LLMs in the Mental Health Context: Development of the Rafira Platform and the Use of Speech Recognition and AI in Therapy Sessions

Elze Pinheiro Lima Neto[1], Josivan Mesquita da Conceição[1], Nádia Félix Felipe da Silva[1], Thiago da Silva Fagundes[1]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
74.690-900 – Goiânia – GO – Brazil

elzeneto@gmail.com, jo.mesquita20@gmail.com, nadia@inf.ufg.br, kasfagundes@gmail.com

Abstract. *This paper presents the development and evaluation of the Rafira platform, an assistant for psychologists that uses speech recognition and Large Language Models (LLMs) for automated transcription and summarization of therapy sessions. This study aims to develop and evaluate an LLM-based platform considering criteria of accuracy, clinical applicability, usability, and ethics. A case study was conducted in three stages: prototype development, testing with psychologists from different approaches, and analysis of the results. Key findings demonstrate technical accuracy, clinical applicability, and the tool's scalability, contributing to the optimization of clinical practice and therapeutic follow-up.*

1. Introduction

Currently, technological innovations across different Information Technology (IT) sectors have permanently transformed the relationship between humans and machines. This relationship has become even more complex with the arrival of Intelligent Virtual Assistants (IVAs) linked to Large Language Models (LLMs). Nowadays, the interaction between people and IVAs, defined here as AI-driven interfaces designed to simulate human-like conversation to perform tasks or provide information, has become much more natural and humanized, which is only possible due to the Natural Language Processing (NLP) behind these assistants (JUNIOR; FARINA, FLORIAN, 2024).

However, this growing relationship also brings social impacts and concerns in different sectors, such as healthcare, especially regarding mental health (FREIRE, 2025). The use of AI in processes that depend on human interaction, such as psychological care, raises debates about the consequences of its use as a therapy tool, given the increase in global demand for more accessible psychological services (ZANQUETTA et al., 2024).

Based on this context, the following question arises: how can automated transcription with LLMs contribute to the therapeutic follow-up of patients and the optimization of the clinical practice of psychologists?. This study aims to develop and evaluate an LLM-based platform for the automated transcription and summarization of therapy sessions, considering criteria of accuracy, clinical applicability, usability, and ethics.

In psychological care settings, session documentation is an essential part of clinical follow-up, but it is also a time-intensive task and one that is prone to the loss of relevant details. This challenge becomes even more critical in high-demand scenarios, where professionals must balance attentive listening, record-keeping, and continuity of care.

Although recent advances in speech recognition and Large Language Models have expanded the possibilities for automating this process, their adoption in mental health contexts still faces critical requirements related to accuracy, privacy, anonymization, and clinical usefulness. In this context, this work presents the Rafira platform, conceived as a tool to support psychological practice, with a focus on the automated transcription and summarization of therapy sessions under human supervision, and discusses preliminary results from its technical and applied evaluation.

The clinical implementation of LLMs requires addressing the risk of misinformation and technical reliability. As noted by Asgari et al. (2025), a specialized framework is essential to assess 'clinical safety and hallucination rates of LLMs for medical text summarisation', ensuring that stochastic outputs remain within safe parameters. Furthermore, recent studies demonstrate that 'adapted large language models can outperform medical experts in clinical text summarization' (VAN VEEN et al., 2024), suggesting that, with proper technical grounding and ethical oversight, these tools can significantly optimize documentation workflows without compromising patient safety.

2. Proposed Solution: The Rafira Platform

The Rafira platform was developed as an assistant for psychologists, aimed at optimizing the documentation process of therapeutic sessions through automated transcription and summarization. The system's architecture is divided into specialized modules, with the **Rafira_IA** module being the core responsible for processing clinical data.

