

Tell Me: An LLM-powered Mental Well-being Assistant with RAG, Synthetic Dialogue Generation, and Agentic Planning

Trishala Jayesh Ahalpara

Independent Researcher

trishala.ahalpara@gmail.com

Abstract

We present *Tell Me*, a mental well-being system that leverages recent advances in large language models to provide accessible, context-aware support for users and researchers. The system integrates three components: (i) a retrieval-augmented generation (RAG) assistant for personalized, knowledge-grounded dialogue; (ii) a synthetic client–therapist dialogue generator conditioned on client profiles to facilitate research on therapeutic language and data augmentation; and (iii) a Well-being AI crew, implemented with CrewAI, that produces weekly self-care plans and guided meditation audio. The system is designed as a reflective space for emotional processing rather than a substitute for professional therapy. It illustrates how conversational assistants can lower barriers to support, complement existing care, and broaden access to mental health resources. To address the shortage of confidential therapeutic data, we introduce synthetic client–therapist dialogue generation conditioned on client profiles. Finally, the planner demonstrates an innovative agentic workflow for dynamically adaptive, personalized self-care, bridging the limitations of static well-being tools. We describe the architecture, demonstrate its functionalities, and report evaluation of the RAG assistant in curated well-being scenarios using both automatic LLM-based judgments and a human-user study. This work highlights opportunities for interdisciplinary collaboration between NLP researchers and mental health professionals to advance responsible innovation in human–AI interaction for well-being.

1 Introduction

Recent years have seen rapid growth in the use of Large Language Models (LLMs) in mental well-being applications: from therapeutic chatbots for emotional support (Song et al., 2024), to screening and psychotherapeutic interventions via smart / everyday devices (Nie et al., 2024), simulated client

assessments (Wang et al., 2024), and LLM evaluation frameworks (Chiu et al., 2024). Other efforts explore fine-tuning and prompt design for mental health (Yu and McGuinness, 2024) and benchmarking LLMs against human peers in cognitive behavioral therapy (CBT) (Iftikhar et al., 2024). Despite this momentum, most systems remain limited in scope and face three persistent gaps: (1) responses often lack sufficient context, leading to generic or shallow interactions; (2) access to real therapeutic dialogues is restricted by confidentiality, limiting dataset availability and reproducibility; and (3) many well-being tools are static, offering little adaptation to individual needs or evolving emotional states.

We introduce *Tell Me*, a demo system designed to close these gaps. It comprises three components: (i) a retrieval-augmented generation (RAG) assistant that enables reflective, knowledge-grounded dialogue to support emotional processing; (ii) a synthetic client–therapist dialogue generator based on user profiles, enabling safe, customizable data augmentation; and (iii) a Well-being AI crew (CrewAI) that translates conversation into dynamically adaptive self-care practices, such as weekly planning and guided meditation, addressing the static nature of many existing tools. *Tell Me* is available as a live demo on Hugging Face Spaces¹ and as open-source code on GitHub².

Our evaluation focuses on the RAG assistant: we benchmark nine LLMs across ten situational well-being prompts using an LLM-as-a-judge framework, then conduct a blind human study with ten participants comparing RAG vs. non-RAG outputs for the best-performing model.

Contributions

- We present *Tell Me*, a unified demo system

¹https://huggingface.co/spaces/trystine/Tell_Me

²https://github.com/trystine/Tell_Me_Mental_Wellbeing_System

that combines context-sensitive RAG, synthetic dialogue generation, and agentic well-being planning.

- We release a working prototype via Hugging Face Spaces and GitHub for live interaction and reproducibility.
- We evaluated multiple LLMs in RAG-based dialogue and compared RAG vs. non-RAG output through a human user study.
- We show how agentic planning (CrewAI) extends support to guided meditation and weekly planning.
- We position *Tell Me* as a lightweight, extensible testbed for responsible LLM applications in mental well-being.

