What is Generative AI?

What is Gen AI?

Generative AI, often shortened to GenAI, is any type of artificial intelligence capable of creating new content by itself. Generative AI content includes text, images, videos, music, translations, summarizations and code. It can also complete certain tasks, such as answering open-ended questions, executing near-arbitrary instructions and participating in chats.

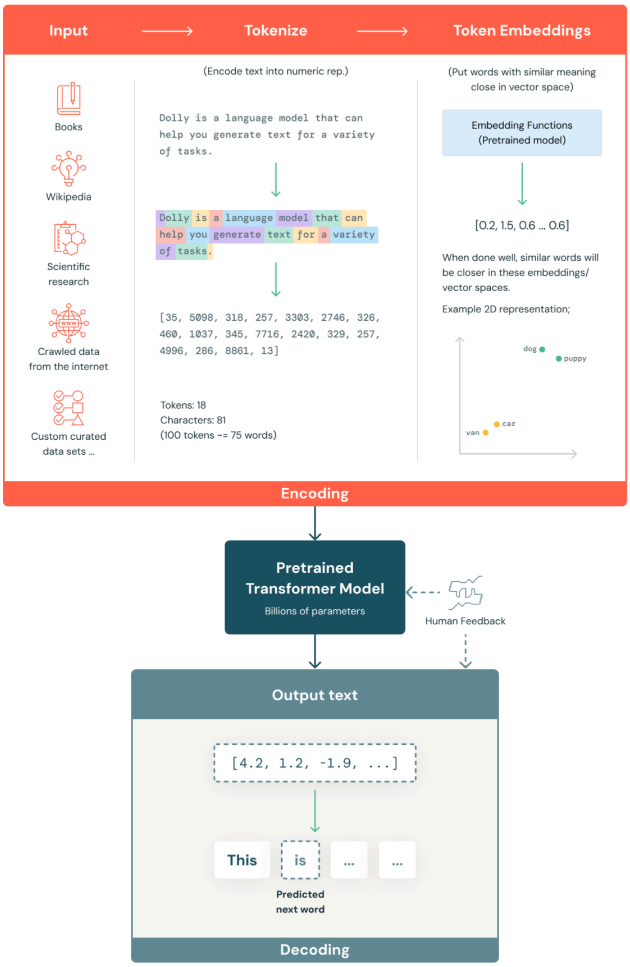
The general public was introduced to the meaning of GenAI by services like ChatGPT and DALL-E, which also greatly raised the popularity of the technology.

How does generative AI work?

GenAI models use deep learning to identify and analyze patterns within existing datasets. Similar to human brain behavior, they use transformers and other deep learning architectures to process and "learn" from datasets. These AI models are trained on huge amounts of data to create new and original content.

You can give the AI model a "prompt" once it's trained by inputting something like text, an image or a sequence of musical notes. The algorithms then generate new content in return. They can even work across media, for example, by using an image to create a text caption or generating an image from a text description.

A common type of generative AI model is large language models (LLMs), which are trained on text. These models learn to recognize words that are used sequentially. They can then form a sentence by predicting which word is most likely to


Input
Tokenize
Token Embeddings
(Encode text into numeric rep.)
(Put words with similar meaning close in vector space)
Books
Wikipedia
Scientific research
Crawled data from the internet
Custom curated data sets ...
Embedding Functions (Pretrained model)
[0.2, 1.5, 0.6 ... 0.6]
Example 2D representation;
dog
puppy
car
van
Encoding
Pretrained Transformer Model
Billions of parameters
Human Feedback
Output text
[4.2, 1.2, -1.9, ...]
This
is
...
...
Predicted next word
Decoding

```
Dolly is a language model that can
help you generate text for a variety
of tasks.
```

```
Dolly is a language model that can
help you generate text for a variety
of tasks.
```

```
[35, 5098, 318, 257, 3303, 2746, 326,
460, 1037, 345, 7716, 2420, 329, 257,
4996, 286, 8861, 13]
```

Tokens: 18
Characters: 81
(100 tokens ~= 75 words)

When done well, similar words will be closer in these embeddings/ vector spaces.

Examples of generative AI models

There are several types of generative AI models currently in use. Their methods and use cases differ, but they all combine various algorithms to process and create content.

Generative adversarial networks (GANs)

A GAN model contains two neural networks, which are trained at the same time. These networks are called the generator and the discriminator, and they compete against each other in a game-style scenario.

The generator creates new outputs, such as an image based on a prompt. The discriminator then evaluates this new content for authenticity and provides feedback to the generator to help improve its output. The generator is always trying to trick the discriminator into accepting generated content as "real," One well-known example of a GAN model is Midjourney (a text-to-image GenAI tool). However, GANs are not limited to image creation; they have also produced text and video content.

