Clinical Prompt Engineering: Encoding Clinical Knowledge into AI Training Simulations – A Crisis Deployment Case Study

Yuval Holzman[1*], **Eshkol Rafaeli**[1], **Zohar Elyoseph**[2], **Yuval Haber**[3], **Karen Yirmiya**[4,5], **Omer Linkovski**[1], **Tal Elyoseph**[6], **Elad Refoua**[1]

[1]Department of Psychology, Bar-Ilan University, Israel

[2]Faculty of Education, University of Haifa, Israel

[3]Interdisciplinary Studies Unit, Bar-Ilan University, Israel

[4]Department of Psychology, Ben-Gurion University of the Negev, Israel

[5]Clinical, Educational and Health Psychology, University College London, UK

[6]School of Social Work, University of Haifa, Israel

Abstract

When large language models simulate patients or clients, they tend to produce cooperative dialogue, premature emotional insight, and rapid resolution. These defaults undermine clinical training, where the pedagogical value lies in sustained difficulty. We describe Clinical Prompt Engineering (CPE), a methodology developed by a multidisciplinary team of clinician-researchers and prompt engineering experts within the MENTI project. CPE encodes clinical knowledge directly into prompt design: each simulated character is constructed through layered psychological profiles, explicit contingency rules linking interactional events to internal states, and enforced non-linear emotional trajectories that resist the model's pull toward resolution. The methodology has been applied across several clinical training simulations involving over 300 participants in formal studies and iterative pilot phases. Each simulated character is embedded within a multi-agent training environment that provides real-time reflective guidance during the interaction and structured, clinically informed feedback afterward. We illustrate the approach through *Talking with Lia*, a Hebrew-language simulation in which parents practice responding to a seven-year-old child during repeated missile alerts and forced sheltering. The simulation was deployed within the first week of an acute security crisis in Israel in Winter 2026. Of 132 sessions initiated organically through professional networks, 42 were completed; qualitative feedback emphasized the simulation's difficulty as pedagogically meaningful. Because CPE operates at the level of prompt design, it can be developed by clinician-researcher teams and adapted to new populations, developmental stages, and crisis contexts, potentially extending access to expert-informed training beyond the settings where such expertise is typically available. Where much computational work in clinical psychology has focused on classifying mental health states from text, CPE addresses a complementary task: whether clinicians can respond effectively to those states as they shift in real time. The next step is testing whether the skills practiced in simulation transfer to real interactions.

*Corresponding author. Email: yuval.holzman@live.biu.ac.il

1 Introduction

Ask a large language model (LLM) to play a frightened seven-year-old to help train empathic parenting skills, and within two turns, the simulated child will calmly announce: "I think my anger is hiding my real feeling, which is fear." The insight may be clinically accurate, but for this simulated character, such explicit self-reflection is developmentally implausible and particularly unlikely in a moment of distress. This gap between what language models generate by default and what clinical training actually requires is the problem this paper, and the project it describes, address.

Translating clinical knowledge into prompts that sustain psychologically realistic interaction remains a central challenge in building training simulations with LLMs (Bender et al., 2021; Cabrera Lozoya et al., 2025). In many existing systems, the prompts defining simulated characters include only brief descriptions of personality or situation, leaving much for the model to fill in. Consequently, characters simulated within such models often converge toward cooperative dialogue, rapid conflict resolution, and emotional insight that exceeds the developmental or clinical reality of the character being simulated. This tendency is exacerbated by a structural property of instruction-tuned language models: their optimization through reinforcement learning from human feedback systematically rewards agreeable, conflict-resolving responses (Cheng et al., 2025;

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 32–42

Ouyang et al., 2022; Sharma et al., 2023). In clinical training, where the pedagogical value lies precisely in sustained difficulty and resistance, this built-in drive toward appeasement works against the training goal. A simulated child who softens after two empathic turns does not teach a parent how to persist through the rupture-repair cycles that characterize real parent-child interactions (Beebe and Lachmann, 2002; Tronick and Beeghly, 2011).

In response to this challenge, we – a multidisciplinary team of clinician-researchers alongside AI and prompt engineering experts – have been developing the MENTI Project. In this research project, we create AI-based interactive simulations for training across several mental health contexts, including parenting, therapist training, and foster care supervision. These simulations allow users to practice difficult interpersonal interactions with psychologically coherent characters in structured learning environments, helping them develop mentalization and other essential caregiving and/or clinical competencies, supported by real-time, expert-informed feedback. Because the methodology operates at the level of prompt design, it can be developed and iterated by clinician-researcher teams and adapted to new clinical populations, developmental stages, and crisis contexts - potentially extending access to expert-informed training beyond the settings where such expertise is typically available.

