Using Llama-3 to Refine Psychotherapy in Silico

Kristin Witte* (kristin.witte@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Milena Rmus* (milena.rmus@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Elif Akata (elif.akata@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Eric Schulz (eric.schulz@helmholtz-munich.de)
Helmholtz Munich
Munich, Germany

Abstract

Psychotherapy is deeply personal and often time-intensive. Although therapeutic interactions crucially impact treatment outcomes, strategies to improve them often rely on trial and error, often resulting in a long and costly process with minimal improvement for the client. This project explores the potential of Large Language Models (LLMs) as low-risk, cost-effective tools for enhancing therapy. Using Llama-3, we simulated therapist-client dialogues, supervised by an expert LLM. The Expert LLM iteratively refined the Therapist LLM's responses, which were subsequently rated by the Client LLM. To extend this framework to real-world data, we applied LLM therapy revision to real therapy transcripts. For each segment of a recorded session, we compared the Client LLMs rating of the real therapist's last response to ratings of an LLM generated response and an LLM-based revision of the actual therapist's response, assessing satisfaction based on prior conversation context. These comparisons revealed that LLM-generated responses were often rated more favorably than human responses, with responses revised based on LLM feedback rated most favorably. This suggests that LLMs could meaningfully support and enhance therapeutic interactions, and improve quality of treatment.

Keywords: Large language models; psychotherapy; computational psychiatry

Introduction

Psychotherapy plays a vital role in shaping individuals' well-being and their ability to function and thrive in daily life. Therapeutic success often depends on the personalization of general treatment approaches to meet the unique needs of each client Norcross & Wampold (2011). However, this process is both time- and resource-intensive, typically relying on a trial-and-error approach as therapists work to tailor interventions effectively Bremer et al. (2018).

Recent advances in generative AI—particularly Large Language Models (LLMs), which can simulate, evaluate, and adapt conversations—have sparked growing interest in their potential role within therapeutic settings Stade et al. (2024). Researchers have begun exploring various ways of integrating LLMs into the therapy process to support mental health professionals such as through LLM-based therapy chatbots Song et al. (2024), by evaluating psychiatric functionality Galatzer-Levy et al. (2023), or as tools for assessing potential future risks Acharya et al. (2024). Some have discussed the possibility of AI fully automating the therapy process Stade et al. (2024). Nonetheless, while AI certainly has potentials in augmenting therapy, the high stakes and potential risks demand that researchers in this domain proceed with caution Obradovich et al. (2024). There is evidence suggesting AI is not capable of handling unanticipated user responses Chan et al. (2022), or providing harmful advice to clients De Choudhury et al. (2023); Song et al. (2024). At the current state of LLMs, it appears evident that oversight of the mental health professionals in therapy is still essential Yuan et al. (2025). Nonetheless, LLMs can plausibly be incorporated to augment the traditional therapy - for instance by helping therapists tailor the treatment to the client.

We conducted a proof-of-concept study using Llama-3-70B to explore how LLMs might enhance therapeutic conversations. We simulated sequential interactions between a therapist and a client, as well as an expert feedback for the therapist to incorporate into subsequent sessions. The Client LLM provided numerical treatment satisfaction ratings following each session as an outcome variable. We further extended this framework to real-world data by extracting conversations from real therapy transcripts, and using the Expert LLM to generate baseline and revised therapist responses. The Client LLM then rated both the real and revised responses using preceding conversation as context.

Methods

LLM setup

We used the open-source Llama 3.1 (iterative improvement) and 3.3 (comparison to human therapist) Instruct models with 70 billion parameters (Meta Platforms (2024)). All experiments were conducted in-context, without any fine-tuning. For response generation, we set the temperature to 0.1 to allow for more exploration in the model's outputs, while keeping all other generation parameters at their default values.

