Prompt Engineering in Mental Health: A Conceptual Review and a Proposal for Layered Socratic Dialogue in AI Therapy

A S Chethan Varma
Consultant Clinical Psychologist & Founder
Psychethan's Mind Garage, Hyderabad, India
International Conference on Creativity & Innovation in Research, BESTIU

ABSTRACT

This paper explores the use of prompt engineering to enhance mental health AI tools, focusing on large language models (LLMs) in digital therapy. It reviews classification, generation, and retrieval-based methods, alongside ethical challenges like AI hallucinations, cultural bias, and user dependency. A novel model, Layered Socratic Prompting, is proposed to emulate cognitive behavioral therapy techniques using structured prompts. This framework supports more reflective, adaptive AI interactions. The review concludes with a call for culturally sensitive design and ethical oversight to maximize the clinical utility of prompt driven AI in mental health.

Keywords: prompt engineering, mental health AI, large language models, Socratic prompting, digital therapy

Introduction and Literature Review:

Introduction

In recent years, the intersection of artificial intelligence (AI) and mental health has opened unprecedented possibilities for expanding access to psychological support and enhancing the effectiveness of digital interventions. Among these technological advancements, large language models (LLMs), such as OpenAI's GPT-4, have shown remarkable potential in understanding and generating human-like text (Brown et al., 2020). Central to maximizing the effectiveness of LLMs is a process known as prompt engineering—the deliberate crafting of input prompts to elicit specific, relevant, and context-sensitive outputs from AI systems.

Prompt engineering is particularly significant in mental health, a domain that demands not only linguistic precision but also emotional nuance, cultural sensitivity, and ethical responsibility. With global mental health services under strain, especially in underserved regions, prompt-engineered AI solutions offer scalable and low-barrier methods to augment clinical care, support early detection of disorders, and improve psychoeducation (Priyadarshana et al., 2024; Topol, 2019).

While numerous mental health chatbots and digital tools exist, many rely on rule-based systems or generic responses that fall short of therapeutic standards. Prompt engineering, by contrast, allows for the dynamic generation of tailored responses, therapeutic dialogues, and diagnostic cues by leveraging the depth of LLMs (Wu et al., 2024). This paper reviews the emerging landscape of prompt engineering in mental health, covering current use cases, technical methodologies, and ethical considerations. In doing so, it identifies a crucial innovation gap and proposes a novel framework: Layered Socratic Prompting—a method to guide LLMs in emulating Socratic questioning, a core component of cognitive behavioral therapy (CBT).

This paper aims to synthesize cutting-edge developments in the field and introduce a new avenue for enhancing the therapeutic utility of AI-driven mental health tools, particularly through culturally responsive and ethically grounded prompt strategies.

Literature Review:

Prompt engineering has emerged as a pivotal tool in optimizing the performance of large language models (LLMs) within mental health applications. Recent advancements demonstrate the versatility of prompt engineering across diagnostic, therapeutic, and educational use cases.

Classification Tasks: Prompt-based strategies enable LLMs to perform mental health screening by classifying disorders such as depression, anxiety, and post-traumatic stress disorder. By designing symptom-specific prompts, models can be steered towards accurate diagnostic impressions without extensive retraining (Brown et al., 2020; Priyadarshana et al., 2024).

Generative Therapeutic Dialogues: Several systems have utilized prompt engineering to simulate supportive therapy-like conversations. For instance, the Illuminate system employs layered prompting aligned with DSM-5 criteria to assess and respond empathetically to user inputs (Wu et al., 2024).

Retrieval-Augmented Generation (RAG): Hybrid models combining LLMs with external clinical knowledge bases have improved the reliability of AI outputs. The SouLLMate system exemplifies how retrieval-augmented generation enhances AI mental health support by integrating structured psychological knowledge into generated responses (Zhang et al., 2024).

