

Project Charter

Local LLM Platform with Retrieval-Augmented Generation (RAG)

Prepared for: James Keller, CISO

Prepared by: Jim Keller II, MBA

Table of Contents

Executive Summary	3
Project Purpose	3
Business Case	3
Business Objectives	4
Project Description	4
Project Objectives and Success Criteria (SMART)	4
Requirements	4
Constraints	5
Assumptions	5
Preliminary Scope Statement	5
Risks and Mitigations	5
Project Deliverables	5
Post-MVP Strategic Options (Future State Pathways)	5
Option A: Scale to a Full Production System	6
Option B: Transition to a Larger Custom Model or Multi-Model Architecture	6
Option C: Hybrid Strategy (Local + Cloud)	6
Summary Milestone Schedule	6
Summary Budget	6
Stakeholders and Approval Requirements	6
Project Approval Requirements	7
Project Manager	7
Authorization	7

Project Charter

Local LLM Platform with Retrieval-Augmented Generation (RAG)

Executive Summary

Organizations and individuals increasingly require secure, private, on-premise alternatives to cloud based AI platforms. This project aims to evaluate and demonstrate the value, feasibility, and security of deploying a fully on-premise local large language model (LLM) platform enhanced with retrieval-augmented generation (RAG). This minimum viable product (MVP) addresses challenges such as confidentiality, governance, compliance, external infrastructure dependency, and cost unpredictability.

The MVP will ingest up to 100GB of internal documents and provide real time, contextually relevant generative text assistance to leadership and high trust users. Deployed entirely on the firm's local hardware (e.g. high-performance servers such as the AMD GMK-Tech server or equivalent) the platform ensures that no data leaves the secured environment.

This charter is structured in alignment with the PMBOK® protocols and practices, ensuring that the project adheres to industry-standard project management practices.

Project Purpose

The purpose of this project is to provide organizations and individuals with a secure on-premise generative AI platform that enables reliable internal decision support while ensuring full data sovereignty. Cloud based AI platforms pose confidentiality and compliance risks. By deploying an on-premise LLM with RAG, organizations and individuals gain private access to AI assisted insights using approved data sources.

This initiative supports long-term strategies in cybersecurity, privacy, governance compliance, and responsible AI adoption.

Business Case

The need for this solution stems from the increasing demand among organizations and individuals for:

- Real-time access to internal business knowledge.
- Testing and validating AI capabilities without exposing sensitive data.
- Cost-predictable, dependency-free AI infrastructure.
- Solutions that meet regulatory and compliance constraints.

Expected benefits include improved data assisted decision making, reduced time spent on knowledge retrieval, and the creation of a secure foundation for future AI-driven capabilities.

Business Objectives

- Enable private, secure, governed generative AI.
- Demonstrate on-premise viability for sensitive workflows.
- Provide value validation through an executive demonstration.
- Increase AI literacy and adoption.

Project Description

The project will deploy a local LLM (8–14B parameter model) running entirely on internal, high-performance hardware such as the AMD GMK-Tech server or an equivalent GPU-enabled system. The RAG architecture will include:

- Ollama for LLM serving.
- LlamaIndex for orchestration.
- ChromaDB is the vector store.
- SQLite for metadata and document tracking.
- A secure browser based internal user interface.

The system will ingest up to 100 GB of internal documents and provide contextual responses using approved (and curated) datasets and documents.

Project Objectives and Success Criteria (SMART)

- Deploy the MVP in 8 weeks.
- Support ≥ 3 concurrent authenticated users.
- Ingest ≤ 100 GB of documents.
- Achieve ≤ 3 -second response time.
- Maintain $\geq 95\%$ contextual accuracy.
- Receive positive feedback during leadership demonstration.

Requirements

Functional Requirements:

- Provide generative responses grounded in internal data.
- Use LlamaIndex for orchestration.
- Enable authenticated, logged sessions.
- Internal browser-based user interface.

Non-Functional Requirements:

- ≤ 3 -second response time.
- Full offline operation after setup.
- Compliance with cybersecurity and governance policies.

