










Getting Started with Generative AI on AWS

eBook



Table of Contents



	Introduction
	Selecting and customizing foundational models
	Issues to consider with foundational models
	The opportunities ahead with generative AI
	Build with generative AI on AWS
	About Magic Beans
	Case study: EfectoLED



Introduction

Large-language models (LLMs) and multi-modal models are enabling new capabilities, from code generation to the creation of images based on natural language descriptions. Together these new capabilities are transforming existing machine learning (ML)-enabled capabilities, such as web search and chatbots. These models are known as foundation models (FMs) because customers can easily customize them for their specific use cases (such as summarization, translation, code generation) without having to build a new ML model from scratch for each use case.

The main step of creating an FM (known as pre-training) involves training a model with terabytes of unlabeled text and/or multi-modal data (such as images, audio, video). This unlabeled data used for pretraining is usually obtained by crawling the web and contains information from publicly crawled sources (such as Wikipedia and other sites) and proprietary data (if available). A large model with billions of parameters can better capture this knowledge as it is able to store richer and deeper context across large amounts of data in its memory compared to a smaller model trained on a smaller data set. Upon completion of the pre-training step, the resulting model can deliver an impressive out-of-the-box performance on a wide range of tasks across multiple domains.

To help you get started with generative artificial intelligence (AI) and foundation models, this ebook presents a broad overview of the FM landscape, outlines opportunities and risks for LLMs, and shows how you can leverage these LLMs using various AWS technologies to build highly differentiated solutions in a secure manner.



Selecting and customizing foundational models



ML practitioners usually choose a particular type of FM based on their needs and then either use it out of the box or customize it for their specific use case.

- **Zero-shot learning** While FMs are typically used by ML practitioners, many FM providers offer a web playground or a chat interface (for example, ChatGPT) to make interacting with an FM easy, even for non-ML experts. Customers simply provide different natural language commands (known as prompts) such as “list all the action items from this meeting transcript,” or “translate this doc into German.”
- **In-context learning** Customers (developers or ML practitioners) can also include a handful of examples as part of their input prompts to improve the relevancy of the model’s outputs on the fly. For example, a customer can provide the prompt: “create a social media ad for the 2023 Polygon Xtrada 7 mountain bike based on its product description” along with a few examples of their company’s past social media ads for similar products.
- **Fine-tuning** FMs can also be customized for specific tasks. Customers can further train the FM by using a small number of labeled examples. This approach is efficient and cost-effective because the amount of labeled data required for fine-tuning is orders of magnitude smaller than what is required to develop a task-specific model from scratch. For example, a professional recruiting firm can customize the FM to automatically process new incoming resumes and generate summarized resumes at scale by fine-tuning the model with a few examples of candidate resumes.



Issues to consider with foundational models

Responsible AI

Foundation models raise new issues in defining, measuring, and mitigating responsible AI concerns across the development cycle including accuracy, fairness, intellectual property considerations, toxicity, and privacy, among others. These new challenges stem from the vast size of foundation models, trained by billions of parameters and their open-ended nature compared to traditional uses of machine learning, which are typically more focused and narrow.

With foundation models we also see emerging concerns like toxicity, the possibility of generating content (whether it be text, images, or other modalities) that is offensive, disturbing, or otherwise inappropriate, and intellectual property considerations, where LLMs occasionally produce text or code passages that were verbatim regurgitations of parts of their training data. While these challenges may seem daunting, new research, science, and policies are already being created to address these challenges from end user education to filtering to more technical concepts like watermarking and differential privacy.

Hallucinations

LLMs can suffer from hallucination (i.e., they make up inaccurate responses that are not consistent with the training data). They are typically a byproduct of the way these FMs represent their inputs, often causing them not to distinguish among different numeric values or names, to “invent” facts to be consistent with the requested output format (such as making up citations and author names when asked to provide evidence to an answer), and to conflate facts that are presented by multiple sources in their input.

Costs

While you typically train a model periodically, a production application can be constantly generating predictions, known as inferences, potentially generating millions per hour. And these predictions need to happen in real-time, which requires very low-latency and highthroughput networking. Customizing models to specific use cases can result in smaller, more fine-tuned models that are more accurate and scale better.

The opportunities ahead with generative AI



Generative AI has the potential to bring about sweeping changes to the global economy. According to Goldman Sachs, generative AI could drive a 7% (or almost \$7 trillion) increase in global GDP and lift productivity growth by 1.5 percentage points over a 10-year period. Much of this growth is driven by spend on generative AI cloud services which are estimated by Bloomberg to reach over \$109B by 2030, a CAGR of 34.6% from 2022 to 2030. At AWS, we have played a key role in democratizing ML and making it accessible to anyone who wants to use it, including more than 100,000 customers of all sizes and industries. This is why customers like Intuit, Thomson Reuters, AstraZeneca, Ferrari, Bundesliga, 3M, and BMW, as well as thousands of startups and government agencies around the world, are transforming themselves, their industries, and their missions with ML leveraging AWS capabilities from our infrastructure through our managed services and access to a variety of FMs.

Sources: Generative AI could raise global GDP by 7%, Goldman Sachs, April 2023
Generative AI market to be worth \$109,37 billion by 2030, Bloomberg, January 2023



Generative AI will play a transformational role in industries like:

- Healthcare and Life Sciences
- Financial Services
- Media and Entertainment
- Education
- Automotive and Manufacturing

Build with generative AI on AWS



Today, many AWS customers are seeing an impact from generative AI. Here are the top reasons why customers choose AWS to build generative AI applications:

Innovate with generative AI

With enterprise grade security and privacy, a choice of leading foundation models, a data first approach, and the most performant, low-cost infrastructure, organizations trust AWS to deliver generative AI fueled innovation at every layer of the technology stack.

Securely build and scale generative AI applications

Customers trust AWS to build generative AI services and capabilities responsibly and securely. AWS also offers the most comprehensive set of services, tooling, and expertise to help you protect your data, so it remains secure and private when you customize and fine tune foundation models.

The most performant, low-cost infrastructure

Train your own models and run inference at scale. With AWS, you get the most performant and low-cost infrastructure for generative AI and the broadest choice of accelerators in the cloud.

Data as your differentiator

With AWS, it's easy to use your organization's data as a strategic asset to customize foundation models and build more differentiated experiences. Securely customize a foundation model on AWS with your data and build generative AI applications that truly know your business and customers.



About Magic Beans

Who are we in the Magic of the Cloud?

- **Magical Transformation:** We lead the Digital Transformation, agilely navigating the vast universe of cloud services.
- Certified partners of **AWS, Google Cloud, Azure, Oracle OCI** and other relevant partners.
- More than **300 projects** in 150 customers with **Migration, Transformation, Modernization, Efficiency, Data Management, Interoperability, Methodology projects**.
- A team of more than 70 consultants with multiple certifications, dedicating more than 30% of their time to training.
- Founded in **2017 in Portugal**. Expansion to **Spain in 2020** and in **2022 in Belgium and Italy**.

Case Study

Challenge

EfectoLED integrates in their websites Zendesk as chat application, with this they can offer detailed assessment of the products to the end-users.

The main challenge was produced because of the need of chat automation since it was a real person who was in charge of answering the doubts customers had. Other key points to be considered was the formality answering and detection of sensitive data.

Solution

The MagicBeans work around solution was to implement a GenAI chatbot using Amazon Bedrock. Bedrock contains several generative models like LLMs, for the solution Anthropic Claude v2 was implemented due to it inherit resistance against sensitive data and conversational tone.

With this, the speed of answering questions was improved from minutes to just very few seconds.

Benefits

- **Cloud Deployment:** Elasticity can be achieved through an implementation of a serverless event-based architecture with AWS Lambda.
- **Cost Optimization:** As AWS infrastructure was implemented using serverless services there is no need to maintain and update them, also the pay-as-you-go feature allows to pay for only what is consumed.
- **Agility:** With Amazon Bedrock, a solution that could be addressed in several months, was speed up and running in few days thanks to the use of Foundational Models.
- **Visibility:** Amazon offers several visibility real-time metrics in order to keep trace of usage with custom dashboards



EfectoLED is more than just an LED lighting company.

It is an ambitious, innovative and forward-thinking business group dedicated to offering high-quality lighting solutions, avant-garde design and, above all, a focus on sustainability.

Learn more with Magic Beans

For more information on how Magic Beans can help your business, visit our website www.magicbeans.pt or email us at team@magicbeans.pt





Take your Business to the Cloud