Ethics Report: Automated Mental Health Pre-Consultation Support Chatbot

1. Ethical Considerations in Automated Mental Health Support

1.1 Role Definition

The chatbot's role must be clearly and consistently defined as a supportive pre-consultation system, not a replacement for licensed mental health professionals. Misrepresenting its capabilities could lead to dangerous user overreliance and misinterpretation, potentially causing users to forgo necessary professional care.

1.2 Safety and Crisis Management

The chatbot system should prioritize user safety, especially when detecting self-harm, suicidal ideation, or expressions of violence. Automated crisis detection with referral to hotlines or emergency services is essential. Ethical deployment requires setting low thresholds for crisis detection to minimize the risk of false negatives.

1.3 Boundaries of Care

It is ethically unacceptable for the chatbot system to provide medical diagnoses, prescribe medication, or recommend treatment protocols. Instead, it should:

- Redirect medical or diagnostic queries to professionals.
- Offer supportive, empathetic, and reflective communication.
- Provide general, evidence-based coping strategies, such as grounding techniques, mindfulness exercises, or self-care reminders.

1.4 User Autonomy

Users should be empowered to make their own decisions. The chatbot should avoid any language that is directive, coercive, or manipulative. Dialogue should be collaborative and open-ended, focusing on helping the user explore their own feelings and solutions.

1.5 Transparency and Disclaimer

Before a user begins any interaction, a clear and conspicuous disclaimer is essential to establish informed consent. This disclaimer must explicitly communicate:

- The chatbot's status as a non-professional AI.
- Its specific boundaries and limitations.

2. Information on when and how to seek immediate professional or emergency help.

3. Potential Risks and Mitigation Strategies

3.1 Misdiagnosis or Misleading Information

Risk: Users may misinterpret the chatbot's empathetic responses as diagnostic confirmation, or the system might accidentally generate misleading information.
Mitigation: This risk is mitigated through a multi-layered approach: a strict system prompt that enforces role boundaries, robust output moderation to block diagnostic or treatment-related phrases, and a consistently reinforced disclaimer that clarifies the AI's non-professional status..

3.2 Crisis Escalation

Risk: A delayed or completely missed detection of suicidal ideation or self-harm could result in tragic real-world harm.
Mitigation: This is addressed with a conservative crisis threshold, which is intentionally set low to catch even subtle or indirect signs of distress. It is further supported by keyword and pattern-based detection algorithms that immediately trigger an emergency referral template.

3.3 Overreliance on AI

Risk: Users might substitute the chatbot for ongoing, professional therapeutic care, leading to a delay in receiving proper treatment.
Mitigation: The chatbot system regularly and gently reminds users that it is a pre-consultation tool, not a replacement for therapy. It should also consistently encourage users to seek a licensed professional for long-term care, reinforcing this message at key points in the conversation.

3.4 Exposure to Harmful Content

Risk: Users may attempt to elicit, share, or request harmful, illegal, or violent content.
Mitigation: The implementation of a strong harmful content filter is needed to cope with this situation. The filter use pre-defined template responses to set clear, firm boundaries, refuse to engage with the harmful topic, and immediately redirect the conversation to safe and appropriate subjects.

4. Limitations

4.1 Lack of Nuanced Human Judgment

The chatbot lacks the nuanced judgment and living experience of human counselors. It cannot perceive and interpret non-verbal cues like tone of voice, body language, or facial expressions, which are fundamental to a human therapeutic relationship and often signal critical, unspoken information.

4.2 False Positives and False Negatives

The system is susceptible to two types of errors:

- False positives: Benign or metaphorical statements ("I'm dying of boredom") could mistakenly trigger a crisis filter, leading to user frustration.
- False Negatives: On the other hand, a user's subtle or culturally specific expression of suicidal ideation may go undetected, which is a catastrophic failure.

Both highlight the limitations of rule-based and probabilistic moderation.

4.3 Cultural and Linguistic Bias

The AI model that the system uses is trained on vast datasets that may not be culturally or linguistically representative. This can lead to algorithmic bias where the system fails to accurately understand or respond to expressions of distress that differ from standard Western phrasing. There is the risk that the system might not be effective for all the diverse user groups.

4.4 Scope of Support

The chatbot's support is strictly limited to pre-consultation emotional support. It cannot provide the depth, continuity, or long-term therapeutic relationship that is crucial for managing complex mental health conditions. It is not a replacement for professional care, psychiatric intervention, or long-term mental health management.

4.5 Technical Constraints

AI-generated responses are probabilistic. There is a small but non-zero chance of producing an inappropriate or nonsensical output, despite all safeguards. Since there is no human oversight of every interaction, errors cannot be fully eliminated.