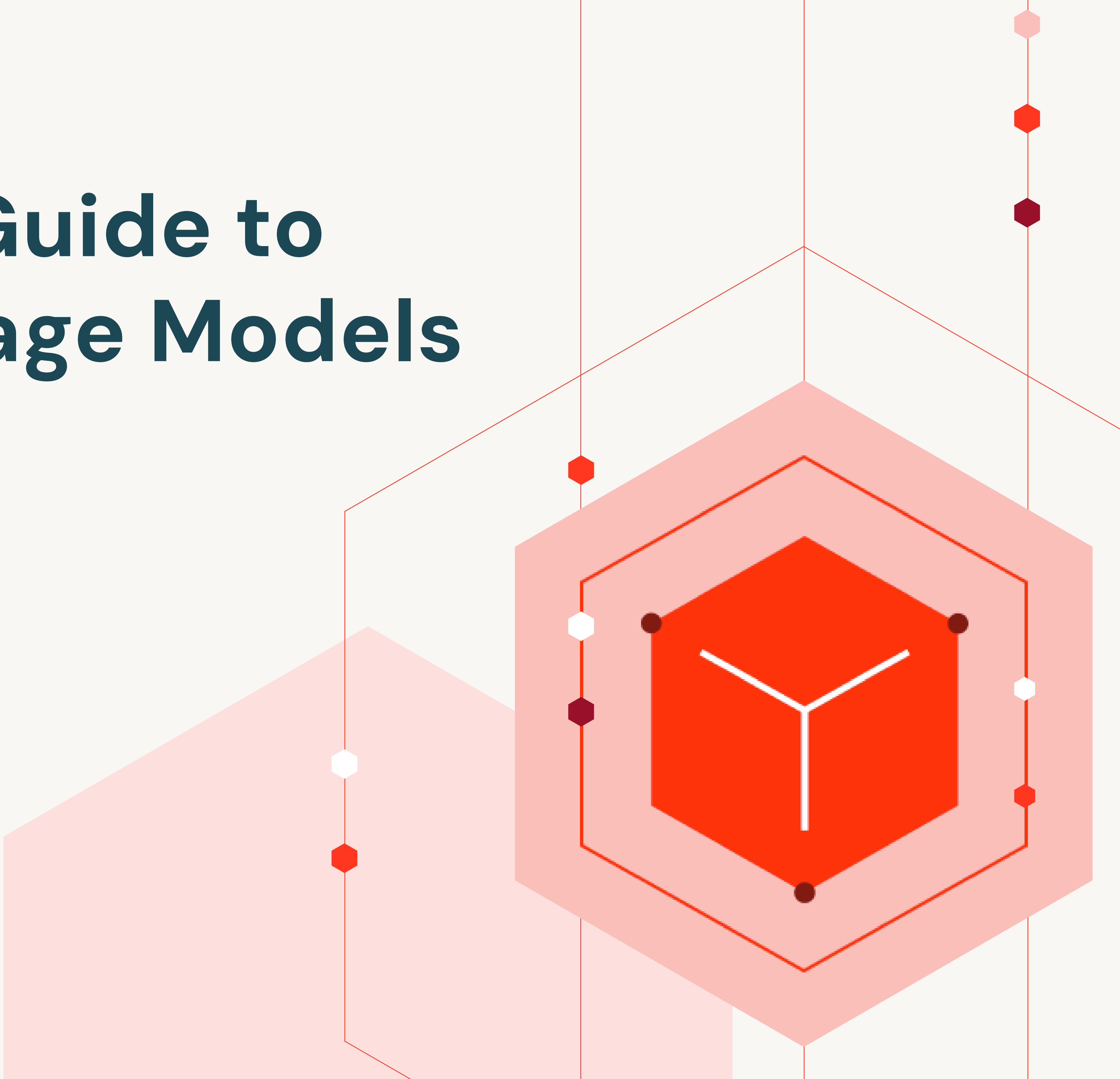
databricks

EBOOK

A Compact Guide to Large Language Models



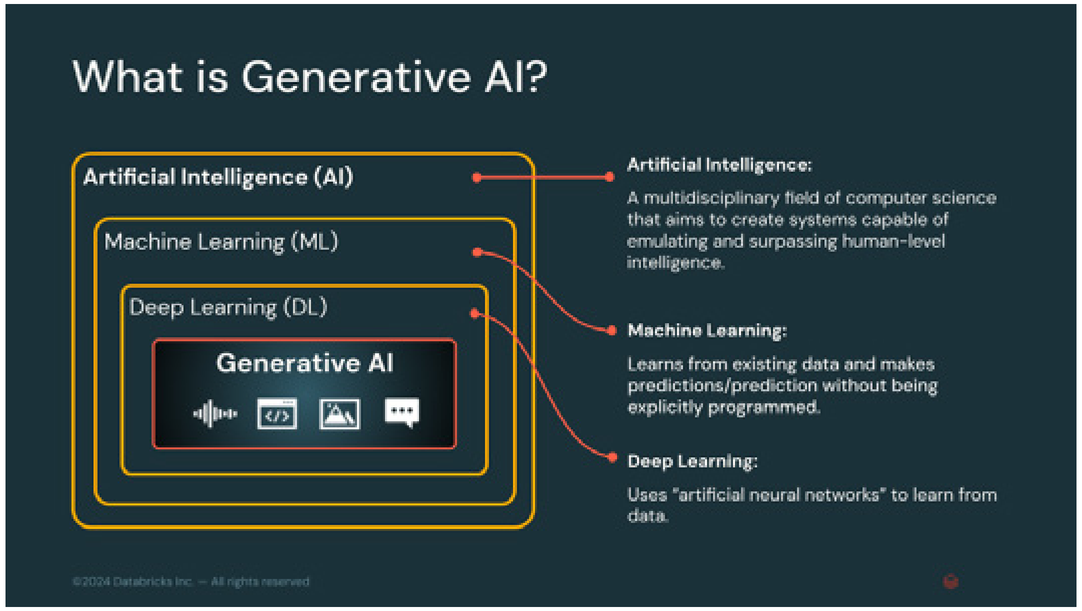
SECTION 1

Introduction

Definition of large language models (LLMs)

Large language models are AI systems that are designed to process and analyze vast amounts of natural language data and then use that information to generate responses to user prompts. These systems are trained on massive datasets using advanced machine learning algorithms to learn the patterns and structures of human language, and are capable of generating natural language responses to a wide range of written inputs. Large language models are becoming increasingly important in a variety of applications such as natural language processing, machine translation, code and text generation, and more.

While this guide will focus on language models, it's important to understand that they are only one aspect under a larger generative AI umbrella. As AI technologies progress, LLMs are becoming crucial components within larger compound AI systems that are helping to push the boundaries of AI capabilities.