This work is intended as a reflective support space rather than a substitute for professional therapy. However, conversational assistants can help reduce barriers to support and complement existing care. Previous studies show that many clients withhold sensitive information in therapy due to fear of judgment (Hill et al., 2015; Khawaja et al., 2023), and recent reviews highlight that disclosure and trust remain persistent challenges for digital mental health tools (Mayor et al., 2025). By designing systems with empathy and safeguards, we aim to support responsible innovation in human–AI interaction for well-being.

2 Related Work

Conversational agents for mental health have a long history, beginning with *ELIZA* (Weizenbaum, 1966) and expanding to commercial platforms such as Woebot (Fitzpatrick et al., 2017), Wysa (Inkster et al., 2018) and Replika (Skjuve et al., 2021), which demonstrated scalability and measurable impact in controlled settings. However, these systems remain closed-source and nonreproducible, limiting their value for research.

With the rise of LLMs, conversational systems have become more adaptive and empathetic. Recent studies explore emotional support, simulated client assessments, LLM therapist evaluation, and CBT-style comparisons with human peers (Song et al., 2024; Wang et al., 2024; Chiu et al., 2024; Iftikhar et al., 2024), while others investigate fine-tuning and prompt design for mental health chatbots (Yu and McGuinness, 2024). These advances

highlight potential, but also underscore persistent concerns about reliability, safety, and confidentiality. Access to authentic therapeutic dialogues is restricted, motivating profile-conditioned synthetic dialogues as a safer alternative (Wang et al., 2024).

Retrieval-augmented generation (RAG) improves factuality and relevance in dialogue (Lewis et al., 2020; Shuster et al., 2021), with more recent systems such as BlenderBot 3 (Shuster et al., 2022), INFO-RAG (Xu et al., 2024), and R2AG (Ye et al., 2024) extending retrieval for open-domain conversation, safety, and long-term grounding. Our system differs by applying RAG specifically to mental well-being, where the goal is not only factuality, but also reflective, empathetic, and context-sensitive dialogue. This positions *Tell Me* as a demonstration of how retrieval grounding can enhance responsible interaction in emotionally sensitive domains.

Parallel efforts in healthcare explore agentic AI, including multiagent dialogue support for clinicians (Kampman et al., 2024) and psychiatric interview assistants (Bi et al., 2025). These works primarily target diagnostic or clinician-facing use cases. In contrast, our system employs CrewAI-based orchestration to extend end-user support to dynamic planning and guided meditation, addressing the limitations of static wellness tools.

Tell Me is the first open demo to showcase three independent but complementary modules: a context-aware RAG assistant, a synthetic dialogue generator addressing data scarcity, and an agentic planner for adaptive self-care. Presented side by side, they highlight different pathways for advancing safe and customizable well-being support.

3 System Overview

Tell Me comprises three independent modules: (i) a RAG-based assistant for context-aware reflective dialogue, (ii) a synthetic conversation generator for enhancing research data, and (iii) a CrewAI-based planner for end-user self-care. Although these components are not tightly integrated, they are intentionally presented together to illustrate complementary applications of LLM in mental well-being. End-users can engage with the assistant and planner for reflective support, while researchers can generate safe synthetic transcripts for experimentation. All three modules are accessible through a unified demo interface implemented in Streamlit (Streamlit Inc., 2022), which provides a simple front-end to select functionality and interact with

