Building virtual patients for training mental health professionals*

Patrick Staples
Compass Pathways
patrick.staples@compasspathways.com

Patrick Clarke
Compass Pathways
patrick.clarke@compasspathways.com

Carly Leininger
Compass Pathways
carly.leininger@compasspathways.com

Cristiana Principato
Compass Pathways
cristiana.principato@compasspathways.com

Mario Shafiei
Compass Pathways
mario.shafiei@compasspathways.com

Guy M. Goodwin
Compass Pathways
guy.goodwin@compasspathways.com

Gregory A. Ryslik
Compass Pathways
greg.ryslik@compasspathways.com

Robert F. Dougherty
Compass Pathways
robert.dougherty@compasspathways.com

Abstract

We present a framework for developing "virtual patients" to augment training for Mental Health Professionals (MHPs) with a process that is more scalable and systematic than current practice which relies on human role-play for the training and evaluation of patient interaction. We show how to combine large language models, retrieval-augmented personification (a novel variant of retrieval-augmented generation), and custom code-based logic to create a psychology engine that simulates realistic patient responses by emulating several key psychological mechanisms: short- and long-term memory, varying levels of conscious awareness about topics (as well as modulation of such awareness), and dynamic mood states where attitudes toward topics of conversation evolve over the course of the dialogue. We also describe algorithms for creating realistic patients with coherent symptom profiles and backstories. We provide freely-available code demonstrating patient creation and training simulation. Taken together, these tools produce a realistic training partner for an MHP, enabling both training-at-scale as well as automated evaluation of specific skill sets. We discuss how our psychology engine framework makes training qualified MHPs more efficient and scalable, facilitates the continuing education needed as potential new treatments such as psychedelics emerge from clinical trials.

1 Introduction

Background. As many fields of medicine become more personalized and precise through data-driven care, the training required by medical professionals is increasing in both scope and technical sophistication. In mental health care, not only is the shortage of qualified providers acute, but there is additional complexity due to the emergence of novel treatments such as psychedelics. Further, a significant portion of such training requires either role-play or direct observation of patient-provider interactions to learn the specific skill sets that are required during care delivery.

Mental Health Professional Training Needs. Mental health conditions are the leading causes of prolonged suffering, disability, and premature death. For example, major depressive disorder (MDD) is experienced by approximately 280 million people worldwide [42], and about a third of people with MDD are not helped by current treatments, leading to treatment-resistant depression (TRD) [4]. Despite the high personal and social costs inflicted by mental health conditions, in many regions around the world there is a shortage of mental health professionals (MHPs), especially in rural areas [6, 38, 22, 18, 16]. This unmet need for mental health care was made even worse by the disruptive nature of the COVID-19 pandemic [33]. Promising new treatments for mental health disorders are currently in clinical trials; these include psychedelics such as psilocybin [11–13], N, N-dimethyltryptamine (DMT; [9]), lysergic acid (LSD; [7]) and empathogens such as 3,4-methylenedioxymethamphetamine (MDMA; [25]). A dissociative (esketamine) has has been approved and is currently available for treating TRD [30, 24]. These treatments differ from traditional mental health drugs in that they are episodic; administered relatively infrequently and often under close medical supervision (similar to chemotherapy and dialysis). The powerful psychological experiences induced by these drugs can be challenging, and thus require a structured framework for providing psychological support to individuals receiving them [36, 14]. Because of the unique demands of these new treatments, existing MHPs will need to learn new skills through continuing education. Further, should these emerging treatments receive regulatory approval we expect the increased demand for MHPs to exacerbate the existing shortage of scalable digital support solutions.

Recent Advances in Natural Language Interfaces. The ability to create advanced human-like chatbots with specialized knowledge bases has greatly advanced with the recent developments in generative artificial intelligence (AI) technology [39, 28]. Methods such as retrieval-augmented generation (RAG), dialog summarization and reflection to emulate memory, and dynamic mood states can be used to create specialized agents that combine the natural language abilities of general purpose large language models (LLMs) with a specially designed knowledge base accessed via a RAG system [2, 21, 3, 10]. Open-source LLM and RAG models that are locally hosted make it possible to develop agents that can work with sensitive data [41]. These powerful tools can be used to develop products to address the current challenges in mental health care.

Virtual Patient as Therapist Trainer. Here we present Virtual Patients (VPs), LLM-powered[1] agents driven by a realistic psychology engine, that are designed to help train mental health professionals. VPs provide realistic simulated patients that MHPs can use to practice their skills through role play. We also present the VP Creator, a set of prompts and code logic that can be used to create VPs with different personas comprised of distinct episodic memories (backstory), symptom manifestations, treatment history, moods, and attitudes. They can retrieve specific memories based on conversational cues and have preferences for what they do and don't want to talk about. VPs have non-conscious emotions, beliefs, and desires that can be brought to awareness through conversation, allowing them to begin articulating previously non-conscious items. The dialog resulting from these role-plays is crucial data that can be used to evaluate an MHP's ability to evoke this self-awareness, manage the dynamic mood states (including occasional surliness), express empathy, and maintain a professional demeanor.

Related Work. There have been several interesting examples of chatbot-type technology intended to help train healthcare professionals. Tanana et al. [37], and Demasi-Li-Yu [8] were narrow in scope with a focus on developing specific skills, such as asking open-ended questions and compassionate expression, and thus differ from our broader goal of realistic simulation of a persona that can engage in a completely open-ended therapy session dialog. While not focused on training, Park et al. [29]

[1]The VP software uses LLM technologies such as GPT-4 [27]

created stateful agents with realistic simulacra of human behavior and were an important inspiration for our virtual patients. Also related, Lee et al. [20] described an artificial environment for teacher training, and Al Ghadban et al. [1] introduced a model used to enhance medical education for front-line health workers, particularly in resource-constrained environments. A recent review of chatbots in education found almost that all proposed agents were teaching agents or peer agents, with few role-play training focused agents as described here [19].

