AWS Bedrock – deploy and consume a LLM

Go to AWS Console → Lambda → Create Function

- **Name**: bedrock-gateway

- **Runtime**: Python 3.10

- **Execution Role**: Create a new role with basic Lambda permissions

Paste following code:

```python
import json

import boto3


def lambda_handler(event, context):

    body = json.loads(event['body'])

    prompt = body.get("prompt", "")


    client = boto3.client("bedrock-runtime", region_name="us-east-1")


    response = client.invoke_model(

        modelId="anthropic.claude-3-sonnet-20240229-v1:0",

        contentType="application/json",

        accept="application/json",

        body=json.dumps({

            "prompt": f"\n\nHuman: {prompt}\n\nAssistant:",

            "max_tokens_to_sample": 300,

            "temperature": 0.7

        })

    )
```

```
result = response['body'].read().decode("utf-8")

return {

  "statusCode": 200,

  "headers": {"Content-Type": "application/json"},

  "body": result

}
```

Test Lambda in Console

1. Go to **"Test" tab**

2. Configure a new test event (name: testPrompt)

3. Paste this as the test input:

```
{

  "body": "{\"prompt\": \"What is quantum computing?\"}"

}
```

4. Click **Test**

You should see 200 status and Cllaude's answer in output

Create API Gateway to Expose Lambda as REST API

Go to AWS Console → API Gateway → Create API

- Choose **HTTP API**

- Integration: **Lambda Function** → bedrock-gateway

- Route: POST /generate

- Deploy → Name the stage: prod

Copy the **Invoke URL**, e.g.:

https://abc123xyz.execute-api.us-east-1.amazonaws.com/generate

est API Gateway from Console

1. Go to **API Gateway > Your API > Stages > prod**

2. Click **"Invoke URL"** → it will open in browser (it won't work directly, it's a POST API)

3. Go to **Postman** or use the **Test Client in Console**:

   o In **API Gateway,** go to **Routes > /generate**

   o Click **"Test"**

   o Set Method: POST

   o Headers:

      Content-Type: application/json

      Body:

      {

        "prompt": "Tell me about black holes."

      }

      Click **Send Test Request**

      🟢 You should get a 200 response with the Claude model's output.