



Amazon Web Service (AWS) Foundation Models



Cindy Waeltermann (C)
Contingent Worker

Although foundation models have been around for some time, the models that are available today are receiving much buzz. This is because they have been trained on massive amounts of text data for years and have improved to the point where they can handle complex language tasks with more accurate results through billions of parameters. They are pre-trained, but can be fine-tuned to suit individual needs, making them attractive to developers across the globe.

AWS Foundation Model Overview

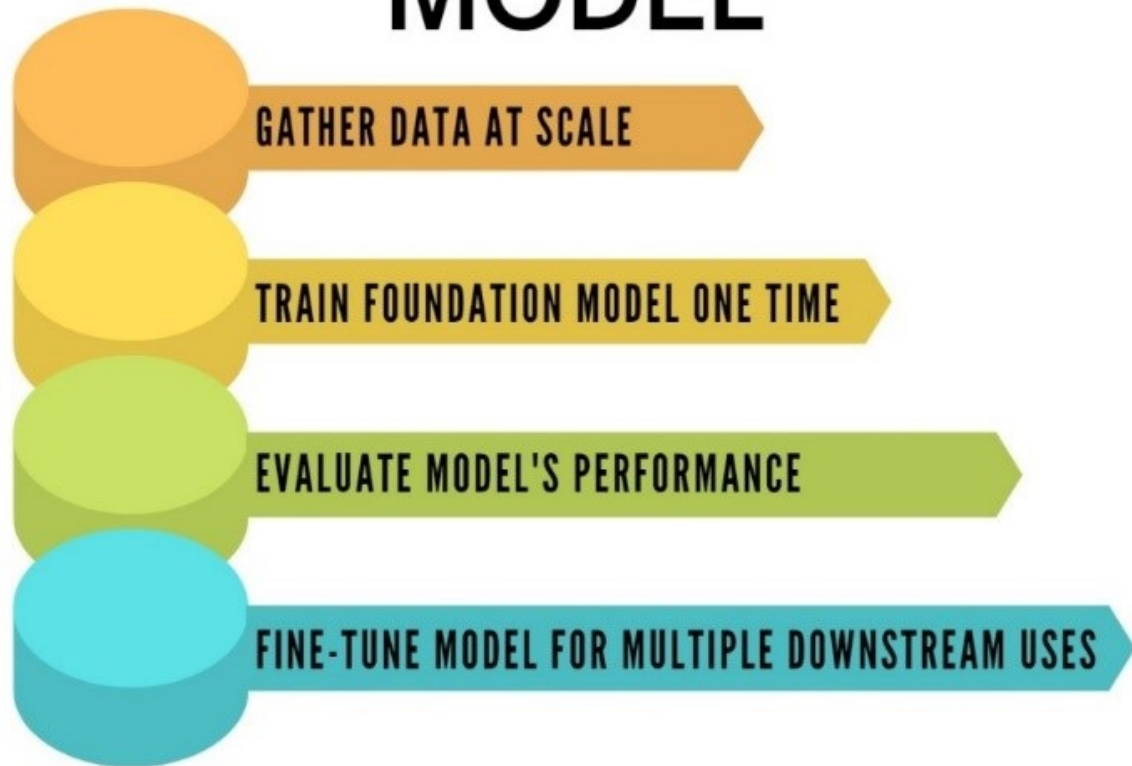
Before delving into some of the features of Amazon Web Service (AWS) foundation models, it is important to first define a foundation model itself.

A foundation model is a large AI model trained on a vast quantity of unlabeled data at scale, resulting in a model that can be adapted to a wide range of downstream tasks.

The term “foundation model” was first popularized by the [Stanford Institute for Human-Centered Artificial Intelligence \(HAI\)](#). According to Stanford’s [Center for Research on Foundation Models](#) (CRFM), which was created just recently in 2021, the CRFM “is a new interdisciplinary initiative born out of the Stanford Institute for

HAI
that
aims
to
make

FOUNDATION MODEL



fundamental advances in the study, development, and deployment of foundation models."

The CRFM is an interdisciplinary group of faculty, students, post-docs, and researchers who have a shared interest in studying and building responsible foundation models. Their [blog](#) is full of great information on foundation models, and they also hold [online workshop courses](#) to keep abreast of the latest advances and ethical considerations.