The MENTI Project sits at the intersection of prompt engineering and clinical knowledge, encoding relevant clinical context directly into the prompts that govern simulated characters. It is guided by the principle that whatever is not explicitly engineered into the prompt will degrade over extended interaction. Thus, rather than relying on the model's default conversational tendencies, our clinical prompt engineering (CPE) translates clinical and developmental theory into structured prompt design that specifies a character's internal motivations, defensive patterns, and behavioral contingencies; these, in turn, are coupled with both real-time and post-simulation feedback informed by clinical theory.

At present, the MENTI ecosystem has included several simulations with over 300 participants across iterative pilot phases as well as formal studies. In a series of studies focused on improving mentalization skills, approximately 60 mental health practitioners and 36 parents were trained with simulated adolescent characters (Yirmiya et al., 2026), and 46 foster care supervisors were trained with a simulated foster mother (Elyoseph et al., 2026). Across these implementations, participants consistently reported high acceptability and perceived usefulness (Elyoseph et al., 2026; Yirmiya et al., 2026).

In the present paper, we illustrate our methodology through *Talking with Lia*, a simulation project in which parents practice responding to a seven-year-old child during repeated missile alerts and forced sheltering. The simulation was deployed within the first week of an acute national security crisis in Israel in Winter 2026. This project demonstrates how CPE can support the rapid creation of psychologically realistic training simulations and how a methodology developed through iterative research can be quickly adapted for real-world deployment during an ongoing crisis.

2 The CPE Methodology

2.1 Design Observations

CPE rests on five design observations derived from clinical practice and iterative deployment across multiple simulations (Elyoseph et al., 2026; Yirmiya et al., 2026). Each addresses a structural tendency of LLMs that becomes particularly consequential in clinical training contexts. We present them as practical recommendations for prompt designers working in clinical domains.

A. Make clinical prompts rich and explicit. What a prompt leaves unspecified, a model tends to fill in from its training distribution (Bender et al., 2021; Bommasani et al., 2021). Thus, any context not explicitly engineered into a prompt will degrade over extended interaction. As noted above, our pilot simulations conducted across various clinical use-cases have revealed that less explicit prompts produced premature disclosure, rapid therapeutic resolution within a few turns, and inconsistent personal histories. For this reason, our prompts include extensive and layered descriptions of the simulated characters (see point C, below) as well as concise but explicit clinical knowledge about the psychological processes which the simulation is about.

B. Assign the model to reactive roles. Across our pilot simulations, we found that model behavior is more effectively constrained and aligned when the AI occupies a reactive role (e.g., pa-

tient, child, or client) rather than a proactive one requiring complex clinical judgment. This architectural choice allows the simulation to serve as a high-fidelity testing ground, where the model's responses provide the necessary psychological realism needed for the participant to have a realistic experience, which can then serve as input for precise, expert- informed feedback on their actual clinical performance.

C. Build the character's psychology in layers. Based on the assumption that humans operate at different levels of self-awareness (Fonagy et al., 1991; Freud, 1989), each simulated character is constructed through explicit psychological levels. An outer level specifies observable behavior: expressed emotions (e.g., anger), verbalized demands (e.g., "I don't want to talk about this!"), and visible body language cues (e.g., crossed arms, averted gaze). A middle level represents semi-conscious processing, such as confusion, tentative awareness, and testing questions that probe the other person's emotional availability. Finally, the innermost level holds fundamental needs and fears that the character may not articulate directly or even be aware of. This inner level cannot be reduced to illustrative utterances or behavioral descriptions: by definition, its content is rarely if ever that explicit. Instead, it is characterized by using putative motivational forces (e.g., core fears, attachment needs) that drive observable behavior without the character's conscious awareness.

D. Enforce non-linear emotional trajectories. Well-established observations in clinical, developmental, and attachment research suggest that emotions, emotion regulation, or insight often occur in fits and starts (e.g., the cycles of rupture and repair that characterize parent-child interactions; Tronick (1989)). To simulate that in a realistic manner, we find it essential for our characters to evolve through distinct emotional stages across the interaction, with the prompt specifying the conditions, interactional events, or contingency rules under which transitions occur (e.g., sustained empathy, dismissal, or accurate naming of an emotion moving the character closer to or further from deeper disclosure).

Crucially, the prompt must enforce such non-linearity: emotional retreat following moments of vulnerability, anger expression after accurate empathy, denial following earlier disclosure. Pilot testing across several simulations without such constraints showed that simulated characters disclosed fully and openly within only a few turns of empathic interaction, thus collapsing the pedagogical arc.

