https://doi.org/10.36315/2026inpact023
Psychological Applications and Trends 2026

EMOTIONALLY RESPONSIVE VIRTUAL PATIENTS FOR PSYCHOTHERAPY TRAINING A PROOF-OF-CONCEPT USING FINE-TUNED LARGE LANGUAGE MODELS

Niels Netten[1], Thierry Desot[1], & Joël du Fosse[2]
[1]*Knowledge Centre Creating 010, Rotterdam University of Applied Sciences (The Netherlands)*
[2]*Erasmus School of Social and Behavioural Sciences, Erasmus University (The Netherlands)*

Abstract

Traditional soft-skills training in psychology and mental healthcare depends strongly on role-play with peers or actors. While valuable, it is resource-intensive, offers limited scenario diversity, and cannot fully capture the complexity of real patients. Consequently, students have insufficient exposure to realistic patient interactions. Virtual patients provide a scalable and flexible alternative, but existing systems often fail to produce consistent, realistic clinical interactions. As part of the TherAvatars project, this explorative study investigates the potential of large language models (LLMs) to generate interactive patient dialogues by fine-tuning open-source generative LLMs on therapist–patient conversations. We further examine retrieval-augmented generation (RAG) as a mechanism to enhance contextual grounding across dialogue turns. Model performance was evaluated using both expert assessments by psychology professionals and NLP-based similarity metrics. Our initial findings suggest that fine-tuned open-source LLMs have the potential to simulate virtual patients capable of supporting psychotherapy training, but substantial challenges remain in achieving robust dialogue structure, consistency, and long-term conversational coherence for autonomous clinical training scenarios.

Keywords: *Artificial Intelligence, psychotherapy training, LLM fine-tuning, evaluation.*

1. Introduction

Virtual patients have increasingly been explored in healthcare education as a means to provide safe, repeatable, and scalable experiential learning. Students can practice challenging clinical scenarios without risking patient safety, while benefiting from personalised learning paths (Kenny et al., 2007; Prescott et al., 2024; Tanana et al., 2019; Khan et al., 2023). However, many existing systems still rely on scripted dialogue and rigid turn-taking, resulting in limited emotional expressiveness and insufficient variation in communication style or personality (Combs & Combs, 2019). These limitations reduce their usefulness in psychotherapy training, where nuanced affective behaviour and interpersonal responsiveness are essential.

Recent advances in transformer-based large language models (LLMs) (Vaswani et al., 2017) offer new opportunities to enhance these simulations. When fine-tuned on domain-specific dialogues, generative LLMs can yield context-sensitive and emotionally congruent interactive patient behaviour (Yang et al., 2023). Fine-tuning on task-specific data aligns model outputs with desired conversational behaviours while preserving general language capabilities, making LLMs particularly suitable for interactive, role-consistent dialogue generation. Such models may complement traditional training modalities (lectures, case studies, role-play) by offering scalable, individualized practice opportunities for training diagnostic reasoning and therapeutic communication.

This study is part of the Theravatars exploratory research project, which aims to develop a prototype interactive virtual patient leveraging advanced language models. Earlier work within this project focused on semi-automatic sentiment corpus generation (Desot et al., 2025). Building on this foundation, the present study focuses on fine-tuning LLMs to simulate interactive patient behaviour in psychotherapy conversations. Unlike commercial APIs (e.g., ChatGPT), we use open-source LLMs that can be run locally, enabling full control over model behaviour and fine-tuning, flexible integration of custom datasets, reproducible experiments, and enhanced privacy for sensitive clinical data. Our proof-of-concept focuses on depression, a highly prevalent disorder with well-defined diagnostic criteria, making it suitable for modelling psychologically consistent virtual patient behaviour.

In this paper, we describe the fine-tuning and evaluation of three open-source LLM configurations that differ in architectural complexity and contextual grounding: a lightweight model (TinyLlama), a larger

transformer-based model (DeepSeek), and a retrieval-augmented generation (RAG) variant of TinyLlama. All models were fine-tuned on psychotherapy dialogues to generate patient responses. Model outputs were evaluated using human expert ratings, complemented by lexical and embedding-based NLP similarity metrics that compare generated responses to original patient utterances on a turn-by-turn basis.

