Meeting Intelligence Copilot (GenAI with RAG)

Owner: Saurabh Dubey (Senior Software Development Engineer)

Objective

Build a reliable, hallucination-safe GenAI system that allows users to upload meeting notes, retain long-term memory, and ask natural language questions with grounded answers derived only from their meeting data.

Use Cases

1. Meeting Notes Ingestion
2. Knowledge Retrieval using semantic search (RAG)

High-Level Architecture

User → API Gateway → Lambda → (S3, Bedrock, OpenSearch)

Upload Flow

User submits meeting notes → Lambda stores raw text in S3 → text is chunked → embeddings generated via Bedrock → vectors stored in OpenSearch.

Query Flow

User submits question → question embedded → vector similarity search → relevant chunks retrieved → LLM generates grounded answer.

Dry Run Example

Meeting: Operational Excellence Planning – Q2

Decision: Reduce P95 latency by 20% next quarter.

Query: What operational decision was made for next quarter?

Answer: Reduce P95 latency by 20% next quarter.

Background Concepts Clarity

Embeddings convert meaning into numbers. Vectors represent these embeddings.

Chunking ensures precision. Vector search retrieves meaning-based context.

RAG ensures hallucinations are prevented by grounding LLM responses in retrieved facts.

Hallucinations

Hallucinations occur when LLMs are asked to answer without context. This system prevents them by retrieving exact factual context before generation.