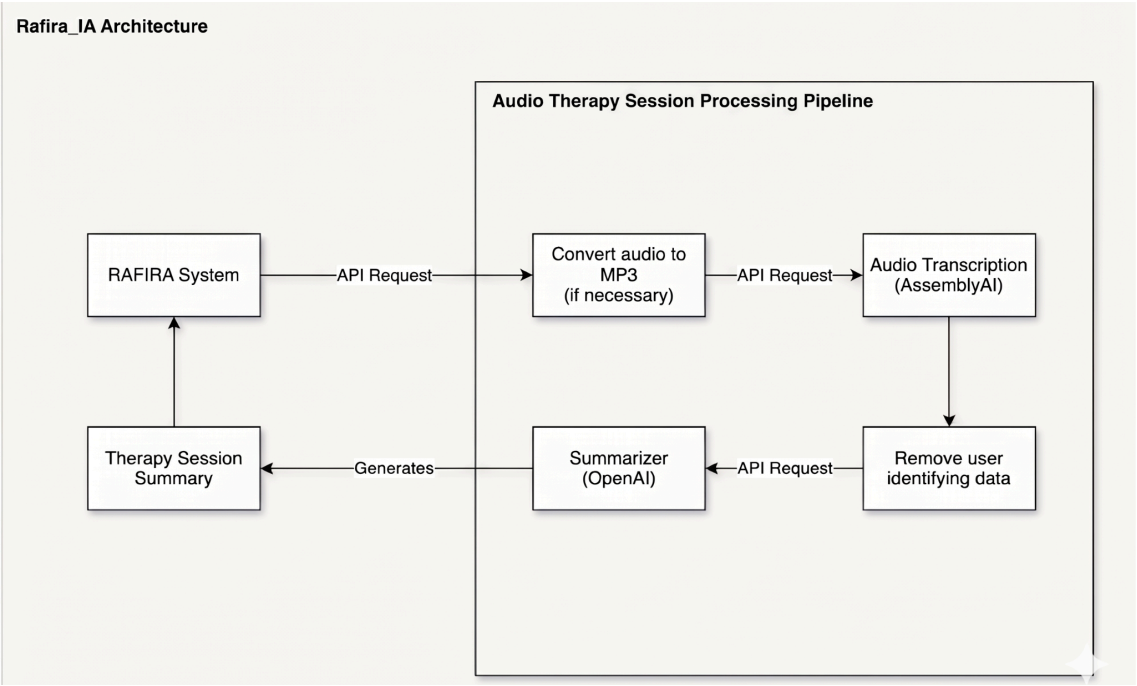

Figure 1. Rafira_AI Pipeline Architecture

The processing pipeline follows a rigorous flow to ensure data quality and security:

1. **Transcription:** Utilization of the *AssemblyAI* API for converting session audio into text, chosen for its high accuracy in Portuguese.
2. **Anonymization:** A critical step where sensitive information (names, locations, specific dates) is identified and masked to comply with the General Data Protection Law (LGPD), following the premises discussed by **Pereira et al. (2025)**.
3. **Structured Summarization:** Using LLMs (such as GPT-4o and Phi-4) to generate summaries organized into five clinical axes: Main Themes, Key Moments, Progress, Difficulties, and Session Plan.

The platform was built using a *backend* in Python (FastAPI) and a *frontend* in React, ensuring a responsive and secure interface for the professional. As highlighted in the study, the integration of these technologies aims not to replace the therapist, but to provide a robust tool for clinical support.

To mitigate clinical hallucinations, the LLMs (GPT-4o and Phi-4) were configured with a low temperature (T = 0.1) and top-p = 0.9. These hyperparameters prioritize deterministic and factual outputs, which are essential for clinical documentation. Regarding ethical compliance, the study was conducted following the approval of the psychologists and their respective patients, and all participating psychologists provided written Informed Consent (TCLE).

3. Validation Benchmarks and Results

To validate the Rafira platform, the study employed a twofold approach: a technical accuracy benchmark using the CORAA dataset (NILC, 2023), a public Brazilian Portuguese speech corpus with aligned transcriptions and samples from different accents and speaking contexts, and a qualitative evaluation with mental health professionals.

While CORAA is a general-purpose dataset, it was selected as a primary benchmark due to its robust representation of Brazilian Portuguese phonetic diversity, providing a reliable baseline for WER (Word Error Rate) before moving to domain-specific clinical corpora in future stages

3.1. Technical Accuracy (WER)

The use of the CORAA dataset serves as an initial phonetic baseline for Brazilian Portuguese. However, the authors acknowledge that a domain-specific clinical corpus is required for further validation. The qualitative evaluation (Section 3.2) involved three psychologists representing Psychoanalysis, Psychodrama, and Behavioral Therapy.