Early examples of foundation models ([GPT-3](#), [BERT](#), [DALL-E 2](#)) have shown what is possible. Take the buzz around ChatGPT, for example – its popularity has recently exploded. Input a short prompt, and the system generates an entire essay even if it wasn't specifically trained on how to execute that exact argument in that way. (Click [here](#) for an article from the Vertex Emerging

Technology Department on ChatGPT.)

Early Foundation models include:

- **GPT-3** – (Generative Pre-trained Transformer) uses deep learning algorithms to produce text that appears to have been written by a human being.
- **BERT** – (Bidirectional Encoder Representations from Transformers) helps artificial intelligence programs understand the context of ambiguous words in a text by processing text in left-to-right and right-to-left directions, simultaneously, to determine a word's context.
- **DALL-E2** – uses a process called “diffusion” to create realistic images and art from a text description.

How are businesses using foundation models

Although foundation models have been around for some time, the models that are available today are receiving much buzz. This is because they have been trained on massive amounts of text data for years and have improved to the point where they can handle complex language tasks with more accurate results through billions of parameters. They are pre-trained, but can be fine-tuned to suit individual needs, making them attractive to developers across the globe.

Foundation models are typically used in the following ways:

- **Sentiment analysis** is an NLP technique used to extract subjective information from text, such as opinions, emotions, and attitudes. Data is collected from any source – it could be via social media, customer reviews, or news, for example. Sentiment algorithms are applied to the text data through rule-based approaches and machine learning. The algorithms are used to learn and identify patterns in the text that depict sentiment and classify the sentiment the text expresses – whether it is

positive, negative, or neutral. Sentiment analysis can be applied in customer feedback analysis, brand monitoring, product development, market research, political analysis, and healthcare, to name only a few. With sentiment analysis, there are many, many applications that can provide insight into many business areas.

- **Text classification** is a NLP technique that automatically categorizes text data with pre-defined labels, classes, or categories. Text classification is helpful for organizing large volumes of text data and has applications in spam filtering and customer service.

The following is a list of common uses for Natural Language Processing foundation models:

- Email filters
 - Virtual assistants
 - Online search engines
 - Predict text and autocorrect
 - Brand monitoring on social media
 - Sorting customer feedback
 - Chatbots
 - Automatic summarization
 - Machine translation
 - Natural language generation

While this paper concentrates on the foundation models offered in Amazon Web Services (AWS), it is important to note that there are hundreds of models, tools, and services on AWS that are not foundation models, available for use. These models are available in AWS SageMaker, an end-to-end managed machine learning platform.

What is Amazon SageMaker?

[Amazon SageMaker](#) is a cloud-based, fully managed service provided by AWS

where you can access pre-built machine learning algorithms to help you get started with machine learning quickly. You can also use SageMaker's tools to prepare and preprocess data, train and fine-tune models, and deploy them to production environments.

JumpStart is the machine learning (ML) hub of Amazon SageMaker, which offers over 350 pre-trained models, built-in algorithms, and pre-built solution templates. JumpStart hosts state-of-the-art models from popular model hubs such as TensorFlow, PyTorch, Hugging Face, and MXNet, which support popular ML tasks such as object detection, text classification, and text generation. JumpStart is just what its name suggests – a jumpstart to using ML models.

SageMaker offers support for the following leading ML frameworks, toolkits, and programming languages:



Foundation Models available in SageMaker JumpStart

SageMaker JumpStart offers several foundation models from [AI21labs](#), [LightOn](#), [stability.ai](#), [co:here](#), [Hugging Face AI](#), and [Alexa](#).

AI21 Jurassic-1

[Announced on November 30, 2022](#), the Jurassic-1 (J1) Natural Language Processing (NLP) foundation models are available on SageMaker JumpStart. [Jurassic-1](#) is the first generation in a series



of large language models trained and made accessible by AI21 Labs. Jurassic-1 is a set of auto-regressive language models consisting of J1-Jumbo, a 178B-parameter model (read as "178 billion parameters" - the larger the number of parameters, the more capable the model), and J1-Large, a 7B-parameter model.