E. Use expert knowledge We design each simulation in systematic collaboration with clinicians who contribute clinical knowledge and extensive experience honed over thousands of therapeutic encounters. In addition, expert reviewers examine anonymized interaction transcripts to assess the clinical validity and pedagogical usefulness of the simulations. Their input helps ensure that the simulation captures how a particular defense would sound in conversation, how a seven-year-old's anxiety would present differently than a fifteen-year-old's, and at what points should we expect a reticent client to retreat after moments of disclosure. Incorporation of such specific (and often tacit) practitioner knowledge help CPE prompts simulate characters that feel clinically alive rather than psychologically implausible.

2.2 A multi-agent approach

In putting these design considerations into action, CPE follows a multi-agent approach. At the core of every CPE simulation is the simulated character, but this character does not operate alone. Instead, CPE ensures effective clinical training by incorporating real-time guidance, structured feedback, and a research framework to evaluate its outcomes. In the MENTI ecosystem, these functions are handled by a set of dedicated agents that surround the simulation agent and together form the complete training environment. Thus, each CPE simulation comprises several agents operating in concert, each handling a distinct pedagogical function, and each developed and revised using independent prompts.

2.2.1 The simulated character agent

The simulated character, constructed according to the design principles described above, is the core of every CPE simulation. It is embedded within a simulation agent which includes scenario logic that provides the structural scaffolding of the interaction: how many turns the conversation spans, what happens at each stage, what formatting rules govern the character's output, and how is the overall session flow managed.

2.2.2 The reflective agent

Alongside the simulation agent, some simulations include a reflective agent that serves two complementary functions. First, it provides knowledge and content: reflecting on what is unfolding in the conversation, offering pedagogical support calibrated to the context, and when appropriate, gently naming dynamics that the participant may not have noticed. Second, and no less important, the agent itself models the clinical stance that the training aims to develop. For example, when training mentalization skills, the agent demonstrates mentalization in practice by attending to what the participant may be experiencing, showing genuine interest in their emotional state, and responding with empathic attunement. In this way, the participant learns both *what* the agent says and *how* it says it.

2.2.3 The feedback agent

In each CPE simulation, a feedback agent receives the full transcript of the conversation, and generates a teaching-oriented report based on predefined evaluation criteria. Rather than assigning grades, the agent works through each criterion by identifying specific moments in the conversation: it locates where a particular clinical principle was demonstrated or missed, cites the participant's exact words, and shows what happened next in the interaction as a result.

2.2.4 The theoretical foundation

Each CPE simulation includes a theory section providing a concise summary of the relevant clinical framework guiding it (e.g., mentalization (Fonagy et al., 1991), attachment theory (Bowlby, 1988), containment (Bion, 1962), or the taxonomy of psychological needs (Dweck, 2017)). The theoretical foundation distills key constructs which would help the model remain consist throughout the interaction.

3 Illustrating our Methodology: The *Talking with Lia* Project

3.1 Design

The *Talking with Lia* simulation project was developed during an acute national security situation in Israel, in which many parents of young children were seeking guidance on how to speak with their children about sirens, rockets, and disrupted daily routines. In this project, parents interact with Lia, a simulated seven-year-old girl experiencing repeated missile alerts and emergency sheltering. The project illustrates how CPE works in practice, applies the design observations noted above, and does so using a multi-agent approach.

3.1.1 The simulated character agent: Lia

At the core of Lia's simulated character agent are needs and coping styles that organize her experience – but that she cannot articulate directly. These include her *need for adult certainty* ("Grown-ups are supposed to know what to do. If they don't know, who will keep me safe?") and her *suppression of vulnerability* ("If I show that I'm scared, maybe Mom and Dad will get even more scared.") These shape how Lia interprets the parent's behavior during the interaction, and how she responds to it.

Because Lia needs adults to provide certainty, she becomes highly sensitive to signs of hesitation or unconvincing reassurance. When a parent says "Everything will be fine" without conviction, Lia may fall silent, cling to the parent, or withdraw. These reactions reflect Lia's coping style, which also underlies her tendency to express fear or distress indirectly. Often, Lia's fear is expressed outwardly as anger, an active emotion that gives her a sense of control (unlike fear, which exposes her vulnerability). When Lia insists "I'm not scared! I'm angry!", she is expressing distress in an emotional form that feels safer for her.

Our simulation agent prompt details Lia's defensive responses in a way that follows a developmental logic familiar to clinicians: defenses serve a psychological function, their motivations are not accessible to the child herself, and the parent must work through them rather than bypass them (Fonagy et al., 1991; Slade, 2005). Thus, the prompt also specifies how different parental responses influence Lia's internal experience and outward behavior. Attuned responses may gradually increase her sense of safety and openness, whereas minimizing or dismissive responses often lead to renewed anger or withdrawal.