The base models of the Rafira platform were defined through accuracy tests aimed at identifying the best options by balancing cost and technical effectiveness. The models selected for testing and analysis were based on the following criteria:

Open-source or proprietary: both open-source and proprietary models were used in order to determine which type of licensing offered the best fit;

Number of parameters: model size was also taken into account, since smaller models may offer advantages in efficiency and deployment.

The main metric adopted was the **Word Error Rate (WER)** (Jurafsky; Martin, 2000), which is a standard metric for evaluating the performance of speech recognition systems. WER is used to quantify transcription errors by comparing the output generated by an LLM with a reference text.

Table 1. Transcription Model Evaluation Results

Modelo	Licença	Parâmetros	WER(%)
Phi-4-multimodal-instruct	Open	5,57B	23,41
Gemini-2.5-flash	Proprietary	Proprietary	23,61
Voxtral-Mini-3B-2507	Open	4,68B	24,15
AssemblyAI-best	Proprietary	Proprietary	24,26
Whisper-large-v3	Open	1,54B	32,57
GPT-4o-transcribe	Proprietary	Proprietary	32,86
Granite-speech-3.3-8b	Open	8,65B	63,57

The results indicate that **Phi-4**, an open-source model with a reduced parameter count, achieved the lowest WER, slightly surpassing robust proprietary models. This suggests high viability for specialized clinical applications with optimized maintenance costs.

3.2. Qualitative Feedback and Clinical Utility

During the development of the Rafira platform, a qualitative validation phase was conducted with the participation of three psychologists from different theoretical backgrounds: psychoanalysis, psychodrama, and behavioral therapy. The aim of this stage was to assess the applicability and perceived usefulness of the platform in real therapeutic practice.

As part of this evaluation process, the professionals were asked to complete a feedback form regarding their main impressions of the platform, and each also participated in in-depth listening interviews (Young, 2019), through which their main needs, expectations, and concerns were identified. The collected data were then analyzed to establish an initial professional assessment of Rafira's reception among mental health practitioners.

The platform's structured summarization feature generated automated insights into key aspects of each session, including Main Themes, Key Moments, Progress, Difficulties, and Next Session Plans. From a qualitative perspective, the results indicate important strengths as well as opportunities for improvement.

Overall, the feedback was highly positive. The professionals highlighted the automatic generation of session summaries as the platform's main benefit, particularly because it

supported the longitudinal tracking of patient progress and improved the organization of clinical information. In addition, the tool was perceived as reducing administrative burden, allowing therapists to dedicate more attention to the patient during the session. One participant, working in a hospital setting with a high volume of consultations and limited time between sessions, emphasized that the feature was especially valuable for supporting more efficient documentation and more assertive decision-making throughout the therapeutic process.

All participating professionals stated that they would incorporate the platform into their work routine and also expressed willingness to pay for the service. The psychologists were also asked to fill out a feedback form. The usability reached an average score of 4.7/5 (Likert Scale) across all categories, with 100% agreement on the clinical utility of the longitudinal tracking feature.

Among the main suggestions for improvement were the development of a mobile application to facilitate everyday use and the recommendation that, if session summaries are to be shared with patients, any suggestions related to referrals or next steps should be removed, thereby preserving the psychologist's autonomy in the technical conduct of treatment.

The initial feedback was highly positive:

- **Clinical Benefit**: Professionals highlighted the automatic generation of summaries as the most significant benefit, aiding in the longitudinal tracking of patient progress.
- **Efficiency**: The tool reduced administrative burden, allowing therapists to focus more on the patient during the session.
- **Acceptance**: All participating professionals expressed interest in incorporating Rafira into their routine and demonstrated a willingness to pay for the service.

4. Conclusion

This work presents initial evidence that the integration of Large Language Models with automated speech recognition can support psychological documentation in a clinically meaningful way. The preliminary results suggest that the Rafira platform has potential to reduce administrative burden, improve the organization of therapeutic information, and assist longitudinal follow-up in psychological practice. At the same time, the findings reinforce that human supervision remains essential, particularly in tasks involving interpretation, ethical judgment, and clinical decision-making. In this sense, rather than replacing the therapist, **Rafira** is positioned as a supportive tool for documentation and information management in mental health care. Future work will focus on expanding the evaluation with a broader set of professionals, refining the summarization outputs, and further investigating issues related to usability, privacy, and clinical reliability.

