Objective

A client needs to answer customer reviews on a webpage. They have people doing this manually. They will continue doing this, but to use Gen AI to help them write a response according to customer comments and rating about a product

Currently they have a webpage like this one:

A screenshot of a computer

Description automatically generated

And a human will Respond the feedback.

Sometimes the feedback will be in a publica page, sometimes will be in a private page.

When it is in a private page we will know it. They want an LLM to propose answers based on the feedback, including client name. Not always the client will provide information.

The attached file is a Sample Data base.



They want to develop a button like the one below, that will connect to a Bedrock LLM and generate a response:

A screenshot of a chat

Description automatically generated

The future webpge will provide the option to ask modifiations to the proposed answer to the model:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Everything should be done by sending the data to an API, we will receive a json or a string of text from the API with the web page info, and we will send it back to the API page, and will receive another one with the requested modifications.

All this should be done with AWS Bedrock and maybe Aws Sagemaker too. All files sent/received will be stored in S3 buckets.

A nice to have is to have a stream response from bedrock, like this one:

<https://docs.aws.amazon.com/bedrock/latest/APIReference/API_runtime_InvokeModelWithResponseStream.html>

For the stream part, I asked a guy that did a similar project in the past:

we have one customer asking for a Streaming response from Bedrock. I have found there is a method that returns a stream called InvokeModelWithResponseStream (<https://docs.aws.amazon.com/bedrock/latest/APIReference/API_runtime_InvokeModelWithResponseStream.html>)  
Has anyone ever used it??? Any lessons learnt?

And this guy answered me this:

Yeah this outputs the response as streams. I have a function for Claude v2 with both options (regular invoke model and response stream) that I used in a previous project.If I’m not mistaken I was not streaming the output, even with invoke model with response stream, I was just getting the whole output at once. I implemented both options because I was facing some throttling exceptions, and when I saw those I switched to the other API call.Hope it helps.