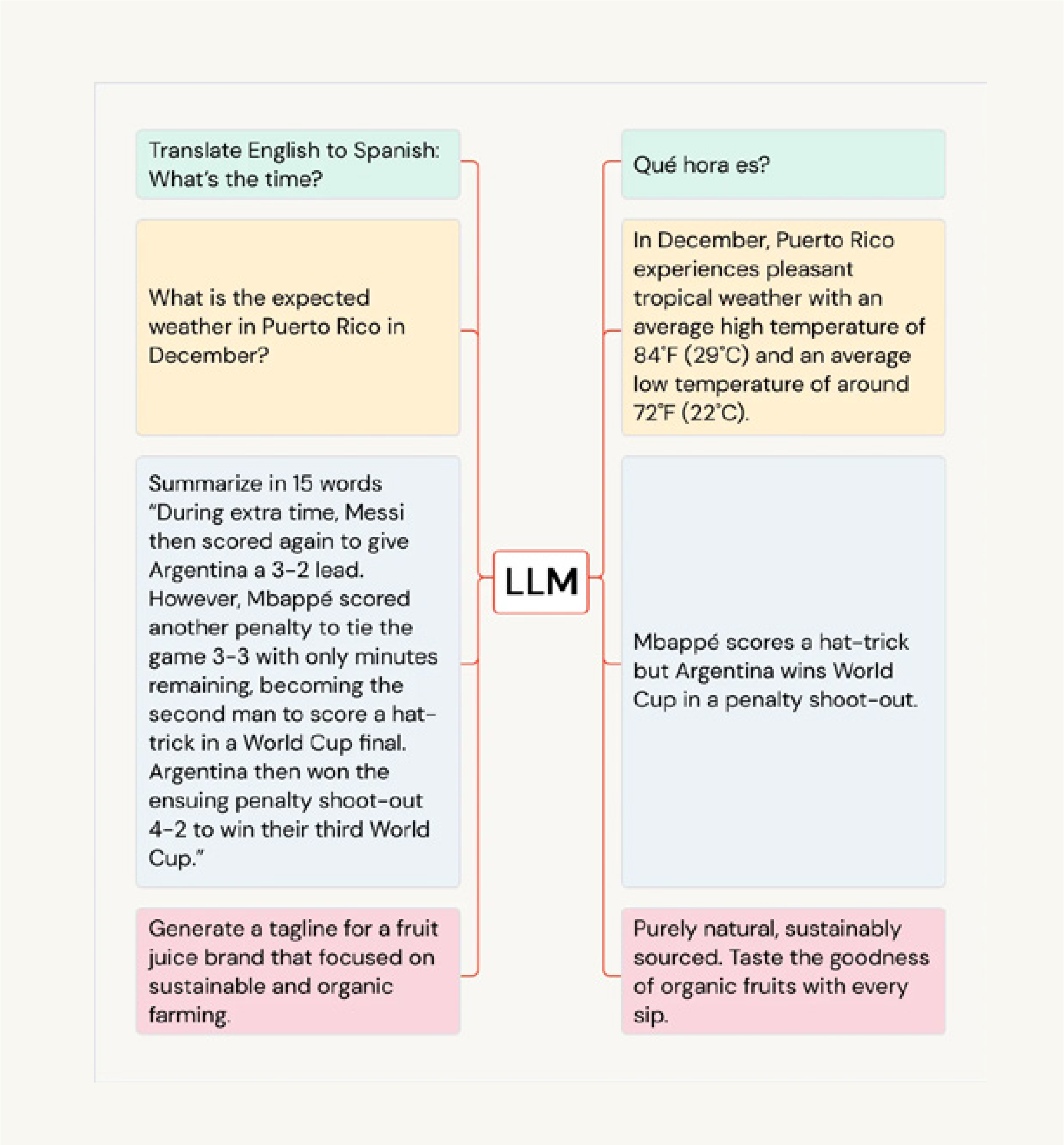


What is Generative AI?

Artificial Intelligence:
A multidisciplinary field of computer science that aims to create systems capable of emulating and surpassing human-level intelligence.

Machine Learning:
Learns from existing data and makes predictions/prediction without being explicitly programmed.

Deep Learning:
Uses "artificial neural networks" to learn from data.



Translate English to Spanish: What's the time?

Qué hora es?

What is the expected weather in Puerto Rico in December?

In December, Puerto Rico experiences pleasant tropical weather with an average high temperature of 84°F (29°C) and an average low temperature of around 72°F (22°C).

Summarize in 15 words "During extra time, Messi then scored again to give Argentina a 3-2 lead. However, Mbappé scored another penalty to tie the game 3-3 with only minutes remaining, becoming the second man to score a hat-trick in a World Cup final. Argentina then won the ensuing penalty shoot-out 4-2 to win their third World Cup."

Mbappé scores a hat-trick but Argentina wins World Cup in a penalty shoot-out.

Generate a tagline for a fruit juice brand that focused on sustainable and organic farming.

Purely natural, sustainably sourced. Taste the goodness of organic fruits with every sip.

Extremely brief historical background and development of LLMs

1950s–1990s

Initial attempts are made to map hard rules around languages and follow logical steps to accomplish tasks like translating a sentence from one language to another.

While this works sometimes, strictly defined rules only work for concrete, well-defined tasks that the system has knowledge about.

1990s

Language models begin evolving into statistical models, and language patterns start being analyzed, but larger-scale projects are limited by computing power.

2000s

Advancements in machine learning increase the complexity of language models, and the wide adoption of the internet sees an enormous increase in available training data.

2012

Advancements in deep learning architectures and larger datasets lead to the development of GPTs (generative pretrained transformers).

2018

Google introduces BERT (Bidirectional Encoder Representations from Transformers), which is a big leap in architecture and paves the way for future large language models.

2020

OpenAI releases GPT-3, which, at 175B parameters, becomes the largest model and sets a new performance benchmark for language-related tasks.

