Scope, Approach, and Objectives Document

# 1. Objectives

## Business Objectives

State the strategic or operational goals driving the project.  
  
Examples:  
- Improve customer support through a GenAI-powered assistant.  
- Automate internal knowledge access via conversational search.  
- Accelerate report generation and data interpretation.

## Technical Objectives

Define specific, measurable AI-related goals.  
  
Examples:  
- Develop a retrieval-augmented generation (RAG) system.  
- Achieve < 2-second latency for user queries.  
- Integrate secure model serving into existing Databricks pipelines.

# 2. Scope

## In Scope

- Curate and prepare training and knowledge data using Unity Catalog.  
- Embed data using pre-trained embedding models and Databricks Vector Search.  
- Implement GenAI model using APIs (OpenAI, MosaicML, etc.) or fine-tuned LLMs.  
- Serve models with Databricks Model Serving.  
- Log model inputs, outputs, and user feedback with MLflow.  
- Enforce governance and access control using Unity Catalog and IAM.  
- Evaluate model quality and user satisfaction.

## Out of Scope

- Production-scale monitoring pipelines (future phase).  
- Fine-tuning LLMs from scratch on proprietary infrastructure.  
- Integration with downstream business systems (ERP, CRM).

# 3. Approach

## 3.1 Methodology

- Agile delivery with iterative model development and evaluation.  
- Human-centered design to ensure alignment with user expectations.  
- Incorporate prompt engineering and feedback loops early and often.

## 3.2 Databricks Services to Be Used

|  |  |
| --- | --- |
| Databricks Service | Purpose |
| Unity Catalog | Data discovery, access control, governance |
| Delta Lake | Store and manage structured/unstructured data with ACID transactions |
| Databricks Vector Search | Semantic search over embedded documents or data |
| Model Serving | Real-time serving of LLMs and other generative models |
| MLflow | Track experiments, log prompts/outputs, register models |
| Databricks Notebooks | Rapid development, prototyping, and testing of GenAI workflows |
| Lakehouse AI | Framework for unifying LLMs with lakehouse data |
| Databricks Workflows | Schedule and automate data/model pipelines |
| Inference Tables (Preview) | Log and analyze model inference across prompts and outputs |

## 3.3 Phased Approach

* Phase 1: Discovery: Define personas, use cases, data sources
* Phase 2: Design: Build prompts, design RAG pipeline, select models
* Phase 3: Build: Data prep, embeddings, vector search, model integration
* Phase 4: Evaluate: User testing, model performance, prompt tuning
* Phase 5: Deliver: Deploy with Model Serving, log inferences with MLflow

## 3.4 Evaluation Metrics

|  |  |
| --- | --- |
| Metric | Target/Threshold |
| Response latency | < 2 seconds |
| Answer accuracy (manual review) | > 85% |
| User satisfaction score | > 4.0/5 |
| Feedback integration time | < 2 sprint cycles |
| Model hallucination rate | < 10% |