The continuous competition between the GAN generator and discriminator means that they can quickly generate high-quality outputs. However, it's important to ensure the two networks are balanced to avoid issues like overfitting, mode collapse and diminished gradients.

Variational autoencoders (VAEs)

Autoencoder models also use two networks to interpret and generate data. In this model, the networks are called the encoder and the decoder. The encoder network is trained to compress data into a simplified, or latent, format that captures key features. Meanwhile, the decoder model is trained to reconstruct content from latent data.

VAEs use continuous latent spaces to enable local variation between training data points. By decoding the slightly modified compressed information, the VAE model outputs similar, but ultimately original, content.

This model is often used for image generation and anomaly detection, but can also create text and audio. VAEs are quick at generating outputs like images, but they can lack detail compared to some other models.

Autoregressive

Autoregressive generative AI models create new samples by considering the context of elements that were generated previously. They model the conditional probability distribution of each data point and generate new data by predicting the next element in the sequence.

These models generate data sequentially, one element at a time, allowing for the generation of complex sequences.

Diffusion models

Diffusion models are sometimes also called de-noising diffusion probabilistic models (DDPMs). They are trained with a two-step process that involves forward diffusion and reverse diffusion.

During forward diffusion, random Gaussian noise is gradually added to training data, effectively destroying it. The AI then learns to reconstruct the samples through reverse diffusion. Once they are fully trained, diffusion models can create new data from completely random noise.

Transformers

Transformers use a specific type of machine learning that helps them process the long-term relationships between sequential input data. This requires the models to be trained on larger datasets.

This concept, known as “attention,” enables transformers to figure out which parts of the input influence other parts, i.e., to understand the context. This makes them ideal for text-generation tasks involving natural language processing (NLP), which requires an understanding of context. The majority of well-known generative AI programs are examples of transformer-based models.

Transformers have proven to be very powerful text generators. This is because they only need text as the training input, and there are billions of pages available to use. Aside from NLP, other uses of transformer AI models include tracking connections and identifying relationships within code, proteins, chemicals and DNA.

What is the role of deep learning in generative AI?

data, such as natural language. The majority of generative AI models will all be using deep learning under the hood.

The name deep learning comes from the large number of processing layers used for these models. The first layer of interconnected nodes is provided with training data. The output from this layer is then used as the input for the next layer. As each layer builds on the knowledge gained from the previous layer, complexity and abstraction increase and the fine details of datasets can contribute to understanding larger-scale patterns.

While programmers need to perform feature extraction during traditional machine learning, deep learning programs can build useful representations of data internally with less supervision.

Furthermore, deep learning techniques allow AI models to handle complex and abstract concepts, such as natural language understanding and image recognition.

There are a number of ways to improve AI performance, such as data augmentation, transfer learning, and fine-tuning. Data augmentation uses generative models to create new synthetic data points for training data. This is then added to existing data to increase the size and diversity of datasets and, consequently, the accuracy of the model.

Transfer learning involves using a pretrained model for a second, related task. By leveraging the output from the existing model as the input for a different learning problem, the model can apply the knowledge gained from the first instance of training. An example of transfer learning would be using a model trained to identify cars to train a model for identifying other vehicles. Transfer learning is useful as it reduces the amount of data needed to train a new model.

Finally, fine-tuning is a technique for customizing an AI model by training it with more specific data. This allows pretrained models to be refined for use on specific domains or tasks. High-quality datasets that are representative of the final task are needed for fine-tuning.

Generative AI technology has a huge range of applications in the real world, from text and image generation to software development. Let's take a look at some of the current common use cases.

Image generation

Tools like DALL-E enable users to create new imagery (photos, illustrations and even videos) by inputting visual or written prompts. Multimodal models can create images from text instructions, so users can be as vague or specific as they like.

For example, you could just ask for a drawing based on "animals" and "rainbows," and see what it comes up with. Or you could give detailed instructions, such as "a baby rhino wearing sunglasses, looking at a rainbow through a window with purple curtains."

Another option is style transfer, where the content of one image is combined with the visual style of another. You input a content image (a photo of a rhino) and a style reference image (a Picasso painting), and the AI is able to blend them together to create a new, Picasso-style rhino image.

Text generation

While one of the most well-known text-based GenAI use cases is chatbots, the technology can now be applied to many other tasks. For instance, tools like GrammarlyGo can help with writing and responding to emails in a business-like style.