2. Related work

Advances in NLP have expanded the potential for emotionally responsive virtual systems capable of addressing the limitations of traditional scripted virtual patients. Early virtual patient systems relied on rule-based question-answering and template matching (Epstein et al., 2013), but the transition to deep learning and transformer-based models (Vaswani et al., 2017) has enabled potentially richer contextual understanding and more robust emotion detection. Recent studies demonstrate that large language models can infer and express emotional states within virtual agents (Yang et al., 2023; Ng et al., 2023) and can identify linguistic markers associated with anxiety and depression for detecting a range of affective states (Tao et al., 2023; Ray, 2023). These insights support emerging efforts to automatically evaluate student–patient interactions through machine learning-based dialogue analysis.

Within the domain of healthcare, recent work has applied LLMs to clinical reasoning and mental health support (Omiye et al., 2024), yet these systems primarily emulate therapist-like behaviour rather than modelling the diverse, emotionally dynamic responses of patients. In contrast, our focus is explicitly on modelling patient-side dialogue behaviour, including affective variability, role consistency, and longitudinal coherence in psychotherapy sessions. As a result, a gap remains in creating virtual patients capable of psychologically consistent and adaptive emotional behaviour at the dialogue level, highlighting the need for methods that integrate emotion detection, corpus generation, and LLM-driven dialogue generation. Desot et al. (2025) explored zero-shot and semi-supervised techniques for generating affective training data in psychotherapy contexts and demonstrated that LLMs can automatically label emotional states in therapeutic dialogue. Combining zero-shot inference with human validation reduces annotation cost while maintaining reliability, enabling downstream fine-tuning for emotion-aware dialogues.

3. Data

The DAIC-WOZ (Distress Analysis Interview Corpus – Wizard of Oz) served in our research as the primary dataset, a publicly available multimodal dataset of 189 clinical interview sessions conducted by a virtual interviewer (Ellie) via a Wizard-of-Oz protocol. With synchronized audio, video, and transcripts, it provides natural interactions mimicking real clinical encounters, making it well suited for LLM-based virtual patient simulations (DeVault et al., 2014; Gratch et al., 2014). While previous research on DAIC-WOZ primarily focused on session-level depression assessment (e.g., PHQ-8 regression, binary classification, or multimodal fusion) often summarizing the session into a single score or set of features, our study operates at the utterance level, generating dialogue turn by turn, capturing moment-to-moment emotional and contextual changes.

4. Method: LLM fine-tuning and evaluation

Fine-tuning adapts a pre-trained generative LLM to a specific task by continuing training on a smaller, domain-specific dataset (Wang & Chen, 2023). This will guide the model to generate more accurate, helpful, and context-sensitive outputs for the intended application. However, these LLMs may still hallucinate, contradict themselves, or vary responses depending on context. For our experiments we used two open-source pretrained generative chat models:

- **TinyLlama (2024)**: A lightweight model designed for fast inference and low computational cost. Fine-tuning is performed using supervised learning with LoRA (Low-Rank Adaptation) adapters, which efficiently update only a low-rank decomposition of the model's weights rather than the full parameter set, significantly reducing memory and compute requirements. The model captures basic turn-taking and responses efficiently, but has limited capacity for complex reasoning or long-term context tracking.
- **DeepSeek (2025):** A Mixture-of-Experts (MoE) model using instruction-following and chain-of-thought reasoning capabilities. In MoE architectures, only a subset of specialized expert networks is activated for each input, allowing the model to scale capacity and reasoning ability without proportionally increasing computational cost. This higher effective capacity enables more nuanced emotional reasoning, consistent patient persona maintenance, and improved long-context coherence across multi-turn dialogues. Fine-tuning involved high-level behavioral prompts instructing the model to respond as a depressed veteran, combining role tracking with emotional

guidance.This architecture enables greater contextual fidelity and role consistency compared to smaller models.