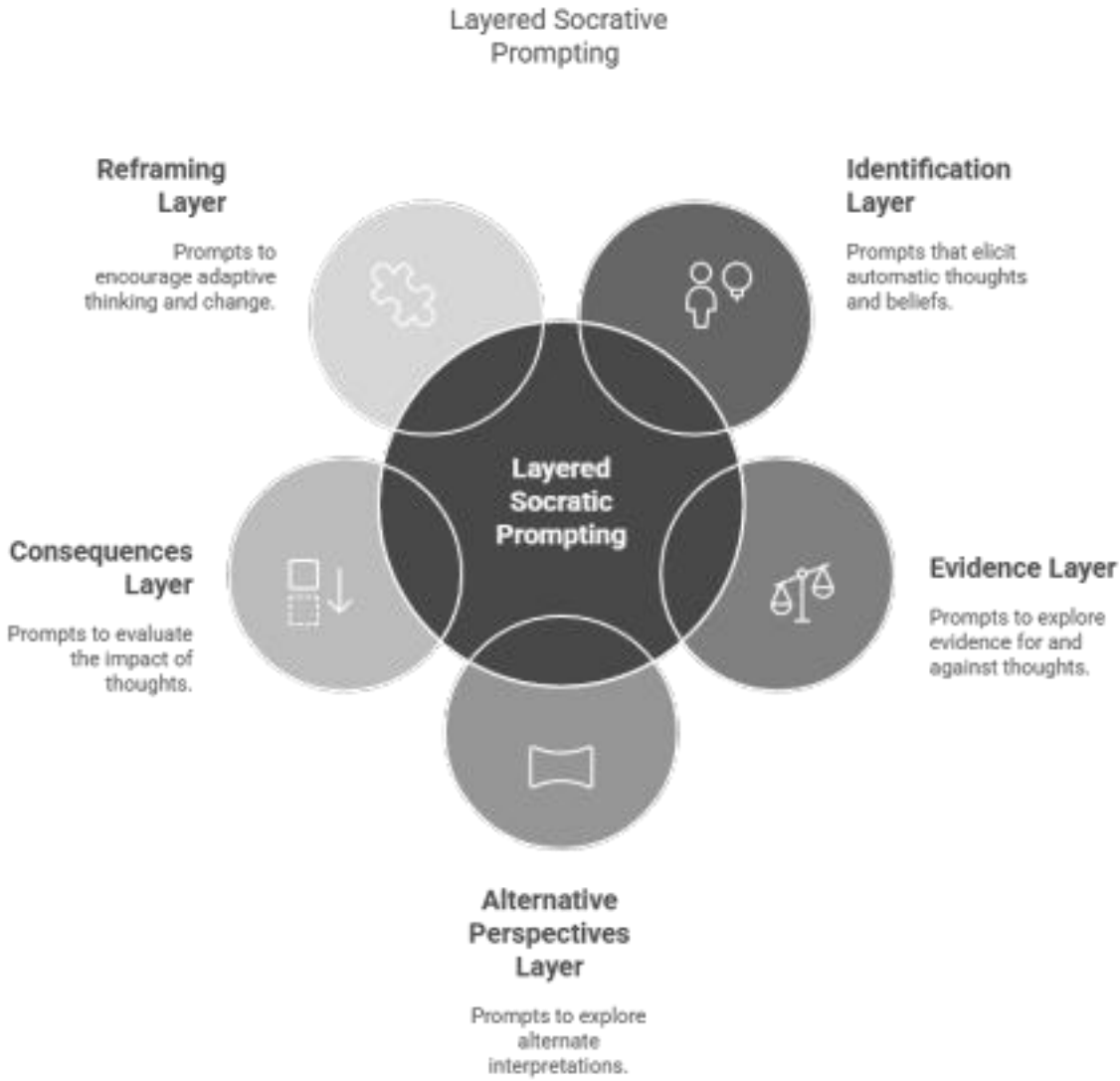
Proactive Questioning Strategies (PQS): Prompt engineering has also been used to design proactive questioning models that guide users through reflective exercises. These techniques promote self-awareness, enabling early identification of mental health symptoms (Priyadarshana et al., 2024).

Emotional Adaptivity and Context Sensitivity: Emerging research focuses on tuning LLMs' emotional expressiveness through prompt design. Techniques such as emotional tagging and dynamic tone adjustment help LLMs maintain therapeutic sensitivity during emotionally charged conversations (Schramowski et al., 2023).

Overall, literature demonstrates that prompt engineering is not just a technical method but a critical bridge between AI capabilities and the nuanced demands of mental health care delivery.

Proposed Framework: Layered Socratic Prompting

Socratic questioning, a foundational technique in cognitive behavioral therapy (CBT), involves a series of structured, reflective questions designed to help clients challenge maladaptive thoughts and arrive at healthier conclusions (Beck, 2011). Translating this dialogical technique into AI systems powered by LLMs presents a novel opportunity for digital therapeutic tools.



I propose a framework called "Layered Socratic Prompting," which emulates the logical progression of Socratic questioning by guiding the model through a sequenced prompt architecture:

1. ***Identification Layer:*** Prompts that help elicit automatic thoughts or beliefs. Example

prompt: "What thought went through your mind when you felt that way?" 2. ***Evidence Layer:*** Prompts that guide the user to explore evidence for and against their thought.

