A Privacy-Preserving Intelligent Assistant for Clinical Psychology Practice

Aaron Pico[1] , **Joaquin Taverner**[1] , **Emilio Vivancos**[1] , **Ana Garcia-Fornes**[1] and **Vicent Botti**[1,2]
[1]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Valencia, Spain
[2]Valencian Graduate School and Research Network of Artificial Intelligence (valgrAI)
{apicpas, joataap, vivancos, agarcia, vbotti}@upv.es

Abstract

This paper describes a fully local, privacy-preserving intelligent system designed to assist in clinical psychology practice. The system automatically transcribes therapy sessions performing speaker attribution. Beyond transcription, the tool enhances clinical reasoning by detecting cognitive distortions and emotional patterns utilizing specialized deep learning classifiers and Large Language Models (LLMs). By guiding an LLM locally through a multi-step analysis process, the assistant synthesizes the enriched data and generates a series of analysis reports and clinical documentation of the session. As a result, the assistant reduces the administrative burden on professionals while preserving privacy with an edge computing approach in which the data never leaves the therapist's device. Finally, the assistant uses human-in-the-loop validation so that the professional always remains in control, ensuring clinical accuracy and trust.

1 Introduction

Psychotherapy requires establishing a therapeutic alliance with the patient while properly monitoring their emotional and cognitive progress [Wampold and Imel, 2015]. To achieve this, professionals rely on clinical notes and documentation to help them maintain continuity across all their patients' sessions, but the time burden of creating and reviewing it is so high [Schwartz-Dillard *et al.*, 2024; Shanafelt *et al.*, 2016] that it leads to professional burnout. As a result, session data is often not collected or used for in-depth analysis; or it consumes valuable time that could otherwise be spent on patient care.

While Large Language Models (LLMs) currently offer excellent capabilities for natural language processing and skills such as information extraction and synthesis [Brown *et al.*, 2020], data sensitivity and privacy requirements in mental health care have slowed their adoption by professionals. Most solutions are cloud-based, mainly due to the large computational resources required by large models. Using these solutions is incompatible with strict data protection frameworks such as GDPR [European Parliament and Council of the European Union, 2016]. In addition, general-purpose models do not have the training or skills necessary for the tasks required in this field, such as the detection and understanding of cognitive distortions, which requires expert level skills [Beck, 1976]. For all these reasons, the adoption of such tools in private clinical settings remains limited.

To address these challenges, we propose an intelligent assistant for clinical psychology practice that preserves privacy through an edge computing approach. The system operates entirely locally on the therapist's device, performing the different tasks necessary to alleviate the administrative burden on professionals. To do this, it transcribes the session with speaker attribution while analyzing it in real time. From this analysis, the assistant obtains structured information enriched with the detection of cognitive distortions, emotions, and conversation topics, which it then uses to generate reports and documentation. Finally, rather than replacing clinical judgment, the assistant relies on a human-in-the-loop validation mechanism that ensures security and control by the therapist. In this way, this tool enhances clinical analysis while reducing the workload of professionals and ensuring security and trust.

2 System Architecture

The system architecture consists of the three modules explained below to provide the assistant with the necessary capabilities to capture the session, analyze it, and generate clinical documentation.

2.1 Real-Time Transcription Module

The Real-Time Transcription Module is designed to collect the session data. It is therefore responsible for transcribing the session with speaker attribution. To do this in real time, we combine three components. First, a voice activity detection model [Silero Team, 2025] detects when someone is speaking. Next, a speaker recognition model [Ravanelli *et al.*, 2021] compares the embedding of each audio fragment to a pre-recorded sample from the therapist. Each fragment is then assigned to either the therapist or the patient based on similarity to the reference. Finally, a Whisper model [Radford *et al.*, 2023] produces the transcription of the audio content. This operating scheme allows detecting and sending the patient's messages during the session to the analysis module as they occur, saving post-processing time.