In addition, we used a technique called Retrieval-Augmented Generation (RAG) (Gao et al., 2023) in combination with the fine-tuned TinyLlama model. At runtime, relevant prior dialogue utterances are retrieved dynamically and appended to the prompt as contextual evidence. This improves response relevance, reduces hallucination, and supports session-level continuity without increasing model size. RAG was applied only to TinyLlama and not to DeepSeek's MoE architecture, because its limited parameter capacity and context window make it more prone to context loss in multi-turn dialogue. Limiting RAG to TinyLlama allows us to isolate the effect of retrieval and assess whether it elevates a lightweight model toward the performance of larger architectures.

Modern LLM evaluation frameworks distinguish between closed-ended tasks, with clearly defined correct outputs, and open-ended dialogue generation, where valid responses are diverse and subjective. NLP metrics for open-ended tasks therefore primarily capture surface similarity. Our evaluation uses both syntactic and semantic NLP metrics and also human expert evaluation. Syntactic metrics include BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to quantify n-gram overlap between generated virtual patient responses and reference patient utterances in the held-out test set. For semantic evaluation we used BERTScore (Zhang et al., 2020), which compares contextual embeddings from a pretrained BERT transformer model with the model output and reference text, and reports precision, recall, and F1 to capture meaning similarity beyond surface tokens. Because NLP metrics often correlate weakly with human judgments in open-ended clinical dialogue, we treat syntactic and semantic metrics as complementary indicators rather than definitive measures of model quality. While BLEU, ROUGE, and BERTScore quantify surface-level or embedding-based similarity to reference responses, they do not directly assess clinical plausibility, or role coherence. Consequently, human expert evaluation serves as the primary basis for comparative assessment in this study.

Consistent with recent recommendations (Awasthi et al., 2025), we explicitly distinguish between turn-level and session-level assessment to capture both local and global conversational quality. Turn-level evaluation focuses on the appropriateness of individual responses to immediate prompts, while session-level evaluation reflects longitudinal properties such as coherence, role consistency, and information retention across multi-turn clinical interactions.

5. Experiments and results

Section 5.1 outlines model fine-tuning and evaluation setup of the experiments, with the results presented in Section 5.2.

5.1. Data preparation, fine-tuning and model performance rating

To prepare DAIC-WOZ corpus for model fine-tuning, all session transcripts were extracted and merged into a single text corpus, composed of two-column prompt–response format suitable for LLM fine-tuning. Each turn in the dialogue of the therapist serves as the prompt, and the corresponding patient reply the response, allowing the models to learn turn-by-turn behavioural patterns characteristic of psychotherapy dialogue. This dataset was split 80/20: training (174) and test (15) sessions.

TinyLlama and DeepSeek were trained for 120 epochs with LoRA adapters using the AdamW optimizer, gradient clipping, and a linear learning-rate scheduler with warm-up on a CUDA-enabled GPU in mixed-precision (fp16). Input prompts were tokenized and truncated to the models' maximum context window (2048 tokens). During inference, PyTorch-based generation employed sampling with temperature 0.7 and top-p of 0.9. For RAG, SentenceTransformer embeddings were used with cosine similarity to select relevant context using content from a random selected holdout test session.

Evaluation of both models was conducted on the DAIC-WOZ hold-out testset via two steps: a manual rating by psychology professionals on the output and using NLP metrics to automatically rate similarity of output versus refence output in the testdata. Human evaluators followed a two-level rating protocol: (1) turn-level assessment, and (2) session-level assessment. At the turn level, evaluators stepped through each session chronologically; for every therapist prompt, the three models generated a response accompanied by a predicted emotion from the classifier of Desot et al. (2025), which assigns one of six basic emotions (anger, disgust, fear, happiness, neutral, sadness). Emotion accuracy was rated categorically (correct/incorrect/unknown), whereas response relevance was rated with discrete scores from 0 to 5, with higher values indicating stronger topical alignment. In total 2055 turns were evaluated by the human evaluators. At the session level, evaluators provided overall ratings per model using discrete scores from 0 to 5 to assess quality on coherence, completeness, role adherence, and knowledge retention. All annotations by the evaluators were collected via a custom web-based tool.

