PhD Open Days

Speech Analysis with Large Language Models for Mental Health Assessment

Biomedical Engineering

Weiming Li (weimingli9999@gmail.com)

PÓLO DO I.S.T

Abstract

Depression and anxiety are highly prevalent and often under-detected in clinical interviews. This project develops a Large Language Model (LLM)-based Mental Health Interview Support System that uses real-time transcription, role recognition, and emotion detection to flag key patient statements. A fine-tuned DeepSeek Reasoning Model 1 (DeepSeek-R1) with Retrieval-Augmented Generation (RAG) scores utterances according to the Patient Health Questionnaire-9 (PHQ-9) and the Generalized Anxiety Disorder 7-item scale (GAD-7), linking text to diagnostic items for transparent results. A working prototype shows the feasibility of real-time, explainable symptom analysis, with plans to expand to multimodal signals such as facial expressions and heart rate.

Research Objectives

- Apply LLMs for mental health assessment.
- Combine PHQ-9 and GAD-7 with interview analysis.
- Provide dynamic, contextual symptom insights.
- Support transparent clinical evaluation.

Methodology

- Transcribes interview speech with Whisper and translates when needed.
- Uses role recognition to separate patient and clinician speech.
- Detects negative/affective content with emotion classification.
- Analyzes flagged sentences using fine-tuned DeepSeek-R1 (LoRA) trained on PHQ-9/GAD-7.
- Integrates DSM-5, ICD-11, and questionnaires through RAG for knowledge grounding.
- Outputs structured scores and explanatory text linking utterances to PHQ items.
- Enables transparent clinical assessment (Figure 1).

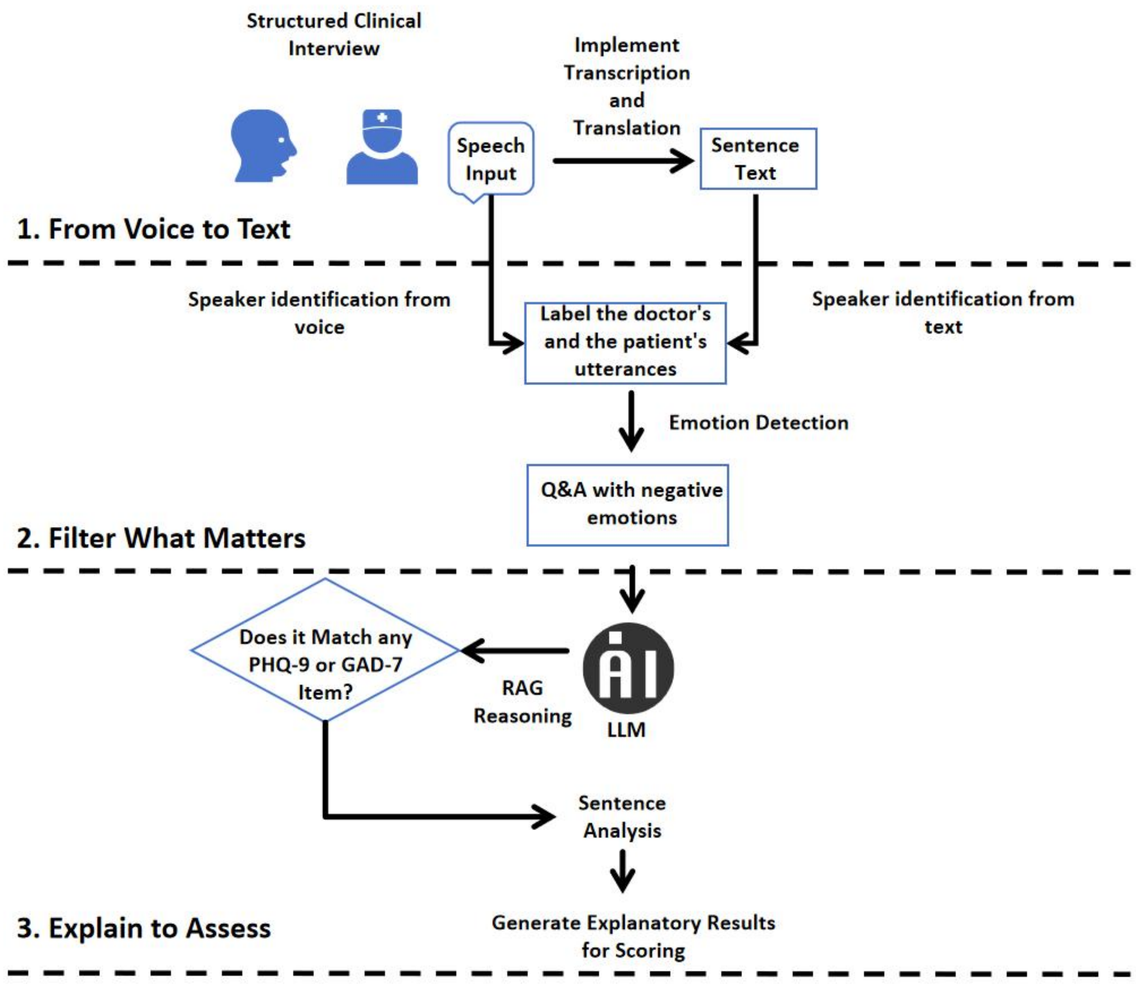

Figure 1. System Architecture for LLM-based Clinical Interview Analysis

Prototype Demo

A prototype demonstration of the proposed system was developed, as shown in Figure 2. The demo interface integrates real-time speech transcription, role recognition, and negative sentence detection. Flagged patient utterances were processed by a fine-tuned DeepSeek-R1 model with RAG support, generating sentence-level analyses linked to PHQ and GAD items. The system produced a structured dialogue history and an explanatory summary report. These results illustrate the feasibility of combining LLM fine-tuning and retrieval-augmented reasoning to deliver transparent outputs, while serving as an initial proof-of-concept for clinical applications.