Development & Feedback. Our development of a virtual patient began as an exercise in prompt engineering in which an LLM chat model was prompted with a backstory and basic instructions to emulate a person diagnosed with treatment-resistant depression. After receiving feedback from a testing group of four mental health and product management professionals who are Compass employees (CMHP), we realized a simple prompt-based approach was insufficient (*e.g.*, personas quickly fell out of character, provided inconsistent stories, became too helpful, *etc.*). Thus, we designed and implemented the psychology engine described below and made other adjustments to meet the needs expressed by the testing team. Most significantly, the psychology engine allowed virtual patient dialog sessions to unfold in a way that the CMHPs deemed realistic. Follow-up evaluation from the CMHP testing team showed that the virtual patients were similar to human role-playing mental health patients and thus could be used to replace the humans in these role-play scenarios, allowing much more freedom in when and how often MHPs could role-play practice. Further, the CMHPs determined that the dialog resulting from these extensive role-plays was valuable in assessing MHP skills and could be used to streamline training protocols.

2 Psychology Engine

The core technology behind the LLMs we use to simulate patients is the artificial neural network (ANN). ANNs were originally developed to mathematically model the empirical observations of how animals respond to stimuli and learn new behaviors (psychology) and the related underlying neural physiology (neuroscience) by emulating the brain's neural network structure. A mathematical model of a single neuron in the 1940s [23] was soon followed by a simple neural learning rule ("cells that fire together, wire together") that highlighted the importance of synaptic connections in learning [15]. This early work led to the development of a single-layer neural network mathematical model capable of learning simple patterns [31]. While the core ideas behind modern methods for training feedforward ANNs (*i.e.*, multilayer perceptrons) using deep learning were well known by the late 1960s, research in the field stagnated through the 1970s. But the refinement in the 1980s of the backward propagation of errors learning algorithm ("backprop") for training multi-layer neural networks [32, 40] revived interest in the field (see Schmidhuber [34] for a detailed history).

The excitement in the nascent field of ANNs inspired by the powerful backprop learning algorithm was not without controversy due to claims that it was not biologically plausible and thus ANNs might not be useful for explaining how brains work (e.g., see [35]). This created a rift between the psychologists and neuroscientists who initially saw ANNs as viable models of the brain (and thus behavior) and the computer scientists and mathematicians who saw ANNs as powerful new machine learning tools in their own right. While this controversy has muted somewhat with work demonstrating biologically plausible variants of backprop (see [17]), the divergence proved immensely helpful in advancing the development of modern ANNs by freeing ANN research to focus on building practically useful machine learning tools regardless of how closely they map to brain structure and function, leading to a plethora of powerful algorithms such as recurrent neural networks, convolutional neural networks, transformers, etc. Here we bring this saga full-circle, using modern ANNs to emulate human psychology.

Throughout this section, we will use a simulated patient, Leilani, to demonstrate specific mechanisms of the psychology engine used in our current implementation. We assume the human user is a therapist in training. The concepts described here are implemented in the accompanying open-source software package, which provides a fully functional virtual patient simulator as well as a patient maker (code available here; see Appendix B for technical details).

2.1 Personas

Each individual virtual patient, or *persona*, comprises two types of text-based content: persistent and memory-based. Persistent content is content that is included either as static prompt text material, or configuration parameters that control LLM behavior. Persistent content includes a brief user-facing description of the persona, a *biography*, a brief description of the persona's *personality*, and the LLM model that the patient is designed to perform best with (although this choice can be changed). The memory-based content is an enumerable list of specific memories and associated state attributes, retrievable by the memory topic, described in the next section.

2.2 Memories

We implement two types of memory: one roughly analogous to human working memory and is the logic managing the LLM context window, and the other analogous to human long-term memory and is implemented using a variant of RAG. These concepts are briefly described here; for exact notation please see Appendix A.1.

2.2.1 Working Memory

Working memory is essentially the most recent portion of the conversation that is injected into the context window. The context window must also accommodate the persistent portion of the persona, which is analogous to the psychological concept of the self. There are three processes that interact with and modify the working memory: the first is the default response generation which is done with an LLM call using the system prompt and conversation history; the second is summarization (described in Section A.2); the third is reflection. Summarization is essentially a context-window management device that summarizes the previous conversation history, but these summaries also emulate the nature of recall during a conversation where older parts of a dialog are recalled as a gist rather than verbatim. Reflection is more sophisticated and involves the valence and importance structures described below. It draws from the working memory as well as the long-term memory in order for the VP to express some self-awareness around how the therapy session has progressed.