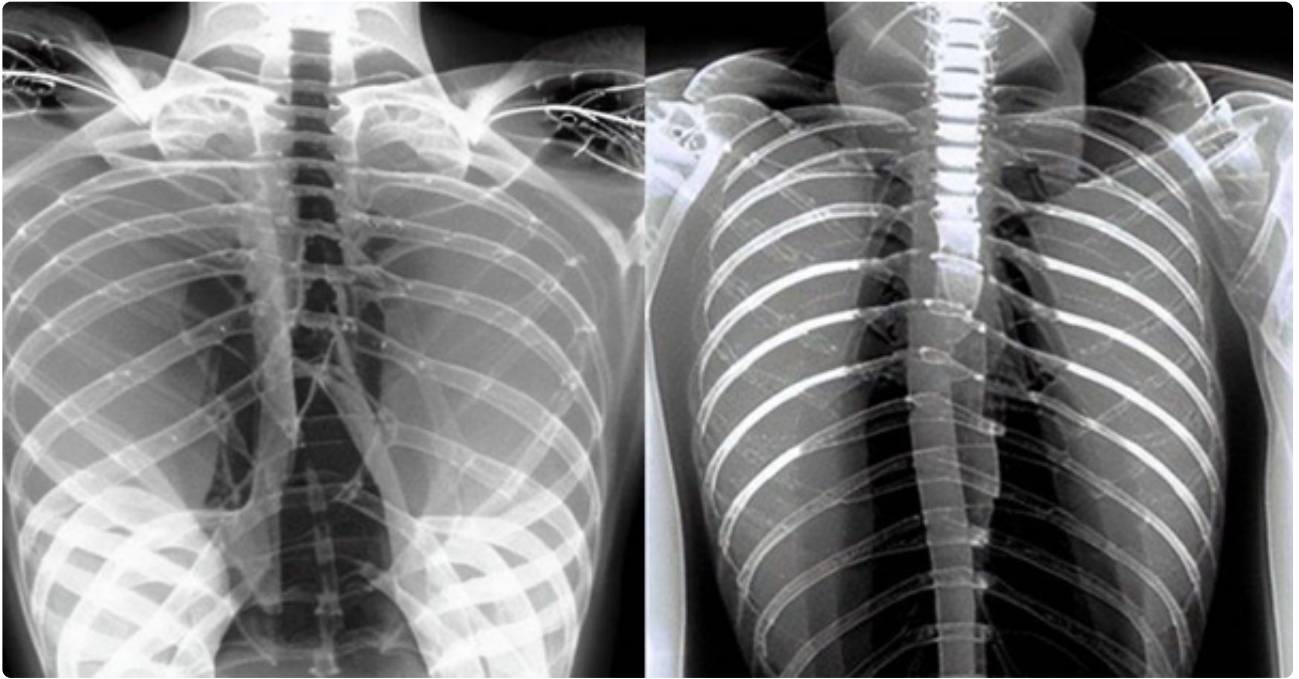
[Latitude](#), a startup gaming company, uses the J-1 model for their wildly popular [AI Dungeon](#) (which recently garnered \$3.3 million in seed funding). AI Dungeon is a text-based adventure game that uses artificial intelligence to generate unique and unpredictable stories, depending on the choices that a player makes. In the past, computer adventure games were limited in terms of the amount of actions and scenarios they offered. By using Large Language Models, AI Dungeon grants players the flexibility to perform virtually any action and have the game respond to that action. (Basically, it's Dungeons and Dragons with AI.)

Stable Diffusion 2

[Stable Diffusion](#) is a deep learning, text-to-image model where you can generate realistic images with text input. Developed by Stability.ai, Stable Diffusion allows you to run the program on your personal computer and generate as many images as you want, without ever connecting to a cloud. Diffusion Models are generative models, meaning that they are used to generate data similar to the data on which they are trained. Stable Diffusion works by destroying training data through the successive addition of Gaussian noise and then learning to recover the data by removing the noise.

While the obvious use cases for stable diffusion include video games, product and architecture design, marketing, facial recognition, etc., Stanford has found a way to use it to create medical images that accurately depict clinical context for training purposes. In other words, medical schools are now using Stable Diffusion to create images that are applicable to certain conditions, so students can see what they look like on an x-ray.