2022

ChatGPT is launched, which turns GPT-3 and similar models into a service that is widely accessible to users through a web interface and kicks off a huge increase in public awareness of LLMs and generative AI.

2023

Open source LLMs begin showing increasingly impressive results with releases such as Meta's Llama, Alpaca and Vicuna. GPT-4 is also released, setting a new benchmark for both parameter size and performance.

2024

Model architectures start changing as scaling single models larger than GPT-4 proves to be economically challenging. Mixture of experts (MoE) models and chain of thought (COT) techniques help to drive overall AI performance. Open source also begins to bridge the gap in performance with proprietary models and systems. Agent architectures begin to promise new avenues for AI to further enhance its usefulness by actively retrieving data or performing actions on a user's behalf.

SECTION 2

Understanding Large Language Models

What are language models and how do they work?

Large language models are advanced artificial intelligence systems that take some input and generate humanlike text as a response. They work by first analyzing vast amounts of data and creating an internal structure that models the natural language datasets that they're trained on. Once this internal structure has been developed, the models can then take input in the form of natural language and approximate a good response.

If they've been around for so many years, why are they just now making headlines?

A few recent advancements have really brought the spotlight to generative AI and large language models:

- **ADVANCEMENTS IN TECHNIQUES:** Over the past few years, significant advancements have been made in the techniques used to train these models, resulting in big leaps in performance. Notably, one of the largest jumps in performance has come from integrating human feedback directly into the training process.

- **INCREASED ACCESSIBILITY:** The release of ChatGPT opened the door for anyone with internet access to interact with one of the most advanced LLMs through a simple web interface. This brought the impressive advancements of LLMs into the spotlight, since previously these more powerful LLMs were available only to researchers with large amounts of resources and those with very deep technical knowledge.

- **GROWING COMPUTATIONAL POWER:** The availability of more powerful computing resources, such as graphics processing units (GPUs), and better data processing techniques allowed researchers to train much larger models, improving the performance of these language models.

- **IMPROVED DATA AND CONTEXT:** As we've gotten better at collecting and analyzing large amounts of data, the model performance has improved dramatically. Advancements in the ability of LLMs to retrieve relevant information in real time make domain-specific questions much easier to answer.

So what are organizations using large language models for?

Here are just a few examples of common use cases for large language models:

- **CHATBOTS AND VIRTUAL ASSISTANTS**
  Some of the most common implementations that organizations use LLMs for are to provide help with things like customer support, troubleshooting or even having open-ended conversations with user-provided prompts.

- **CODE GENERATION AND DEBUGGING**
  LLMs can be trained on large amounts of code examples and give useful code snippets as a response to a request written in natural language. With the proper techniques, LLMs can also be built in a way to reference other relevant data that it may not have been trained with, such as a company's documentation, to help provide more accurate responses.

- **SENTIMENT ANALYSIS**
  LLMs can help take a piece of text and gauge emotion and opinions — often a hard task to quantify. This can help organizations gather the data and feedback needed to improve customer satisfaction.

- **TEXT CLASSIFICATION AND CLUSTERING**
  The ability to categorize and sort large volumes of data enables the identification of common themes and trends, supporting informed decision-making and more targeted strategies.

- **LANGUAGE TRANSLATION**
  Globalize all your content without hours of painstaking work by simply feeding your web pages through the proper LLMs and translating them into different languages. As more LLMs are trained in other languages, quality and availability will continue to improve.

- **SUMMARIZATION AND PARAPHRASING**
  Entire customer calls or meetings can be efficiently summarized so that others can more easily digest the content. LLMs can take large amounts of text and boil it down to just the most important points.

- **CONTENT GENERATION**
  Start with a detailed prompt and have an LLM develop an outline for you. Then continue on with those prompts and LLMs can generate good first drafts for you to build off. Use them to brainstorm ideas, and ask the LLM questions to help you draw inspiration from them.

Note: Most LLMs are not trained to be fact machines. They know how to use language, but they might not know who won the big sporting event last year. It's always important to fact-check and understand the responses before using an LLM as a reference.

SECTION 3

Applying Large Language Models

There are a few paths that one can take when looking to apply large language models for their given use case. Generally speaking, you can break them down into two categories, but there's some crossover between each. We'll briefly cover the pros and cons of each and what scenarios fit best for each.