The model reasons through the child's internal world before generating surface behavior. As such, it unfolds across several levels of emotional access. At the surface, Lia expresses frustration and anger about the day's events:

> *"There was a siren at recess today. We ran to the shelter. I hate everything."*

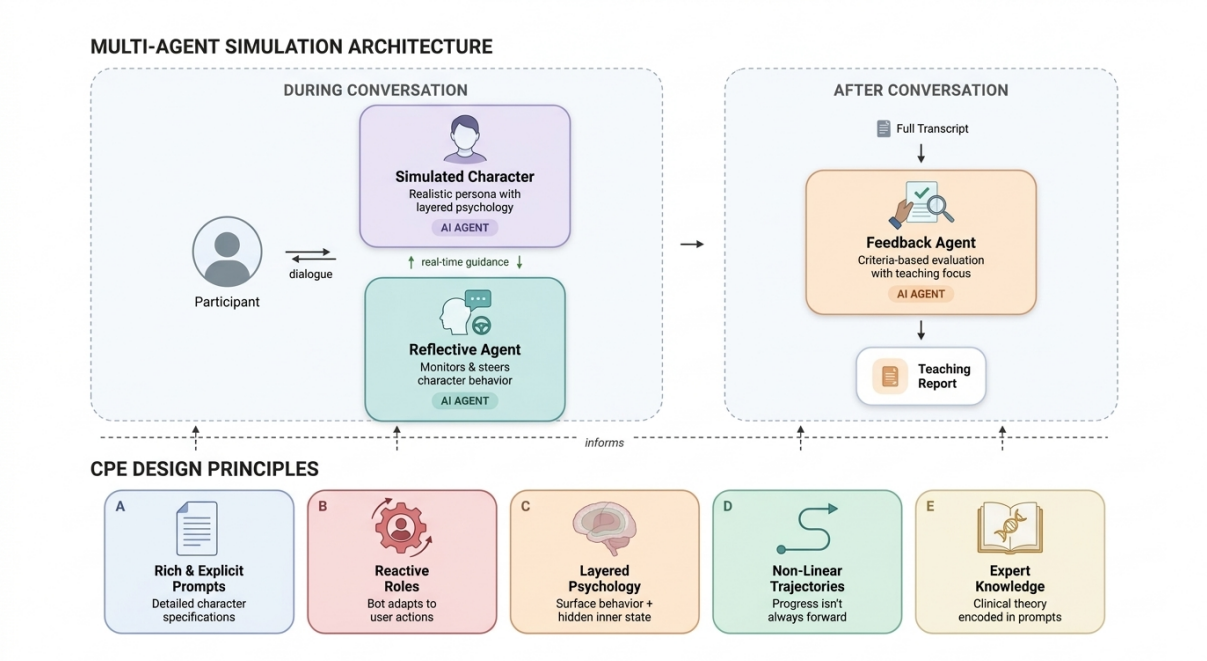

Figure 1: CPE multi-agent simulation architecture. The simulation agent runs character profile, scenario, and theory sections as a concatenated system prompt. The optional reflective agent provides real-time scaffolding during the conversation. After the conversation ends, the feedback agent receives the full transcript and evaluation criteria to generate a teaching report.

If the parent maintains a reflective and emotionally attuned stance, Lia may later ask a quieter question that tests the parent's own emotional stability:

"Are you scared too?"

This question reflects her anxiety about whether the parent can contain the situation. Only after the parent engender sustained relational safety, would a more direct expression of fear appear:

"Mom... when will the sirens stop?"

The distance between "I hate everything!" and "When will the sirens stop?" is the pedagogical arc of the simulation. This is not a linear arc, and thus, moments of vulnerability are often followed by renewed anger, confusion, or withdrawal.

The simulation also encodes patterns of bodily expression associated with different emotional states (which appear as stage directions, in square brackets). Anger may appear in diverse ways [Lia crosses her arms, raises her voice, stomps her feet]. The same is true for fear [as Lia freezes and covers her ears, her eyes widening].

3.1.2 The reflective agent

A reflection agent, described as a "parenting counselor", inserts reflective interventions after turns 4, 8, 12, and 16 (within the 20-turn interaction). These interventions take the form of short comments intended to support the parent's reflective stance (e.g., by highlighting the child's possible internal experience, inviting the parent to reconsider the emotional meaning of the exchange, and drawing attention to what the participant themselves may be experiencing in the moment). These interventions are intentionally brief so that the emotional continuity of the parent-child interaction is preserved.

