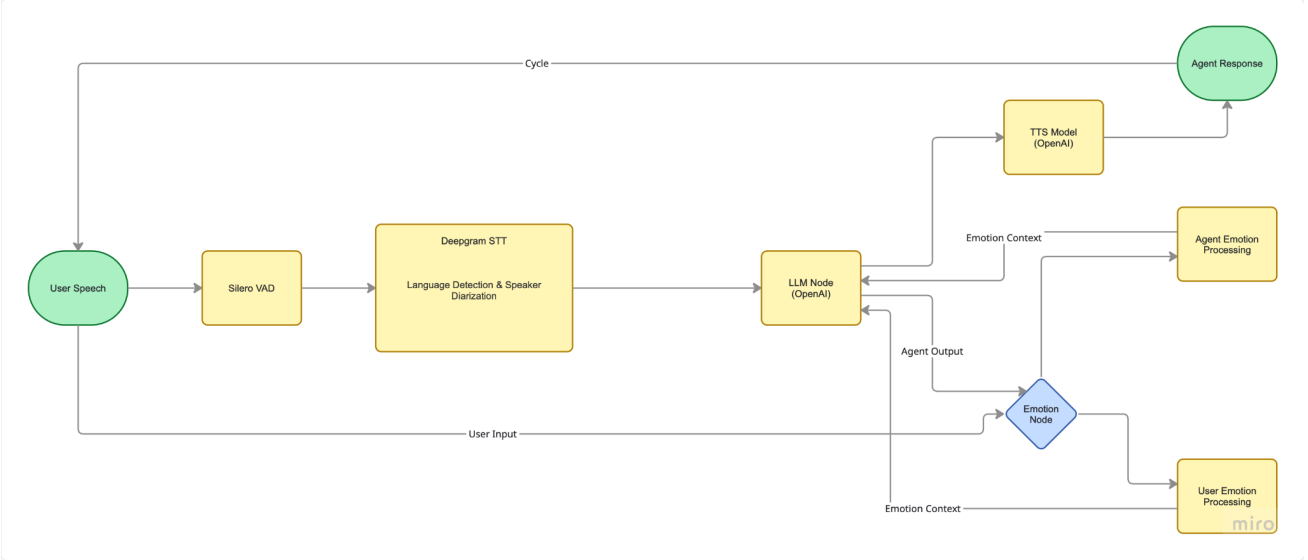
Medical Follow Up Agent

In Depth Pipeline

The overall architecture has been built over the livekit framework. I have chosen livekit primarily due to its extensive support for VAD, Turn Detection Model (Semantic VAD), Extensive support for numerous tts, stt and llm model support and also have support in agentic task builder which is a key essential for build an agent which have workflow based use cases. As per the requirements of the assessment, the system is equipped with steaming ASR, deep learning modules such as Emotion Detection, Language Detection and also Speech Diarization.



Latency Optimizations

- **Preemptive Generation** - Allows the agent to begin generating a response before the user's end of turn is committed. The response is based on partial transcription or early signals from user input, helping reduce perceived response delay and improving conversational flow.
- **ONNX Emotion Detection Model** - Rather than using a resource intensive model, onnx model is lightweight and also I have implemented multi threading method so that the model will run in background and update the emotion on the fly rather than waiting for the response from the model before passing to the stt model.
- **Language and Emotion Data Injection -** Used prompt injection method for feeding the emotion and language detection result just before sending the user query to the llm to avoid additional llm calls to the system.

Overall Model Analysis

- Despite having a Streaming ASR model, the system is still having a subsequent latency and this can be avoided using On Prem deployments for STT and if possible LLM.
- Emotion Detection Model Accuracy is very low and also the number of options in the market is also less. Building a product on this has a clear market scope considering the agent call monitoring is on the rise.