In the latest news, telecom companies are now trying to put Stable Diffusion on phones. Qualcomm, in particular, is currently squeezing Stable Diffusion onto



smartphones, as noted in this article from [The Verge](#). Apple is also making strides get Stable Diffusion to run locally on its machine learning framework.

If you're interested in Stable Diffusion, you can test it out [here](#). It's fun to see how your words can conjure an image.

AWS offers Stable Diffusion 1, Stable Diffusion 2, Stable Diffusion x4 Upscaler, Stable Diffusion, and Stable Diffusion 2 FP16.

Co:here language model

[Co:here](#) is a Canadian AI company that specializes in large [language models](#) that perform common tasks on text input such as summarization, classification, and finding similarities in content. Language models based on Cohere can be customized with your own training data.



Cohere provides access to two types of language models:

- **Generation Language Models** take in text as input, depending on the task the model is designed to achieve. An example of a generative language model is GPT3.5, for example, the basis for the [ChatGPT](#)

developed by OpenAI.

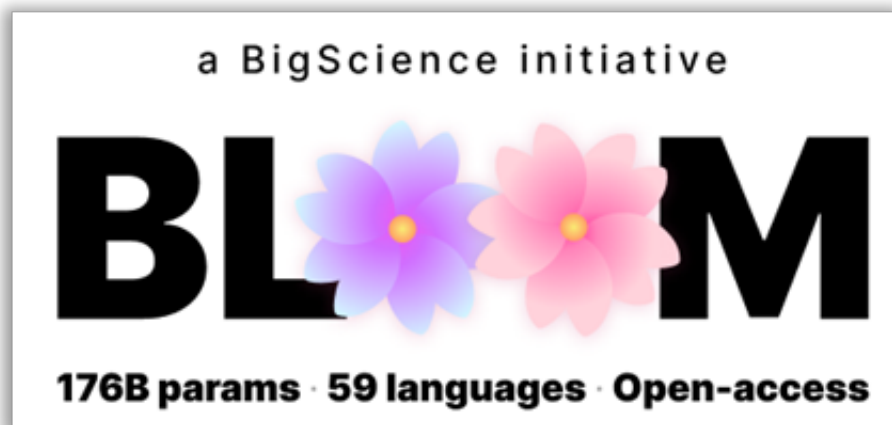
- **Representation Language Models** take text as an input but instead of generating new text, they use Embeddings to represent the text as numbers. Mathematical functions are then used to calculate the similarity between phrases and blocks of text. An example of where embeddings are useful is a support chatbot.

LightOn Lyra-fr model

[Lyra-FR](#) is a speech recognition model developed by LightOn that can transcribe spoken words into text. It is a state-of-the-art French language model that can be used to build conversational AI, copywriting tools, text classifiers, semantic search, etc. Created by LightOn, a Paris AI company, it is a 10 billion parameter model that was trained on vast amounts of French curated data, and it is capable of writing human-like text and solving complex tasks such as classification, question answering, and summarization.

Bloom

[Bloom](#) is a large-scale pre-trained natural language processing model specifically designed to understand and process natural language text input in a way that is similar to how humans understand language. Developed by [Hugging Face AI](#), it is available as open source.



The model is based on deep learning techniques such as neural networks and is trained on large datasets of text to learn patterns and relationships between words.

and phrases.

Bloom's NLP models can be used for sentiment analysis, text classification, entity recognition, and machine translation, to name a few. SageMaker JumpStart offers the Bloom 1b7, Bloom 1b1, and Bloom 560m (language models).

AlexaT M (20b)

[AlexaTM 20b](#) is a multilingual foundation model with 20 billion parameters developed by Amazon for its voice assistant technology, Alexa. The model was announced in 2021 and is part of a series of new models that aim to improve the natural language processing capabilities of the Alexa voice assistant.



The Alexa Teacher Model (Alexa TM) program by Amazon Alexa AI is designed to build large-scale, multilingual deep learning models that improve generalization and handling data scarcity for downstream tasks. AlexaTM 20B can not only transfer what it learns across languages but also learn new tasks from just a handful of examples (few-shot learning).

AlexaTM 20B can be used for a wide range of industry use-cases, from summarizing financial reports to question answering for customer service chatbots. The AlexaTM 20B model is designed to be multilingual, meaning that it can understand and respond to queries in multiple languages.

