Empowering Enterprises with Serverless Generative AI: Amazon Bedrock

OCTOBER 12, 2023

Matt Carey, **aleios**

As the number of Large Language Models (LLMs) continues to grow and enterprises seek to leverage their advantages, the practical difficulties of running multiple LLMs in production is becoming evident. Established hyper-scale cloud providers, such as AWS, are in a favourable position to facilitate the adoption of Generative AI, due to their existing computing infrastructure, established security measures, and modern cloud-native patterns like Serverless.

AWS's introduction of Bedrock stands out as a poignant reaction to these trends and is well positioned through it's platform centric, model agnostic and serverless operating model to be the facilitator for this next era of GenAI. Last week, Bedrock became Generally Available (GA), giving AWS customers their first look at the tools which allow them to integrate the GenAI into all aspects of their operations.

Serverless Advantage

Infrastructure management of LLMs at high scale, especially for those not entrenched in the ML/AI domain, can be daunting. Managing compute, load balancers and exposed APIs requires platform teams and few businesses are willing to make that up front investment.

Amazon Bedrock alleviates these concerns. As a serverless service, businesses only pay for the tokens consumed and generated by the LLM. These scaling challenges become things of the past.

For instance, you are building a customer support chatbot: It has 100x the users on Black Friday compared to a Monday in March. You do not have to provision extra servers. Bedrock handles the scale out and back in to meet demand.

API request Copy code

```
{
  "modelId": "cohere.command-text-v14",
  "contentType": "application/json",
  "accept": "*/*",
  "body": {
    "prompt": "Write a LinkedIn post about starting a career in tech:",
    "max_tokens": 100,
    "temperature": 0.8,
    "return_likelihood": "GENERATION"
  }
}
```

Example API Request from the AWS Console

Data Security

With an increasing emphasis on ensuring good data governance and audit trails, Bedrock provides peace of mind for enterprises seeking to adopt GenAI. All data provided to the Bedrock LLMs is encrypted at both rest and in transit and customers are free to use their own keys.

Amazon Bedrock has achieved HIPAA eligibility and GDPR compliance and provided data is never used to improve the base models or shared with third-party model providers.

Enterprises can even use AWS PrivateLink with Amazon Bedrock to establish private connectivity between LLMs and their VPC's to avoid exposing traffic to the public internet. This gives businesses the security to create tools using LLMs that can use their own sensitive data archives as context.

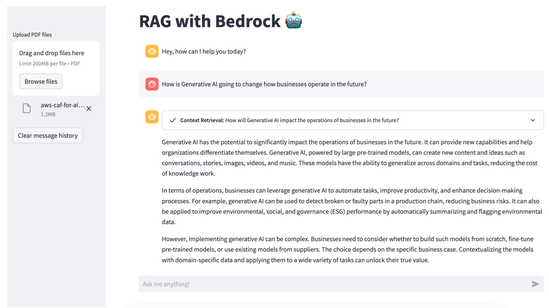
Imagine having a tool which provides enhanced search capabilities for diverse data types, from research papers and medical journals to company meeting notes and tech standards. Like your own personal Google on steroids. We have seen the beginnings of this with Bing Chat for web browsing and tools like Quivr for personal data, but now imagine searching through your company's internal files with the same volition. It would be pretty incredible right?

> Find the provisional projections that me and Jeff came up with last week for Q2.

> What is fonts do we use for external product pitches?

This is made possible by using Retrieval Augmented Generation (RAG) techniques. RAG allows you to provide Language Models (LLMs) with external knowledge that they were not trained on.

When using Bedrock, whether you are ingesting data directly from S3 or storing it in vector databases like OpenSearch, Aurora, or RDS, the data remains within AWS data centres. This allows for greater security and easier compliance with relevant data governance requirements.



