Milestone 1: Problem Definition & Literature Review

1. Problem Definition & Objectives

The objective of our project is to build a **Multimodal AI Medical Assistant** that integrates **speech, vision, and text** data to support healthcare practitioners and patients.

- **Speech:** Transcribe patient descriptions, extract symptoms, urgency, and emotional tone.
- **Text:** Process doctor's notes, prescriptions, and medical literature using a domain-specific LLM.
- **Vision:** Analyze medical images (X-rays, MRI, CT scans) to detect potential abnormalities. (extension to the first phase of the project)

- **Fusion:** Combine all three (or two) modalities into a unified representation that enables reasoning.

- **Output:**
  - **For doctors:** Structured, detailed diagnostic summary.
  - **For patients:** Simplified explanation in plain English.

Objective Summary: Provide an AI assistant that reduces diagnostic time, supports clinical decision-making, and improves patient understanding.

---

2. Literature Review

Existing Solutions:

- **Speech-to-Text & Symptom Extraction:** Tools such as the *Google Medical Speech-to-Text API* and *ClinicalBERT* can transcribe and extract symptoms from spoken inputs.
- **Medical Imaging (CV):** CNN and Transformer-based models (e.g., *CheXNet*, *MedViT*) achieve state-of-the-art performance in X-ray and MRI classification tasks.
- **Medical Text (LLMs):** Models such as *BioBERT*, *PubMedBERT*, and *MedPaLM* are fine-tuned on medical corpora to provide clinical reasoning.
- **Multimodal Models:** Recent research (e.g., *BioViL*, *MedCLIP*) explores combining vision and text but often excludes patient speech as an input.

Baselines & Benchmarks:

- **CheXpert** and **MIMIC-CXR** datasets for imaging benchmarks.
- **i2b2** clinical NLP dataset for text-based symptom/diagnosis extraction.
- **Wav2Vec2 + ClinicalBERT** pipelines for speech-to-symptom baselines.

---

3. Gaps & Opportunities

- Most existing systems are **unimodal** (focus on speech *or* vision *or* text), not integrated.

- Few models provide **patient-friendly explanations**; most are designed only for doctors.

- Current multimodal approaches do not effectively combine **speech, vision, and medical text** into a unified diagnostic reasoning framework.

- Opportunity to design a system that:
  - Uses **multimodal embeddings** for richer diagnostic context.
  - Delivers **dual outputs** (doctor-oriented & patient-oriented).
  - Improves **accessibility** in telemedicine and resource-limited settings.

---

4. Datasets

- The datasets that will be used for this project will be from open source because manual data collection for our objective will take more time and resources.

- We will try to use Hugging Face and Kaggle datasets as much as possible, we are also looking into some datasets that have been released by some universities.

- We assure that all datasets will be from open source which are under MIT or Apache 2.0 License.