Business Case: Implementation of a Local LLM Platform MVP with Retrieval-Augmented Generation (RAG)

1. Executive Summary

The organization proposes the development of a minimum viable product (MVP) for a local large language model (LLM) platform enhanced with retrieval-augmented generation (RAG). This initiative enables secure, cost-effective exploration of AI-driven knowledge assistance while maintaining data privacy, compliance, and control.

Deployed entirely within the company's secure infrastructure, the MVP will demonstrate how local generative AI can deliver real-time, context-specific insights for C-suite and need-to-know users, without depending on external cloud services. The system will be capable of supporting at least three concurrent users, validating both scalability and user experience in a controlled environment.

2. Problem Statement

While cloud-based AI platforms provide convenience, they raise major concerns about:

- Data confidentiality and third-party access.
- Vendor lock-in and unpredictable operational costs.
- Latency and dependency on external networks.

A local LLM platform with RAG addresses these challenges by ensuring that sensitive data remains on-premises, leveraging internal infrastructure to deliver AI capabilities safely and efficiently.

3. Proposed Solution

The MVP will:

- Host an 8–14B parameter model on secure, high-performance hardware such as the ADM GMK-Tech workstation or similar systems.
- Constructed on a fully open-source, locally hosted stack consisting of Ollama (for LLM inference), LangChain/LlamaIndex for orchestration, and Chroma DB as the vector store.
- Ingest and index up to 100 GB of confidential internal documents in a secure vector database, synchronized to a sequestered tracking database (i.e., to avoid duplicates, etc.).
- Provide a web browser chat interface, which is accessible through internal networks.
- Support at least three concurrent users with secure, authenticated sessions.
- Maintain full compliance with internal data governance standards.

4. Strategic Alignment

This project aligns with strategic goals by:

- Strengthening data protection and reducing external dependencies.
- Improving data-driven decision making by enabling AI systems to securely access and analyze institutional knowledge or proprietary data that would otherwise be prohibited.
- Showcasing an on-premises solution capable of supporting multiple simultaneous users.
- Demonstrating innovation in responsible AI adoption.

5. Expected Benefits

Tangible Benefits:

- Cost savings via open-source software (such as Ollama, LangChain, Chroma DB).
- Enhanced data privacy, regulatory compliance, and decreased risk exposure.
- Verified multi-user performance, supporting three or more concurrent users in real time.

Intangible Benefits:

- Increased executive confidence in internal AI capabilities.
- Faster knowledge access for strategic decision support.
- Improved governance through secure, traceable operations.

6. Technical Overview

Category	***Description***
Hardware	Workstations such as the ADM GMK-Tech or equivalent, equipped with GPUs and sufficient local storage to host 8–14B parameter models and support three or more concurrent users.
Software	Open-source tools such as Ollama, LangChain, and Chroma DB; lightweight frameworks such as FastAPI or Streamlit for user interface.
Security Controls	Comprehensive audit logging and user authentication.
Deployment Model	Fully on-premises network segment, isolated from external services, designed for concurrent user operation.

7. Cost and Resource Summary

Category	*Estimate*
Hardware (such as ADM GMK-Tech)	$2,500
Software (open source)	$0
Labor (80 hours × $100/hr)	$8,000
Contingency (10%)	$800
Total Cost	*$11,300*

8. Implementation Plan

Milestone	*Target Period*	*Notes*
Charter Approval	Week 1	Project sponsor authorization.
Project Management Plan	Week 2-3	Defines scope, schedule, resources, communications, and governance for the project.
Cyber Review	Week 4	Ensures cyber risk assessment, security controls, and compliance requirements are validated before project execution.
Procurement	Week 5-7	Confirms acquisition or verification of required hardware, software, and licenses.
Quality Assurance & Software Quality Assurance Review	Week 7	Confirms project readiness, documentation completeness, and pre-implementations quality standards prior to execution.
Hardware Setup	Week 8	Configure server capable of supporting ≥ 3 concurrent users.
MVP Deployment (LLM + RAG)	Week 8	Implement and test core RAG pipeline.
Document Ingestion	Week 8	Securely ingest and index up to 100 GB of internal content.
User Acceptance Testing (UAT)	Week 9-10	Validate performance and functionality for multiple users.

Executive Demonstration	Week 11	Showcase capabilities to C-suite and Need-to-Know users.
Go/No-Go Decision	Week 11-13	Evaluate pilot outcomes and determine production readiness.

9. Risks and Mitigation

Risk	***Mitigation***
Performance constraints on limited hardware	Optimize model quantization and GPU utilization; limit active sessions to 3 users.
Incomplete data coverage	Curate a representative subset of internal documents.
Model inaccuracies (hallucination)	Integrate retrieval grounding and confidence scoring.
Security exposure	Enforce controls.
Low user engagement	Conduct structured demonstrations and executive briefings.

10. Evaluation and Success Criteria

- The MVP must support at least three concurrent users with stable performance.
- Average response time ≤ 3 seconds.
- ≥ 95% accuracy for contextual responses based on indexed content.
- 100% adherence to internal security and compliance standards.
- Majority of executive participants confirm perceived value and usability.

11. Conclusion and Recommendation

The proposed MVP provides a secure, low-cost, and high-value pathway to validate the benefits of local generative AI for the organization. Leveraging open-source components and in-house hardware, development of the MVP will safely evaluate scalability and real-world application potential before investing in production deployment.

Recommendation: Approve and proceed with the 13-week MVP implementation, followed by evaluation and decision on enterprise rollout.