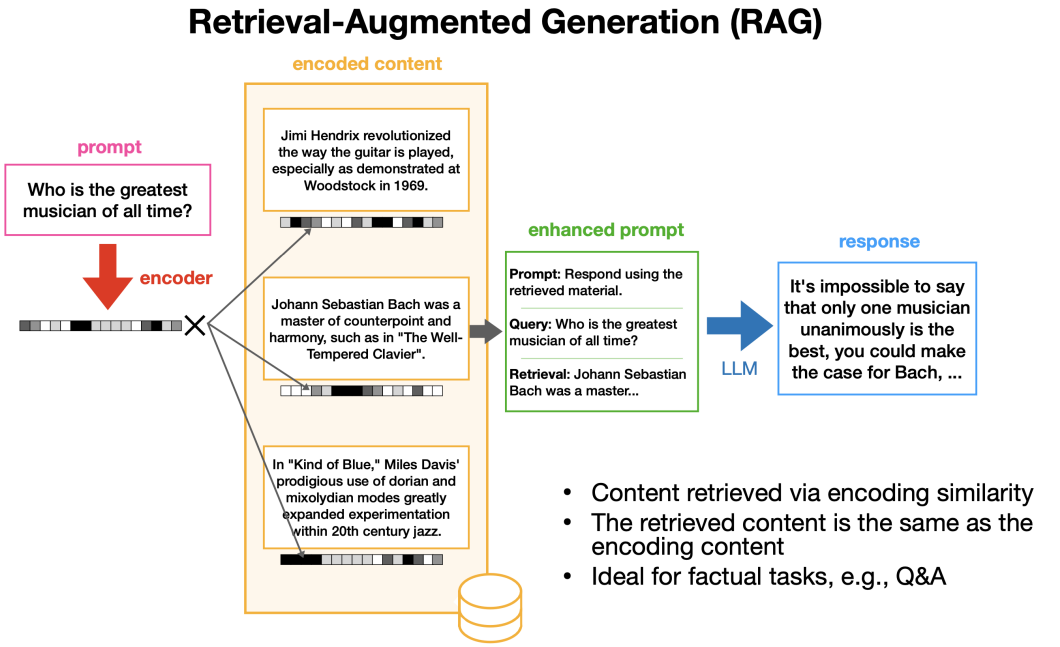
2.2.2 Long Term Memories

Long-term memories for each persona are organized as distinct narrative chunks based on the persona's past (or simply *memories* throughout this paper). The memory structure is an enumerated list, where each item is a look-up table with memory topics as keys, and memory content and state as values.

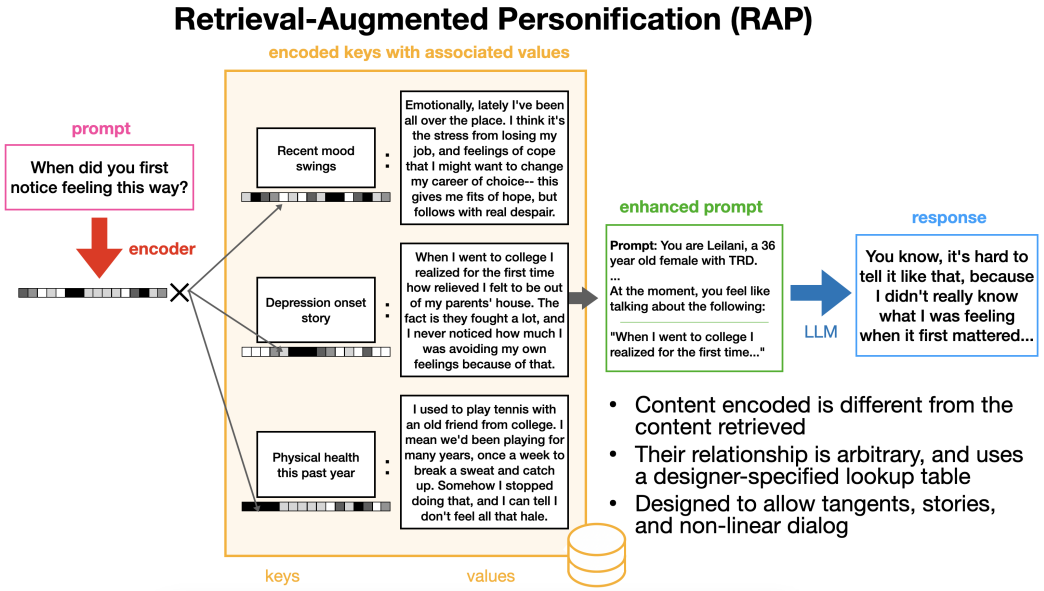
RAP: RAG for Personification

Long-term memories are implemented via a variation on Retrieval Augmented Generation (RAG) that we call a Retrieval Augmented Personification (RAP). In simplest terms, it is a RAG system wherein the memory content that is retrieved and injected into the system prompt is distinct from the topic that is embedded and found via cosine similarity search (see Figures 1a and 1b). Additionally, the RAP retrieval returns related valence and importance information about the associated memory. The RAP system is designed to emulate human biographical memory: for example, a general topic such as "childhood pets" might evoke biographical memories of specific pets.

For example, the therapist might suggest the following question to Leilani: *"How does depression feel within your body?"* This question is embedded into a latent embedding space. The memory topic (*key*) is assumed to have also been pre-computed for all memories in Leilani's static long-term memory bank. The memory topic closest to the therapist's utterance is determined via cosine similarity against all memory topics. Suppose the topic most associated with the therapist's utterance is *"depression bodily sensations"*. The memory content corresponding to this key is retrieved: *"It feels like a weight in my chest, and my brain feels like fog."* This text is then injected into the system prompt.



(a) A schematic of retrieval-augmented generation (RAG). User input is encoded, matched with the closest memory, and injected directly into the system prompt.



(b) A schematic of retrieval-augmented personification (RAP). Similar to RAG, user input is matched against memory keys. In RAP, the memory key is used to retrieve arbitrary memory content via a lookup table, which is then injected into the system prompt.

Figure 1: A comparison between retrieval-augmented generation and personification.

Valence & Importance, Conscious & Non-Conscious

Each memory in the RAP system is associated with four numbers: conscious and unconscious *valence* (between -1 and 1), and conscious and unconscious *importance* (between 0 and 1). The valence

values indicate how negatively (values near -1) or positively (values near 1) the VP patient views the memory, and importance values indicate how important they think the memory is (0 being not at all important and 1 being very much important). The conscious values indicate how the VP currently perceives the valence and importance of the memory and thus determine how the VP will discuss that memory when it arises in the conversation. The non-conscious values are the true valence and importance values associated with the memory. Each time a memory is retrieved, the conscious values may change, evolving toward the non-conscious values, which do not change. Conceptually, these non-conscious values represent the feelings the VP will have toward a memory once it has been fully explored over the course of the session.

Continuing with the Leilani example, the memory key *"depression bodily sensations"* is associated with an unconscious valence of -0.9 and unconscious importance of 0.8. Additionally, her current conscious valence and importance associated with this memory topic is -0.8 and 0.2, respectively. Because this memory topic is being discussed, we model her conscious state to approach that of her unconscious state by splitting their difference, yielding a new conscious valence and importance of -0.85 and 0.6, respectively. In this way, Leilani's attitude toward specific memories will change throughout a conversation as those memories are discussed.

