**Future Roadmap**

This document outlines areas for improving the system over time, including scalability, feature extension, and deeper localization.

**1. Model Improvements**

* Explore fine-tuning GPT or Claude using actual Omani dialect data (e.g., JARO corpus or domain-specific chat data).
* Switch to hosted open-source models if budget/API becomes limiting (e.g., Mistral with Arabic adapter).
* Integrate real-time streaming response generation.

**2. Full Mobile Support**

* Port the interface to a PWA (Progressive Web App) or native mobile app using Flutter or React Native.
* Integrate native TTS (e.g., Android’s SpeechSynthesis API) to reduce latency.

**3. Advanced LLM Features**

* Add memory for multi-turn dialogue continuity.
* Integrate agent-like goal tracking (e.g., maintain session summaries).
* Support multi-modal inputs in the future (e.g., facial expressions via webcam for emotion estimation).

**4. Safety and Escalation**

* Integrate real-time suicide risk assessment with hotline triggers.
* Add consent modal and encryption for logs.
* Build escalation dashboard for professional review if needed.

**5. Metrics and Analytics**

* Automatically analyze logs for:
  + Avg. response quality
  + Latency per model
  + Fall-back rates
  + Usage patterns

**6. Scalability**

* Containerize with Docker for easy deployment.
* Add Redis or SQLite for lightweight state storage and logging.
* Move backend to a FastAPI server for robust API exposure.

**7. Cultural Enhancement**

* Add Quranic encouragement options if flagged.
* Train with local idioms and emotion expressions unique to Oman.
* Collaborate with local therapists for prompt curation.