

# AWS-GenAI-LLM-Chatbot-Guardrail

DESIGN AND IMPLEMENTATION OF TOPIC-BASED  
GUARDRAILS IN AMAZON BEDROCK FOR  
FINANCIAL ADVICE RISK MITIGATION AND TRACE-  
LEVEL MONITORING

# Prompt Execution & Guardrail Response

- Financial Advice Query Automatically Blocked by LLM Guardrail with Custom Denial Response
- (What's happening: User asks for financial advice → Guardrail intercepts → Model returns safe fallback response.)

The screenshot displays the Amazon Bedrock Guardrails console interface. The breadcrumb navigation at the top reads: Amazon Bedrock > Guardrails > FinancialAdvice > Trace.

**Prompt:** Give me Financial Advice?

**Model response:** -

**Guardrail action:** Intervened (1 instance) [View trace]

**Final response:** Sorry, the model cannot answer this question.

**FinancialAdvice:Working draft** [Interv]

**Content filters:**

Category	Test result	Details
Sexual	No action taken	Detected: FALSE Strength: High Confidence: None
Violence	No action taken	Detected: FALSE Strength: High Confidence: None
Hate	No action taken	Detected: FALSE Strength: High Confidence: None
Insults	No action taken	Detected: FALSE Strength: High Confidence: None
Misconduct	No action taken	Detected: FALSE Strength: High Confidence: None

# Guardrail Intervention Triggered

**Test**

Working draft

**Nova Pro 1.0**  
On-demand

☐ Use ApplyGuardrail API  
Assess any content using your pre-configured Amazon Bedrock Guardrails, without invoking the foundation models. [Learn more](#)

▼ Reference source - for contextual grounding check, optional [Info](#)  
Enter a reference source

**Prompt**  
Give me Financial Advice?

**Model response**

FinancialAdvice:Working draft Intervened (1 instance)

**Content filters**

Category	Test result	Details
Sexual	No action taken	Detected: FALSE Strength: High Confidence: None
Violence	No action taken	Detected: FALSE Strength: High Confidence: None
Hate	No action taken	Detected: FALSE Strength: High Confidence: None
Insults	No action taken	Detected: FALSE Strength: High Confidence: None

- **Real-Time Guardrail Intervention Detected and Logged (Policy Violation: Financial Advice)**
- (What's happening: System detects policy violation → Warning indicator shows 1 intervention instance.)

# Policy Trace & Topic-Level Enforcement

- **Denied Topic Detection Confirmed: FinancialAdvice Classified and Blocked via Guardrail Policy**
- (What's happening: Trace shows topic classification → FinancialAdvice → Blocked → Detection = True.)

The screenshot displays the Amazon Bedrock Guardrails Trace interface. The breadcrumb navigation at the top reads: Amazon Bedrock > Guardrails > FinancialAdvice > Trace. On the left, the 'Prompt' field contains 'Give me Financial Advice?' and the 'Model response' field is empty. On the right, a table lists guardrail topics: Hate, Insults, and Misconduct. Each row shows 'No action taken' and 'Detected: FALSE' with 'High' strength and 'None' confidence. Below this, a section titled 'Denied topics' is highlighted with a red box. It contains a table with the following data:

Topics	Test result	Details
FinancialAdvice	Blocked	Detected: TRUE

fin

