Harness the Power of Generative AI in Healthcare with Amazon AI/ML Services

Sherry Ding
World-wide Public Sector
Amazon Web Services
Herndon, USA
xiaoydin@amazon.com

Veda Raman
World-wide Public Sector
Amazon Web Services
Herndon, USA
vedashrr@amazon.com

2024 IEEE 12th International Conference on Healthcare Informatics (ICHI) | 979-8-3503-8373-7/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHI61247.2024.00070

***Abstract*—As a transformative and innovative technology, generative AI enables us to solve really complex problems and re-imagine how we do things. There are big opportunities in how healthcare companies and organizations will use it to transform the whole industry and deliver amazing experience for their customers. As a leading cloud computing company with over twenty years innovation in machine learning (ML), AWS has developed a set of AI/ML services that allow healthcare companies and organizations to unleash the power of generative AI. In this paper, we will introduce AWS generative AI service stack, highlight some commonly used AWS AI/ML services in building generative AI applications for the healthcare field. We will discuss architectures of two popular applications in healthcare: chatbot and intelligent document processing (IDP), to showcase how different services work together in generative AI applications.**

***Index Terms*—Generative AI, AWS AI/ML services, Chatbot, intelligent document processing (IDP)**

I. Introduction

Healthcare and life science organizations are reinventing how they collaborate, make data-driven clinical and operational decisions, enable precision medicine, and decrease the cost of care. To help healthcare and life science organizations achieve business and technical goals, AWS for Healthcare and Life Sciences provides an offering of AWS services and AWS Partner solutions used by thousands of customers globally.

Additionally, since its inception, generative AI has been a disrupting and transformative technology. While healthcare has been using AI and ML for years, generative AI is bringing new possibilities to accelerate innovation, increase efficiencies and improve outcomes across healthcare. From generating new therapeutic candidates, to better matching patients with the right clinical trials, to powering patient engagement applications, AWS makes it easier to access the services, data, models, and secure infrastructure needed to scale generative AI across any healthcare organization.

Fig. 1 shows the AWS generative AI stack. On the bottom layer is the infrastructure available on AWS used for generative AI model (also called foundation model (FM)) training and running these models in production. The middle layer(Amazon Bedrock) provides API based access to leading pre-trained FMs from various providers. It also provides tools needed to build and scale generative AI applications. The top layer has applications available as services to customers. These applications are built leveraging the FMs in the bottom layers.

Fig. 1. Generative AI stack on AWS.

II. Key AI/ML Services for Generative AI

AWS provides a set of services with latest technologies to support your development of generative AI applications in healthcare. These services can be categorized into two groups: One group involves popular AI/ML services applying to all industries including healthcare, for instance, Amazon SageMaker, Amazon Bedrock, Amazon Q, and Amazon CodeWhisperer. The other group includes services such as AWS HealthScribe, AWS HealthLake, AWS HealthImaging, and AWS HealthOmics, which are created specifically for healthcare domain. In this paper, we focus on the AI/ML services for generative AI and how to architect them with some other services into generative AI applications for healthcare. We introduce three services: Amazon SageMaker, Amazon Bedrock, and Amazon Q. These three services allow you to leverage FMs in generative applications with different levels of controllability and flexibility.

A. Amazon SageMaker

Amazon SageMaker is a fully managed service with scalable infrastructure that brings together a broad set of tools to enable high-performance, low-cost ML for any use case. With SageMaker, users can build, train and deploy ML models

2575-2634/24/$31.00 ©2024 IEEE
DOI 10.1109/ICHI61247.2024.00070

at scale. SageMaker provides tools and features covering the whole ML lifecycle - all in one integrated development environment (IDE), from data preparation and model building to model training and model deployment, as well as ML operations (MLOps) like automation and monitoring. It also supports governance requirements with simplified access control and transparency over your ML projects. It provides a platform for data scientist, ML engineer, and business analysts to collaborate together by leveraging different features.

Among all features, SageMaker JumpStart is the ML Hub of SageMaker that consists hundreds of built-in content including algorithms, pre-trained models, and solutions templates. It allows users to easily use these contents through UI without writing any codes, or through APIs for programmatic access. SageMaker JumpStart includes a FM hub where users can deploy or fine-tune these FMs with the click of a button. These FMs also come with pre-built training and inference scripts that are compatible with SageMaker training and inference features so that users can further configure them to meet their requirements. We've partnered with well-known FM providers to offer a variety of models in SageMaker JumpStart.

SageMaker provides a platform for model builders and tuners to develop, train, and tune FMs, with or without using SageMaker JumpStart. Developers also can bring public available FMs to host on SageMaker, or deploy directly from SageMaker JumpStart, to serve in their generative AI applications.

B. Amazon Bedrock

Amazon Bedrock is a fully managed service that provides you an easiest way to build and scale generative AI applications. It offers a choice of high-performing FMs from leading AI companies like AI21 Labs, Anthropic, Cohere, Meta, Stability AI, and Amazon via a single API, along with a broad set of capabilities you need to build generative AI applications with security, privacy, and responsible AI. You can easily experiment with and evaluate top FMs for your use case, privately customize them with your data. With fine-tuning and continued pre-training, Bedrock makes a separate copy of the base FM that is accessible only by you, and your data is not used to train the original base models.

Bedrock allows you to extend the power of FMs with your data using techniques such as Retrieval Augmented Generation (RAG) by using Bedrock Knowledge Bases, and build Bedrock Agents that execute complex business tasks by dynamically invoking APIs. Since Amazon Bedrock is serverless, you don't have to manage any infrastructure, and you can securely integrate and deploy generative AI capabilities into your applications using the AWS services you are already familiar with.

