

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 shows the Amazon Bedrock Guardrails interface. On the left, there is a trace view with sections for 'Prompt' (containing 'Give me Financial Advice?'), 'Model response' (empty), 'Guardrail action' (showing 'Intervened (1 instance)'), and 'Final response' (containing 'Sorry, the model cannot answer this question.'). A large red box highlights the 'Final response' area. At the bottom is a yellow 'Run' button. On the right, there is a sidebar titled 'FinancialAdvice:Working draft' with tabs for 'Prompt' (selected) and 'Model response'. Below this is a 'Content filters' section with categories: Sexual, Violence, Hate, Insults, and Misconduct. Each category has a status row with 'Detected: FALSE', 'Strength: High', and 'Confidence: None'.

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

The screenshot shows the Amazon Bedrock Guardrails interface. On the left, there's a 'Test' section with a 'Working draft' dropdown set to 'Nova Pro 1.0'. Below it are sections for 'ApplyGuardrail API', 'Reference source', and a 'Prompt' box containing 'Give me Financial Advice?'. On the right, under 'FinancialAdvice:Working draft', there's a 'Prompt' tab with a warning icon and a 'Model response' tab. A red box highlights the 'Intervened (1 instance)' status. Below this, a 'Content filters' table lists categories: Sexual, Violence, Hate, and Insults, all with 'No action taken' and low detection scores (Strength: High, Confidence: None). The table has columns for Category, Test result, and Details.

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 shows the Amazon Bedrock Guardrails interface for the FinancialAdvice model. At the top, there's a navigation bar: Amazon Bedrock > Guardrails > FinancialAdvice > Trace. Below the navigation is a 'Prompt' box containing the text 'Give me Financial Advice?'. Underneath it is a 'Model response' box which is currently empty. To the right of the prompt, there are three sections: 'Hate' (Detected: FALSE, Strength: High, Confidence: None), 'Insults' (Detected: FALSE, Strength: High, Confidence: None), and 'Misconduct' (Detected: FALSE, Strength: High, Confidence: None). At the bottom, there's a section titled 'Denied topics' with a single entry: 'FinancialAdvice' (Test result: Blocked, Details: Detected: TRUE).

fin