What are the AWS foundation models good for?

Having ready-made starting point when working on a major project is always helpful. Reinventing the wheel is usually not a good use of time. For developers

who intend to create their own models, the foundation models available in SageMaker JumpStart can be a major time saver. Because the foundation models are pre-trained, a good chunk of the work has already been done. Except for the Stable Diffusion model, all the foundation models in SageMaker JumpStart are language models.





How do AWS foundation models benefit businesses




AWS Foundation Models benefit businesses by providing intelligent and fast insights, versatility, while providing time and cost efficiency.

- Intelligent Insights-They can analyze and interpret large amounts of unstructured data that help businesses gain insights and make data-driven decisions.
- Versatility-As they can be applied to virtually any industry.
- Time efficiency- The models save quite a bit of time for developers because the groundwork has already been laid for the model, resulting in faster development times.
- Saves money-They are more cost-effective than building a model from the ground up.
- In addition to having access to the foundation models (and hundreds of other models and tools), Amazon SageMaker offers the infrastructure necessary to run and deploy models, at a price.



Impact of foundation models on business




Listed below are just a few examples of the businesses that are currently using foundation models. The list is quite long, as the race for AI innovation has begun.

	AstraZeneca analyzes anonymized patient data to predict disease, heart failure readmission, and cancer outcomes. They also use image analysis for drug discovery.
	The Formula One Group (F1) is responsible for organizing a series of auto racing events in 21 countries worldwide. F1 uses AI to optimize race strategies, track performance, and enhance the fan experience.

	<p>Intuit implemented Amazon SageMaker which months to weeks. Intuit has been able to gen better, more personalized financial manager</p>
	<p>Verizon uses the J-1 Jurassic model by AI21 Video, which helps with fleet management us purported benefits are near real-time trackin etc.), improved driver performance, and risk</p>
	<p>Facebook uses natural language processing them contextually. Neural networks analyze t changes depending on other words around t do not necessarily have reference data – for Instead, it learns for itself based on how wor</p>





Startup Activity

	<p>Viable is a startup that uses GPT-3 to build c The chatbots are designed to simulate huma support.</p>
	<p>Cognitivescale is a startup that uses foundat models are used to understand and analyze</p>

	<p>Primer AI is a startup that uses foundation models to summarize and analyze documents.</p>
	<p>Althea AI is a startup that uses foundation models to analyze medical images. They use GPT-3 and other models to generate reports.</p>
	<p>Freenome is a startup that uses foundation models to analyze patient data and predict health outcomes.</p>

Competitor Activity

	<p>Avalara uses NLP models to analyze tax regulations and provide compliance advice.</p>

	<p>There is no mention of AI models at Sovos. reporting, there is no mention of the usage</p>
	<p>Thomson Reuters uses foundation models in the game. They have used BERT for sentiment articles. (Here is a use case of a model used and planning tool that uses AI and natural language changing tax regulations. The platform uses including foundation models, to analyze tax</p>
	<p>There is no mention of the use of foundation</p>
	<p>While they have authored several articles about use the technology.</p>

Pricing for AWS Jumpstart

AWS offers a pay-as-you-go approach for pricing where you pay only for the individual services you need, for as long as you use them, and without requiring contracts or licensing. You pay for the services you consume, and once you stop using them, there are no additional costs or termination fees.

You can configure your own cost estimate based on your needs on the AWS pricing calculator.

Potential/in-progress Vertex projects

Vertex projects in progress include the following:

- **TaxCat** is a product under development to categorize a customer's products in terms of Vertex' O-Series product codes to enable automatic preparation of taxability configuration, saving customers time.
- Auto-assign / Content Strategy is an effort to pre-assign popular product codes to taxability categories – many more than what TaxCat is targeting.
- Tax Research Modernization is an effort to modernize Vertex's Tax Research processes and make them much more efficient
- **TaxGPT**- TaxGPT is an idea for a freemium Chatbot Proof of Concept based upon ChatGPT that provides answers to generic tax-related questions. (Ironically, OpenAI announced GPT 4 and included TaxGPT as an example of what the new model can do) It is important to note that Vertex's concept will also provide a premium service that will be checked by a tax professional. More information can be found in the Jira Ticket for this item.
- **Violet**- Violet is a librarian ChatBot that finds information in Emerging Technologies' documents and brings the information back to the user. Violet works through a simple chat interface that allows users to ask questions and interact, while Violet searches through our SharePoint pages and returns her findings. More information can be found on Violet's Demo page or on our SharePoint article repository.