5.2. Results

In Table 1, human evaluator scores are shown on the left and NLP similarity metrics on the right. Based on the results we see that RAG produces the highest proportion of highly relevant responses (score 4 or 5) and highest average relevance score, followed by DeepSeek and TinyLlama, highlighting the effectiveness of retrieval augmentation. RAG also demonstrates clear positive semantic alignment with the reference utterances, as confirmed by BERTScores F_1 and cosine similarity, whereas TinyLlama and DeepSeek show lower overlap, though DeepSeek performs slightly better than TinyLlama according to SBERT-cosine similarity. BLEU and ROUGE scores remained low for TinyLlama and DeepSeek, reflecting the high lexical variability inherent in open-ended dialogue generation. RAG substantially improved BLEU and ROUGE scores, indicating stronger verbatim alignment with reference turns; however, higher lexical overlap alone does not imply superior conversational or clinical quality. Overall, retrieval augmentation produced more contextually grounded, patient-like responses, while the non-RAG models generated paraphrastic but still semantically coherent outputs.

Table 1. Turn-level analysis of human ratings and NLP metrics.

	Turn level scores (0 to 5)							
	Human evaluator ratings		***NLP metrics***					
Model	**Relevance (prop. score ≥ 4)**	**Avg.**	**BLEU**	**Rouge-1**	**Rouge-2**	**Rouge-L**	**Bert F_1**	**SBERTCosine**
TinyLlama	0.48	3.04	0.00	0.11	0.02	0.10	0.22	0.23
DeepSeek	0.64	3.70	0.01	0.13	0.02	0.11	0.07	0.27
RAG	0.76	4.13	0.23	0.35	0.21	0.33	0.19	0.45

In Table 2 we show the average rating scores for the session quality assessment categories. TinyLlama shows the lowest average scores overall. These quantitative trends align with the qualitative observations from the evaluators, who noted that TinyLlama frequently tends to produce irrelevant, repetitive, or nonsensical answers and struggles to maintain context, making it the least reliable model. DeepSeek generally keeps the conversation on track and tracks context well, though it can occasionally repeat answers or provide minimal responses. RAG is consistently strong, maintaining relevance and coherence with fewer reported issues, aligning with its higher quantitative scores.

Table 2. Session-level average scores per model.

	Session level scores (0 to 5)			
Model	**Coherence**	**Role Adherence**	**Completeness**	**Knowledge Retention**
TinyLlama	2.62	2.73	2.54	2.42
DeepSeek	2.85	3.04	2.92	2.65
RAG	3.17	2.96	3.17	2.96

6. Conclusion and discussion

Our findings demonstrate that fine-tuned open-source LLMs can generate locally plausible responses in psychotherapy-style dialogue, supporting their potential use as experimental virtual patients for early-stage training. RAG improves contextual relevance, enabling lightweight models to approach performance of larger architectures, while DeepSeek shows stronger role consistency and emotional fidelity than TinyLlama. At the same time, the results reveal clear limitations. Across models, sustained dialogue structure, long-term consistency, memory management, and reliable dialogue state tracking remain unresolved. Human evaluators observed repetition, loss of focus, and generic behaviour, indicating that current LLM-based systems cannot yet support realistic, fully autonomous clinical interactions. Overall, this prototype demonstrates promising potential, but significant methodological and architectural advances are required before LLM-driven virtual patients can reliably approximate the complexity of real clinical dialogue. To address these limitations, future work will integrate the models into a 3D multimodal digital human interface via PEX (Platform for Empathic eXperiences, n.d.), incorporating multimodal cues (e.g., prosody, facial expressions), and case-based mental health profiles and multi-annotator datasets. We hypothesize that this multimodal integration will improve virtual patient dialogue interactions by enabling more accurate modelling of affective states and conversational intent.

Acknowledgments

This work is part of the TherAvatars project funded by Regieorgaan SIA (The Netherlands), SVB/HT. KIEM.01.063.

References

Awasthi, R., Bhattad, A., Ramachandran, S. P., ... & Mathur, P. (2025). Human evaluation of large language models in healthcare: gaps, challenges, and the need for standardization. *npj Health Systems, 2*(1).