Example prompt: "What makes you think this thought is true? Is there any evidence that challenges it?"

3. ***Alternative Perspectives Layer:*** Prompts to explore alternate interpretations.

Example prompt: "Could there be another way to look at this situation?" 4. ***Consequences Layer:*** Prompts that help the user evaluate the impact of the thought.

Example prompt: "How does believing this thought affect your mood or actions?" 5. ***Reframing Layer:*** Prompts that encourage adaptive thinking and behavioral change.

Example prompt: "If you were helping a friend with the same thought, what would you say to them?"

Each layer is presented sequentially, with transitions informed by the user's previous responses, mimicking the therapeutic dialogue flow. This prompt structure encourages deeper self-reflection and supports cognitive restructuring, aligning AI-generated responses more closely with therapeutic best practices.

Layered Socratic Prompting can be adapted to different cultures, languages, and diagnostic categories, making it a versatile component in AI-driven therapy tools.

Challenges and Ethical Concerns

As LLMs are increasingly deployed in mental health contexts, several challenges and ethical considerations emerge. While prompt engineering holds the potential to shape AI responses in clinically meaningful ways, the variability and unpredictability of LLM-generated content raise questions about reliability and safety (Wu et al., 2024).

Hallucinations and Misleading Outputs: Large language models are prone to producing "hallucinations," where responses may appear confident but are factually incorrect or misleading. In mental health contexts, such inaccuracies can have serious consequences, especially if users act on incorrect psychological guidance without professional oversight (Priyadarshana et al., 2024).

Emotional Volatility and AI Mood Mimicry: Recent studies suggest that LLMs may display state-dependent behavior—responding differently to emotionally charged prompts or trauma related narratives. This pseudo-emotional variance introduces concerns about consistency,

emotional safety, and unintended reinforcement of user distress (Schramowski et al., 2023).

Informed Consent and AI Autonomy: Users often interact with LLM-based tools without fully understanding their limitations or the AI's non-human nature. Transparent disclosure, informed consent mechanisms, and the clear positioning of these tools as adjuncts—not replacements—for human therapy are essential (Topol, 2019).

Cultural and Linguistic Bias: Prompt engineering practices, unless carefully localized, can inadvertently reproduce biases embedded in the training data. This includes cultural stereotyping, gender bias, or the exclusion of non-Western mental health frameworks. Ethical prompt engineering must involve inclusive design and multilingual testing (Devlin et al., 2019).

Dependency and Dehumanization Risks: Over-reliance on AI tools for emotional support may lead users to substitute genuine interpersonal connection with automated interactions. This carries the risk of deepening social withdrawal or emotional isolation, particularly among vulnerable populations (Topol, 2019).

Conclusion

Prompt engineering is a powerful interface between clinical intent and AI capability. When applied with clinical insight and ethical rigor, it can transform LLMs into therapeutic allies.

This paper's proposed Socratic layering framework offers a structured and evidence-informed way to scale mental health support via AI.

Future work should involve interdisciplinary collaboration, empirical testing of prompt-based interventions, and development of culturally adaptive models. In doing so, prompt engineering could move from being a technical tool to a cornerstone in the digital mental health revolution (Topol, 2019).

References

Beck, J. S. (2011). *Cognitive behavior therapy: Basics and beyond* (2nd ed.). The Guilford Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Priyadarshana, C., Fernando, B., Bandara, D., Jayawardena, S., & Goonetilleke, R. (2024). Large language models for mental health text classification and question answering: A prompt engineering approach. *Frontiers in Digital Health, 4*, 1410947. https://doi.org/10.3389/fdgth.2024.1410947

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., & Kersting, K. (2023). When AIs get anxious: The psychological impact of trauma prompts on large language models. *LiveScience AI Behavior Series*.

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*, 44–56.

Wu, A., Deng, K., & Xu, B. (2024). Illuminate: Layered prompting for depression detection and mental health assessment with LLMs. *arXiv preprint arXiv:2402.05127*.

Zhang, X., Li, Y., & Lee, C. (2024). SouLLMate: A retrieval-augmented conversational agent for personalized mental health support. *arXiv preprint arXiv:2410.16322*.

Acknowledgment

I extend my heartfelt gratitude to all my clients and patients, whose journeys of growth and resilience have continually inspired my work. I sincerely thank my Gurus, whose wisdom and guidance have shaped my professional and personal path. I am deeply grateful to my family and friends for their unwavering support, encouragement, and belief in me. Lastly, I bow in reverence to Nature, the ultimate teacher, whose rhythms and resilience constantly remind me of the profound interconnectedness of life and healing.