Interim Report:

Task 1 - EDA and Preprocessing

Exploratory Data Analysis

The CFPB Consumer Complaint Database was analyzed to understand its structure and suitability for the RAG system. The dataset contains columns such as Product, Consumer complaint narrative, Issue, Company, and Date received. Key findings include:

- **Complaint Distribution**: The dataset includes various financial products, with Credit Card and Checking or Savings Account having the highest complaint volumes, reflecting their widespread use. Money Transfers and BNPL-relatedwerpen

Task 2: Text Chunking, Embedding, and Vector Store Indexing

- **Chunking Strategy**
- Narratives were split into chunks of 500 characters with a 50-character overlap using LangChain's RecursiveCharacterTextSplitter. This size ensures semantic coherence while keeping embeddings manageable, as longer texts can dilute meaning in vector representations. The overlap preserves context across chunk boundaries, critical for complaints with sequential details. The separators parameter prioritizes natural breaks (paragraphs, sentences) for meaningful splits.
- **Embedding Model Choice**
- We selected sentence-transformers/all-MiniLM-L6-v2 for its balance of efficiency and performance. This model generates 384-dimensional embeddings, suitable for semantic similarity tasks on complaint narratives. It is lightweight, enabling fast processing of large datasets, and performs well on short, unstructured texts. Larger models like all-mpnet-base-v2 were considered but deemed too resource-intensive for CrediTrust's scale.
- **Vector Store**
- The embeddings were indexed in a FAISS IndexFlatL2 for exact nearest-neighbor search, appropriate for our dataset size (500,000+ complaints). Each chunk's metadata (chunk ID, complaint ID, product, text) was stored to enable traceability during retrieval. The vector store is persisted in vector_store/ as faiss_index.bin and metadata.pkl.