# Model Evaluation Report

## 1. Overview

This document outlines the evaluation process and rationale behind the selection of each AI model and API used in the OMANI-Therapist-Voice ("Elile") project. The primary goal was to choose a stack that balanced high performance, accuracy, and cultural relevance with real-world constraints, particularly the need for a cost-effective, sustainable solution.  
Each component of the Speech Processing Pipeline was evaluated based on a specific set of criteria tailored to its role in the system.

## 2. Speech-to-Text (STT)

Model Chosen: OpenAI Whisper API (whisper-1 model).

### Evaluation Criteria:

- Accuracy: High transcription accuracy for Arabic, specifically with dialectal variations.  
- Performance: Low latency to support real-time conversation.  
- Ease of Integration: Simple and reliable API integration.

### Justification:

The Whisper API is industry-renowned for its state-of-the-art accuracy across a vast number of languages. It performs exceptionally well with non-standard Arabic, making it ideal for the Omani dialect. By providing a specific prompt ("هذه محادثة باللهجة العمانية."), we further guide the model to achieve even higher accuracy. Using the API offloads the significant computational cost of running a large STT model, which would be too slow on a standard CPU and is a critical factor in meeting the project's su...

### Alternatives Considered:

Local Whisper Models: While free, running the base or small Whisper models locally on a CPU would introduce significant latency (10-20 seconds just for transcription), making real-time conversation impossible. The API was the superior choice for performance.

## 3. Emotion & Intent Detection

Model Chosen: bhadresh-savani/bert-base-go-emotion (from Hugging Face).

### Evaluation Criteria:

- Performance: Fast inference speed on a CPU.  
- Resource Footprint: Low memory usage suitable for a free-tier deployment server.  
- Cost: Must be free to use.

### Justification:

The purpose of this component is to provide a quick emotional "hint" to the primary LLM, not to perform a deep clinical analysis. Therefore, speed and low resource usage were prioritized. bert-base-go-emotion is a well-optimized, BERT-based model that is small enough to load quickly and run efficiently on a CPU. Its integration via the transformers pipeline is straightforward and, most importantly, it adds no operational cost to the project.

## 4. Dual-Model Response Generation

### 4.1. Primary LLM

Model Chosen: OpenAI gpt-4o.

#### Evaluation Criteria:

- Conversational Quality: State-of-the-art ability for nuanced, empathetic, and coherent conversation.  
- Instruction Following: Ability to adhere strictly to the complex system prompt regarding persona, cultural sensitivity, and safety protocols.  
- Reasoning: Strong reasoning capabilities to handle complex user inputs.

#### Justification:

gpt-4o was selected as the primary model because of its superior ability to generate high-quality, therapeutic-grade responses. For a mental health application, the quality of the conversation is paramount. Its advanced instruction-following capabilities ensure it reliably maintains the "Elile" persona and adheres to the critical safety rules defined in the prompt.

### 4.2. Fallback LLM

Model Chosen: Google gemini-1.5-flash.

#### Evaluation Criteria:

- High Availability & Speed: Must be extremely fast to serve as a reliable fallback.  
- Cost-Effectiveness: Lower cost to ensure the fallback mechanism is sustainable.  
- Competence: Strong enough to provide a coherent and safe response if the primary model fails.

#### Justification:

The dual-model strategy was implemented to build a robust, production-ready system. gemini-1.5-flash is an excellent fallback choice due to its very low latency and high efficiency. Its function is to prevent a conversational dead-end in the rare event of an OpenAI API outage, ensuring the user always receives a timely and helpful response.

## 5. Text-to-Speech (TTS)

Model Chosen: facebook/mms-tts-arb (from Hugging Face).

### Evaluation Criteria:

- Cost: Must be completely free to run, with zero API charges.  
- Voice Quality: Must sound natural, clear, and empathetic, avoiding a robotic tone.  
- Arabic Support: High-quality synthesis for Arabic text.  
- Performance: Acceptable latency on a CPU.

### Justification:

This was a key strategic decision to ensure the project's long-term viability and cost-effectiveness. Initial consideration of paid APIs like Google Cloud TTS was rejected because they require an active billing account, which contradicts the goal of building a sustainable, low-cost solution.  
The facebook/mms-tts-arb model was chosen because it is a state-of-the-art, open-source model that produces a remarkably natural and high-quality Modern Standard Arabic (MSA) voice.

### Acknowledged Trade-offs:

- Performance: Running a large TTS model on a CPU is inherently slower than a dedicated API call. This was mitigated by implementing a "warm-up" routine on server startup to reduce initial request latency.  
- Accent: The model produces an MSA accent, not a specific Omani one. This was deemed an acceptable trade-off to gain a high-quality, natural-sounding voice at zero cost. The clarity and empathetic tone of the MSA voice are well-suited for a therapeutic context.