Let's say you had to produce a brochure advertising a technical product. As a human, you'd take time to read about the features and specifications, make detailed notes, and come up with a draft narrative. A generative AI program can do all that in seconds after you provide the information, creating ready-to-go content quickly. Text generation is also useful for dubbing movies, providing closed captions for video content or translating content into various languages.

Music composition

generators, you can provide details or allow full creative freedom, for example, “A song about rainbows” or “a three-verse children’s song about rainbows in waltz time, accompanied by ukulele and kazoo.”

You can also ask the AI to blend two different pieces together with style transfer, for example, the Happy Birthday song in the style of Gershwin, or create a remix. Amper Music creates musical tracks from prerecorded samples, while other tools can create a soundtrack by recognizing objects in video footage.

What are the industry applications of generative AI?

With so many ways that this technology can be used, it’s no wonder that generative AI for retail, financial services, healthcare and more is becoming the standard rather than the exception.

Retail

Many retail companies already use chatbots to automate customer service and, as generative AI advances, these chatbots will become more sophisticated. In the future, AI could provide further personalization for customers with virtual fitting rooms, product development and proactive marketing. Retail businesses could also benefit from using GenAI for inventory and demand planning as well as the identification of phishing or fraud for stronger security.

Financial services

Businesses within the financial services industry (FSI) are already investing in GenAI to analyze large amounts of data. One example of this is the BloombergGPT LLM, which was announced earlier this year. The 50-billion parameter AI is purpose-built for FSI reporting and forecasting.

improving operational efficiency and increasing customer personalization.

Manufacturing

Since the Industrial Revolution, the manufacturing industry has aimed to optimize efficiency through automation. Generative AI provides a new tool that will push this industry into the future yet again.

AI can provide automated reports on continuous manufacturing operations, identifying performance gaps or bottlenecks and using data-driven prioritization to boost efficiency. As well as monitoring operations, AI can also monitor equipment and reduce downtime with predictive maintenance and troubleshooting.

Finally, LLMs in manufacturing can personalize the customer experience, both during customer service and in certain products, such as vehicles or smart technology.

Media

How the entertainment industry will use AI has been the subject of much recent debate. However, there are many ways that generative AI can be used without affecting industry jobs.

The analysis of user preferences, consumption patterns and social media signals by AI models can be used to optimize media recommendations from entertainment services. GenAI models could also improve targeted advertisements. However, the most exciting development regarding LLMs in entertainment is the potential for immersive, interactive storytelling where the viewer's decisions would shape the narrative.

Healthcare

In healthcare, generative AI models can help with the discovery of new drugs by creating graphs to show new chemical compounds and molecules. AstraZeneca already uses AI for technology.

These models can also suggest new compounds to test, identify suitable trial candidates and fine-tune medical image analytic applications with synthetic images. Furthermore, AI can be used to generate personalized treatment plans or to transcribe consultations for upload to electronic health records.

How the generative AI application landscape can benefit businesses

We’ve seen some of the real-world applications, but what does generative AI mean for businesses? Here are some of the key benefits.

Revenue streams

This technology enables businesses to create and launch new products quickly by coming up with fresh new designs and accelerating the R&D process. It can analyze trends and customer behavior in order to present new ideas for extra revenue streams.

As well as product innovation, AI can help you produce new marketing plans and create promotional materials. Analyzing customer preferences allows it to generate targeted ads, tailor recommendations and personalize products and services. AI data analysis also helps businesses spot opportunities that keep them ahead of their rivals and retain a competitive edge.

Finally, using GenAI improves business chatbot performance, which increases customer satisfaction, sales and retention.

Productivity

Another major benefit is productivity, as GenAI can be used to automate time-consuming manual tasks, such as data entry, routine emails and meeting or call transcriptions.

also analyze your data and suggest ways to improve existing workflows for maximum efficiency.

In customer support, businesses can deploy AI-powered chatbots and virtual assistants to reduce the burden on support agents. Advantages include shorter response and resolution times, as well as allowing agents to handle other tasks while AI takes care of common queries.

Risk mitigation

Generative AI platforms can give you deeper visibility into your data as well as quickly identify financial or security vulnerabilities. Advanced AI programs can even simulate potential business risks, enabling you to assess compliance and implement protocols to avoid or mitigate problems.

Meanwhile, data compression means organizations only need to retain essential data, which lessens the risks of holding a lot of personal information.

What are the differences between the most common LLMs?

LLMs form a crowded field, and the number of options to choose from will only continue to increase. However, you can generally group LLMs into two categories: proprietary services and open source models. Let’s take a closer look.