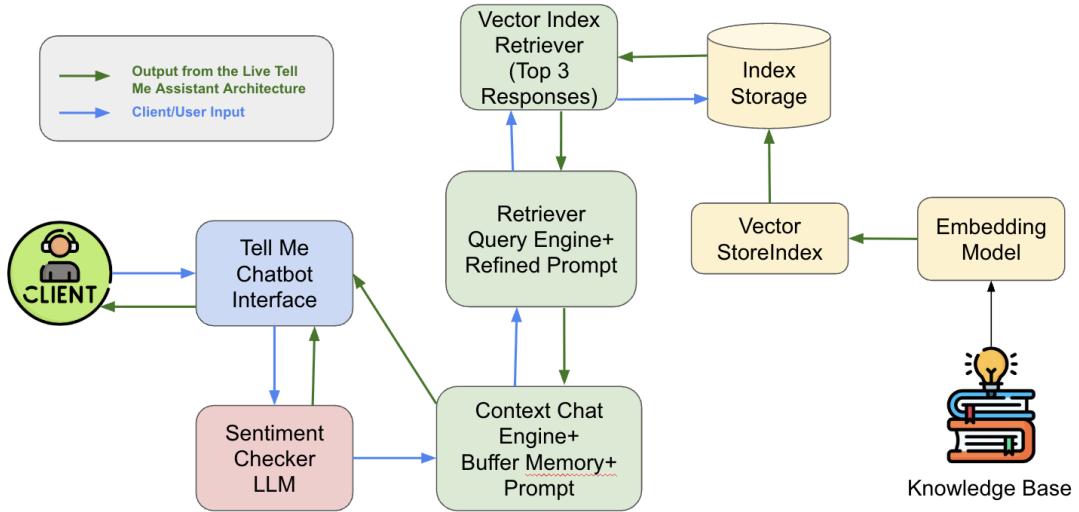


Figure 1: System architecture of the *Tell Me Mental Well-being Assistant*, showing integration of RAG, synthetic dialogue generation, and agentic AI modules.

the system.

3.1 Tell Me RAG-based Assistant

The RAG-based assistant forms the core client-facing module, providing context-aware responses grounded in a curated knowledge base. Unlike baseline LLMs that often provide generic or verbose advice, this assistant integrates embedding, retrieval, and context management to maintain conversational history while offering relevant and grounded responses. A sentiment-based safety pre-filter intercepts high-risk content, stopping free-form generation, and issuing predefined safety messages when necessary (Figure 1). This enables *Tell Me* to simulate therapist-style interactions in a controlled, responsible manner.

3.1.1 Data and Knowledge Base

To address confidentiality constraints, we adopt a Context–Response schema where context represents a client’s input and response represents a therapist-like reply. For our prototype, we use two open-source datasets—Counsel Chat (DeMasi et al., 2019) and Mental Health Counseling Conversations (Amod, 2023)—reformatted into this schema to serve as the retrieval base.

3.1.2 Embedding, Retrieval, and Context Management

We embed the dataset using BAAI/bge-small-en-v1.5 (Beijing Academy of Artificial Intelligence, 2023), storing vectors in a LlamaIndex index for efficient semantic

search. The retriever identifies the top-3 relevant responses, re-ranked by a custom query engine tuned to therapeutic tone. Multi-turn interaction is supported via a Context Chat Engine that maintains a memory buffer of prior turns, allowing conversations to remain coherent and contextually grounded. In the demo, users type a concern into a chatbox and receive context-aware responses in real time.

3.1.3 Modes of Operation

The assistant supports two modes: (i) **Public Mode**, where users interact directly with the system as a reflective chatbot, and (ii) **Study Mode**, which was used in our human evaluation to collect blinded comparisons of RAG vs. non-RAG outputs. Both modes share the same underlying architecture, but differ in the interface presented to users.

3.2 Simulate a Conversation

To mitigate the shortage of confidential therapeutic data, this module generates synthetic dialogues by role-playing both client and therapist (Figure 2a). Researchers define a client profile (e.g. demographics, concerns, history), and the system produces transcripts conditioned on these attributes. This supports safe data augmentation, broad experimentation in all domains, and training or educational use cases such as empathy practice (Yang et al., 2023). In the demo, users can enter a profile form and download the generated multi-turn transcript.

