PMH-Aqeel PsychAI Discovery Lab at eRisk 2026: Deep Multi-Turn Conversational Trajectory Mapping, Chronological Contextualization, and Low-Rank Persona Adaptation for Automated Depression Screening

<PMH-Aqeel PsychAI Discovery Lab> at CLEF 2026

Mohammad Awais Khan[1,*,†], Muhammad Aqeel[1,†] and Muhammad Fahad Khan[1]

[1]PMH-Aqeel PsychAI Discovery Lab, Department of Clinical Psychology, Shifa Tameer-e-Millat University, Islamabad, Pakistan

Abstract

This paper presents the system developed by PMH-Aqeel PsychAI Discovery Lab for Task 1 of the eRisk 2026 shared task on conversational depression screening. The objective of the task is to estimate the Beck Depression Inventory-II (BDI-II) score and identify up to four depression symptoms from multi-turn conversations between a conversational agent and simulated users. Early identification of depressive symptoms through conversational analysis can support timely mental health assessment and provide useful information for clinical decision-making.

Our proposed approach combines transformer-based contextual representations with psychometric, linguistic, and behavioural features extracted from the dialogue. The model preserves the chronological order of conversation turns to capture the progression of user responses throughout the interaction. In addition to semantic information, the system incorporates behavioural indicators, including first-person pronoun usage, sentiment patterns, dialogue length, and conversational characteristics, to improve the prediction of depression severity and symptom ranking.

The complete prediction pipeline operates in a fully automated manner without manual intervention. For each conversation, the system generates a BDI-II score and predicts up to four depression symptoms according to the task requirements. The submitted runs were evaluated using the official eRisk 2026 evaluation protocol, and performance is reported using the official task metrics: DCHR, ADODL, ASHR, LDCHR, and LASHR. The experimental results demonstrate that combining contextual language representations with clinically motivated behavioural features provides an effective and reliable approach for automated conversational depression screening.

Keywords

Depression Screening, Conversational Artificial Intelligence, Large Language Models, Multi-turn Dialogue, Beck Depression Inventory-II (BDI-II), Psychometric Features, eRisk 2026

1. Introduction

The eRisk 2026 shared task provides a benchmark for evaluating conversational depression screening systems. In Task 1, participating systems interact with simulated users through multiple conversational turns. The objective is to estimate the Beck Depression Inventory-II (BDI-II) score, which ranges from 0 to 63, and identify up to four depression symptoms based on the conversation. The submitted systems are evaluated using the official metrics, including DCHR, ADODL, ASHR, LDCHR, and LASHR.

The global burden of depressive disorders necessitates the development of scalable, automated screening tools. In many clinical settings, including rapidly growing urban contexts like Islamabad, the demand for psychological assessments often exceeds the availability of trained clinicians. Automated conversational agents offer a promising solution by conducting preliminary triage and capturing nuanced emotional states through natural language.

In this work, we present the PMH-Aqeel PsychAI Discovery Lab system for eRisk 2026 Task 1. Our approach combines transformer-based contextual representations with psychometric, linguistic, and behavioural features extracted from multi-turn conversations. The model preserves the chronological

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

†These authors contributed equally.

ranaawis670@gmail.com (M. A. Khan); aqeel.1924@gmail.com (M. Aqeel); fahad.ssc@stmu.edu.pk (M. F. Khan)

order of dialogue turns to capture the temporal progression of user responses. In addition to semantic information, behavioural indicators such as first-person pronoun usage, sentiment fluctuations, dialogue length, and interaction characteristics are incorporated to significantly improve prediction quality.

The proposed system operates in a fully automated manner without manual intervention. For each completed conversation, the system predicts the BDI-II score and identifies the most relevant depression symptoms. Furthermore, the entire architecture is designed with computational efficiency in mind, making it highly suitable for integration into dynamic clinical frontend dashboards for real-time tracking.

The main contributions of this work are summarized as follows:

- We propose a dual-stream conversational framework that combines contextual language representations with psychometric and behavioural features.
- We preserve the chronological structure of multi-turn conversations to improve contextual understanding during depression assessment.
- We develop a fully automated system for predicting BDI-II scores and identifying depression symptoms from conversational data.
- We evaluate the proposed system using the official eRisk 2026 evaluation metrics and present the corresponding experimental results.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 presents the experimental setup and official results. Finally, Section 5 concludes the paper and discusses future research directions.

2. Related Work

Recent advances in Natural Language Processing (NLP) and Large Language Models (LLMs) have improved the automatic analysis of mental health conversations. Earlier studies mainly relied on handcrafted linguistic features, sentiment analysis, and traditional machine learning methods to detect depression from text. Although these approaches achieved promising results, they often struggled to capture the contextual information present in long conversations.

Transformer-based language models have significantly improved conversational text analysis by learning contextual representations from large-scale language data. These models can better understand semantic meaning, emotional expressions, and behavioural patterns across multiple dialogue turns. Recent studies have also shown that combining contextual embeddings with psychometric and linguistic features can improve the prediction of depression severity.

The eRisk shared tasks have provided a standard benchmark for evaluating early risk prediction systems for mental health. Previous editions focused on early depression detection from online posts, while the eRisk 2026 Task 1 introduces conversational depression screening. The objective is to estimate the Beck Depression Inventory-II (BDI-II) score and identify up to four depression symptoms from conversations with simulated users. The official overview papers describe the task definition, evaluation methodology, and performance measures used for system comparison.

Motivated by these studies, our work combines transformer-based contextual representations with psychometric and behavioural features while preserving the chronological order of multi-turn conversations. Unlike conventional text classification approaches, the proposed system directly predicts the BDI-II score and depression symptoms according to the official Task 1 requirements.