Simulating conversations

We simulated six sequential exchanges between a therapist and a client. The Therapist LLM was prompted with the prior session dialogue and feedback from the Expert therapist LLM. The Client LLM received the previous conversation along with its own earlier satisfaction ratings to maintain contextual continuity. The Expert therapist LLM was prompted with the full therapist-client dialogue and a set of evaluation guidelines. These guidelines focused on qualities such as actionable advice, empathetic understanding, and clear therapeutic goals (Bordin, 1979). We repeated this process across five independent runs to evaluate the robustness of our results.

Therapy transcripts

We used the Alexander Street therapy transcript database (ESS (2008)). Transcripts were carefully filtered to include only coherent conversations with complete exchanges and minimal transcription noise. For the Client LLM ratings, we only used therapist responses that were at least 15 words long.

Results

Iterative improvements in therapy satisfaction

We began by examining how the Client LLM's satisfaction with therapy evolved across sessions when Therapist LLM received feedback from the Expert LLM, and in the absence of feedback. To ensure a fair comparison across these conditions, we initialized all sessions with the same baseline satisfaction rating of 30 by explicitly including this value in the Client LLM prompt. Overall, we observed a significant increase in satisfaction ratings over the course of iterative sessions, suggesting that feedback-driven refinement meaningfully improved the perceived quality of therapy(feedback$>$no feedback t(5): 4.73, p=002, Fig.1B)

To rule out alternative explanations, such as that the presence of feedback alone was sufficient to drive we conducted another control experiment where the Therapist LLM received non-specific or unhelpful feedback (e.g., "pay closer attention to the client's needs"). This condition yielded less of an improvement than the helpful feedback (feedback$>$generic feedback t(5): 4.03, p=004 Fig.1B). This suggests that it is not repetition or feedback per se, but the quality of the feedback that drives meaningful improvements in therapeutic interactions.

Comparison against human therapist

To evaluate our approach on real-world data, we extracted 189 conversation snippets from therapy transcripts involving 21 clients. Each snippet included a portion of client-therapist dialogue, which we used as context in prompts to the Client LLM. For 50 randomly sampled snippets, the Client LLM was then asked to rate three types of therapist responses: (1) the actual human therapist's next response, (2) a response generated by an LLM conditioned only on the prior conversation, and (3) Expert-guided LLM revision of the human response.

Our results showed that the Client LLM consistently rated the Expert-informed LLM responses more favorably than both the human and baseline LLM responses (LLM $>$human $t(49) = 8.07, p < .001$; Human revised by LLM $>$LLM $t(49) = 4.51, p < .001$). This suggests that expert-guided LLMs may enhance therapeutic dialogue, offering a promising tool for augmenting therapist communication and training.

Discussion

Our results demonstrate the potential of LLMs to adjust therapy in a client-specific way, leading to higher satisfaction ratings. Rather than replacing therapists, this approach aims to augment existing practices with LLM-generated support, with mental health professionals' oversight present - addressing the risk and safety concerns.

While current evaluations rely on LLM-based ratings, future work will incorporate human assessments of therapy quality. These experiments were conducted fully in-context, with next steps including fine-tuning of LLMs, to better align with therapeutic standards.

Acknowledgments

This work was supported by Helmholtz Munich, a Jacobs Research Fellowship, and an ERC Starting Grant to E.S