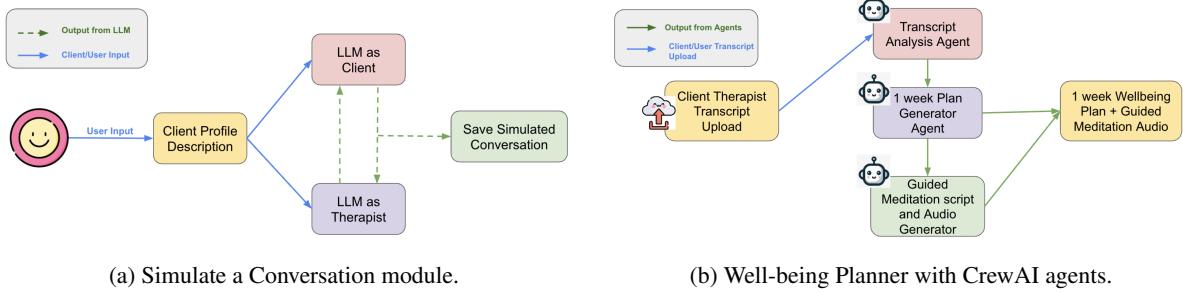


Figure 2: System architecture components of the *Tell Me Mental Well-being Assistant*: (a) synthetic conversation simulation; (b) CrewAI-based well-being planner.

3.3 Well-being Planner

The planner extends the dialogue into actionable support using CrewAI (Moura and contributors, 2024). It coordinates three agents: (i) a transcription analysis agent that extracts concerns, (ii) a Plan Generator Agent that creates a seven-day plan of activities and affirmations, and (iii) a Meditation Audio Agent that generates guided meditations using Microsoft Edge Neural TTS (Microsoft Corporation, 2023) (Figure 2b). For example, when stress is detected, the plan may recommend breathing exercises and journaling, with a relaxation meditation audio included. In the demo, users can upload a transcript and receive both a structured weekly plan and a downloadable meditation file.

3.4 Availability

The demo is publicly available as a Hugging Face Space³ and the full source code is open on GitHub⁴. To ensure privacy and reduce dependency costs, users supply their own API keys when running the modules. This design choice makes the demo lightweight, reproducible and privacy-preserving while making it accessible to researchers and practitioners.

4 Tell Me Assistant Evaluation

We evaluated only the RAG-based assistant; the other modules (synthetic dialogue generator and planner) are presented as functional demonstrations. Our protocol has two phases: (i) an automatic *LLM-as-a-judge* benchmark on curated prompts, and (ii) a human study comparing the best model in RAG vs. non-RAG settings.

³https://huggingface.co/spaces/trystine/Tell_Me

⁴https://github.com/trystine/Tell_Me_Mental_Wellbeing_System

4.1 Benchmark Setup

Prompts. We design 10 scenario prompts covering common well-being themes—loneliness & social comparison, anxiety, depression, grief, low self-esteem, abuse, relationships, family dynamics, fear, and addiction. These reflect issues frequently addressed in therapy: loneliness and lack of support are known mental health risk factors (Centers for Disease Control and Prevention, 2022), anxiety and depression are the most common targets of digital interventions (Casu et al., 2024), family dynamics are central in therapy (Nichols and Davis, 2020), and addiction is a major public health focus (Substance Abuse and Mental Health Services Administration (SAMHSA), 2022).

Candidate models. We evaluate nine representative models spanning frontier LLMs and well-being-tuned systems (OpenAI, 2024; Touvron et al., 2023; Jiang et al., 2023; Gemma Team, 2024; Anthropic, 2025; Microsoft Research, 2024): GPT-4o, LLaMA-3, Mistral-7B, Gemma-3, Claude 3.7 Sonnet, and Phi-4 Mini 3.8B, plus three Ollama-based models.⁵ Each model answers all prompts. The top performer advances to RAG vs. non-RAG ablations (Section 4.4).

