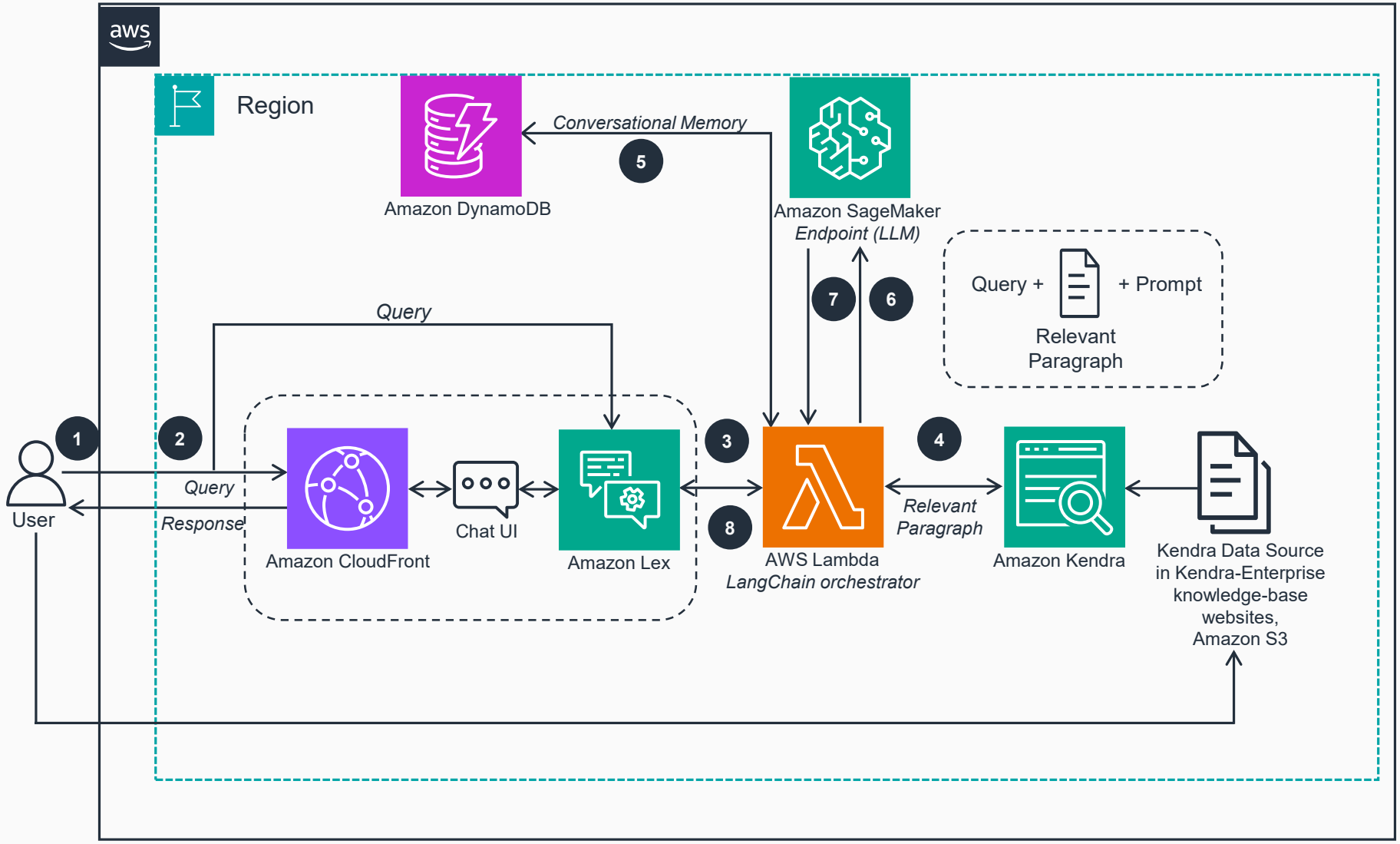
Guidance for Conversational Chatbots Using Retrieval Augmented Generation on AWS

This architecture diagram demonstrates how to implement a Retrieval Augmented Generation (RAG) workflow by combining the capabilities of Amazon Kendra with large language models (LLMs) to create generative artificial intelligence (AI) applications.



1. The user initiates a request. The request ingests content from the web, serving as a data source in **Amazon Kendra**, and an index ID is created in **Amazon Kendra**.
2. The user interacts with the **Amazon Lex** conversational chatbot using the **Amazon Lex** chat window or, optionally, through the **Amazon Lex** web user interface (UI), an open-source project, to submit a query or request. **Amazon Lex** is responsible for understanding and interpreting users' intent and extracting relevant information from the input.
3. **Amazon Lex** invokes the **AWS Lambda** function. This **Lambda** function handles user interactions. **Lambda** receives the request from the user either through the **Amazon Lex** UI, distributed by **Amazon CloudFront**, or an **Amazon Lex** chat window, and returns the responses.
4. Using **Amazon Kendra**, relevant passages are identified and extracted when there is a user request sent through the **Lambda** function. **Amazon Kendra** identifies and extracts the relevant passages when a request is submitted.
5. The conversation is also stored in **Amazon DynamoDB** to serve subsequent user requests, and is used for conversational memory.
6. A query, along with context from **Amazon Kendra**, is forwarded to the large language model (LLM) by the **Lambda** function with the help of a LangChain orchestrator, an open-source framework for developing applications powered by language models. The LLM processes this query and produces a response.
7. The generated response from the LLM is returned to the **Lambda** function.
8. The **Lambda** function formats and delivers the response back to the user through the **Amazon Lex** chat UI distributed by **CloudFront**, or through the **Amazon Lex** chat window in the console.

AWS Reference Architecture