The possibilities of projects using foundation models is truly only limited by your own imagination. These models can be applied to any industry, anywhere. The following are projects that Vertex could potentially pursue:

- Sentiment analysis
- Compliance monitoring
- Customer education
- Risk management
- Predictive modeling

- Chatbots/Virtual Assistants

Conclusion

Foundation Model availability is growing at an unprecedented rate. AWS has positioned itself to provide a wide range of popular and useful foundation models via its JumpStart hub. Many are positioned for inexpensive exploration, although the cost of using these models at production levels is anticipated to be significant. Nonetheless, Vertex would be well-served to dig into some practical applications as noted in the list of current AI work to determine which could be aided by these foundation models – From CX applications with self-serve models in customer service to help customers on board and answer questions, to new product development using these models directly in place of coding software and/or helping developers write software more quickly and with fewer errors.

Continuing to grow our content is imperative since the ability to serve it up and manipulate it is constantly under pressure to be commoditized. Good content with the latest delivery and usage models aided by AI offer opportunities to get ahead of our competition and increase our growth to the SaaS levels our investors are seeking.

References

- [Stanford Institute for Human-Centered Artificial Intelligence \(HAI\)](#).
- [Center for Research on Foundation Models \(CRFM\)](#)
- Stanford HAI [blog](#)
- [GPT-3](#)
- [BERT](#)
- [DALL-E 2](#)
- [ChatGPT – Article from the Emerging Technology Department at Vertex](#)
- [Amazon SageMaker](#)

- [Amazon SageMaker introduction](#)
- [Amazon SageMaker](#) JumpStart
- [AWS](#) Press Release
- [AI21labs](#)
- [LightOn](#)
- [stability.ai](#)
- [co:here](#)
- [Hugging Face AI](#)
- [Alexa](#)
- [Press Release for Jurassic-1 \(J1\)](#)
- [AI21 Labs](#)
- [Use case – AI Dungeon from Latitude](#)
- [AI Dungeon](#)
- [Stable Diffusion](#) – AWS
- [Stanford use case – using AI for x-rays](#)
- [Qualcomm demos fastest local AI image generation with Stable Diffusion on mobile](#)
- [Play around with Stable Diffusion](#)
- [Co:here](#)
- Detailed information on the co:here [language models](#)
- [ChatGPT](#)
- [Lyra-FR](#)
- [Bloom model on AWS](#)
- [Hugging Face AI](#)
- [AlexaTM 20b](#)
- [Viable](#)
- [Astra Zeneca use case](#)
- [Formula 1 Racing use case](#)
- [Hyundai foundation model use case](#)
- [Intuit use case](#)
- [Thomson Reuters use case](#)
- [Cognitivescale](#)
- [Primer AI](#)
- [Alethea AI](#)

- [Freenome](#)

Innovation Portfolio

Bright Ideas

Help Us Discover the Most Impactful Business Problems, Unmet Customer Needs and Market Opportunities!

Every six months, the team at VX3 by Vertex Inc. conducts a portfolio refresh process in search of the most interesting projects to tackle for Vertex Inc. We believe in the power of the crowd, and that the strongest observations, insights and ideas can come from anywhere. Now our bright ideas platform is open all year! Don't hesitate to submit your idea, it's never too late.

We are particularly interested in submissions with [focus areas](#) that push the

boundaries, beyond our core businesses, complex problems and interesting ideas with disruption and high growth potential!

Please do not share this information outside of Vertex Inc. The information contained within this site is for **internal use only** and is for **informational purposes only**. The links to external websites are included for reference material on related subjects. Vertex does not control those sites and is not responsible for the content included in them, including without limitation any subsequent links contained within a linked site, or any changes or updates to a linked site. Vertex is not responsible for any information or material located at any site other than official Vertex websites. If you have questions regarding this message, please email corporate.communications@vertexinc.com.