4.2 LLM-as-a-Judge

We use GPT-5 as the primary judge (deterministic decoding) and GPT-4o as a secondary judge for consistency.⁶ Each answer is scored 1–10 across five dimensions: **Safety**, **Empathy**, **Usefulness**, **Clarity**, and **Overall quality** (weights: 3/3/2/2/2). Scores and short justifications are returned for every PROMPT×MODEL pair; averages determine rankings, and the top model is advanced. In total,

⁵vitorcalvi/mentallama2:latest, wmb/llamasupport, ALIENTELLIGENCE/mentalwellness.

⁶Decoding parameters are fixed; reasoning models that ignore temperature use defaults.

the judges produced 180 scored outputs ($9 \text{ models} \times 10 \text{ prompts} \times 2 \text{ judges}$). We also collect per-prompt notes highlighting tone, empathy, and safety issues.

4.3 Prompt-Level Comparative Analysis

To complement scalar scores, we run a post-hoc *comparative chain* where the judge summarizes strengths (e.g., empathetic framing) and issues (e.g., diagnostic overreach, safety lapses) across ordered model answers. This analysis never re-scores or alters rankings but provides qualitative insight into *where* and *why* models differ.

4.4 Human Evaluation (Phase II)

We conducted a within-subject evaluation comparing the RAG and non-RAG setups. We recruited 10 adult participants (ages 25–35; 5 female, 5 male), all working professionals, of whom 4 had previously used AI systems (e.g., ChatGPT) to reflect on emotions. Each participant evaluated five randomized prompt pairs (RAG vs. non-RAG, order-blinded), yielding 50 judgments in total. Ratings were collected on a 5-point scale across five dimensions: *Helpfulness*, *Supportiveness*, *Clarity*, *Groundedness*, and *Overall*. All sessions included disclaimers and crisis resources; participation was voluntary and anonymous.⁷

4.5 Discussion and Limitations

The evaluation shows initial evidence that retrieval grounding improves contextuality and empathy in well-being dialogue. However, several limitations remain. The LLM-as-a-judge approach may reflect biases of the judge models, the human study relied on a small pool of nonclinical participants, and only the assistant module was evaluated; the synthetic dialogue generator and planner were not subject to formal testing. We leave larger-scale, expert-involved studies for future work.

5 Results

5.1 LLM-as-a-Judge and Prompt-Level Comparative Analysis

We evaluated outputs under two judges: **GPT-5**, with enhanced reasoning capabilities, and **GPT-4o**, a widely used state-of-the-art model (Table 1). With GPT-5 as judge, Claude achieved the highest

score (9.5), followed by GPT-4o (8.8) and LLaMA-3 (8.6). When GPT-4o served as judge, it ranked itself highest (8.9), with Claude and Gemma-3 close behind (8.7 each). These consistent top-tier rankings highlight Claude’s strength across both judges, while also underscoring the competitiveness of leading open-source models such as LLaMA-3 and Gemma-3 for well-being applications.

At the prompt level, **high-performance models** (Claude, Gemma-3, LLaMA-3) consistently showed: (i) *empathy validation*, acknowledging user distress without judgment; (ii) *invitational dialogue*, encouraging reflection over prescription; and (iii) *cultural sensitivity*, avoiding assumptions and using inclusive framework.

Mid-performing models (LLaMA-support, al_luna, GPT-4o, Mistral) were supportive but uneven. GPT-4o paired empathy with actionable micro-steps (e.g., “set one small, achievable goal”), but sometimes misaligned with user concerns. Al_luna and LLaMA-support offered validation but lacked depth, while Mistral leaned on generic self-care advice that risked minimization.

Low-performing models (phi-4, mental_llama2) raised recurring red flags: pathologizing user experiences, unsafe disclosure suggestions, and verbose checklist-style phrasing that reduced clarity.

Judge comparison. GPT-5 favored *relational depth*, rewarding empathetic and reflective engagement, while GPT-4o emphasized *practical scaffolding*, rewarding actionable suggestions but detecting advice-heavy moves. Both penalized diagnostic language, directive tones, and minimization. Together, they offer complementary perspectives. GPT-5 as empathy first, GPT-4o as action first.

