**Safety Protocol Documentation**

**Purpose**

This document outlines the safety measures and crisis response protocols implemented in the Omani Arabic mental health voice assistant. The goal is to ensure ethical, secure, and context-aware conversations that prioritize user wellbeing.

**Crisis Identification and Escalation**

I implemented simple safety heuristics within the LLM prompts to help detect phrases that may indicate emotional distress or crisis. Examples of flagged signals include Arabic equivalents of:

* "I want to give up"
* "I can't take this anymore"
* "I’m thinking about hurting myself"

When such phrases are detected, the assistant responds with de-escalating, empathetic messages and encourages seeking help from family or professionals. The response avoids judgment, uses supportive language, and prompts the user to speak with someone they trust.

**Suicide Risk Mitigation**

While I did not train a dedicated classifier for suicide risk, I added fallback behaviors to the LLM responses:

* Emphasize hope, faith, and support systems
* Avoid overly optimistic or dismissive replies
* Offer general advice to pause and reflect, rather than act impulsively

These were achieved through carefully structured prompt templates, instructing the model to prioritize safety and empathy.

**Privacy and Session Consent**

As a prototype, the assistant does not yet log identifiable user data. However:

* Audio inputs are temporarily saved during processing and can be optionally logged for evaluation purposes.
* The demo interface could include a disclaimer informing users their voice may be processed for research/demo.

Users are not required to provide names, contact info, or any personal details.

**Content Filtering**

To reduce the chance of harmful, offensive, or triggering replies:

* LLM responses are guided by a curated prompt
* The assistant avoids giving medical advice or making diagnoses
* The tone is neutral, respectful, and emotionally attuned

**Limitations**

* There is no real-time human-in-the-loop monitoring or escalation.
* The LLM can still hallucinate or misinterpret nuance.
* Cultural distress signals in Omani Arabic may not be fully captured by current models.

**Ongoing Improvements**

I plan to continue refining the safety features by:

* Testing with more realistic scenarios
* Collecting anonymized logs (with consent)
* Exploring fine-tuning or moderation APIs for sensitive content