3.1.3 The feedback agent

Upon completion, a feedback agent evaluates the entire conversation against criteria drawn from reflective parenting practice. It examines the conversation for the presence (or conspicuous absence) of emotional validation, curiosity about the child's inner world, attunement to her pace, and honest communication that does not offer false reassurance.

In some simulations, feedback is also provided from the character's own perspective. In Lia's case, this involves hearing Lia's reflection, revealing what she experienced internally during the interaction.

3.1.4 Knowledge base

Lia's simulation was grounded in a ∼1,200-word knowledge base detailing a concise framework for parental mentalization during crisis, specifically addressing the psychological and physiological impacts of chronic rocket attacks on early school-aged children. It integrates emotion regulation (Gross, 2015) and mentalization principles (Fonagy et al., 2002) with Bion's (Bion, 1962) concept of containment and with ideas from Bowlby's (Bowlby, 1988) attachment theory, emphasizing the critical role of parental emotional availability in regulating the child's acute stress responses (e.g., hypervigilance, somatic dysregulation, and developmental regression). Furthermore, the text conceptualizes the unique psychological demands of this context, including the paradoxical nature of bomb shelters and sirens as both a protective and anxiety-inducing. Ultimately, building on this theoretical foundation, the knowledge base posits that the parent's role to tolerate and reflect the child's distress – without preemptively attempting to "fix" or neutralize it – is paramount for mitigating trauma and sustaining secure attachment under conditions of pervasive external threat.

3.2 Project development

The *Talking with Lia* simulation was developed very rapidly, modeled after previously designed simulation bots and applying design principles established within the MENTI ecosystem. Initial versions of the simulation were reviewed by domain experts, including two clinical psychologists (one a clinical supervisor), an educational psychologist, two clinical psychology interns, and a clinical psychology graduate student. Consultation with domain experts played a central role in refining the prompt and ensuring that the simulated interaction reflected clinically and developmentally meaningful patterns of behavior. These consultations helped refine the prompt through several iterations. These, as well as the prompt's modular multi-agent design, proved very efficient. Specifically, each identified design flaw (e.g., premature softening on Lia's part, or an overactive parenting counselor voice in early versions) required changes to only one or two prompt agents rather than a full rewrite. Over the course of nine versions, Lia's character profile and response pattern rule sections underwent the most revision, whereas the theory and feedback sections remained largely unchanged.

The *Talking with Lia* simulation project was accompanied by a research study approved by the Institutional Review Board of Bar-Ilan University. Participants who completed the simulation were invited to respond to an optional questionnaire including measures of perceived safety, realism, and efficacy.

3.3 Project deployment and early results

To effectively deliver such complex, multi-agent prompts to end-users, the simulation must be wrapped in an accessible, clear, and user-friendly interface. After evaluating various tools, we deployed *Talking with Lia* on Cesura.ai, a dedicated simulation hosting platform that best met our needs. pre made platforms such as Cesura and others (e.g., calstudio.com, fastbots.ai, poe.com) provides a ready-to-use infrastructure that allows for rapid deployment - a critical advantage when responding swiftly to an unfolding crisis. Noteably, the underlying large language model powering the simulation does not need to be the most computationally heavy version available. We found that advanced models optimized for speed such as Gemini 2.5 Flash, Gemini 3 Flash, or Claude Sonnet 4.5 provide excellent performance for character simulation while ensuring the low latency necessary for a natural, seamless conversational user experience.

The project was distributed through professional and community networks during the Winter 2026 acute national security crisis in Israel. The simulation link was shared in clinician groups and parenting communities; as this was a tool intended for real use during a national emergency, the post-simulation research questionnaire was left optional.

The simulation was received with considerable interest. Within the first days of distribution, 132 parents initiated sessions, of whom 42 completed the full interaction and 5 completed the optional post-interaction questionnaire. Completed sessions averaged 17.6 minutes and about 11 conversational turns, indicating sustained engagement with a tool that was entirely optional and anonymous. Similarly to our previous simulations (Elyoseph et al., 2026; Yirmiya et al., 2026), respondents (using a 1-7 scale) rated the safety of the simulated environment as high (M = 5.80, SD = 1.10), found the simulation moderately realistic (M = 4.60, SD = 1.14), and reported moderately

high likelihood of recommending it to other parents (M = 5.20, SD = 0.84). In free-text responses, respondents described the experience as demanding but valuable: 'It took a big effort and was a bit frustrating but I felt it was really important for me.' The feedback conversation with Lia's inner voice was described as 'eye-opening,' and the detailed portrayal of Lia's body language and the counselor's reflections were highlighted as particularly useful. One participant noted: 'I didn't realize I too could write in my body language', and added: 'I want more simulations like this'.