Proprietary services

The most well-known LLM service is ChatGPT, which was released by OpenAI toward the end of 2022. ChatGPT offers a user-friendly search interface that accepts prompts and typically provides fast and relevant responses. The ChatGPT API is also accessible to developers, allowing them to integrate the LLM into their own applications, products or services.

Other proprietary service generative AI examples are Google Bard and Claude from Anthropic.

The other type of LLM is open source and available for commercial use. The open source community has quickly caught up to the performance of proprietary models and their models, which can be self-hosted or provided via cloud service APIs, can be customized via fine-tuning.

Popular open source LLMs include Llama 2 from Meta and MPT from MosaicML, which has been acquired by Databricks.

Choosing the best generative AI LLM

Understanding the differences between proprietary and open source LLMs is the first step, but there is still a lot to consider when selecting an LLM for GenAI applications. Future-proofing, cost and leveraging data as a competitive advantage should all be taken into account when choosing between a closed third-party vendor API or an open source (or fine-tuned) LLM.

While proprietary service LLMs are often very powerful, they can also raise governance concerns due to their “black box” style, which permits less oversight of their training processes and weights. Another risk is that proprietary models can be deprecated or removed, which will break any existing pipelines or vector indexes.

On the other hand, open source models are accessible to the buyer indefinitely. These models also offer more customization and oversight, which can result in better performance-cost trade-offs. Finally, with future fine-tuning of open source models, organizations can leverage their data as a competitive advantage to build better models than are available publicly.

Why are there concerns about generative AI ethics?

Any form of AI tends to raise ethical concerns, as humans grapple with the implications of intelligent machines. So, what are the ethics of generative AI? To start with, this technology is relatively new and is also evolving very rapidly. Even developers in this field aren’t quite sure of where it will end up, but as

One problem that has been identified in GenAI models is "hallucinations." This is when chatbots essentially make stuff up. This can have serious consequences if a model is being used for things like medical advice or accurate reporting.

Additionally, if unconscious or deliberate biases, such as racism or homophobia, are contained in training datasets, they can become encoded in models and influence an AI's output.

Apart from misinformation or potentially harmful content, there's a common concern about "deepfakes" — digitally forged images or videos. Cyberattackers can also use generative AI to mimic the style of a trusted sender and write messages that ask for passwords or money.

Furthermore, it's often difficult to trace the outputs from models back to authors, which creates copyright and plagiarism issues. This is complicated further by the lack of information around certain datasets — for instance, with image generation tools, users can request something "in the style of artist X," when "artist X" has never consented to their images being part of a dataset.

Other concerns surrounding GenAI and ethics include sustainability, as the tech requires huge computational power and electricity, as well as how it can give credence to claims that genuine reporting is fake by encouraging distrust of words and images online.

How can you test AI quality?

As we mentioned in the previous section, generative AI can sometimes produce inaccurate or low-quality outputs. When you're researching generative AI tools and frameworks, performance metrics will be displayed in the sales materials, but it's always best to check for yourself.

Developers and engineers should test AI-generated content for quality and diversity to ensure that the model is behaving as it's been trained to do. It's relatively simple to see if there's

Tools like Databricks' MLflow evaluation API can be used to track GenAI parameters and models to see if the outputs are sufficient for your needs. This can also be combined with human evaluation — for example, using your own judgment to assess whether a generated piece of music or art is appealing. More objective evaluation metrics include inception score, Fréchet Inception Distance (FID) and ground truth.

Ground truth

This evaluation method involves identifying the ground truth on which the generative AI has been trained. Ground truth is essentially the "correct" response to a query, based on information that is known to be factually true. It should be contained in the training datasets that teach the AI how to arrive at a reliable output.

For example, if you were training a model to recognize inaccurate content, you'd need a large dataset of text and images that have been classified as true or false. Developers can measure the accuracy of the answers and predictions by taking this dataset as the standard.

However, as AI designers are the ones that construct ground truth, you're reliant on their diligence in ensuring that information is correct. Ideally, your ground truth will come directly from your users as feedback.

Databricks Lakehouse Monitoring can help AI professionals ensure that their assets are high quality, accurate and reliable. Proactive reporting and unified tooling provide complete visibility of data and models for simple detection of anomalies, and built-in model-as-a-judge metrics can be augmented with your custom quality metrics.

AI quality metrics

AI quality metrics are measurements that are used to determine the performance of a generative AI model. In addition to traditional ML metrics like accuracy and recall, For example, the Fréchet Inception Distance (FID) metric assesses the quality of images created by generative AI. By comparing the distribution of generated images with that of the real images used to train the tool, the distance between the distribution of activations for some deep layers in a classifier can be calculated. A score of 0.0 is the best result for the FID.