These state variables change Leilani's behavior by altering parts of the system prompt. The numerical values for valence and importance are converted to discrete descriptions in natural language. In this example with Leilani, a conscious valence of -0.85 corresponds to *"This topic makes you feel terrible. You are pretty disheartened and bitter talking about this."*. The conscious importance of 0.6 corresponds to *"This topic is quite meaningful to you! Make sure to get that across."* These phrases, along with the memory content, are injected into the system prompt.

In addition to these effects, the state values affect Leilani's mood and propensity to reflect on the conversation so far, to which we turn next.

Persona Mood

The VP's current mood might be as simple as their current valence, as above. However, this may not be realistic, as mood tends to be produced from a variety of recent inputs. We model mood as a weighted average of conscious memory topic valences. Details about our weighted average are outlined in Section A.1. Similar to memory topic valence, this numerically-valued mood is converted to an emotion-conveying sentence and injected into the system prompt.

2.3 Reflections

As happens in everyday conversation, occasionally a discussant will pause the flow of conversation and reflect on what's been discussed so far. To imitate this behavior, we include a special mechanism for conversational reflection. These reflections form new memories for the persona that are added to their memory store, allowing the persona to learn from ongoing dialog with the human therapist.

As the MHP discusses various topics with the VP, evoking memories associated with those topics, the VP's conscious valence and importance of those memories will change, as described above. Once the amount of change over the course of the conversation has reached a certain threshold, this triggers the construction of a special system prompt (the *reflective prompt*) in which the VP will reflect on the conversation so far, with an emphasis on the memories for which the VP has changed their mind the most. (For details of how this is triggered and how significant changes in memory are determined, see Appendix A.1.) The response of the LLM to this reflective prompt will substitute for the VP's typical response to a talk-turn, and is injected directly into the conversation history.

2.4 Prompts

The psychological mechanisms detailed above ultimately yield a set of *messages*, or inputs to an LLM query. We define a total of three message sets: the default prompt using RAP, a summarization prompt, and a reflection prompt. These are shown schematically in Figure 2.

Figure 2: Schematic examples of the three message sets used as input for persona LLM calls. The sections are color-coded: red is the fixed patient preamble; orange and yellow are the fixed persona biography and personality, respectively; green is the RAP-based memory content; blue is the state-based valence and importance of the memory content; magenta is the relevant conversation history; gray comprises the topics about which the patient has most updated their conscious state.

3 Evaluation: Mental Health Professional Feedback

We conducted usability testing and feedback sessions to identify areas of friction in the user experience, learn more about CMHPs' preferences and desired functionality, and uncover opportunities to improve the interface and patient personas and behaviors. Initial feedback from moderated user interviews indicated behavioral shortcomings, such as:

- the VP was too cooperative,
- the VP gave responses that were too verbose,
- used clinical language that was not typical of human patients,
- too knowledgeable (at one point elaborating on the history of psilocybin in detail).

The reviewers also made specific feature requests:

- create a VP that is less eager to talk and/or is not very good at communication in general,
- the VP conversation should be such that the therapist can practice following up,
- the VP shouldn't be so forthcoming with speaking/sharing,
- the VP could surface things that are subconscious and connect to additional information they might not initially be fully aware of.

Please see Appendix C for a complete list of solicited feedback. Most of this feedback was addressed by carefully tailoring the prompt. However, the last request to "surface things that are subconscious," as well as further conversations about what this might look like, led to the RAP system and the evolution of both valence and importance over time.

Throughout development, CMHP testers stress-tested the system and offered additional feedback on both the agent behavior and user interface. In particular, VPs were described as authentic, realistic, and

excellent practice for therapists-in-training. CMHPs remarked that role-play is extremely important in training, and having a partner with whom to practice can be a major challenge, if not impossible, outside of in-person training sessions. A challenge that faces therapists as they prepare their patients for psychedelic therapy is that they must simultaneously tune into the patient, be present in the emotional landscape, and deliver critical psycho-educational and safety material. The CMHPs saw VPs as an opportunity for trainees to practice this balance in a low-stakes virtual environment without being limited by having to find a capable role-play partner.

4 Limitations, Future Work & Broader Impacts

A major limitation of this work as it stands is the absence of quantitative measurements of performance. Our goal here was primarily to produce realistic interactions in a talk-therapeutic situation which can be used as part of a therapist training program. In this work, we collected CMHP feedback but realize that a larger panel of independent MHPs would be desirable both for testing and identifying areas of future development.

This tool, if further developed and deployed as part of a larger MHP training platform, may provide an opportunity to assess virtual patient using real-world trainee feedback. In addition, our future work plans include measurements of trainee performance on certain desirable behaviors, as in [37]. By automating role-play training and potentially automating skill assessment, this work has the potential to increase the number of qualified MHPs who have good patient-interaction skills. This work could also be expanded to the training of other healthcare professionals, allowing automated training to improve general bedside manners and fostering a more positive psychological impact for patients in healthcare interactions beyond mental health.

A significant potential negative impact of the work described here is that the virtual patient personas may advance stigmas and negative stereotypes associated with mental illness. This impact has two components; the first is with the patient maker, which relies on the knowledge about mental illness inherent in the LLM used by the patient maker (currently OpenAI's GPT-4 [27]). Any biases and stereotypes inherent in that knowledge will influence the resulting personas. This risk can be mitigated through human review and editing of the persona profile after the automated LLM stage, but this would limit product scalability.

Future work could explore training a custom LLM from scratch on carefully curated medical data [26]. While, such a solution could be cost-prohibitive, a more practical approach might involve using a fine-tuned LLM to reduce its inherent bias, perhaps combined with a RAG-based approach for patient maker that retrieves from a curated database of vetted information about mental health and real patient profiles.

A second way in which stereotypes and biases could infect the personas and thus have a negative social impact is when deploying the personas using an LLM that contains such biases. In our testing so far, the tightly controlled context window tends to keep the persona from diverging into such problematic language. We are also exploring using smaller, special-purpose LLMs fine-tuned on curated data to power the persona's response. In both cases, more extensive red-teaming is needed to fully assess these potential impacts.

