps and mitigation measures. One good practice is to filter various training data to remove harmful content or personal data wherever possible before training, which reduces the chance the model will respond with toxic speech or personal information. Another good practice is to add additional steps such as fine-tuning, classifiers, and guardrails to help the model avoid responding with harmful patterns.

What type of human oversight or input might be involved?

Human feedback and evaluations are important in developing LLMs responsibly. At the outset, those creating the LLM should create policies for these systems to outline prohibited use, including various forms of abuse and harm. During development, a good practice is to perform adversarial testing – encouraging test users to actively find problems and problematic requests, so that they can be fixed. After launch, users should be able to flag content which might be unsafe or harmful.

For example, before Google Bard's launch, thousands of Trusted Testers were invited to use it and give feedback on their experience. This feedback helped improve the overall experience before public launch.

Mitigating risks and preventing misuse

What is a "hallucination" and why do LLMs "hallucinate?"

A hallucination is a response from an LLM that may be coherent and presented confidently but is not based in factuality. Among other reasons, hallucinations can occur if that response is not grounded in its training data or real-world information. Hallucinations can be reduced, but very difficult to eliminate altogether.

As explained above, generative models do not retrieve information, but predict which words will come next based on user inputs. For this reason, there is no guarantee that the LLM's prediction will contain factual information – nor that their outputs to a given prompt will remain stable over time.

For example, if you ask an LLM-based interface to give information about a person who is not well known, it might reply that the person has a degree in a field they never studied, from a university that they never attended. This can occur largely because the model is predicting an output about something it does not have enough training data to learn from. When there's limited or no information about the person, it is more likely the model will hallucinate in its response.

This is why users may see disclaimers when engaging with LLMs, alerting them to the risk of relying on the output of these systems without verifying the responses' underlying accuracy.

Can we prevent hallucinations?

Hallucinations can be reduced in an LLM, but inaccuracies cannot be 100% prevented since responses are created via a prediction mechanism. During fine-tuning, models can be optimized for recognizing correct patterns in their training data, which will reduce the number of factual mistakes. Another technique for reducing hallucinations is to connect LLMs to other systems to provide verified information in the response.

For example, if a user requests a mathematical calculation from an LLM that is connected to a calculator service, the LLM can pass part of the request to that calculator to perform the task. The LLM then returns the calculator's response to the user in its answer.

Why is it sometimes difficult for generative AI to attribute and cite sources?

Generative AI models are usually designed to generate original outputs based on their underlying prediction mechanisms. For example, when it runs, a generative image model creates a new, unique image based on concepts it has picked up across its training data. This makes it difficult for generative models to attribute specific parts of their responses to any one source. A good analogy might be an artist studying multiple other artist's styles and then creating their own.

How can we build additional guardrails for generative AI models?

Generative AI models are intended to respond to a wide variety of input prompts (e.g., "draw me a photo...", "write me a post..."). For this reason, it is also important to take a multifaceted approach to guardrails for generative AI. This can include using training data that aligns models as closely as possible to policies; implementing prompt-and-response detection tools that intercept possible violations; and otherwise reducing the potential for harmful or offensive outputs with additional systems such as classifiers.

Technical guardrails help ensure that these technologies respond to prompts that adhere to a set of policies governing acceptable use. However, no automated systems are able to catch all possible violations all of the time.

How can bias be prevented or mitigated when developing generative AI models? Is generative AI necessarily biased?

Importantly, the way that generative AI models are trained means that they are not able to identify information that's non-factual, biased or potentially harmful on their own. That's why building such models responsibly is important and necessary.

For a number of reasons, a generative AI model might produce responses that reflect gaps, biases or stereotypes, as it tries to predict a plausible response. For example, a model is more likely to generate low-quality or inaccurate information if its training data includes an insufficient amount of reliable information or examples. In addition, biases or stereotypes in training data – if not addressed responsibly during the development process – might be reflected in the model's responses.

One way to reduce bias is to continue improving the model via fine-tuning, as issues are flagged and reported. Another mitigation measure is to train generative AI models on data that represents a more balanced view of the world.

For example, these models can be trained on many images of weddings, from a wide array of cultures and settings, so that they produce a diverse set of images for the prompt "photo of a wedding."

Yet another method is to train the generative AI model to represent a wide range of viewpoints for subjective topics, without endorsing one or another.

For example, if prompted to reply with the "best cat," a model could be trained to respond by stating that the "best" is a matter of opinion, followed by a range of possible cat breeds.

Google