C. Amazon Q

Amazon Q, announced at 2023 AWS ReInvent, is your generative AI–powered assistant designed for work that can be tailored to your business. It can help you get fast, relevant answers to pressing questions, solve problems, generate content, and take actions using the data and expertise found in your company's information repositories, code, and enterprise systems. Its areas of expertise involves in your business, building on AWS, Amazon QuickSight, Amazon Connect, and AWS Supply Chain.

Amazon Q Business can be tailored to your business by connecting it to company data, information, and systems, made simple with more than 40 built-in connectors. It is aware of which systems they can access, so users can ask detailed, nuanced questions and get tailored results that include only information they are authorized to see.

Amazon Q Developer is not only excellent at code generation and guidance, but also can do much more tasks such as scanning your code for hard-to-find vulnerabilities and suggesting remediations that improve the quality and security of the code. You can also upgrade applications with Amazon Q Code Transformation by automating the end-to-end process of upgrading code. Amazon Q's game-changing capabilities allow you to offload many time-consuming tasks and help you innovate faster.

III. Build Generative AI Applications for Healthcare Using AI/ML Services

Above mentioned ML and generative AI services often don't work solely. Instead, they work together with other AWS services in well-designed architectures for building generative AI applications, so as to achieve operational excellence, security, reliability, performance efficiency, cost optimization, and sustainability. In this part, we provide two examples of generative AI applications that have been widely adopted by healthcare companies and organizations.

A. ChatBot

Intelligent chatbots and conversation assistants has been the most implemented use case in all domains including healthcare. Generative AI powered chatbots not only improve patient engagement but also overcome staffing shortages for providers. Chatbots can also help administrators, clinicians and other staff get access to information quicker and hence help ease clinician and people burnout. They also make the staff more productive and giveback time for them to innovate in other areas. Fig. 2 shows the architecture of such a chatbot built using services like amazon Lex, Bedrock and Kendra. It uses the RAG technique for grounding the large language models responses and providing a reference back to the source of the information.

B. Intelligent Document Processing (IDP)

In healthcare field, documents such as doctor's notes, prescriptions, discharge summary, surgical reports, and medical insurance claims, etc, contain valuable information and come in various shapes and forms. Manually processing these documents to extract information and insights is time consuming, error prone, expensive, and difficult to scale. To help overcome these challenges, AWS IDP provides you choices when it

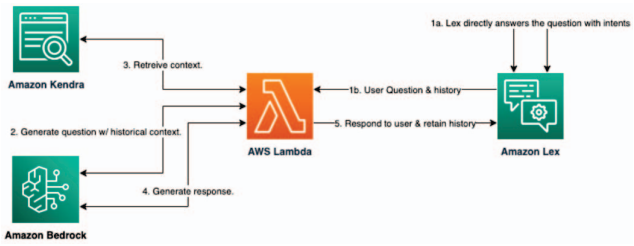

Fig. 2. Architecture of a Generative AI powered chatbot built on AWS.

comes to extracting information from complex content in any document format.

Generative AI complements IDP services like Amazon Textract and Amazon Comprehend to further automate document processing workflows. When you use FMs in each IDP stage, your workflow will be more streamlined and performance will improve. To achieve this, you can leverage FMs from above mentioned AI/ML services like SageMaker and Bedrock. A detailed architecture illustrating the IDP workflow with generative AI is in Fig. 3.

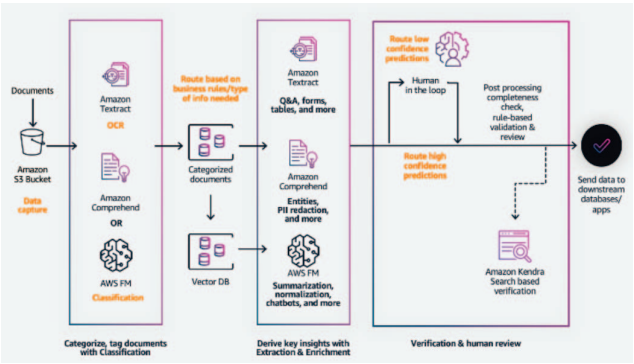

Fig. 3. General IDP workflow enhanced with generative AI on AWS.

In extraction stage of the IDP pipeline, when FMs can't directly process documents in their native formats (such as PDFs, img, jpeg, and tiff) as an input, you can use Textract to extract lines and words and then pass them to downstream FMs for further processing. The IDP pipeline can also be seamlessly automated using AWS serverless services. Serverless services help provide the mechanism to build a solution for IDP quickly. Services such as AWS Lambda, AWS Step Functions, and Amazon EventBridge can help build the document processing pipeline with integration of FMs, to achieve tasks such as medical document summarization.

IV. Conclusion

In this paper, we introduced how AWS brings generative AI technology into its cloud computing offerings by providing a comprehensive set of services and solutions. We highlighted some services that are specifically developed for healthcare use cases, and discussed two common solutions that have been widely applied in healthcare companies and organizations. This paper gives you ideas and insights about how you can harness the power of generative AI to innovate applications and tramsform business in healthcare industry on AWS.

References

[1] AWS generative AI for every business: https://aws.amazon.com/generative-ai/.
[2] AWS for Healthcare and Life Sciences: https://aws.amazon.com/health/.
[3] Intelligent Document Processing on AWS: https://aws.amazon.com/solutions/ai-ml/intelligent-document-processing/.
[4] QnABot on AWS: https://aws.amazon.com/solutions/implementations/qnabot-on-aws/.
[5] AWS solutions for Healthcare, Life Sciences, and Genomics: https://aws.amazon.com/solutions/health/.