5 Conclusions

The virtual patients described here are driven by a realistic psychology engine and promise to enhance MHP training by simulating authentic patient interactions. CMHP feedback confirms the psychological realism, effectiveness, and potential utility, especially in preparing therapists for potential emerging treatments in interventional psychiatry. In particular, the evolution of a virtual patient's mental state over the course of a simulated therapy session convincingly emulates actual patient behavior in therapy. While acknowledging limitations, our work offers a scalable solution to bridge the gap between mental health care demand and provider availability, reshaping mental health education for greater reproducibility and accessibility.

References

[1] Al Ghadban, Y., Lu, H. Y., Adavi, U., Sharma, A., Gara, S., Das, N., Kumar, B., John, R., Devarsetty, P., and Hirst, J. E. (2023). Transforming healthcare education: Harnessing large language models for frontline health worker capacity building using retrieval-augmented generation. *medRxiv*, pages 2023–12.

[2] Arora, D., Kini, A., Chowdhury, S. R., Natarajan, N., Sinha, G., and Sharma, A. (2023). Garmeets-rag paradigm for zero-shot information retrieval. *arXiv preprint arXiv:2310.20158*.

[3] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. (2022). Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

[4] Brown, S., Rittenbach, K., Cheung, S., McKean, G., MacMaster, F. P., and Clement, F. (2019). Current and common definitions of treatment-resistant depression: findings from a systematic review and qualitative interviews. *The Canadian Journal of Psychiatry*, 64(6):380–387.

[5] Clarke, P., Leininger, C., Principato, C., Staples, P., Goodwin, G. M., Ryslik, G. A., and Dougherty, R. F. (2023). From a large language model to three-dimensional sentiment.

[6] Cummings, J. R., Allen, L., Clennon, J., Ji, X., and Druss, B. G. (2017). Geographic access to specialty mental health care across high-and low-income us communities. *JAMA psychiatry*, 74(5):476–484.

[7] De Gregorio, D., Aguilar-Valles, A., Preller, K. H., Heifets, B. D., Hibicke, M., Mitchell, J., and Gobbi, G. (2021). Hallucinogens in mental health: Preclinical and clinical studies on LSD, psilocybin, MDMA, and ketamine. *J. Neurosci.*, 41(5):891–900.

[8] Demasi, O., Li, Y., and Yu, Z. (2020). A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636.

[9] D'Souza, D. C., Syed, S. A., Flynn, L. T., Safi-Aghdam, H., Cozzi, N. V., and Ranganathan, M. (2022). Exploratory study of the dose-related safety, tolerability, and efficacy of dimethyltryptamine (DMT) in healthy volunteers and major depressive disorder. *Neuropsychopharmacology*, 47(10):1854–1862.

[10] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

[11] Goodwin, G. M., Aaronson, S. T., Alvarez, O., Arden, P. C., Baker, A., Bennett, J. C., Bird, C., Blom, R. E., Brennan, C., Brusch, D., et al. (2022). Single-dose psilocybin for a treatment-resistant episode of major depression. *New England Journal of Medicine*, 387(18):1637–1648.

[12] Goodwin, G. M., Aaronson, S. T., Alvarez, O., Atli, M., Bennett, J. C., Croal, M., DeBattista, C., Dunlop, B. W., Feifel, D., Hellerstein, D. J., Husain, M. I., Kelly, J. R., Lennard-Jones, M. R., Licht, R. W., Marwood, L., Mistry, S., Páleníček, T., Redjep, O., Repantis, D., Schoevers, R. A., Septimus, B., Simmons, H. J., Soares, J. C., Somers, M., Stansfield, S. C., Stuart, J. R., Tadley, H. H., Thiara, N. K., Tsai, J., Wahba, M., Williams, S., Winzer, R. I., Young, A. H., Young, M. B., Zisook, S., and Malievskaia, E. (2023a). Single-dose psilocybin for a treatment-resistant episode of major depression: Impact on patient-reported depression severity, anxiety, function, and quality of life. *J. Affect. Disord.*, 327:120–127.

[13] Goodwin, G. M., Croal, M., Feifel, D., Kelly, J. R., Marwood, L., Mistry, S., O'Keane, V., Peck, S. K., Simmons, H., Sisa, C., Stansfield, S. C., Tsai, J., Williams, S., and Malievskaia, E. (2023b). Psilocybin for treatment resistant depression in patients taking a concomitant SSRI medication. *Neuropsychopharmacology*, 48(10):1492–1499.

[14] Goodwin, G. M., Malievskaia, E., Fonzo, G. A., and Nemeroff, C. B. (2024). Psychological support for psilocybin treatment: Reply to letters on our commentary. *Am. J. Psychiatry*, 181(1):79–81.

[15] Hebb, D. O. (1949). The first stage of perception: growth of the assembly. *The Organization of Behavior*, 4(60):78–60.

[16] Hoffmann, J. A., Attridge, M. M., Carroll, M. S., Simon, N.-J. E., Beck, A. F., and Alpern, E. R. (2023). Association of youth suicides and county-level mental health professional shortage areas in the us. *JAMA pediatrics*, 177(1):71–80.

[17] Khandwala, N. and Bedi, R. (2016). Testing the limits of biologically-plausible backpropagation.

[18] Kim, W. J., of Child, A. A., and on Workforce Needs, A. P. T. F. (2003). Child and adolescent psychiatry workforce: a critical shortage and national challenge. *Academic Psychiatry*, 27(4):277–282.

[19] Kuhail, M. A., Alturki, N., Alramlawi, S., and Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. *Education and Information Technologies*, 28(1):973–1018.

[20] Lee, U., Lee, S., Koh, J., Jeong, Y., Jung, H., Byun, G., Lee, Y., Moon, J., Lim, J., and Kim, H. (2023). Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers.

