Local LLM Psychological Pre-Consultation Chatbot

Ethics Report

1. Introduction

This project presents a locally hosted conversational agent for psychological pre-consultation. The assistant is scoped to provide empathic listening, light coping suggestions and referrals to qualified professionals. The design is anchored in the principal of non-maleficence. It prioritizes user safety, respects autonomy and clearly outlines its limitations. This report articulates the underlying ethical rationale, identifying potential risks and mitigation strategies.

2. Ethical Considerations in Automated Mental Health Support

2.1. Beneficence and Non-Maleficence

The application aims to reduce distress without inadvertently causing harm. Because formulaic "empathy" can alienate users and improvisational advice can be unsafe, the assistant is prompted to listen first → reflect user feeling → offer coping suggestions only after clarifying needs. It refrains from diagnosis, prescription, or speculation about clinical status. Detected crisis or medical request are redirected to pre-approved fallbacks emphasizing contract with human supports and professional services.

2.2. Autonomy and Informed Use

An initial disclaimer clarifies that the agent is not a therapist and delineates boundaries, escalation pathways and available support. Throughout interactions, the assistant frames suggestions as option rather than directives, invites collaborative decision making (e.g., "Would it help if we explored...?") and acknowledges uncertainty so that users can make informed choices about seeking human care.

2.3. Privacy

The application runs locally. Conversation history is held only in memory during the execution and discarded when the session ends. Moreover, the system prompt is designed to minimize personal data collection and avoid fabricating credentials.

2.4. Accountability and Human Oversight

Accountability is advanced through transparent documentation of moderation thresholds, escalation logic and boundary conditions. The system enforces a maximum conversation length, and upon reaching it, recommends taking a break, reiterating that sustained care requires human providers.

3. Potential Risks and Mitigation Strategies

Risk	Description	Mitigation
Missed crisis cues	A user expresses suicidal intent using uncommon language.	Extensive keyword list, regex patterns for implicit phrasing, low crisis threshold (0.3 strict / 0.5 balanced) ensuring BLOCK responses even for borderline matches.
Improper medical advice	The model speculates about medication or diagnosis.	SAFE_FALLBACK triggers on medical keywords/patterns, model-output moderation scans for prescribing language, and fallback template redirects to licensed professionals.
Violent or illegal coordination	Users seek guidance on harming others or committing crimes.	Harmful content filters block violence/illegal/harassment terms; fallback message sets firm boundary and suggests constructive de-escalation.
Over-reliance on AI	Users treat the assistant as therapy.	System prompt and disclaimer reiterate limitations, encourage human help, and enforce turn limits with break suggestions.
Model hallucination	The model fabricates credentials or resources.	Prompt instructs transparency when uncertain.
Inadequate cultural competence	Responses fail to resonate with diverse users.	Neutral, inclusive language in templates

4. Confidence Threshold Strategy

4.1. Confidence Calculation:

For each category, confidence is computed as follows

Category	Keyword	Pattern hit	Keyword + Pattern
Crisis	0.7	0.85	0.9
Medical	0.7	0.85	0.9

For harmful content, confidence reflect breath of violation:

- One harmful subcategory (e.g., violence or illegality or harassment): 0.75
- Two or more concurrently (e.g., violence and illegality): 0.85

4.2. Decision Threshold by Safety Mode

These thresholds are used to determine when to trigger a BLOCK or SAFE_FALLBACK response.

Mode	Crisis	Medical	Harmful
Strict	0.30	0.40	0.50

Mode	Crisis	Medical	Harmful
Balanced	0.50	0.60	0.60
Permissive	0.7	0.75	0.80

4.3. Rationale

4.3.1. Crisis (BLOCK):

Any crisis keyword signal (0.70) meets or exceeds the threshold in all modes (strict 0.30, balanced 0.50, permissive 0.70), therefore triggers a immediate BLOCK response. Its also the same for pattern (0.85) and combined (0.90).

Self-harm and imminent-risk language such as "kill" presents high potential harm. The strategy minimizes false negatives by acting on the earliest reliable cue, independent of additional corroboration.

4.3.2. Medical (SAFE_FALLBACK):

Similarly, for strict (0.40) and balanced (0.60) modes, a keyword (0.7) is sufficient to trigger SAFE_FALLBACK for medical requests. However, for permissive (0.75) mode, a keyword along does not trigger. Pattern (0.85) or keyword + pattern (0.90) are required to trigger SAFE_FALLBACK.

Many users make exploratory health inquiries. In permissive settings, requiring a stronger signal (pattern or combined) reduces over-moderation of benign information-seeking while still intercepting diagnostic/prescriptive requests.

4.3.3. Harmful (BLOCK)

For strict(0.50) and balanced (0.60) modes, a single harmful subcategory (0.75) exceeds threshold and triggers BLOCK. In permissive (0.80) setting, a single subcategory does not trigger, but two or more (0.85) do.

Heated conflict talk can include violent or harassing terms without intent to act. Permissive mode therefore requires broader evidence (multiple subcategories) to reduce false positives, whereas strict/ balanced modes favor precaution and intervene on a single clear violation.

5. Limitations

1. **Scope of support**: The assistant does not replace therapy, crisis counselling, or medical care. Users requiring longitudinal support must engage professionals
2. **Language coverage**: Detection is current English-centric. Expressing crisis or medical needs in other languages will not be moderated by the system
3. **Adversarial inputs**: Users may attempt to bypass filters via obfuscation.
4. **User Privacy**: Although the current system uses local language model, if a third-party service is used, the user data will be sent to the third-party service.

5. **Model limitations**: Hallucinations and outdated information remain risks. This may be mitigated through updates or incorporating model context server server to provide accurate information. Moreover rule-based guardrails and post generation checks can prevent unsafe output.