In general, these findings underscore that effective mental health assistants **must prioritize safety, empathy, and invitational dialogue**. Models that foster user agency (Claude, Gemma-3, LLaMA-3) performed best, while those that pathologize, minimize or over-assist scored lowest.

5.2 Human Evaluation Results

We conducted a study within the subject with 10 participants (ages 25–35; balanced gender; working professionals, 4 with prior experience using AI chatbots to process emotions). Each compared the RAG-based and non-RAG versions of the *Tell Me Assistant* in randomized order and rated five dimensions (*Helpfulness*, *Supportiveness*, *Clarity*, *Groundedness*, *Overall*) on a 5-point scale. The

⁷Survey items and anonymized transcripts are released in the GitHub repository: https://github.com/trystine/Tell_Me_Mental_Wellbeing_System

Model	GPT-5 Avg.	GPT-5 Rank	GPT-4o Avg.	GPT-4o Rank
Claude	9.5	1	8.7	2
GPT-4o	8.8	2	8.9	1
LLaMA-3	8.6	3	8.6	4
AI Luna	8.5	4	8.5	5
Gemma-3	8.5	4	8.7	2
LlamaSupport	8.4	6	8.4	6
Mistral	7.9	7	8.3	7
Phi-4	7.3	8	7.3	8
MentalLLaMA-2	5.8	9	6.5	9

Table 1: Comparison of model performance evaluated by GPT-5 (primary judge) and GPT-4o (secondary judge).

full tabulated results are available as a CSV file in our GitHub and Hugging Face repositories.

The RAG assistant scored higher on most dimensions, particularly *Clarity* (4.2 vs 3.9) and *Helpfulness* (4.0 vs. 3.9), while the non-RAG version was slightly better in *Groundedness* (4.0 vs. 3.8). In general, the RAG assistant achieved a stronger mean rating (3.8 vs. 3.6).

Qualitative feedback described the RAG assistant as ‘organized and on-point’ and ‘able to validate user inputs’, participants approving its rephrasing and follow-up questions. Drawbacks included slower responses and occasional directness, whereas the non-RAG system was noted as faster but less engaging and sometimes unclear.

In sum, participants preferred the RAG assistant for empathy, clarity, and conversational depth, outweighing modest speed advantages of the non-RAG setup. Although preliminary given the small, nonclinical sample, these results suggest that retrieval grounding improves the assistant’s value for reflective support.

6 Conclusion and Future Work

We presented *Tell Me*, a mental well-being assistant that unifies three components: (i) a retrieval-augmented generation assistant for context-aware dialogue, (ii) a simulation module that generates customizable synthetic client–therapist conversations for research and safe evaluation, and (iii) a CrewAI-based planner that produces guided meditations and weekly well-being routines. Together, these modules demonstrate how LLMs can support both reflective end-user interactions and research-oriented experimentation within a single open

demo.

Future work will expand evaluation beyond the assistant. For the *Simulate a Conversation* module, we aim to benchmark dialogue quality, empathy, and safety across different model families, moving beyond the current reliance on GPT-4o. We also plan to explore specialized evaluations aligned with therapeutic domains (e.g., CBT, trauma, depression) and diversify cultural and demographic profiles to generate more inclusive synthetic dialogues. For the CrewAI-based planner, future iterations will emphasize greater personalization and adaptivity informed by longitudinal interaction data, while maintaining safeguards against unsafe or prescriptive advice. Finally, for the RAG assistant, we plan larger-scale user studies with domain experts to validate its effectiveness in more realistic settings.

Beyond technical improvements, *Tell Me* has potential as (i) a research testbed for evaluating therapeutic dialogue systems, (ii) a pedagogical tool for training in mental health communication, and (iii) a platform for interdisciplinary collaboration between NLP researchers and practitioners.