Databricks has shown the value of using LLMs as a judge of the quality of chatbots. These cutting-edge techniques have resulted in the new MLflow model-as-a-judge functionality, which can compare the text output from different AI models to evaluate toxicity and perplexity.

Current generative AI challenges

While the technology is developing fast, there are still some considerable challenges to using generative AI models.

Infrastructure scaling

One of the key challenges in deploying GenAI successfully is scalability. As we've learned, these models need a huge quantity of high-quality and unbiased data in order to generate the desired outputs.

Large-scale computing infrastructure and power are required for the development and maintenance of generative AI models, which in turn requires significant capital expenditure and technical expertise. This has led to a growing demand for scalable solutions.

Optimization complexities

Machine learning practitioners can also face a number of low-level challenges that result from the need to optimize generative AI models. Examples of these complexities can include mode collapse and vanishing gradients.

the discriminator has accepted instead of producing more varied outputs. If the discriminator doesn't learn to reject similar, repeated outputs, every subsequent iteration of the generator will rotate through a small set of output types.

Vanishing gradients can happen when additional layers with certain activation functions are added to neural networks and cause the gradient of the loss function to become too small. A too-small gradient prevents the weights and biases of the initial layers from updating properly, causing key elements of input data recognition to fail and the network to become inaccurate.

The Databricks Fine-Tuning solution provides data scientists with the tools to help counteract these challenges. A simplified interface for users hides techniques for avoiding the above optimization challenges under the hood, handled automatically for the user.

Disjointed data, ML and AI tooling

Separate, poorly integrated tooling for data, classical machine learning, and generative AI can also create challenges for data scientists.

High-quality data is essential for training both machine learning and GenAI models, and the outputs of ML and GenAI models need to be fed back into data pipelines. There is a need to address governance, quality and implementation holistically across data and ML/AI, and separate platforms can result in friction, inefficiency and additional costs for organizations.

The Databricks Data Intelligence Platform supports core data workloads, classical ML and generative AI, and it understands data usage throughout. By combining the open, unified structure of the lakehouse with generative AI, our Data Intelligence Platform optimizes performance, simplifies the user experience and provides strong and secure governance and privacy.

look like?

According to Gartner, generative AI is set to have an impact similar to that of the steam engine, electricity and the internet, eventually becoming a “general-purpose technology.” That’s because there are so many potential applications for the technology.

For example, activities that account for up to 30% of hours currently worked across the U.S. economy could be automated by 2030. We will also see an increase in software providers integrating AI capabilities into their tools.

Understandably, humans are worried about losing their jobs to machines, but the future of AI could also see many new jobs created. For example, humans will still need to develop and train GenAI systems, including choosing the most suitable model for a given task and gathering training data to evaluate outputs.

The fast adoption of technologies like ChatGPT highlights the challenges of using GenAI responsibly. Countries and states are already finding that they need new legal and security protocols to handle issues around copyright and threats to cybersecurity, and these technologies are likely to be further regulated in the future.

Meanwhile, the Databricks Data Intelligence Platform has generative AI built-in — making it easier to maintain data security and governance, as well as track the quality of data and monitor and fine-tune your models.

As architectures and training algorithms become more advanced, generative AI models will become more powerful. Organizations must remember that with power comes responsibility, and they must find the balance between automation and human involvement.

Where can I find more information about generative AI?

Training

- **Generative AI Fundamentals**: Take this free course from Databricks and learn about the basics of generative AI.
- **LLMs**: Learn about Foundation Models From the Ground Up (edX and Databricks Training). This free training from Databricks dives into the details of foundation models in LLMs.
- **LLMs**: Level up your skills with the Application Through Production course (edX and Databricks Training). This free training from Databricks focuses on how to build LLM-focused applications with the latest and most well-known frameworks.

Sites

- Databricks AI and Machine Learning page

eBooks

- The Great Acceleration: CIO Perspectives on Generative AI: Read this report from MIT Technology Review that has insights from 600+ CIOs on generative AI
- Redefine What's Possible With Generative AI: Get started with 52 use cases
- The Big Book of MLOps

Technical Blogs

- Creating High-Quality RAG Applications With Databricks
- Best Practices for LLM Evaluation of Retrieval Augmented Generation (RAG) Applications
- Using MLflow AI Gateway and Llama 2 to Build Generative AI Apps (Achieve Greater Accuracy Using RAG With Your Own Data)
- Deploy Your LLM Chatbot With Retrieval Augmented Generation (RAG), Foundation Models and Vector Search