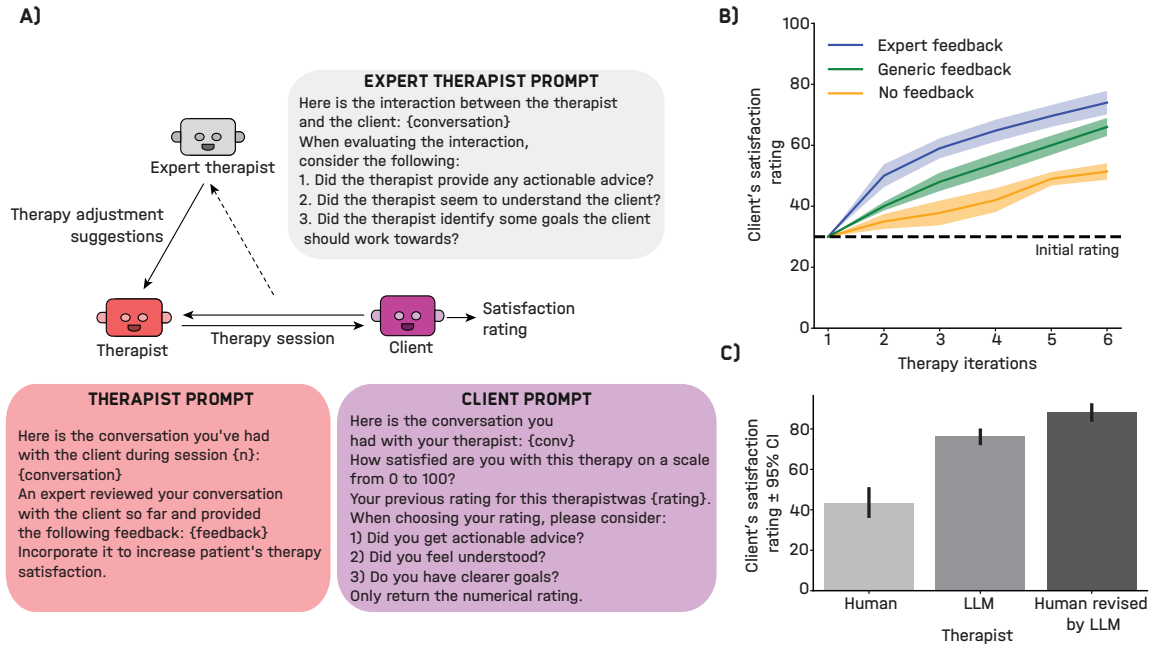

Figure 1: A: Schematic of our approach, B: Improvement of ratings by Client LLM over the course of feedback iterations. C: Average ratings by Client LLM of human therapists' responses, LLM responses and human responses revised by the LLM.

References

Acharya, A., Shrestha, S., Chen, A., Conte, J., Avramovic, S., Sikdar, S., ... Das, S. (2024). Clinical risk prediction using language models: benefits and considerations. *Journal of the American Medical Informatics Association*, *31*(9), 1856–1864.

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, research & practice*, *16*(3), 252.

Bremer, V., Becker, D., Kolovos, S., Funk, B., Van Breda, W., Hoogendoorn, M., & Riper, H. (2018). Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: data-driven analysis. *Journal of medical Internet research*, *20*(8), e10275.

Chan, W. W., Fitzsimmons-Craft, E. E., Smith, A. C., Firebaugh, M.-L., Fowler, L. A., DePietro, B., ... Jacobson, N. C. (2022). The challenges in designing a prevention chatbot for eating disorders: observational study. *JMIR Formative Research*, *6*(1), e28003.

De Choudhury, M., Pendse, S. R., & Kumar, N. (2023). Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.

ESS, S. P. (2008). Counseling and psychotherapy transcripts, client narratives, and reference works.

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., & Malgaroli, M. (2023). The capability of large language models to measure psychiatric functioning. *arXiv preprint arXiv:2308.01834*.

Meta Platforms, I. (2024). *Llama 3.1 70b model*. Retrieved from `https://github.com/meta-llama/llama3`

Norcross, J. C., & Wampold, B. E. (2011). What works for whom: Tailoring psychotherapy to the person. *Journal of clinical psychology*, *67*(2), 127–132.

Obradovich, N., Khalsa, S. S., Khan, W. U., Suh, J., Perlis, R. H., Ajilore, O., & Paulus, M. P. (2024). Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, *2*(1), 8.

Song, I., Pendse, S. R., Kumar, N., & De Choudhury, M. (2024). The typing cure: Experiences with large language model chatbots for mental health support. *arXiv preprint arXiv:2401.14362*.

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., Schwartz, H. A., Yaden, D. B., ... Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, *3*(1), 12.

Yuan, A., Garcia Colato, E., Pescosolido, B., Song, H., & Samtani, S. (2025). Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. *ACM Transactions on Management Information Systems*, *16*(1), 1–26.