Project Brief: Retrieval-Augmented Generation (RAG) for CBP Knowledge Systems

# 1) Project Objective

This project aims to integrate a Retrieval-Augmented Generation (RAG) pipeline into CBP’s new large language model (LLM) deployment. The goal is to ensure the LLM provides grounded, accurate, and policy-compliant answers by connecting it directly with CBP’s structured and unstructured data sources.

# 2) Problem Statement

While the new LLM offers significant improvements in natural language understanding, it does not inherently have access to CBP’s latest operational knowledge, documents, or collaboration systems (e.g., JIRA, Confluence, MS Teams). This leads to three key issues:  
• Answers may be outdated or incomplete without real-time context.  
• Sensitive or restricted information may be mishandled if not tied to policy-aware filters.  
• The LLM cannot explain or trace its answers back to authoritative CBP documents without an external retrieval layer.

# 3) Proposed Solution: RAG

Retrieval-Augmented Generation (RAG) solves these issues by coupling the LLM with a vector database and retrieval pipeline. Instead of relying only on its internal training, the LLM will:  
• Retrieve the most relevant CBP-approved documents, notes, or records in real-time.  
• Use those retrieved snippets as context for generating accurate and policy-compliant answers.  
• Provide citations (document IDs, project numbers) alongside answers for auditability.  
  
Purpose: Ensure every LLM answer is grounded in CBP data, up-to-date, and traceable to a source of record.

# 4) Plan of Attack

Phase 1 – Data Intake & Preparation:  
• Parse and segment CBP data sources (PDFs, JIRA issues, Confluence pages, Teams transcripts).  
• Tag chunks with metadata (project, sprint, classification).  
  
Phase 2 – Embeddings & Indexing:  
• Generate dense embeddings for text and entities using self-hosted models.  
• Store in a vector database (pgvector or Qdrant) with policy-aware metadata filters.  
  
Phase 3 – Retrieval & LLM Integration:  
• Deploy a retrieval pipeline that encodes queries, searches the vector DB, applies ABAC filters, and passes top results to the LLM.  
• Ensure responses include citations and stay within compliance boundaries.  
  
Phase 4 – Governance & Evaluation:  
• Monitor retrieval quality (Recall@10, latency, operator acceptance).  
• Audit with OpenTelemetry and SIEM integration.  
• Conduct red-teaming and drift evaluations.  
  
Timeline: Initial pilot within weeks, with incremental scaling to production-ready deployment.

# 5) Strategic Impact

By integrating RAG, CBP ensures its new LLM deployment is not only powerful but also compliant, explainable, and mission-aligned. This approach will allow analysts and operators to ask natural language questions and receive answers directly grounded in CBP systems of record, dramatically improving both accuracy and trust. This project bridges cutting-edge AI with operational reliability and governance.

# 5) Inclusion of Code and Technical Sources

In addition to operational documents and collaboration systems, this project will extend retrieval to CBP’s technical and developer knowledge assets. This includes SQL queries, Python and Java code, JavaScript and HTML files, as well as mobile application code. By embedding and indexing these artifacts, the RAG system will enable semantic search across CBP’s software repositories and technical documentation.  
  
Purpose: Ensure the LLM can answer not only policy and operational questions but also provide context on technical systems, application behavior, and data pipelines — grounded in actual CBP codebases.

# 6) Strategic Impact

By integrating RAG, CBP ensures its new LLM deployment is not only powerful but also compliant, explainable, and mission-aligned. This approach will allow analysts, operators, and technical staff to ask natural language questions and receive answers directly grounded in CBP systems of record, including both operational documents and code repositories. The result is improved accuracy, trust, and efficiency across both mission and technical domains.