2.2 Cognitive Analysis Module

The Cognitive Analysis Module receives the transcribed messages from the patient as input to perform an analysis of each one and enrich the information. To detect and recognize cognitive distortions, the system uses a combination of two RoBERTa Large models [Liu *et al.*, 2019] that we have fine-tuned on CODIPAD, a novel expert-annotated dataset built for this purpose [Pico *et al.*, 2025]: (1) a Binary Filter that distinguishes between neutral discourse and discourse influenced by distorted thinking; and (2) a Fine-Grained Classifier that discriminates among ten specific cognitive distortion categories. This approach enhances general precision and mitigates false positives. At the same time, the emotion of the messages is recognized using the pre-trained EmoBERTa model [Kim and Vossen, 2021], while conversation topics are identified locally using Gemma 4 E2B LLM [Google, 2026] guided by prompt.

2.3 Clinical Report Generation Module

The Clinical Report Generation Module receives all the information from the previous modules once the session has ended. A first layer is responsible for aggregating all the data obtained, both the transcribed messages and the augmented information from the classifier models. It extracts and structures all combinations of emotions and themes with cognitive distortions to ground the model's reasoning in the next step. This contextualized data is then used by an LLM running locally, Gemma 4 E4B [Google, 2026], which is guided to perform different concatenated tasks to synthesize and analyze the session based on what has been detected and the patterns extracted. The result of these tasks is in turn synthesized and combined for the final construction of reports and clinical documentation. This multi-step reasoning process, based on the information detected in the previous step by specific models, minimizes the risk of hallucinations and ensures a comprehensive final result consistent with the aggregated session data.

3 Clinical Workflow Integration

The developed application utilizes the system architecture detailed in the previous section to provide the therapist with a set of functionalities to enhance clinical analysis and alleviate the documentation burden.

3.1 Real-Time Cognitive Monitoring

During the session, the system displays the real-time analysis of the patient's interventions. This immediate feedback allows the therapist to remain fully present while having a support system that monitors patterns that might otherwise be overlooked.

3.2 Automated Documentation

When the session ends, the application automatically analyzes it and generates reports and documentation: (1) a Clinical Note drafted in a standardized format like BIRP or SOAP; (2) an Emotional-Cognitive Report detailing the cognitive distortions detected along with emotion and topic patterns and triggers; and (3) Inter-Session Exercises personalized for the patient that match the patterns identified and come from a set configured by the therapist. It is important to note that the Clinical Note is fully editable, so the therapist already has a first draft that they can revise and expand or adjust as needed. As for the analytical reports, their content also remains under the professional's control through the interactive validation mechanism detailed in the following subsection.

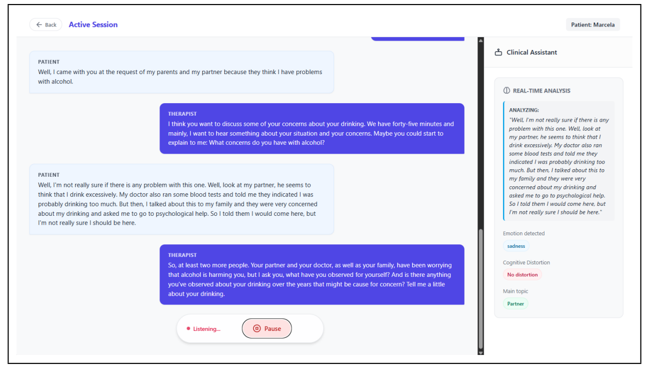

Figure 1: Real-Time Cognitive Monitoring. The interface displays the live session transcript with instantaneous speaker attribution. Detected cognitive distortions and emotions are visualized in-stream.

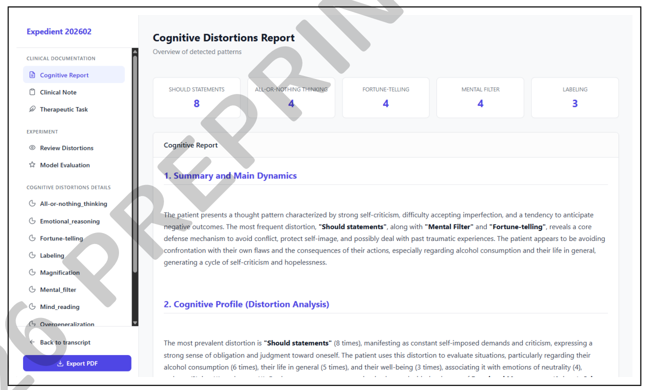

Figure 2: Automated Documentation Generation. The screenshot displays the Analytical Emotional-Cognitive Report. The interface also integrates the generated Clinical Note and suggested Inter-session Exercises.