Constraints

- Fully on-premise deployment.
- 8-week delivery timeline.
- Hardware limitations (RAM/CPU/Storage).
- MVP only (no enterprise-wide rollout).

Assumptions

- Approved internal documents will be provided.
- Hardware will be available on schedule.
- Stakeholders will participate in UAT and demonstration.
- Open-source components remain stable.

Preliminary Scope Statement

In Scope: LLM setup, RAG deployment, ingestion, UI, authentication, UAT, demo.

Out of Scope: Enterprise deployment, mobile apps, cloud integration.

Risks and Mitigations

- Hardware limitations → model quantization.
- Insufficient data → curated document sets.
- Security vulnerabilities → cyber review.
- Low adoption → structured demos and briefings.

Project Deliverables

- Project Management Plan.
- Configured hardware and environment.
- RAG pipeline (LLM + embeddings + vector DB).
- Secure ingestion pipeline.
- Web UI.
- UAT results.
- Executive demo package.
- Go/No-Go report.

Post-MVP Strategic Options (Future State Pathways)

Following successful MVP validation, the organization will determine if and how the “Local LLM Platform” should be expanded. Leadership may choose among several strategic paths:

Option A: Scale to a Full Production System

If the MVP demonstrates clear value, executives may approve:

- Expanding LLM capabilities across additional use cases or business units.
- Hardening the system for high availability, reliability, cyber resilience, and security.
- Integrating the platform with enterprise systems (CRM, ERP, ticketing, HRIS, data lakes).
- Investing in MLOps/DevOps pipelines for model lifecycle management and continuous improvement.

Option B: Transition to a Larger Custom Model or Multi-Model Architecture

If the MVP proves valuable but constrained, leadership may consider:

- Training a domain-specific model using internal organizational data.
- Upgrading to a more capable open-source foundation model with improved reasoning or context windows.
- Deploying multiple specialized models (e.g., routing models, summarization models, reasoning models, and vector-based retrieval models).

Option C: Hybrid Strategy (Local + Cloud)

If performance, compute cost, or model complexity require it, leadership may pursue a hybrid approach:

- Keeping sensitive workloads and data processing on local, isolated infrastructure.
- Utilizing cloud LLMs for non-sensitive or high-complexity workloads.
- Establishing a governance policy defining when workloads are executed locally vs. in the cloud.

Summary Milestone Schedule

Weeks 1–8: Charter, PMP, Cyber Review, Procurement, QA/SQA, Hardware Setup, Deployment, Ingestion, UAT, Demo, Go/No-Go.

Summary Budget

- Hardware: \$2,500
- Software: \$0
- Labor: \$8,000
- Contingency: \$800
- *Total: \$11,300*

Stakeholders and Approval Requirements

Role	Individual	Responsibility
Project Sponsor	James Keller, CISO	Charter and funding approval.
Project Manager	Jim Keller II	Planning and execution.
Technical Lead	Jim Keller II	System engineering and deployment.
Security Officer	James Keller, IT Security	Cybersecurity compliance.
Cyber Lead	William Keller, MS	Oversee cybersecurity architecture, risk assessments, and security controls implementation.
Procurement Lead	James Keller, CISO	Manage vendor selection, purchasing, licensing, and contract coordination.
QA Lead	Matthew Keller	Plan and oversee quality assurance testing to ensure functional correctness.
SQA Lead	William Keller, MS	Define and enforce software quality standards, processes, and documentation.
End Users	C-suite, NTK Executives	Provide UAT insight and feedback.

Project Approval Requirements

- On-premise deployment completed.
- UAT confirms performance and compliance.
- Leadership demonstration validates feasibility.
- Sponsor authorizes Go/No-Go decision.

Project Manager

Jim Keller II, MBA – Responsible for planning, execution, risk management, and deployment.

Authorization

Role	Name	Signature	Date
Project Sponsor	James Keller, CISO	_____	_____
Project Manager	Jim Keller II, MBA	_____	_____