Combs, C. D. & Combs, P. F. (2019). Emerging roles of virtual patients in the age of AI. *AMA Journal of Ethics, 21*(2), E153-E159.

DeepSeek. (2025). *DeepSeek-R1-Distill-Llama-8B-unsloth-bnb-4bit* [Large language model]. HuggingFace. https://huggingface.co/unsloth/DeepSeek-R1-Distill-Llama-8B-unsloth-bnb-4bit

Desot, T., Basharirad, B., Netten, N., & Muijs, R. (2025). Corpus Generation for Emotion Classification in Psychotherapy using Large Language Models. *Proceedings of 9th International Conference on Natural Language Processing and Information Retrieval (NLPIR 2025)*, Fukuoka, Japan.

DeVault, D., Artstein, R., Benn, G. … & Morency, L.-P. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 1061–1068).

Epstein, J. H., Levin, M., & Jowell, M. S. (2013). Agent based simulation for training and assessing students in the field of anesthesiology. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems* (pp. 332-336). IEEE.

Gao, Y., Xiong, Y., Gao, X.... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S. & Morency, L. P. (2014). The distress analysis interview corpus of human and computer interviews. *LREC* (pp. 3123-3128).

Khan, M. N. R., Ahmmed, F., Al Zabir, M. K., & Lippert, K. J. (2023, October). Virtual Patient in Medical Education. *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)* (pp. 1-6). IEEE.

Kenny, P., Parsons, T. D., Gratch, J., Leuski, A., & Rizzo, A. A. (2007). Virtual patients for clinical therapist skills training. *Intelligent Virtual Agents: 7th International Conference, IVA 2007 Paris, France, September 17-19, Proceedings 7* (pp. 197-210). Springer Berlin Heidelberg.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Annual Meeting of the Association for Computational Linguistics*.

Ng, H. W., Koh, A., Foong, A., & Ong, J. (2023). Real-Time Hybrid Language Model for Virtual Patient Conversations. *International Conference on Artificial Intelligence in Education* (pp. 780-785). Cham: Springer Nature Switzerland.

Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J., & Daneshjou, R. (2024). Large Language Models in Medicine: The Potentials and Pitfalls: A Narrative Review. *Annals of Internal Medicine, 177*(2), 210-220.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)* (pp. 311-318). https://doi.org/10.3115/1073083.1073135

Platform for Empathic Experiences (PEX). (n.d.). *Hogeschool Rotterdam – Slimme en sociale stad*. Retrieved from https://www.hogeschoolrotterdam.nl/hogeschool/missie-visie-en-strategie/slimme-en-sociale-stad/hybride-wereld/

Prescott, J., Ogilvie, L., & Hanley, T. (2024). Student therapists' experiences of learning using a machine client: A proof-of-concept exploration of an emotionally responsive interactive client (ERIC). *Counselling and Psychotherapy Research, 24*(2), 524-531.

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121-154.

Tao, Y., Yang, M., Shen, H., Yang, Z., Weng, Z., & Hu, B. (2023). Classifying Anxiety and Depression through LLMs Virtual Interactions: A Case Study with ChatGPT. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2259-2264). IEEE.

Tanana, M. J., Soma, C. S., Srikumar, V., Atkins, D. C., & Imel, Z. E. (2019). Development and evaluation of ClientBot: Patient-like conversational agent to train basic counseling skills. *Journal of medical Internet research, 21*(7), e12529.

TinyLlama. (2024). *TinyLlama-1.1B-Chat-v1.0* [Large language model]. HuggingFace. https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0

Vaswani, A., Shazeer, N., Parmar, N.,... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30.*

Wang, J., & Chen, Y. (2023). *Introduction to Transfer Learning: Algorithms and Practice*. Springer Singapore.

Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., & Liu, N. (2023). Large language models in health care: Development, applications & challenges. *Health Care Science, 2*(4), 255-263.

Zhang, T., Kishore, V., Wu, F., Weinberger, K., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. *International Conference on Learning Representations (ICLR)*.