Phase 1–2: Foundations (Human-Friendly Guide)

This document explains Phase 1–2 of the assignment in a simple, human, and practical way. Think of this phase as building the *mental model* before writing real code. If Phase 3 onwards is construction, Phase 1–2 is understanding the blueprint.

1. What Problem Are We Actually Solving?
We want to build a system that can read large PDF manuals (with text and images) and answer questions intelligently. Normal chatbots fail because they don't know the document. RAG fixes this by first *retrieving* relevant content and then *generating* an answer.

2. What is RAG in Simple Terms?
Retrieval-Augmented Generation (RAG) is a two-step thinking process. First, the system searches for the most relevant information from stored knowledge. Second, it sends that information to an LLM to generate a clear answer. This prevents hallucination and keeps answers grounded in real documents.

3. Why Embeddings Exist
Computers cannot understand raw text or images directly. Embeddings convert text and images into numbers (vectors) that capture meaning. Similar meanings produce similar vectors. This is the backbone of semantic search.

4. Text Embeddings Explained Like a Human
When you read two sentences like 'How to replace a motor belt' and 'Steps for changing the belt in a motor', you know they mean the same thing. Text embeddings help computers do exactly that by placing similar sentences close together in vector space.

5. Image Embeddings (Why Images Matter)
Manuals often explain things using diagrams. Image embeddings allow the system to understand that a wiring diagram or machine sketch is related to certain text explanations. Models like CLIP learn how images and text relate to each other.

6. What is a Vector Database and Why We Need It?
A vector database is like a super-fast memory designed for similarity search. Instead of searching keywords, it searches meaning. When a user asks a question, the question is converted into a vector and matched against stored vectors to find the closest content.

7. Why Milvus?
Milvus is designed for large-scale vector search. It can handle millions of vectors efficiently, supports filtering with metadata, and works well in production environments.

8. Metadata is as Important as Embeddings
Embeddings alone are not enough. Metadata like page number, section name, document source, and content type helps refine search results and build accurate answers.

9. How Everything Connects (Mental Flow)
PDF → Extract text/images → Convert to embeddings → Store in vector DB → User asks question → Convert question to embedding → Retrieve relevant chunks → Send context to LLM → Generate answer.

10. Success Criteria for Phase 1–2
By the end of this phase, you should clearly understand: what RAG is, why embeddings exist, how text and images become vectors, and why a vector database is required. You should be confident enough to explain the system without touching code.