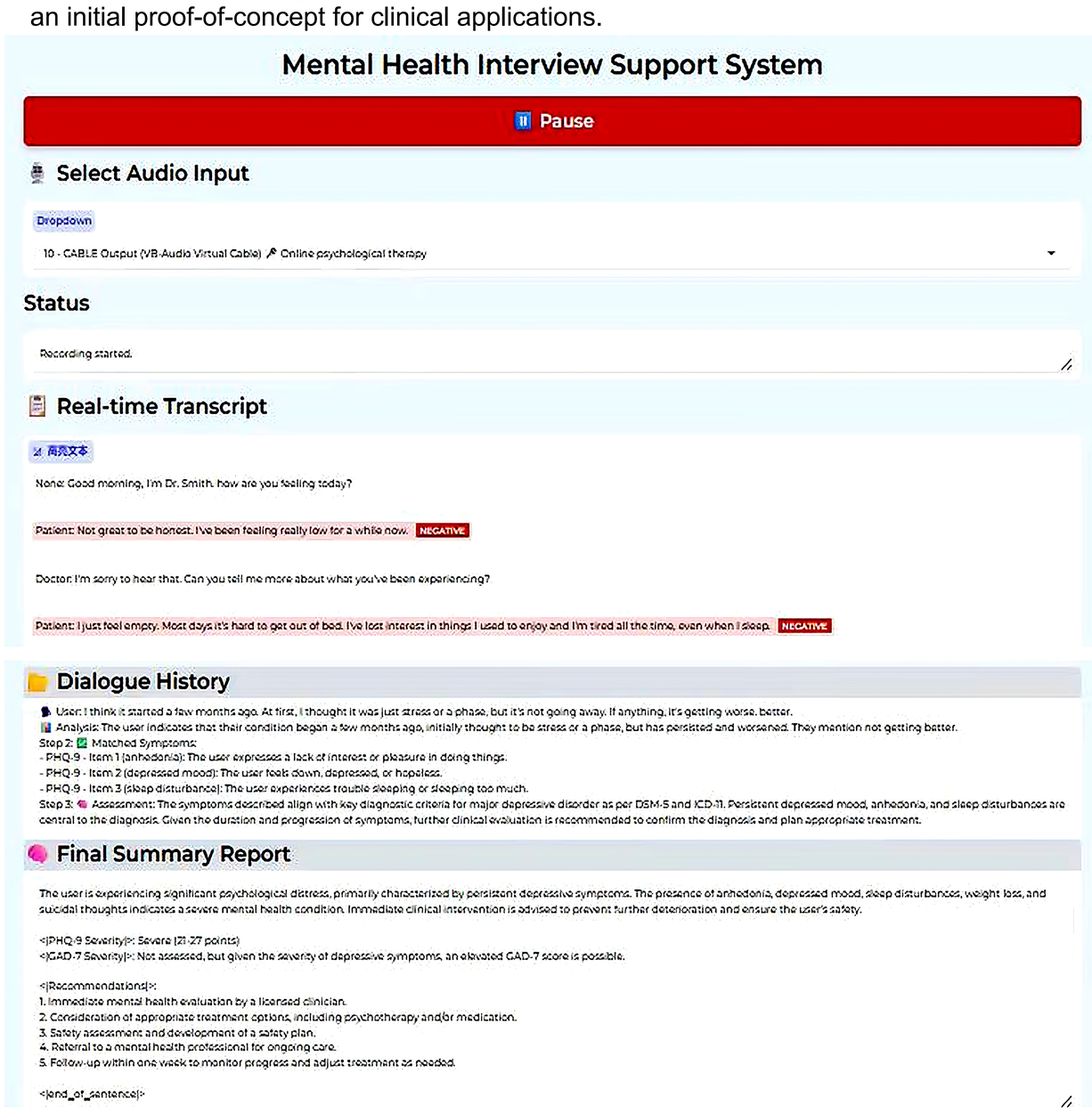

Figure 2. Prototype Demo Interface with Scoring and Explanatory Outputs

Future Work

Future development will extend this system into an LLM-based multimodal assessment platform, combining speech and prosody, text, facial expressions, eye blinks, and heart rate. This integration aims to capture subtle behavioral and physiological signals, providing clinicians with richer, transparent, and personalized mental health evaluations.

REFERENCES

[1] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (pp. 28492-28518). PMLR.
[2] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., ... & Gill, M. P. (2020, May). Pyannote. audio: neural building blocks for speaker diarization. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7124-7128). IEEE.
[3] Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. arXiv preprint arXiv:2005.00547.
[4]Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.
[5]Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.
[5] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
[6] Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ- 9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9), 606-613.
[7] Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of internal medicine, 166(10), 1092-1097.
[8] Vahia, V. N. (2013). Diagnostic and statistical manual of mental disorders 5: A quick glance. Indian journal of psychiatry, 55(3), 220-223.
[9] World Health Organization. (2004). International Statistical Classification of Diseases and related health problems: Alphabetical index (Vol. 3). World Health Organization.