3.3 Human-in-the-Loop Validation

To ensure clinical safety and trust, the "Validation View" presents the professional with the list of patient messages in which cognitive distortions have been detected. These messages are sorted by distortion category, and the professional can click on the one they want to review to display the conversation fragment from the session. The therapist can then confirm the classification, change it by selecting a different distortion from the list, or indicate that the message does not contain distorted thinking. This "Human-in-the-Loop" design allows the professional to refine the system's results. When this interaction is complete, the application regenerates the clinical reports, ensuring accuracy and reflecting expert judgment.

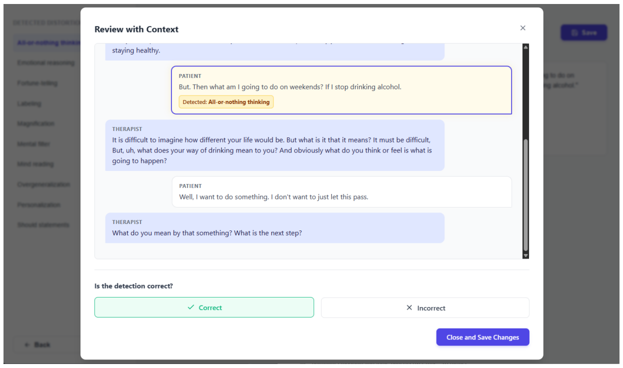

Figure 3: Human-in-the-Loop Validation. Therapists can audit and correct AI-detected patterns in context to ensure accurate, safe final documentation.

3.4 Longitudinal Patient Dashboard

Beyond capturing session data, it is important to be able to review it quickly and easily. Therefore, the application provides therapists with a comprehensive overview of the patient's history. All of the patient's clinical information is aggregated to show visual metrics and insights, such as the patient's predominant emotional and topic patterns over time. To further analyze the patterns and triggers, the interface also allows the professional to filter the data to show those that occurred in messages with distorted thinking. From here, the professional can access the reports generated for each session, but also has a micro summary to recall the context of each session. Finally, using the aggregated information and summaries, the assistant synthesizes the patient's evolution throughout the different sessions, helping to identify long-term trajectories and patterns.

3.5 System Customization and Calibration

The application has been designed to be adaptable to the professional with a customization section. The first thing the therapist must do before integrating the tool into their daily routine is to configure it. This begins with a system requirement: recording a sample of their voice so that the speaker recognition system can use it to attribute the dialogue to the correct speaker in the transcripts. To suit their documentation preferences, therapists can choose from several standardized templates for clinical notes. Finally, they can edit the list of tasks and therapeutic activities from which they want the system to select suggestions for the patient after each session.

4 Preliminary Expert Evaluation

To assess the clinical utility of the assistant, we conducted a pilot evaluation with two licensed clinical psychologists. To do this, we showed the experts a demonstration of how the application works. Then, the experts provided qualitative feedback and rated a series of statements on a 5-point Likert scale (1: Strongly Disagree, 5: Strongly Agree). In terms of privacy, both psychologists gave the highest score to the fact that processing data entirely on-device was a decisive factor for their trust and adoption. Automatic documentation was also highly valued, as both gave the maximum score to the fact that clinical notes are generated with a structure and content suitable for use, and both considered that it would reduce the time they spend on documentation. While the experts showed variance on whether the automated distortion report helps to detect overlooked patterns, both positively assessed these resources as an effective support resource for clinical practice and also positively assessed the longitudinal patient dashboard for revealing valuable patient progress insights that are difficult to capture with traditional documentation methods. However, Human-in-the-loop validation received mixed ratings. One expert's qualitative feedback warned that manually verifying AI detections could limit the time-saving potential offered by the assistant. Regarding the suggestions for Inter-Session Exercises, both experts concurred in remaining neutral, which possibly indicates a preference for maintaining total control over therapeutic interventions. Despite these concerns, the experts indicated that they would adopt the tool in their clinical practice, giving the highest rating for adoption.