Proprietary services

As the first widely available LLM-powered service, OpenAI's ChatGPT was the explosive charge that brought LLMs into the mainstream. ChatGPT provides a nice user interface (or API) where users can feed prompts to one of many models (GPT-3.5, GPT-4 and more) and typically get a fast response. These are among the highest-performing models, trained on enormous datasets, and are capable of extremely complex tasks both from a technical standpoint, such as code generation, as well as from a creative perspective like writing poetry in a specific style.

The downside of these services is the absolutely enormous amount of compute required not only to train them (OpenAI has said GPT-4 cost them over $100 million to develop) but also to serve the responses. For this reason, these extremely large models will likely always be under the control of organizations and require you to send your data to their servers in order to interact with their language models. This raises privacy and security concerns and also subjects users to "black box" models, whose training and guardrails they have no control over. Also, due to the compute required, these services are not free beyond a very limited use, so cost becomes a factor in applying these at scale.

In summary: Proprietary services are great to use if you have very complex tasks, are OK with sharing your data with a third party and are prepared to incur costs if operating at any significant scale.

Open source models

The other avenue for language models is to go to the open source community, where there has been similarly explosive growth over the past few years. Communities like Hugging Face gather hundreds of thousands of models from contributors that can help solve tons of specific use cases such as text generation, summarization and classification. The open source community has been quickly catching up to the performance of the proprietary models but still hasn't matched the performance of the proprietary frontier models.

It does currently take a little bit more work to grab an open source model and start using it, but progress is moving very quickly to make them more accessible to users. On Databricks, for example, we've made improvements to open source frameworks like MLflow to make it very easy for someone with a bit of Python experience to pull any Hugging Face transformer model and use it as a Python object. Oftentimes, you can find an open source model that solves your specific problem that is **orders of magnitude** smaller than ChatGPT, allowing you to bring the model into your environment and host it yourself. This means that you can keep the data in your control for privacy and governance concerns as well as manage your costs.

Another huge upside to using open source models is the ability to fine-tune them to your own data. Since you're not dealing with a black box of a proprietary service, there are techniques that let you take open source models and train them to your specific data, greatly improving their performance on your specific domain. We believe the future of language models is going to move in this direction, as more and more organizations will want full control and understanding of their LLMs.