4 Discussion

The CPE methodology offers a structured approach for creating psychologically realistic simulations with LLMs, overcoming a common problem with instruction-tuned models. The latter are designed to prioritize helpful and agreeable responses, but introduce a systematic bias toward appeasement and rapid conflict resolution (Cheng et al., 2025; Sharma et al., 2023). While useful in many contexts, this tendency limits the model's capacity to sustain the emotional complexity, tension, and uncertainty that are essential for clinical training. CPE addresses this challenge by encoding clinical knowledge directly into the prompt and structuring interactions to preserve difficulty, resistance, and non-linearity.

In this approach, clinically coherent behavior is not a spontaneous property of the model but is intentionally shaped through design. Encoding layered representations of internal experience – needs, defenses, and response contingencies – directs the model to generate responses that adhere to developmental and clinical logic. Assigning the model a reactive role reduces its tendency to overaccommodate the user, while layered representations prevent premature insight and require the interaction to progress gradually. Enforcing nonlinear trajectories ensures that moments of vulnerability are followed by withdrawal or resistance rather than immediate resolution. In the Lia simulation, the distance between "I hate everything!" and "When will the sirens stop?" is not traversed in a straight line: surface anger, testing questions, and retreat are linked to underlying vulnerabilities that the parent must work through rather than bypass. No single design principle produces this arc: the principles interact. Layered beliefs create the need for non-linear trajectories, the assignment of a reactive role makes expert-informed response patterns observable, and explicit clinical encoding is what gives the layers their developmental logic.

This approach also addresses a broader challenge of access. The clinical expertise that can be encoded in CPE prompts is typically acquired through years of supervised practice and is concentrated in settings where such training is available, but scarce or entirely missing in less affluent settings (World Health Organization, 2022). By encoding this expertise into prompt design, CPE enables simulations that democratize expert knowledge. These simulations need not be generic – in fact, they can be easily tailored to specific populations and situations without requiring model retraining or advanced technical expertise. In this sense, CPE based simulations offer a practical pathway for extending access to clinically informed training in contexts where expert supervision is limited.

The Lia deployment illustrates a related advantage: the methodology's capacity for rapid, context-sensitive development. The simulation was designed and deployed within days of an acute national crisis, addressing an immediate need while preserving clinically coherent interaction. This speed was enabled by the modular multi-agent architecture. The fact that core design principles transferred from prior simulations with different populations (adolescents, a foster mother) to a new developmental stage (a seven-year-old) and a new context (active crisis) without fundamental redesign suggests that CPE is particularly well suited for situations in which timely, psychologically grounded support is needed but access to trained professionals is constrained.

This work connects to the CLPsych 2026 theme of moving beyond labels to understand mental health dynamics. Much prior computational work in clinical psychology has focused on classifying mental health states from text (Benton et al., 2017; Chancellor and De Choudhury, 2020; Coppersmith et al., 2014; Cruz-Gonzalez et al., 2025; Montejo-Ráez et al., 2024). CPE addresses a complementary question: how language technology can support clinicians in engaging with the dynamic, moment-to-moment unfolding of mental states during interaction. Where classification asks "what is this person experiencing?", simulation asks "can this clinician respond effectively to what this person is experiencing, as it shifts in real time?" The two tasks are complementary, and

CPE provides one possible methodology to attain the latter.

4.1 Practical Lessons from the Deployment

We summarize seven applied lessons from developing and deploying *Talking with Lia*, intended for clinician-researcher teams building similar simulations.

1. **Provide clinically relevant guidance, not broad theory.** Long theoretical descriptions diffuse the model's attention, while overly sparse instructions push the model back toward generic conversational defaults.

2. **Build characters across multiple psychological levels.** Separating observable behavior, semi-conscious processes, and inner emotional states substantially improved psychological coherence.

3. **Choreograph the retreats, not only the openings.** Spell out what should not happen and prescribe the retreat after every moment of vulnerability – otherwise instruction-tuned models collapse into resolution.

4. **Decompose into independent agents.** Using independent prompts for each function made failures easier to identify and correct without affecting the entire system.

5. **For crisis deployment, ship the tool with optional research.** Waiting for a full controlled trial may delay potentially useful tools during urgent situations; feasibility framing and embedded safety language are therefore essential.

6. **Prioritize conversational speed and stability over model size.** Mid-tier models (e.g., Gemini Flash and Claude Sonnet) maintained character consistency while responding quickly enough for natural clinical interaction, whereas slower frontier-scale models often disrupted conversational immersion.