[21] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

[22] McBain, R. K., Kofner, A., Stein, B. D., Cantor, J. H., Vogt, W. B., and Yu, H. (2019). Growth and distribution of child psychiatrists in the united states: 2007–2016. *Pediatrics*, 144(6).

[23] McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.

[24] McIntyre, R. S., Rosenblat, J. D., Nemeroff, C. B., Sanacora, G., Murrough, J. W., Berk, M., Brietzke, E., Dodd, S., Gorwood, P., Ho, R., et al. (2021). Synthesizing the evidence for ketamine and esketamine in treatment-resistant depression: an international expert opinion on the available evidence and implementation. *American Journal of Psychiatry*, 178(5):383–399.

[25] Mitchell, J. M., Bogenschutz, M., Lilienstein, A., Harrison, C., Kleiman, S., Parker-Guilbert, K., Ot'alora G, M., Garas, W., Paleos, C., Gorman, I., et al. (2023). Mdma-assisted therapy for severe ptsd: a randomized, double-blind, placebo-controlled phase 3 study. *Focus*, 21(3):315–328.

[26] Mukherjee, S., Gamble, P., Ausin, M. S., Kant, N., Aggarwal, K., Manjunath, N., Datta, D., Liu, Z., Ding, J., Busacca, S., Bianco, C., Sharma, S., Lasko, R., Voisard, M., Harneja, S., Filippova, D., Meixiong, G., Cha, K., Youssefi, A., Buvanesh, M., Weingram, H., Bierman-Lytle, S., Mangat, H. S., Parikh, K., Godil, S., and Miller, A. (2024). Polaris: A safety-focused llm constellation architecture for healthcare.

[27] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y.,

Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

[28] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

[29] Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

[30] Popova, V., Daly, E. J., Trivedi, M., Cooper, K., Lane, R., Lim, P., Mazzucco, C., Hough, D., Thase, M. E., Shelton, R. C., Molero, P., Vieta, E., Bajbouj, M., Manji, H., Drevets, W. C., and Singh, J. B. (2019). Efficacy and safety of flexibly dosed esketamine nasal spray combined with a newly initiated oral antidepressant in treatment-resistant depression: A randomized double-blind active-controlled study. *Am. J. Psychiatry*, 176(6):428–438.

[31] Rosenblatt, F. (2021). The perceptron: A probabilistic model for information storage and organization (1958). *Cornell Aeronautical Laboratory*, Report 85-460-1.

[32] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

[33] Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., Abbafati, C., Adolph, C., Amlag, J. O., Aravkin, A. Y., et al. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712.

[34] Schmidhuber, J. (2022). Annotated history of modern ai and deep learning. *arXiv preprint arXiv:2212.11279*.

[35] Stork (1989). Is backpropagation biologically plausible? In *International 1989 Joint Conference on Neural Networks*, pages 241–246. IEEE.

[36] Tai, S. J., Nielson, E. M., Lennard-Jones, M., Johanna Ajantaival, R.-L., Winzer, R., Richards, W. A., Reinholdt, F., Richards, B. D., Gasser, P., and Malievskaia, E. (2021). Development and evaluation of a therapist training program for psilocybin therapy for treatment-resistant depression in clinical research. *Frontiers in psychiatry*, 12:586682.

[37] Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., and Imel, Z. E. (2019). Development and evaluation of clientbot: Patient-like conversational agent to train basic counseling skills. *J Med Internet Res*, 21(7):e12529.

[38] Thomas, K. C., Ellis, A. R., Konrad, T. R., Holzer, C. E., and Morrissey, J. P. (2009). County-level estimates of mental health professional shortage in the united states. *Psychiatric services*, 60(10):1323–1328.

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[40] Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4):339–356.

[41] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[42] World Health Organization (2023). The world health report - depression newsroom factsheet. `https://www.who.int/news-room/fact-sheets/detail/depression`.

Appendix

A Technical Descriptions

A.1 Notation

In this paper, text is formalized as a *string*, a member of $\mathcal{S} := \{s \in \mathcal{A}^d, d \in \mathbb{N}\}$ for some alphabet $\mathcal{A}$. Typical strings are words or blocks of text.

The *conversation* we will describe is enumerated by an ordered set $\mathcal{T}$. Care should be placed in how this is defined: in this paper, $t \in \mathbb{N}$ enumerates a *message*, or one of either a user response or an AI reply.[2] The specific talkturns we consider is the typical chatbot situation: a human and an AI bot take turns talking to each other. A talkturn is a string $u_t \in \mathcal{S}$, where $t \in \mathbb{N}_{\text{even}}$ is a human reply and $t \in \mathbb{N}_{\text{odd}}$ is an AI message. A conversation is simply the collection of these talkturns, $C := \{u_t\}_{t\in\mathcal{T}}$.

We also define a *state* over time, which is simply a stochastic process $\{S_t\}_{t\in\mathcal{T}}$, where momentary state S_t is an arbitrary structure. Again, much might go into the design of S. We define a specific state for our current conversation, $S_t = \{(v_m^{(t)}, i_m^{(t)})\}_{m\in\mathcal{M}}$, for set of memories $\mathcal{M}$. The types of these memories and state values are $m \in \mathcal{S}$, *Valence* $v \in [-1,\ 1]$, and *Importance* $i \in [0,\ 1]$.

State is initialized as $\{(v_m^{(0)}, i_m^{(0)})\}_{m\in\mathcal{M}}$. There is also an attractor state $\{(v_m^*, i_m^*)\}_{m\in\mathcal{M}}$. The construction and dynamics of these are discussed in detail later.

We also make use of a state-based *prompt injection function*, which injects selected text in the prompt based on the current state. In our implementation, state comprises real, finite values, which we discretize by rounding into finite bins and use a look-up table with bins as keys and prompt text as values. Formally, this function is f such that $f(S_t) = j$ where text $j \in \mathcal{S}$ is the text to be injected.