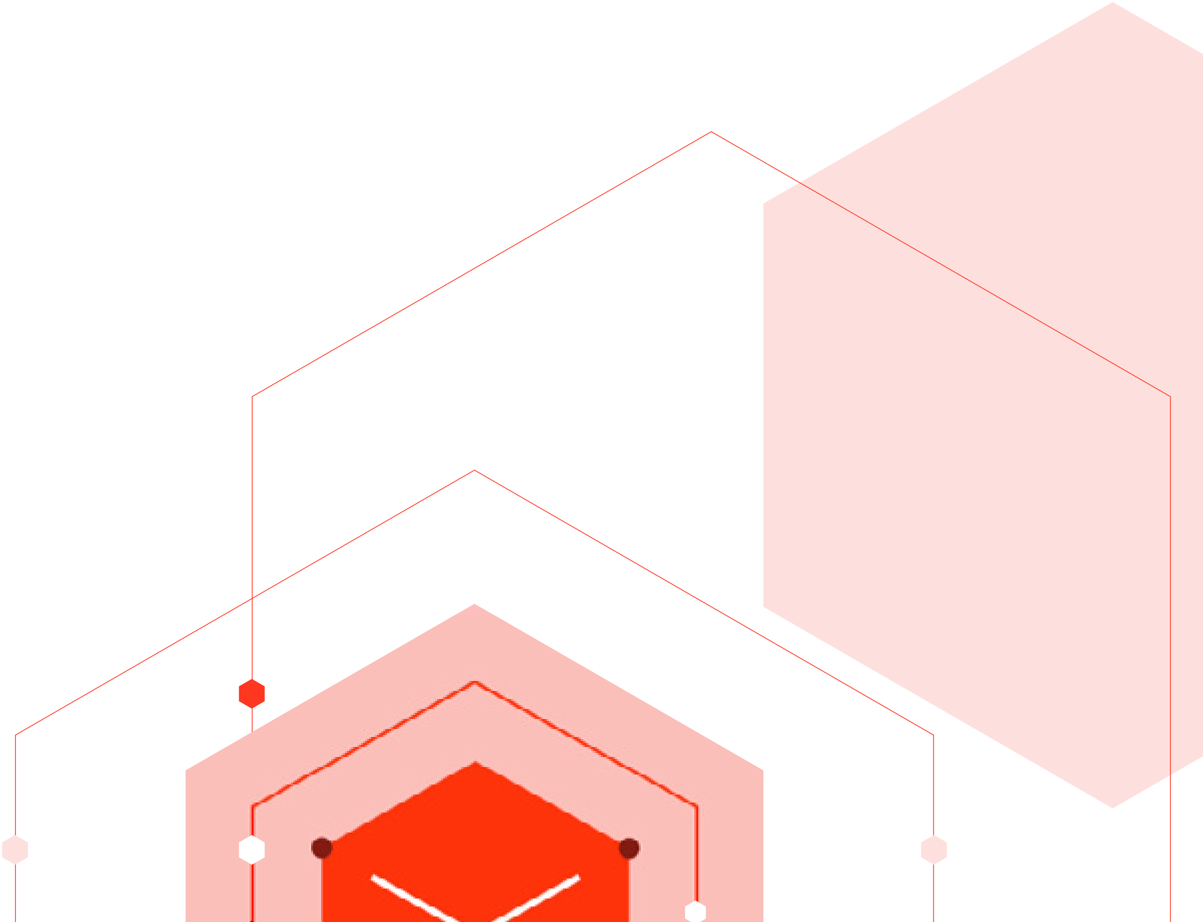
Agentic and compound AI systems

As AI maturity progresses, more complex systems are being created that use not just one LLM but a collection of one or more LLMs in combination with tools or functions that the LLMs can execute. This leads to systems that have *agency*, or the ability to take action, and greatly expands their capabilities. These agent systems leverage the LLM as the "brain" to help with planning and decision-making.

Such systems can be built on both open source and proprietary models, and when architected properly can use tools to reference the latest data and give the highest-quality responses to the end user.

Conclusion and general guidelines

Ultimately, every organization is going to have unique challenges to overcome, and there isn't a one-size-fits-all approach when it comes to LLMs. As the world becomes more data driven, everything — including LLMs — will be reliant on having a strong foundation of data. LLMs are incredible tools, but they have to be used and implemented on top of this strong data foundation. Databricks brings both that strong data foundation as well as the integrated tools to let you use and fine-tune LLMs and build agent systems that can reason over your data.



SECTION 4

So What Do I Do Next If I Want to Start Using LLMs?

That depends on where you are on your journey! Fortunately, we have a few paths for you.

If you just want to dip your toes in the water and see what using LLMs looks like in practice, you can check out a quick video of Databricks Co-founder and CEO Ali Ghodsi giving a recap:

https://www.databricks.com/resources/demos/videos/ali-ghodsis-microsoft-ignite-demo-data-intelligence-platform?itm_data=demo_center

If you're ready to get hands-on but aren't sure where to start, we offer free hands-on generative AI training:

https://www.databricks.com/resources/learn/training/get-started-with-generative-ai

For those who might be more experienced developers and just want some code examples to get them started, we have a code-based tutorial that will get you up and running here:

https://docs.databricks.com/en/generative-ai/tutorials/ai-cookbook/introduction.html

Getting started with NLP using Hugging Face transformers pipelines

Fine-Tuning Large Language Models with Hugging Face and DeepSpeed

Introducing AI Functions: Integrating Large Language Models with Databricks SQL

About Databricks

Databricks is the data and AI company. More than 9,000 organizations worldwide — including Comcast, Condé Nast and over 50% of the Fortune 500 — rely on the Databricks Lakehouse Platform to unify their data, analytics and AI. Databricks is headquartered in San Francisco, with offices around the globe. Founded by the original creators of Apache Spark™, Delta Lake and MLflow, Databricks is on a mission to help data teams solve the world's toughest problems. To learn more, follow Databricks on Twitter, LinkedIn and Facebook.

START YOUR FREE TRIAL

Contact us for a personalized demo:
databricks.com/contact

databricks