

Root Cause Analysis (RCA)

Retrieval-Augmented Generation (RAG) Application

Melting Face
1920 McKinney Avenue
Dallas, TX 75201

Date

Table of Contents

- 1. Introduction
- 2. Event Description
- 3. Chronology of Events / Timeline
- 4. Investigative Team and Method
- 5. Findings and Root Cause
- 6. Corrective Action

Introduction

The purpose of this Root Cause Analysis (RCA) is to determine the underlying causes of the recent system failure in the Retrieval-Augmented Generation (RAG) Application. The failure resulted in inaccurate responses being generated by the AI model due to an issue with external data retrieval. This analysis aims to identify what happened, how it happened, and why it occurred. The findings will be used to implement corrective actions to prevent future occurrences.

Event Description

On October 10, 20xx, at 2:45 PM, the RAG system failed to retrieve external knowledge from integrated knowledge bases, causing the AI model to hallucinate incorrect responses. The issue was first detected by a quality assurance engineer during a routine validation check. Users reported misleading information being generated, negatively impacting decision-making processes. This event affected multiple stakeholders, including product teams and end-users relying on the RAG system for accurate data.

Chronology of Events / Timeline

Time	Event
2:30 PM	RAG system initiates a query to external knowledge base.
2:35 PM	API response time exceeds expected thresholds, triggering latency alerts.

2:40 PM	System logs show repeated failed connection attempts to knowledge base API.
2:45 PM	Quality assurance engineer detects incorrect AI responses.
2:50 PM	Incident reported to the technical team for immediate investigation.
3:10 PM	Technical team confirms external API connectivity failure.
3:30 PM	Investigation reveals incorrect fallback mechanism causing AI to generate misleading responses.

Investigative Team and Method

The investigative team was assembled by the CTO to determine the root cause of the failure. The team included:

- Lead AI Engineer: Responsible for reviewing AI model outputs.
- API Integration Engineer: Investigated external API failures.
- Quality Assurance Lead: Verified test case failures.
- System Architect: Analyzed infrastructure and fallback mechanisms.

The investigation involved reviewing system logs, replicating API failures in a test environment, and conducting interviews with engineers who managed the API integrations. The Ishikawa (fishbone) diagram method was used to analyze potential causes, identifying data retrieval as the primary failure point.

Findings and Root Cause

The investigation led to the following key findings:

- The external API experienced intermittent downtime, leading to connection failures.
- The fallback mechanism designed to handle API failures did not activate as expected.
- Due to missing error-handling logic, the AI model generated responses based on outdated cached data.
- There was no real-time monitoring mechanism to detect API failures before incorrect responses were generated.

Root Cause: The failure was caused by a combination of API downtime, insufficient error-handling mechanisms, and an ineffective fallback strategy. The AI model lacked real-time validation to detect missing or incorrect retrieved data.

Corrective Action

To prevent recurrence, the following corrective actions will be taken:

- Implement a robust fallback mechanism to handle API failures and prevent hallucinated responses.
- Introduce a real-time monitoring system to detect and alert on API connectivity failures.
- Enhance error-handling logic to ensure the AI model does not rely on outdated or missing data.
- Regularly test API response times and implement auto-scaling solutions for improved reliability.
- Conduct quarterly audits on knowledge base integrations to ensure data accuracy.

Sponsor Acceptance

Approved by the Project Sponsor:

Date: _____

Corbin Imgrund
Melting Face