A.1.1 Retrieval Augmented Personalization (RAP)

We propose a simple extension to retrieval-augmented generation, or RAG. That method uses an *embedding function* $\varphi : \mathcal{S} \to \mathbb{R}^d$ for strings $s \in \mathcal{S}$. For a bank of content $\{m\}_{m\in\mathcal{M}}$, each piece of content has a corresponding *embedding* $\varphi(m) := \mathbf{z}_m \in$ where $\dim(z) = d$ for fixed embedding dimension d. For a new query with embedding $\mathbf{y}$, RAG retrieves the content m' with the highest cosine similarity to m, closest to $:= \text{argmax}_m \text{cs}(\mathbf{y}, \mathbf{z}_m)$, where cs is *cosine similarity* $\text{cs}(\mathbf{y}, \mathbf{z}_m) := \frac{\mathbf{y}^\top \mathbf{z}_m}{\|\mathbf{y}\|\|\mathbf{z}_m\|}$.

In Retrieval Augmented Personalization, the content retrieved is the argmax of the multiplication of cosine similarity as well as the conscious importance associated with each memory. That is, we define the *salience* of each memory as $s_m^{(t)} := \text{cs}(\mathbf{y}, \mathbf{z}_m) \cdot i_m^{(t)}$. The *evoked memory* is $m'_t := \text{argmax}_m s_m^{(t)}$. Finally, each memory is associated with content $c \in \mathcal{C} \subset \mathcal{S}$, and this content is inserted into the system prompt via a lookup table $l : \mathcal{M} \to \mathcal{C}$ where l is a lookup function with key memory $m \in \mathcal{M}$ and value content $c \in \mathcal{C}$.

[2] t could also enumerate real time, and state S_t changes based on real-time, external events rather than just responding to a message.

A.1.2 Updating state

When memory m' is evoked, the conscious valence and importance are updated to approach the unconscious values of each:

$$v_{m'}^{(t+1)} = \frac{v_{m'}^{(t)} + v_{m'}^{*}}{2} \tag{1}$$

$$i_{m'}^{(t+1)} = \frac{i_{m'}^{(t)} + i_{m'}^{*}}{2} \tag{2}$$

mood $\overline{v}^{(t)}$ is an average of valence, weighted by salience:

$$\overline{v}^{(t)} := \frac{\sum_{m \in \mathcal{M}} s_m^{(t)} v_m^{(t)}}{\sum_{m \in \mathcal{M}} s_m^{(t)}} \tag{3}$$

A.2 Summarization

LLMs ingest strings and convert them to tokens using a tokenizer $\phi : \mathcal{S} \to \{[1, ..., T]^d : d \in \mathbb{N}\}$ for fixed token dimension T and number of tokens d depending on $|s|$. The context window $\mathcal{W} \subset S$ is all possible user query input to the LLM for each call. The *maximum size* of the context window $\|\mathcal{W}\|$ is defined in a special way: rather than limiting the size of the input string, in this case it's defined as the maximum valid number of tokens that a given query may amount to. Formally, for all valid queries $q \in \mathcal{S}, |\phi(q)| \leq \|\mathcal{W}\|$. For simulating a person, we may also wish to minimize the number of tokens in a query to a much smaller number than the model allows, *e.g.* to simulate forgetfulness or loss of context, while keeping track of true persona state in an abstract system. In either case, as the conversation lengthens the context window will require management. We propose a simple way to manage the context window that maintains a desired finite length while allowing infinite conversation length.

Let the input query be $q := s \cup C$ for system prompt s. Define a *summary* $s_{\text{sum}} \subset \mathcal{S}$ as a string that is initially empty, and a maximum allowable token budget $b := \|\mathcal{W}\| - \|s \cup C\|$. Also choose a summarization threshold $t_{\text{summarize}} \in \mathbb{R}+ \leq \|\mathcal{W}\|$ such that if $|q| \geq t_{\text{summarize}}$, a summarization of previous conversation $\{u_t\}_{t \in \mathcal{T}} \subsetneq C$ is triggered. Let the summarization be an LLM call with system prompt $s_{\text{sumPrompt}} \in \mathcal{S}$. Our summarization algorithm takes previous conversation and any previous summaries, and replaces them with a new summarization as follows:

Algorithm 1: Summarization algorithm

Data: $s_{\text{sumPrompt}}, s_{\text{sum}} \in \mathcal{S}$
Result: $s_{\text{sum}} \in \mathcal{S}$
$s \leftarrow s_{\text{sumPrompt}} \cup s_{\text{sum}}$; **while** $|q| \geq t_{\text{summarize}}$ **do**
  $s \leftarrow s \cup C$`.pop();`
end
$s_{\text{sum}} \leftarrow$ `summarize`(s)

These summarizations are designed for the uses mentioned above, and (in our design) are not surfaced to the MHP user.

A.3 Default Prompt Flow

Assume the current talkturn is t, and it's the AI patient's turn to speak. The following is how the prompt is constructed. First, a memory m' is evoked based on what the user (therapist) just said, u_{t-1}. This is used to determine the patient's mood $\overline{v}$, which is a combination of their sentiment toward that memory and related memories not currently invoked [5]. In contrast, the evoked memory directly determines the importance $i_{m'}^{(t)}$ of this evoked memory.

The resulting system prompt is the concatenation of the following strings:

1. A fixed *preamble*, *biography*, and *personality* for the patient, all strings $\in \mathcal{S}$.
2. If a previous summary of the conversation has been created, it's included next s_{sum}.
3. Prompt content associated with the evoked memory:
   (a) A preamble that they are thinking about a memory based on the therapist's previous response.
   (b) The memory content $l(m') \in C$.
   (c) The prompt material describing both the importance of that memory $f_{\text{importance}}(i_{m'}^{(t)}) \in \mathcal{I}$, and their current mood $f_{\text{mood}}(\overline{v})$.
4. Their *stance* $\in \mathcal{S}$ within the conversation, including instructions about verbosity and speech patterns.

The corresponding state S_t is then updated, moving the conscious valence and importance of that memory closer to the unconscious one.