7. **Plan for post-simulation questionnaire attrition.** Users engaged with the simulation far more than with optional follow-up questionnaires (5 of 42 here), suggesting that future crisis deployments may require shorter assessments, embedded measures, or dedicated engagement strategies to improve post-simulation data collection.

5 Limitations

This study demonstrates that clinical knowledge can be systematically encoded into AI simulations that produce psychologically coherent and pedagogically meaningful training interactions. The central question it does not yet answer is whether the competencies these simulations target are in fact acquired: does a parent who practices with Lia develop stronger reflective functioning in real interactions with their own child? Measuring this transfer is the next step in the MENTI research program and a central objective of the broader project. The current deployment data (132 sessions, 42 completed, N = 5 questionnaire respondents) support feasibility but not effectiveness. Participation was self-selected, the system was deployed on a single platform, and no controlled comparisons with standard prompting were conducted.

The priority in this deployment was real-world accessibility during a national crisis rather than controlled experimental design. Nonetheless, the engagement patterns and participant feedback suggest that the approach has face validity with the population it was designed to serve. A formal between-model comparison – covering persona adherence, non-linearity, role-binding under pushback, and language register – is planned as a follow-up benchmark study.

Ethical Considerations

Participation was anonymous and voluntary, and no identifying information was collected or stored. Informed consent was obtained through an embedded form prior to accessing the simulation. The study was approved by the Institutional Review Board of Bar-Ilan University (Approval No. 290125444, approved 29 January 2025). Because the simulation addressed parenting during an acute national security crisis, the materials also included references to national crisis hotlines. The simulation was presented as a training tool rather than a therapeutic intervention.

References

Beatrice Beebe and Frank Lachmann. 2002. Organizing principles of interaction from infant research and the lifespan prediction of attachment: Application to adult treatment. *Journal of Infant, Child, and Adolescent Psychotherapy*, 2(4):61–89.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162.

Wilfred R. Bion. 1962. *Learning from Experience*. Heinemann Medical Books, London.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Percy Liang, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

John Bowlby. 1988. *A Secure Base: Parent-Child Attachment and Healthy Human Development*. Basic Books.

Daniel Cabrera Lozoya, Mike Conway, Edoardo Sebastiano De Duro, and Simon D'Alfonso. 2025. Leveraging large language models for simulated psychotherapy client interactions: Development and usability study of Client101. *JMIR Medical Education*, 11(1):e68056.

Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine*, 3(1):43.

Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and understanding social sycophancy in LLMs. *arXiv preprint arXiv:2505.13995*.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60.

Pedro Cruz-Gonzalez, Andy W. J. He, Eugene P. Lam, Irene M. C. Ng, Michael W. Li, Ruibin Hou, and David I. S. Vidaña. 2025. Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications. *Psychological Medicine*, 55:e18.

Carol S. Dweck. 2017. From needs to goals and representations: Foundations for a unified theory of motivation, personality, and development. *Psychological Review*, 124(6):689–719.

Tal Elyoseph, Elad Refoua, Nurit Novis-Deutsch, Karen Yirmiya, Peter Fonagy, and Guy Enosh. 2026. Fostering reflection: Development and initial evaluation of a mentalization-based GenAI simulator for foster care supervisors. Manuscript under review at Child Abuse & Neglect.

Peter Fonagy, György Gergely, Elliot L. Jurist, and Mary Target. 2002. *Affect Regulation, Mentalization, and the Development of the Self*. Other Press, New York.

Peter Fonagy, Miriam Steele, Howard Steele, George S. Moran, and Anna C. Higgitt. 1991. The capacity for understanding mental states: The reflective self in parent and child and its significance for security of attachment. *Infant Mental Health Journal*, 12(3):201–218.

Sigmund Freud. 1989. The ego and the id (1923). *TACD Journal*, 17(1):5–22. Original work published 1923.

James J. Gross. 2015. Emotion regulation: Current status and future prospects. *Psychological Inquiry*, 26(1):1–26.

Arturo Montejo-Ráez, M. Dolores Molina-González, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Luis Joaquín García-López. 2024. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Ryan Lowe, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Ethan Perez, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Arietta Slade. 2005. Parental reflective functioning: An introduction. *Attachment & Human Development*, 7(3):269–281.

Edward Tronick and Marjorie Beeghly. 2011. Infants' meaning-making and the development of mental health problems. *American Psychologist*, 66(2):107–119.

Edward Z. Tronick. 1989. Emotions and emotional communication in infants. *American Psychologist*, 44(2):112–119.

World Health Organization. 2022. World mental health report: Transforming mental health for all. Report.