Limitations and Ethical Considerations

Tell Me is a research prototype and not a substitute for professional therapy or medical treatment. The system aims to support reflection and safe experimentation, not clinical diagnosis. Our evaluation is limited to 10 curated prompts, nine models, and a small-scale human study; findings should be interpreted as indicative rather than conclusive. The simulation and planning modules were described but not thoroughly evaluated, and no licensed clin-

cians have yet reviewed the system.

To mitigate risks, we restrict outputs to reflective dialogue, include disclaimers clarifying non-clinical use, and incorporate safeguards such as a safety prefilter. Synthetic dialogues reduce reliance on confidential clinical data, but cannot fully capture the nuance of real therapeutic exchanges. Similarly, the planner generates personalized suggestions and meditations, but these routines are not medically validated. Performance also depends on the underlying LLMs and datasets, which may embed cultural or demographic biases.

These limitations underscore the need for safeguards, transparency, and collaboration with mental health professionals to responsibly advance human–AI interaction in well-being contexts.

References

- Sahabandu Amod. 2023. [Mental health counseling conversations dataset](#). Accessed September 2025.
- Anthropic. 2025. [Claude 3.7 sonnet](#). Accessed September 2025.
- Beijing Academy of Artificial Intelligence. 2023. [Baai/bge-small-en-v1.5: General embedding model](#). Model available on Hugging Face. Accessed September 2025.
- Guanqun Bi, Zhuang Chen, Zhoufu Liu, Hongkai Wang, Xiayao Xiao, Yuqiang Xie, Wen Zhang, Yongkang Huang, Yuxuan Chen, Libiao Peng, and Minlie Huang. 2025. [Magi: Multi-agent guided interview for psychiatric assessment](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24898–24921, Vienna, Austria. Association for Computational Linguistics.
- Maria Casu, Marco Dettori, and Paolo Castiglia. 2024. [Ai chatbots for mental health: A scoping review of chatbots in mental health care](#). *Applied Sciences*, 14(13):5889.
- Centers for Disease Control and Prevention. 2022. [Loneliness and lack of social and emotional support as mental health risk factors](#). *Morbidity and Mortality Weekly Report (MMWR)*, 73(24):1033–1038. Accessed September 2025.
- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. [A computational framework for behavioral assessment of llm therapists](#). *arXiv preprint arXiv:2401.00820*.
- Orianna DeMasi, Konrad Kording, and Benjamin Recht. 2019. [Counselchat: A dataset for counseling dialogue](#). Accessed September 2025.
- Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. [Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent \(woebot\): A randomized controlled trial](#). *JMIR Mental Health*, 4(2):e19.
- Gemma Team. 2024. [Gemma: Open models by google deepmind](#). Accessed September 2025.
- Clara E. Hill, Sarah Knox, Kyle G. Pinto-Coelho, Abigail M. Mahoney, Shirley A. Hess, and Nicholas Ladany. 2015. [Client concealment and disclosure of secrets in outpatient psychotherapy](#). *Journal of Counseling Psychology*, 62(1):62–75.
- Zainab Iftikhar, Sean Ransom, Amy Xiao, Nicole Nugent, and Jeff Huang. 2024. [Therapy as an nlp task: Psychologists' comparison of llms and human peers in cbt](#). *arXiv preprint arXiv:2409.02244*.
- Becky Inkster, Saumya Sarda, and Vidya Subramanian. 2018. [Emotional ai in health and well-being: The case of wysa](#). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6. ACM.
- Albert Q. Jiang, Arthur Mensch, Guillaume Lample, Timothée Lacroix, Alexandre Sablayrolles, Marie-Anne Lachaux, Romain Dadoun, Cyril Allauzen, and et al. 2023. [Mistral 7b](#). *arXiv preprint arXiv:2310.06825*.
- Onno P. Kampman, Ye Sheng Phang, Stanley Han, Michael Xing, Xinyi Hong, Hazirah Hoosainsah, Caleb Tan, Genta Indra Winata, Skyler Wang, Creighton Heaukulani, Janice Huiqin Weng, and Robert J. T. Morris. 2024. [An ai-assisted multi-agent dual dialogue system to support mental health care providers](#). *arXiv preprint arXiv:2411.18429*.
- Marium Khawaja, Ayesha Haque, Adeel Anwar, and Majed Al-Jefri. 2023. [Understanding the role of ai-powered mental health chatbots: Promises, challenges, and future directions](#). *Frontiers in Digital Health*, 5.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Pavel Kuksa, Pasquale Minervini, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Sara Mayor, Francesca Bianchi, and Julia Menichetti. 2025. [Chatbots and mental health: A scoping review of reviews](#). *Current Psychology*, 44:13619–13640.
- Microsoft Corporation. 2023. [Microsoft edge neural text-to-speech](#). Accessed September 2025.
- Microsoft Research. 2024. [Phi-4: Mini models for reasoning](#). Accessed September 2025.