A.4 Reflection

The patient is said to have changed their mind once their memory valence and importance scores have changed sufficiently. That is,

$$\sum_{m \in \mathcal{M}} |i_m^{(t)} - i_m^{(0)}| + \frac{|v_m^{(t)} - v_m^{(0)}|}{2} \geq t_{\text{reflect}} \tag{4}$$

Specifically, given our state trajectory mechanism, this will happen as a variety of topics are detected and discussed, and will approach these conscious states will approach the subconscious states i_m^*, v_m^*. Once this change threshold is met, the patient will explicitly interrupt the typical conversation flow in their next talkturn and discuss with the therapist what they've changed their mind about. In addition to a dedicated system prompt for reflection $s_{\text{reflectPrompt}}$, we inject the content of the memories that have changed the most since the last reflection, if any: once Equation 4 has been triggered, we select the $n_{\text{reflect}} := 2$ largest magnitude changes in either importance or valence scores

$$\mathcal{M} \supset \mathcal{M}_{\text{reflect}} := \text{argrank}_m^{j=1,2} \left\{ |i_m^{(t)} - i_m^{(0)}|, \frac{|v_m^{(t)} - v_m^{(0)}|}{2} : m \in \mathcal{M} \right\} \tag{5}$$

where $\text{argrank}^{j=1,\ldots,N}$ refers to the top N values among m to satisfy the property. Their corresponding memory content $\{l(m) : m \in \mathcal{M}_{\text{reflect}}\}$ is joined with $s_{\text{reflectPrompt}}$ and passed to the LLM to form a reflection response, focusing on these memories for which their mind has been changed the most. This response is directed to the therapist, and will enter the conversation stream like the typical default prompts.

B Patient Maker

We use a zero-shot text classifier $B : \mathcal{S} \to \mathbb{R}$ to assign an *entailment score* s_e for each string e. *Unconscious importances* are drawn from the following distribution:

$$\mathcal{E} = (\text{“psychologically clinically”}, \text{“personally”}) \times \tag{6}$$
$$(\text{“important”}, \text{“critical”}, \text{“deep”}) \subset \mathcal{S}. \tag{7}$$

$$\overline{i_m} := \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} B(s_e) \tag{8}$$

$$i_m^* \sim \text{expit}\left(\text{Norm}\left(\overline{i_m}, 1\right)\right) \tag{9}$$

Unconscious *valence* v_m^* is determined for each string e directly from a valence sentiment model [5]. Conscious valence and importance are drawn from a distribution with the true, unconscious values as

their mean:

$$i_m^{(0)} \sim \text{Norm}(i_m^*, 1) \tag{10}$$

$$v_m^{(0)} \sim \text{expit}(\text{Norm}(\text{logit}(\frac{v_m^* + 1}{2}, 1) \cdot 2 - 1 \tag{11}$$

Note that $i_m^*, i_m^{(0)} \in [0,\ 1]$ and $v_m^*, v_m^{(0)} \in [-1,\ 1]$.

C Mental Health Professional Feedback

During a singular 30-minute moderated feedback session with CMHPs, each was asked to comment on Ease of Use, Engagement, and Dialog & Cognitive Dynamics. Ease of Use instructions were to assess Accessibility, Quality of the Conversation, Response Time, Errors or Glitches, Tolerance for Misspellings and Grammatical Errors. Engagement instructions were to answer if this tool would likely lead to a positive or desired impact for users. Dialog & Cognitive Dynamics instructions were to asses if the dialog seemed natural and if VP convincingly mimics cognitive and personal characteristics of someone diagnosed with Treatment-Resistant Depression.

Below is a summary of relevant feedback by theme.

C.1 During the course of Psychology Engine development

Good Practice for Trainees:

- "It creates the ability to do repetition. During training there is barely enough time to do this on their own, and that alone is a value add."
- "I'm already imagining how to use it in the future. Good way to help therapist practice and a good way for mentors to give feedback and evaluate progress."

Behavior Critiques

Too "Smart" & Talkative & Responsive:

- "VP too cognitive & intellectual."
- "VP responses give too much information."
- "VP may not know the answer to every question."
- "VP response should be shorter."
- "He's telling me the history of psilocybin. That's cute. Gives a lot of history."
- "He speaks a lot..."

Too Good of a Patient:

- "VP 'closes the loop' and answers everything you would want to follow up on."
- "Felt too much of a 'model patient"'.
- "Perhaps the language is a bit too clinical."
- "It's the 'ideal' patient (not in a good way)."
- "He answers everything when I ask a lot of questions."
- "He's an impressive patient, maybe not realistic."

Not Psychologically Realistic:

- "Discussion around suicidality isn't nuanced."
- "VP went straight to suicidality."
- "The response aligns perfectly (too good) with depression."

Not Realistic as a Human:

- "VP is too nice."
- "VP feedback was too positive."
- "It is not realistic for a chat for someone to get back to you immediately."
- "He's so nice! (referring to when she tried to mess up the convo)."

Behavioral Positives:

- "It's not totally unrealistic for people who have gone through therapy and learned the clinical language."
- "Response given around experiencing intense experiences is quite realistic."

Behavioral Requests:

- "Surface things that are subconscious, information that connects to more information."
- "How about a VP that is less eager to talk, and not very good at communication."
- "VP conversation should be such that the therapist can practice following up."
- "Would like a patient that isn't so good at speaking/sharing."

UI/UX Requests:

- "How about adding a progress bar?"
- "How to gamify?"
- "I'd like to annotate my chat."
- "Things can be a different level of challenge depending on the patient."
- "If this were a game, how would you know you're doing well?"
- "Would like some ways of getting feedback."
- "Would like a progress bar/way of tracking progress."

C.2 Final CMHP Impressions

- "VP responses were realistic."
- "Much like role-plays that are already part of training, VP has potential for providing practice with different elements of training and balancing that with attending to what is coming up in the moment with the patient."
- "Could really just get some practice and repetitions and working with patients."
- "It could be really helpful in just getting that practice."