🌸 FertiliCare - AI-Powered Fertility Assistant

An intelligent, multimodal medical assistant designed to help individuals understand their fertility test results and navigate their reproductive health journey with empathy and evidence-based information.

Overview

FertiliCare combines state-of-the-art AI models to provide:

- **Medical Image Analysis** - Extract and interpret hormone panel results from photos
- **Conversational AI** - Answer questions about fertility with empathy
- **Voice Interaction** - Support for audio questions and responses
- **Knowledge Retrieval** - GraphRAG-powered medical context from trusted sources

Key Features

1. Multimodal Input

- **Text**: Type your questions naturally
- **Audio**: Record questions via microphone or upload audio files
- **Image**: Upload photos of medical test results (hormone panels, ultrasounds)

2. Intelligent Analysis

- **Vision-Language Model**: Florence-2-base for OCR and medical image understanding
- **Large Language Model**: Qwen2.5-3B-Instruct for contextual responses
- **Retrieval System**: GraphRAG with vector search for accurate medical information

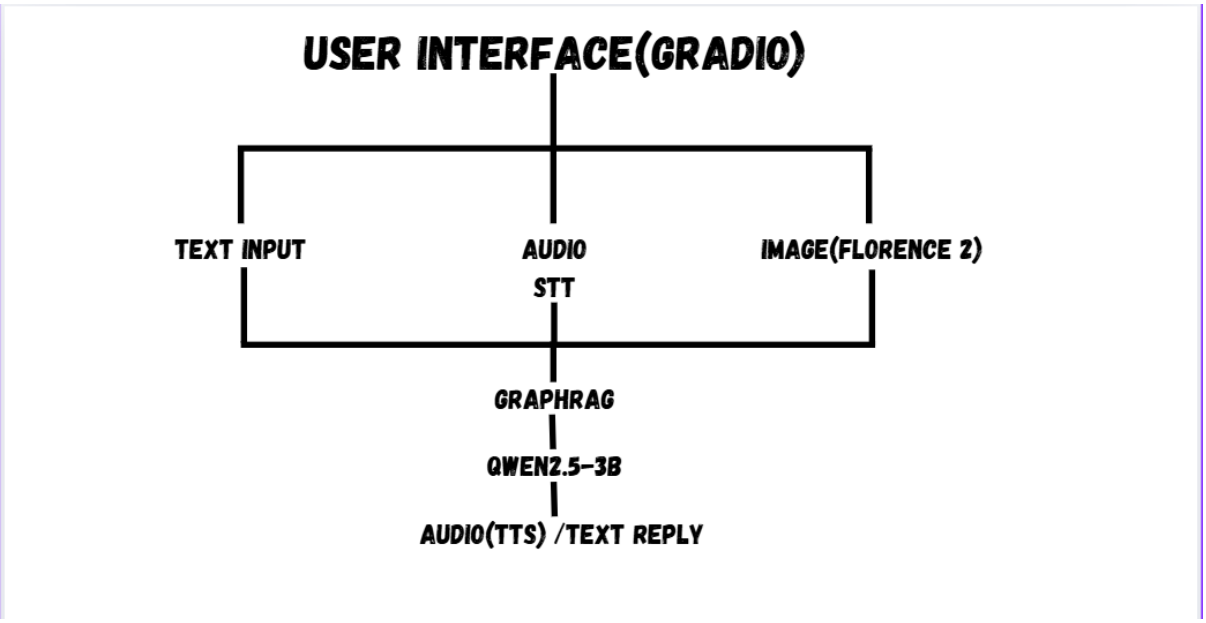
3. Memory-Optimized Architecture

- Sequential model loading (one at a time)
- Automatic memory cleanup after each inference
- Supports systems with 8GB+ GPU memory or CPU-only mode

4. Safety & Ethics

- Always includes medical disclaimers
- Never provides definitive diagnoses
- Encourages professional consultation
- Warm, empathetic tone throughout

Architecture



FertiliCare implements a **sequential multimodal pipeline** designed for memory-efficient inference on consumer hardware. The system accepts three input modalities through a Gradio interface: text queries (direct passthrough), audio recordings (transcribed via Whisper STT), and medical images such as hormone panels (analyzed using Florence-2 vision model for OCR). All three inputs converge into a **GraphRAG retrieval layer** that combines vector similarity search (ChromaDB with sentence-transformers embeddings) and knowledge graph traversal (NetworkX) to extract relevant medical context from a curated database of fertility literature. This enriched context is then fed to **Qwen2.5-3B**, a compact yet powerful instruction-tuned language model, which generates empathetic, evidence-based responses while adhering to strict safety guardrails that prevent medical diagnoses. The output is delivered both as text

and optional audio (via gTTS), with each heavy model (Florence-2, Qwen) being loaded on-demand, executed, and immediately unloaded to maintain a peak memory footprint of just 6GB VRAM. This architecture enables sophisticated medical AI assistance without requiring expensive cloud APIs or high-end hardware, while ensuring user privacy through fully local processing.