Therapy AI – Emotional Support System Design Document

1. Introduction

This document describes the complete design of the Therapy AI project, an emotionally intelligent mental health support system for teens and young adults. The system provides empathetic, non-judgmental conversational support through both text and speech modalities. It is explicitly designed as a support companion and not a replacement for professional therapy.

2. Objectives

• Provide emotional support through empathetic conversation
• Detect emotional states and mental distress safely
• Support text and speech-based interaction
• Encourage healthy coping mechanisms
• Escalate responsibly during crisis scenarios
• Maintain privacy, ethics, and user safety

3. Target Users

• Teenagers (13–19 years)
• Young adults (20–30 years)
• Users experiencing stress, anxiety, loneliness, or emotional distress

4. Explicit Non-Goals

• Diagnosing mental illness
• Replacing licensed therapists
• Providing medical or medication advice
• Handling suicide intervention independently

5. System Overview

The system follows a modular, safety-first architecture. User input is processed through emotion detection and safety layers before response generation. Speech and text channels are treated equally. Memory systems ensure continuity while respecting privacy.

6. High-Level Architecture

User Input → Speech-to-Text (optional) → Emotion Detection → Safety Check → Therapy Logic → Response Generation → Text-to-Speech (optional)

7. Core Modules

7.1 Input Module: Handles text and voice input
7.2 Emotion Detection Module: Classifies emotional states using NLP models

7.3 Therapy Engine: Generates empathetic, structured responses
7.4 Safety & Crisis Module: Detects and escalates risk
7.5 Memory Module: Stores short-term and long-term conversational context
7.6 Output Module: Returns text and speech responses

8. NLP & ML Design

Pre-trained transformer models are used for emotion detection and sentiment analysis. Fine-tuning is restricted to conversational style and empathy alignment. Statistical datasets are never used for dialogue generation.

9. Dataset Usage Strategy

- Therapy conversation datasets for response style reference
- Emotion datasets for classifier training
- Depression and anxiety datasets for risk detection
- Survey and statistics datasets for contextual awareness only

10. Therapy Design Principles

- Emotional validation before advice
- Open-ended reflective questioning
- Optional coping suggestions
- Calm, simple, non-clinical language
- User autonomy and consent

11. Safety & Crisis Handling

The system continuously monitors for self-harm or suicidal ideation. Upon detection, it switches to a crisis-safe mode that encourages reaching out to trusted individuals or professional helplines. No advice or reassurance clichés are given during crisis escalation.

12. Memory Design

Short-term memory maintains conversation flow. Long-term memory stores non-sensitive emotional patterns and preferences using vector embeddings. Sensitive content is never permanently stored.

13. Speech Interface

Speech input uses speech recognition libraries. Speech output uses neural or rule-based TTS engines with calm, neutral voice profiles to avoid emotional manipulation.

14. Ethics & Compliance

- Clear user disclaimers
- Explicit non-therapist positioning
- Privacy-first data handling
- No dark patterns or emotional dependency
- Transparency in AI behavior

15. Privacy & Data Protection

User data is anonymized by default. Memory can be cleared at any time. No data is sold or used for targeted advertising.

16. Testing Strategy

- Unit tests for emotion detection and safety logic
- Scenario-based testing for crisis flows
- Bias and robustness testing across emotional states

17. Deployment Plan

Initial deployment as a local application, followed by API-based web or mobile deployment using FastAPI. Scalable modular design supports future platform expansion.

18. Risks & Mitigation

- Misinterpretation of emotions → mitigated by conservative confidence thresholds
- Over-dependence → mitigated by encouragement of real-world connections
- False negatives in crisis detection → mitigated by multi-layer checks

19. Future Enhancements

- Multilingual support
- Personalization with consent
- Therapist-reviewed response validation
- Wearable or journaling integrations

20. Conclusion

This design prioritizes emotional safety, ethical responsibility, and technical scalability. When built according to this document, the Therapy AI can provide meaningful emotional support while respecting human boundaries.