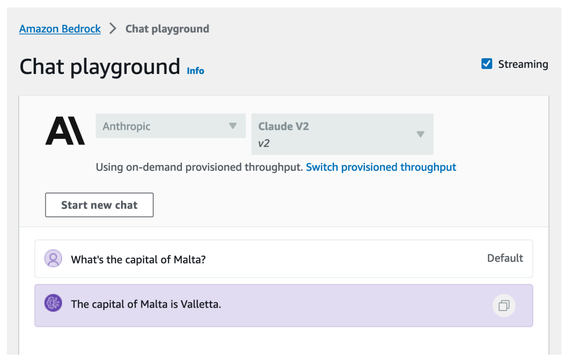
Basic RAG demo using Amazon Bedrock

Features

Amazon Bedrock's offering designed to be a full suite of tools to empower builders:

Easy Model Selection

Bedrock supports a selection of both proprietary and open-source models including Amazon's new Titan models. Depending on the task at hand, whether it's long form text generation, quick summarisation, or back and forth conversation, you will be able to find a model which meets your use-case.
Bedrock also offers a playground, allowing teams to try out various models and work out the best prompts for their chosen model.
The confirmed addition of Meta's Llama 2 model through a Serverless API is definitely a unique selling point of Bedrock. AWS recently partnered with Hugging Face and made a significant investment into their $235 million Series D funding round so it's a safe bet to expect more open-source models to be included with Bedrock in the coming months.
While Amazon are the first cloud provider to react to the need for model federation we are seeing advances in 3rd party libraries. Libraries like LiteLLM standardise calls to other model providers by exposing a common interface compatible with OpenAI.



Bedrock Chat Playground

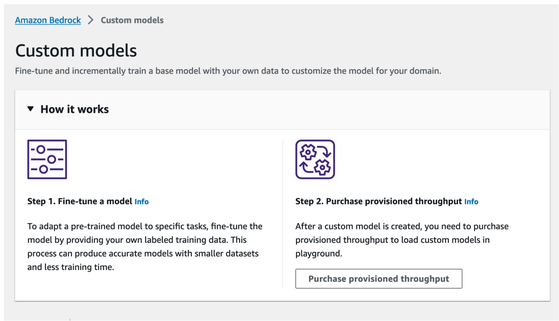
Agents

Autonomous agents for Bedrock are now in preview. Agents are capable of automating tasks normally done by humans; such as pitch deck creation, replying to emails or coding tasks.
Companies like Canva and Adobe have already integrated GenAI to resize images, remove backgrounds and objects, and it won't be long before we can also incorporate external style guides as context for these creations. With just a selection of notes, these tools will be able to create slides, fliers and other materials.
Code generation is also becoming easier with single shot accuracy possible for increasingly more complex use-cases. Progress in this area has been rocketing recently with AI code assistants, code generation through test driven development and autonomous code reviews becoming more common.

Although the output of agents may not be perfect, even at around 70% accuracy, it is a significant time saver for the human operator. The days of paying analysts substantial sums for mundane tasks like colour adjustments on slide decks may soon be seen as nostalgic.

Fine-Tuning

LLMs work well when they have been trained on the general context of the task and are able to rephrase it. If it has no knowledge of the underlying concept, for instance a particular part or process which only happens in your company, you may get better results from fine-tuning your own custom model. Users can fine-tune any of the Bedrock LLMs directly in the console, using data stored in S3.



Fine-tuning Custom Models

Bedrock over OpenAI?

Many of the arguments for using Bedrock are also applicable to OpenAI's platform. Both have a choice of models, a Serverless cost per token pricing model and support fine-tuning in the console.

However Bedrock supports models from a large variety of providers and has much clearer data governance policies. Bedrock also benefits from the huge AWS ecosystem, allowing for closer integrations with other services such as Lambda for compute, OpenSearch for vectors and S3 for object storage.

Pricing is also in favour of Bedrock: 1k tokens using Claude 2 through Bedrock will cost $0.01102 for input and $0.03268 for output whereas the closest offering from OpenAI (GPT4 32k context) will cost $0.06 for input and $0.12 for output.

There is a situation where an individual may opt for using OpenAI's GPT models. If they are particularly invested in their prompts or are making use of function calling where the LLM returns strictly JSON then sticking with OpenAI could be a good option. Otherwise switching to Bedrock is straight forward, especially if your application uses a library like LangChain which has a drop in replacement for Bedrock.

Conclusion

Amazon Bedrock is not just another AWS service; it's the platform that gives leaders confidence in how they are leveraging their data, whilst giving developers the tools to make their best applications without the managing the underlying infrastructure.

As we look towards the future, it's becoming increasingly clear: GenAI is not merely an add-on. It's a necessity, an integral component that will underpin the next generation of products and services. With Amazon Bedrock, the future of GenAI integration is not just possible; it's here.

Matt Carey

Developer at Aleios & AWS Community Builder

GenAI and Serverless London Meetup Organiser

With help from Aleios, Theodo & SICARA