5 Ethical Considerations

The deployment of Artificial Intelligence in mental health demands rigorous adherence to ethical standards, particularly regarding patient privacy and clinical safety. The proposed AI-based assistance system has been designed to comply with the ethical, legal, and security requirements established by the European AI Act [European Parliament and Council of the European Union, 2024] and the General Data Protection Regulation (GDPR) [European Parliament and Council of the European Union, 2016]. In accordance with the AI Act, the system is classified and developed as a high-risk AI application and incorporates human oversight and transparency to ensure reliable clinical support without replacing professional judgment. This Human-in-the-Loop mechanism serves as a guardrail, ensuring that final clinical responsibility and authorship remain exclusively with the qualified professional. GDPR compliance is ensured through data protection by design, including explicit patient consent, purpose limitation, data minimization, data encryption, and secure access control. These measures ensure that the system respects fundamental rights, patient safety, confidentiality, and ethical principles while enabling responsible and trustworthy use of AI in clinical psychological practice.

6 Conclusion

This paper presented an intelligent assistant that automates documentation and analysis tasks in mental health care without compromising data privacy. This is achieved by running deep learning models and LLMs locally, ensuring that data never leaves the therapist's device. In addition, the professional's expert judgment is maintained through editing functionalities and human-in-the-loop validation mechanisms, which also ensures security and trust. Finally, this edge computing approach establishes a viable path for the ethical integration of AI in mental health care.

Acknowledgments

This work was partially supported by Conselleria de Cultura, Educación y Ciencia, Generalitat Valenciana CI-PROM/2021/077, Spanish government projects PID2024-158227NB-C33 and TSI-100130-2024-0014, and First Research Project Support PAID-06-24 by Vicerrectorado de Investigación de la Universitat Politècnica de València.

References

[Beck, 1976] Aaron T. Beck. *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York, 1976.

[Brown *et al.*, 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[European Parliament and Council of the European Union, 2016] European Parliament and Council of the European Union. General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. Accessed: 2026-05-26.

[European Parliament and Council of the European Union, 2024] European Parliament and Council of the European Union. Artificial Intelligence Act. https://eur-lex.europa.eu/eli/reg/2024/1689/oj, 2024. Accessed: 2026-05-26.

[Google, 2026] Google. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed: 2026-05-20.

[Kim and Vossen, 2021] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Pico *et al.*, 2025] Aaron Pico, Joaquin Taverner, Emilio Vivancos, Ana García-Fornes, and Vicent Botti. CODI-PAD: Cognitive Distortions in Patient Discourse. https://doi.org/10.5281/zenodo.20307820, 2025. Dataset. Zenodo. Accessed: 2026-05-26.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, July 2023.

[Ravanelli *et al.*, 2021] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

[Schwartz-Dillard *et al.*, 2024] Jessica Schwartz-Dillard, Travis Ng, Joann Villegas, Derrick Johnson, and Mary Murray-Weir. Electronic documentation burden among outpatient rehabilitation therapists: a qualitative descriptive study and quality improvement initiative. *Journal of the American Medical Informatics Association*, 31(10):2347–2355, 2024.

[Shanafelt *et al.*, 2016] Tait D. Shanafelt, Lotte N. Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P. West. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. *Mayo Clinic Proceedings*, 91(7):836–848, 2016.

[Silero Team, 2025] Silero Team. Silero vad: Pre-trained enterprise-grade voice activity detector. https://github.com/snakers4/silero-vad, 2025. GitHub repository. Accessed: 2026-02-04.

[Wampold and Imel, 2015] Bruce E. Wampold and Zac E. Imel. *The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work*. Routledge, 2015.