Exploring Generative AI in conversational experiences: An Introduction with Amazon Lex, Langchain, and SageMaker Jumpstart

by Marcelo Silva, Kanjana Chandren, Justin Leto, Mahesh Biradar, Ryan Gomes, and Victor Rojo | on 08 JUN 2023 | in Amazon Lex, Amazon SageMaker, Amazon SageMaker JumpStart, Artificial Intelligence, Generative AI, Technical How-To | Permalink | Comments | Share

Customers expect quick and efficient service from businesses in today's fast-paced world. But providing excellent customer service can be significantly challenging when the volume of inquiries outpaces the human resources employed to address them. However, businesses can meet this challenge while providing personalized and efficient customer service with the advancements in generative artificial intelligence (generative AI) powered by large language models (LLMs).

Generative AI chatbots have gained notoriety for their ability to imitate human intellect. However, unlike task-oriented bots, these bots use LLMs for text analysis and content generation. LLMs are based on the Transformer architecture, a deep learning neural network introduced in June 2017 that can be trained on a massive corpus of unlabeled text. This approach creates a more human-like conversation experience and accommodates several topics.

As of this writing, companies of all sizes want to use this technology but need help figuring out where to start. If you are looking to get started with generative AI and the use of LLMs in conversational AI, this post is for you. We have included a sample project to quickly deploy an Amazon Lex bot that consumes a pre-trained open-source LLM. The code also includes the starting point to implement a custom memory manager. This mechanism allows an LLM to recall previous interactions to keep the conversation's context and pace. Finally, it's essential to highlight the importance of experimenting with fine-tuning prompts and LLM randomness and determinism parameters to obtain consistent results.

Solution overview

The solution integrates an Amazon Lex bot with a popular open-source LLM from Amazon SageMaker JumpStart, accessible through an Amazon SageMaker endpoint. We also use LangChain, a popular framework that simplifies LLM-powered applications. Finally, we use a QnABot to provide a user interface for our chatbot.
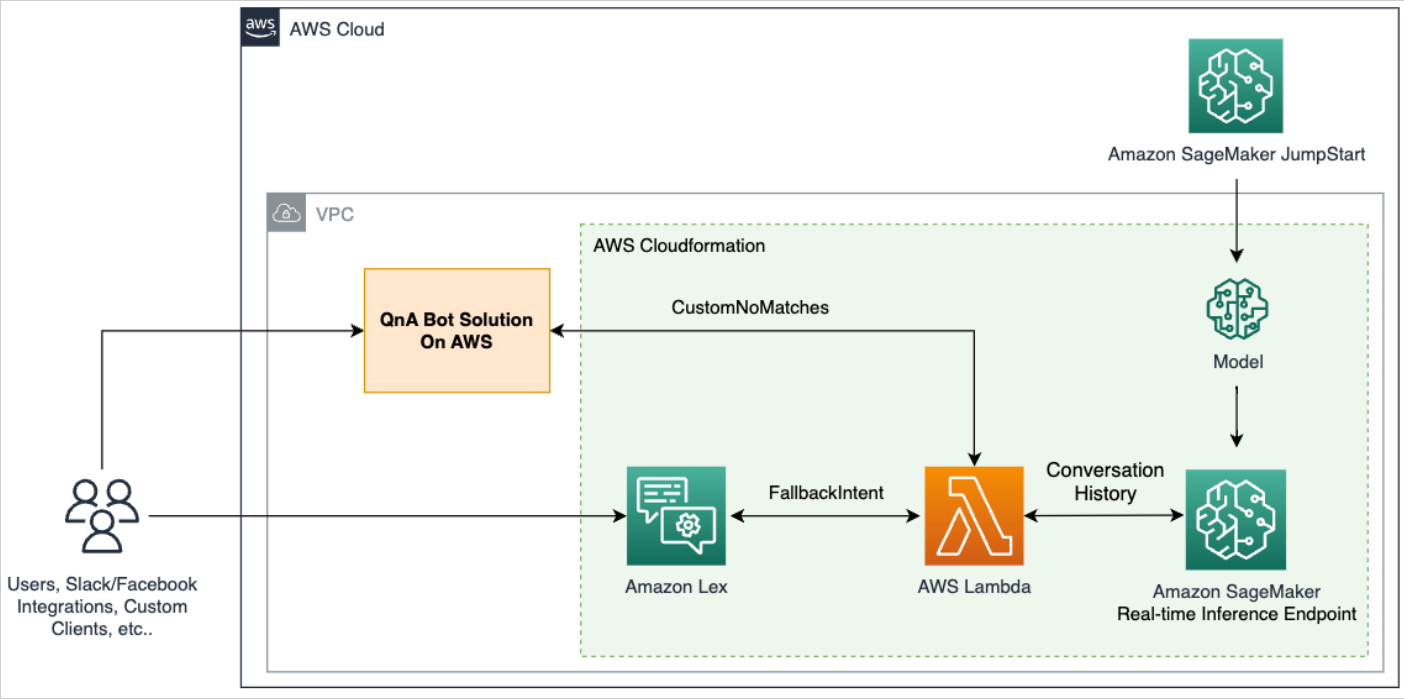
First, we start by describing each component in the preceding diagram: