# Elile Technical Documentation

## 1. Model Evaluation Report: Comparative Analysis of Dual-Model Approach

This section presents a comparative evaluation between GPT-4o and Gemini Flash within Elile's dual-model architecture. The system defaults to GPT-4o but falls back to Gemini Flash when reliability or latency issues arise.  
  
Key Findings:  
- GPT-4o demonstrates higher empathy alignment and contextual awareness, particularly for mental health narratives.  
- Gemini Flash shows lower latency and resource consumption, making it suitable for fallback and mobile-first deployments.  
  
Evaluation Metrics:  
- Response quality (human-rated empathy and coherence)  
- Latency under load (p95 response time)  
- Fallback rate and recovery time  
  
Conclusion:  
While GPT-4o is the preferred model for primary responses due to its alignment and cultural nuance, Gemini Flash adds redundancy and resilience to the conversational pipeline.

## 2. Cultural Adaptation Guide: Omani Arabic Implementation Details

Language Localization:  
- Whisper STT prompts are tuned for Omani dialect features and vocabulary.  
- GPT-4o responses are fine-tuned using system prompts referencing Omani cultural norms and conversational tone.  
  
Voice Output:  
- MMS TTS leverages the Arabic model, preconfigured for Gulf Arabic phonetics.  
- Future plans include voice cloning using local speaker data.  
  
Cultural Filters:  
- LLMs are configured to avoid culturally inappropriate topics and ensure respectful tone during emotional escalation.

## 3. Safety Protocol Documentation: Crisis Intervention and Escalation Procedures

Trigger Detection:  
- Emotion classifier flags severe distress (e.g., sadness + urgency).  
- LLMs are prompted to assess for crisis language (e.g., suicidal ideation).  
  
Response Strategy:  
- If crisis detected, TTS response advises user to seek professional help.  
- System provides pre-scripted emergency referral messages for hotline and in-person care.  
  
Escalation Workflow:  
- Logged sessions are flagged for review by a clinical supervisor.  
- Future: automated alert to registered clinician via secure dashboard.

## 4. Performance Benchmarks: Latency, Accuracy, and Scalability Metrics

Latency Benchmarks:  
- End-to-end turn time: < 20 seconds  
- Whisper transcription: ~5s  
- GPT-4o response: ~7s  
- MMS TTS synthesis: ~4s  
  
Accuracy:  
- Whisper STT accuracy for Omani Arabic: ~88%  
- Emotion classification F1-score: 0.76  
- LLM empathy rating (human eval): GPT-4o = 4.6/5, Gemini = 3.9/5  
  
Scalability:  
- Concurrent users supported (Render baseline): ~25  
- GPU-inference optimization planned for local models  
  
Recommendations:  
- Implement async batch processing for STT and TTS to improve concurrency.  
- Quantize BERT and MMS models for reduced memory footprint.