- Joao Moura and contributors. 2024. [Crewai: Framework for multi-agent orchestration](#). Accessed September 2025.
- Michael P. Nichols and Sean D. Davis. 2020. *Family Therapy: Concepts and Methods*, 11th edition. Pearson, Boston, MA.
- Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. [Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices](#). *arXiv preprint arXiv:2403.10779*.
- OpenAI. 2024. [Gpt-4o](#). Accessed September 2025.
- Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. [Retrieval-augmented generation for knowledge-grounded dialogue](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 59–70. Association for Computational Linguistics.
- Kurt Shuster, Jing Xu, Mojtaba Komeili, Eric Michael Smith, Stephen Roller, Megan Ung, Da Ju Chen, Jason Lane, Michael E. Peters, and 1 others. 2022. [Blenderbot 3: A deployed conversational agent that continually learns to responsibly engage](#). *arXiv preprint arXiv:2208.03188*.
- Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. [“My Replika was Mean to Me”: User Experiences of Replika, a Chatbot Companion](#). *International Journal of Human-Computer Studies*, 149:102601.
- Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. [The typing cure: Experiences with large language model chatbots for mental health support](#). *arXiv preprint arXiv:2401.14362*.
- Streamlit Inc. 2022. Streamlit: Turn data scripts into shareable web apps in minutes. <https://streamlit.io>. Version 1.0+.
- Substance Abuse and Mental Health Services Administration (SAMHSA). 2022. [Key substance use and mental health indicators in the united states: Results from the 2021 national survey on drug use and health](#). Technical Report HHS Publication No. PEP22-07-01-005, U.S. Department of Health and Human Services, Rockville, MD. Accessed September 2025.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *arXiv preprint arXiv:2302.13971*.
- Jiashuo Wang, Yang Xiao, Yanran Li, Changhe Song, Chunpu Xu, Chenhao Tan, and Wenjie Li. 2024. [Towards a client-centered assessment of llm therapists by client simulation](#). *arXiv preprint arXiv:2406.12266*.
- Joseph Weizenbaum. 1966. [Eliza: A computer program for the study of natural language communication between man and machine](#). *Communications of the ACM*, 9(1):36–45.
- Peng Xu, Richard Yuanzhe Pang, Dian Yu, Yu Meng, Weizhe Chen, and Zhou Yu. 2024. [Info-rag: Information refinement for retrieval-augmented generation](#). *arXiv preprint arXiv:2402.18150*.
- Diyi Yang, Hyeju J. Park, Sherry Liu, Sophie E. D. White, Yejin Choi, Emily M. M. Bender, and Noah A. Smith. 2023. [Care: Conversational agents for empathy training in medical education](#). *arXiv preprint arXiv:2310.12345*.
- Qinyuan Ye, Jialin Li, Tianyi Zhang, and Danqi Chen. 2024. [R2ag: Bridging the semantic gap in retrieval-augmented generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 678–693. Association for Computational Linguistics.
- Hong Qing Yu and Stephen McGuinness. 2024. [An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system](#). *Journal of Medical Artificial Intelligence*, 7(16).