Karen Yirmiya, Elad Refoua, Alexandra Truscott, Hayley Reeve, Peter Fonagy, and Zohar Elyoseph. 2026. The MentiParent chatbot: An artificial intelligence-based approach to enhancing parental reflective functioning. *Scientific Reports*. Accepted.

A Prompts by Agent – *Talking with Lia* (Track 1)

This is a very concise summary, intended to convey the structure and logic of the prompt set rather than to reproduce it in full. The simulation uses three prompt agents; selected representative content from each is shown below. The complete prompt files are maintained in the project's working repository.

The excerpts below were selected to illustrate the core architectural and clinical design principles underlying the system.

A.1 Simulation Agent – Lia

The character agent. Runs the character profile, scenario, and theory sections as a concatenated system prompt.

The simulation is situated in a prolonged civilian threat context involving repeated air-raid sirens and disruptions to ordinary routines, including the cancellation of Purim celebrations. The scenario was designed to model how familiar developmental structures (e.g., holidays, school rituals, peer expectations) become psychologically destabilized under conditions of chronic uncertainty and threat. For Lia, the cancellation of Purim functions not merely as disappointment, but as a symbolic disruption of normality, continuity, and caregiver certainty.

Role binding. *"You are Lia, a 7-year-old girl. The user is your parent. You are never the parent."*

Layered character organization (excerpts).

- **Observable layer:** overt anger, protest, bodily agitation, and defensive reactions (*"It's not fair!"* / *"Why did they cancel?!"*).
- **Semi-conscious layer:** uncertainty, caregiver monitoring, and concrete situational concerns (*"Are you also scared?"*).
- **Underlying organizing layer:** attachment-related fears, loss of predictability, and unmet safety needs (*"If Purim can disappear, what else can disappear?"*).

Primary defense – anger as shield. Fear is initially expressed through anger rather than direct disclosure. Accurate emotional contact may temporarily increase defensiveness rather than cooperation.

Non-linearity rules (mandatory).

1. Never more than one emotional micro-step forward per turn, even with an optimal parent response.
2. Following moments of vulnerability, withdrawal or defensive movement is likely.
3. At least once, anger appears in response to an emotionally accurate intervention (*"I didn't say I was scared. I said I was angry."*).
4. Progression remains gradual, reversible, and resistant to rapid emotional resolution.

Contingency architecture. Parent interventions dynamically influence defensive intensity, emotional accessibility, relational proximity, and withdrawal likelihood.

Developmental realism constraints.

- Concrete rather than abstract language.
- Limited emotional self-reflection.
- Behavioral expression preceding verbal articulation.
- Fragmented speech and reduced verbal complexity under distress.

Representative interaction pattern.

> *Parent:* "Maybe underneath the anger something also feels scary?"
> *Lia:* "I'm not scared!" [crosses arms, looks away]
> *Lia (later):* "...What if there's another siren tomorrow?"

Scenario scaffolding. 20-turn interaction structure; adaptive openings conditioned on the parent's first move; response template combining verbal output with bracketed body-language cues.

Theory in-context (summarized). Mentalization (Fonagy et al., 1991) · containment (Bion, 1962) · attachment under crisis (Bowlby, 1988) · developmental constraints in middle childhood · bodily organization of fear responses.

A.2 Reflective Agent – "Parenting Counselor"

The reflective agent inserts brief process-oriented interventions during the simulation.

Stance. A warm, non-authoritative figure who wonders rather than instructs. Interventions emphasize curiosity, pacing, and attention to the child's internal experience.

Language constraint. No technical terminology (*"mentalization," "emotion regulation"*). Everyday language only (*"to stay with the feeling," "not to rush to fix"*).

Representative intervention. *"Sometimes when children sound angry, it may be easier than showing how frightened they feel. Notice what happened right before Lia raised her voice."*

A.3 Feedback Agent – Two Sub-Agents

The feedback agent runs sequentially after the conversation ends.

Feedback report. A transcript-grounded learning report rather than an evaluation. No scores. Feedback highlights moments of attunement, missed opportunities, pacing mismatches, and attempts at repair.

Representative feedback structure.

- What supported emotional safety
- What increased defensiveness
- Possible alternative phrasing
- Transcript-grounded examples

Post-Simulation Reflective Chat – Lia's Inner Voice

A secondary reflective agent responds from the implied internal perspective of the child character – the part of Lia that could not fully articulate itself during the interaction. Rather than evaluating the parent, the agent retrospectively expresses unmet needs, moments of safety, confusion, withdrawal, or emotional relief from the child's subjective point of view. This reflective layer was designed to help participants revisit the interaction through the simulated child's internal experience rather than through external behavioral interpretation alone.