References

ASGARI, E.; MONTAÑA-BROWN, N.; DUBOIS, M.; KHALIL, S.; BALLOCH, J.; YEUNG, J. A.; PIMENTA, D. **A framework to assess clinical safety and hallucination rates of**

LLMs for medical text summarisation. *npj Digital Medicine*, v. 8, n. 1, p. 274, 2025. Available at: https://doi.org/10.1038/s41746-025-01670-7

BENTES, A., SANCHES, D., and FONSECA, P. (2024). Assistentes Virtuais Inteligentes e saúde mental: debates regulatórios no Brasil. *Revista Eletrônica de Comunicação, Informação & Inovação em Saúde*, v. 18, n. 3.

FIGUEIREDO JUNIOR, J. S., FARINA, R. M., and FLORIAN, F. (2024). Inteligência artificial (IA) na reprodução de vozes humanas: explorando vantagens e desafios. *Revista Científica Semana Acadêmica*, ed. 251, v. 12, p. 1-16. DOI: 10.35265/2236-6717-251-13060. Available at: http://dx.doi.org/10.35265/2236-6717-251-13060. Accessed: July 30, 2025.

FREIRE, T. (2025). "Sessão de terapia" no ChatGPT oferece riscos e preocupa especialistas. *Agência Brasil*. Available at: https://agenciabrasil.ebc.com.br/geral/noticia/2025-05/sessao-de-terapia-no-chatgpt-oferece-risco-e-preocupa-especialistas. Accessed: July 27, 2025.

JURAFSKY, D. and MARTIN, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd ed. (draft) . Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf.

NILC – Interinstitutional Center for Computational Linguistics (2023). *CORAA ASR – v 1.1: open speech dataset for Automatic Speech Recognition in Brazilian Portuguese* [GitHub repository]. Available at: https://github.com/nilc-nlp/CORAA. Accessed: August 7, 2025.

PEREIRA, A. M., MARTINS, L. F., SARTES, L. M. A., ALMEIDA, L. F., BERNARDINO, H. S., and SOUZA, J. F. (2025). Anonimização de Textos Clínicos Utilizando LLM. In: *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, 25., Porto Alegre. *Anais...* Porto Alegre: Sociedade Brasileira de Computação, p. 365-376. DOI: https://doi.org/10.5753/sbcas.2025.7150.

SILVEIRA, P. V. R. and PARAVIDINI, J. L. L. (2024). Ética da aplicação de inteligências artificiais e chatbots na saúde mental: uma perspectiva psicanalítica. *Revista Pesquisa Qualitativa*, v. 12, n. 30, p. 01-16. DOI: 10.33361/RPQ.2024.v.12.n.30.717. Accessed: July 27, 2025.

VAN VEEN, D.; VAN UDEN, C.; BLANKEMEIER, L.; DELBROUCK, J. B.; AALI, A.; BLUETHGEN, C.; PAREEK, A.; POLACIN, M.; REIS, E. P.; SEEHOFNEROVÁ, A.; ROHATGI, N.; HOSAMANI, P.; COLLINS, W.; AHUJA, N.; LANGLOTZ, C. P.; HOM, J.; GATIDIS, S.; PAULY, J.; CHAUDHARI, A. S. **Adapted large language models can outperform medical experts in clinical text summarization**. *Nature Medicine*, v. 30, n. 4, p. 1134-1142, 2024. Available at: https://doi.org/10.1038/s41591-024-02855-5

YOUNG, I. (2019). Listening Deeply. *Medium*. Available at: https://medium.com/inclusive-software/listening-deeply-fe55ac43bd32. Accessed: July 30, 2025.

ZANQUETTA, B., OLIVEIRA, M. V. M., OLIVEIRA, L. R., CINTRA, M. E., and FERNANDES, A. F. C. (2024). Avaliação de Assistentes Virtuais Baseados em Inteligência Artificial para Simulações de Atendimento Psicológico. Anais da ERI-ES. Accessed: October 15, 2025.