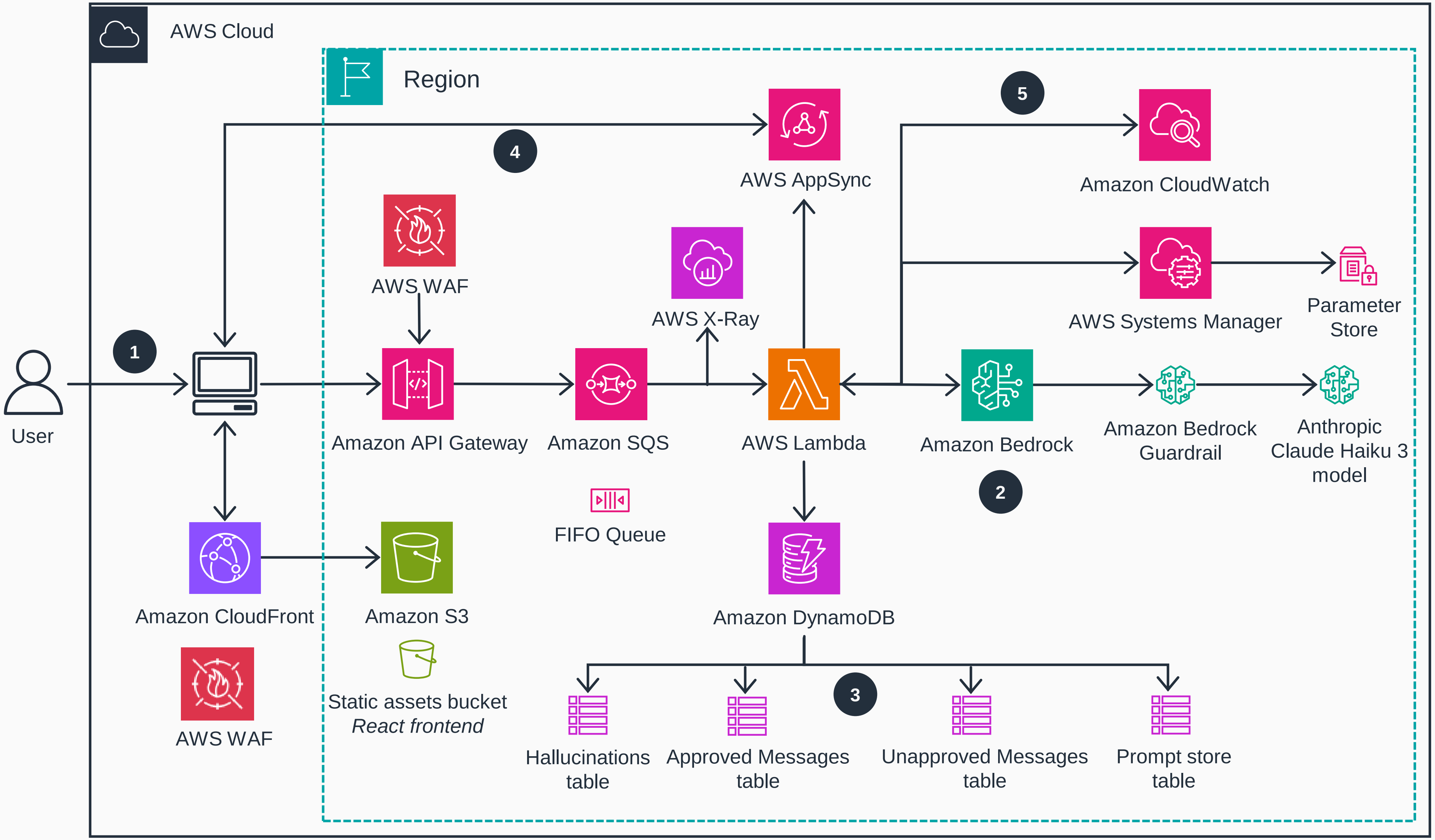
Guidance for Live Chat Content Moderation with Generative AI on AWS

This architecture diagram illustrates a real-time chat moderation system designed for live streaming platforms. It uses AWS services and generative artificial intelligence (AI) to automatically filter and moderate chat messages, creating a safer and more engaging environment for users.



1 **User interaction and message submission:** Users access a web application based on React that is served by **Amazon CloudFront** and **Amazon Simple Storage Service** (Amazon S3). Additional security from **AWS WAF** is used to block requests from potential threats. When a user sends a message, the application sends a POST request to **Amazon API Gateway**. **API Gateway** then routes the message to an **Amazon Simple Queue Service** (Amazon SQS) First-In-First-Out (FIFO) queue for processing.

2 **Message processing and AI moderation:** An **AWS Lambda** function, triggered by **Amazon SQS**, applies **Amazon Bedrock Guardrails** for initial content filtering. The message is analyzed by the Anthropic Claude Haiku model. The model evaluates the content based on the moderation guidelines and responds with either "y" for approved or "n" for rejected messages.

3 **Message handling and storage:** Based on the model's decision, the **Lambda** function routes the message accordingly. Approved messages are stored in the **Amazon DynamoDB** Approved Messages table, while rejected messages go to the Unapproved Messages table. Other messages are stored in a Hallucinations table.

4 **Real-time updates and notifications:** For approved messages, **AWS AppSync** broadcasts the content to all subscribed clients through WebSocket connections. For rejected messages, **AppSync** notifies only the original sender. This system maintains user privacy while enforcing moderation policies.

5 **Monitoring and observability: Amazon CloudWatch** and **AWS X-Ray** log metrics and trace requests, providing insights into message flow. This data is aggregated into a **CloudWatch** dashboard, offering visibility into the chat moderation system's operations.

Reviewed for technical accuracy February 26, 2025

AWS Reference Architecture