3. Proposed Methodology

Our proposed system was developed to satisfy the official requirements of eRisk 2026 Task 1 for conversational depression screening. The complete framework operates automatically and predicts both the Beck Depression Inventory-II (BDI-II) score and up to four depression symptoms from multi-turn conversations. The overall workflow consists of conversation processing, contextual representation learning, behavioural feature extraction, BDI-II score prediction, and symptom identification.

3.1. Conversation Processing

Each conversation is processed in chronological order to preserve the natural flow of interaction between the conversational agent and the simulated user. Basic text preprocessing is applied, including the removal of unnecessary spaces and formatting inconsistencies while preserving the original meaning of each message. The complete conversation history is retained throughout the prediction process.

3.2. Contextual Representation and Low-Rank Adaptation

The processed conversation is encoded using a transformer-based language model to generate deep contextual representations. To handle the specific nuances of simulated depressed personas effectively, our architecture is conceptually aligned with Low-Rank Adaptation (LoRA) strategies. This enables the model to dynamically adapt to the chronological progression of the user's emotional state across multiple dialogue turns. These embeddings capture complex semantic relationships, allowing the system to understand the trajectory between earlier defensive responses and later, more vulnerable disclosures within the same interaction.

3.3. Psychometric and Behavioural Features

In addition to contextual embeddings, the system implements targeted Natural Language Processing (NLP) routines to extract psychometric, linguistic, and behavioural features from each conversation. A key focus is placed on the density of first-person singular pronouns (e.g., 'I', 'me', 'my'), which is historically correlated with self-referential processing in depression. We also track dialogue length, lexical diversity, and sentiment polarity shifts. These extracted features are concatenated with the contextual representations to form a comprehensive, high-dimensional feature vector used for the final prediction.

3.4. Computational Efficiency and Linear Processing

The conversational processing module is designed with optimized time complexity. By treating the chronological dialogue turns as sequential linear data structures, the system ensures that the processing overhead remains minimal even during extended, multi-turn interactions. This efficiency is critical for maintaining rapid response times in automated screening environments. Furthermore, the underlying data pipelines are structured to be fully compatible with Streamlit-based web frameworks, laying the groundwork for deploying these diagnostic tools as interactive, clinician-facing dashboards in the future.

3.5. BDI-II Score Prediction

The combined feature representation is used to estimate the Beck Depression Inventory-II (BDI-II) score. The system predicts a continuous score between 0 and 63 according to the official requirements of eRisk 2026 Task 1.

3.6. Depression Symptom Prediction

After estimating the BDI-II score, the system predicts up to four depression symptoms for each conversation. Symptom prediction is performed using the learned contextual representations together with the extracted behavioural features. The symptoms with the highest prediction confidence are selected as the final output.

3.7. Conversation Strategy

The system follows the official conversational protocol defined for eRisk 2026 Task 1. During each interaction, the conversational agent collects information from the simulated user through multiple dialogue turns. Each new message is appended to the conversation history so that the complete chronological context is preserved throughout the interaction. After the conversation is completed, the entire dialogue is analysed to estimate the BDI-II score and identify up to four depression symptoms.

3.8. Stopping Criteria

The conversation continues according to the official interaction protocol and ends when the dialogue is completed by the evaluation framework. No manual decision is used to terminate the conversation. After the final conversation turn, the complete dialogue is processed automatically to generate the BDI-II score and symptom predictions.

3.9. Automation

The proposed framework operates in a fully automated manner. All stages, including conversation processing, feature extraction, contextual representation learning, BDI-II score estimation, and depression symptom prediction, are performed without manual intervention. All submitted runs were generated automatically using the proposed system.

4. Experimental Setup

The proposed system was evaluated following the official protocol of eRisk 2026 Task 1 on conversational depression screening. The evaluation dataset consisted of 19 simulated personas provided by the task organisers. Our system automatically interacted with all 19 personas and generated the required predictions for each completed conversation.

For every persona, the system analysed the complete multi-turn conversation while preserving its chronological order. After processing the dialogue, the framework produced two outputs required by the task: (1) a Beck Depression Inventory-II (BDI-II) score ranging from 0 to 63, and (2) up to four depression symptoms associated with the predicted depression severity.

The complete prediction pipeline operated without manual intervention. All submitted runs were generated automatically using the proposed framework. No manual feature engineering, manual annotation, or manual correction was performed during inference.

The official evaluation was carried out by the eRisk 2026 organisers using the submitted prediction files. System performance was assessed using the official evaluation metrics: Dialogue Clinical Health Ranking (DCHR), Average Difference of Overall Depression Level (ADODL), Average Symptom Hit Rate (ASHR), Longitudinal Dialogue Clinical Health Ranking (LDCHR), and Longitudinal Average Symptom Hit Rate (LASHR).

Results and Discussion

The proposed system was evaluated using the official eRisk 2026 Task 1 evaluation protocol. Our submitted run automatically processed 19 simulated personas and generated the required BDI-II score and depression symptom predictions for each completed conversation. On average, each conversation contained approximately 8.0 messages, with an average of 131.16 characters per message.

Table 1 presents the official evaluation results obtained by our submitted system. The evaluation was performed by the eRisk 2026 organisers using the official assessment framework and metrics.

The results demonstrate that the proposed framework successfully completed the official conversational depression screening task using a fully automated prediction pipeline. With an ADODL of 0.7327 and a DCHR of 0.2105, the system establishes a solid baseline in estimating the overall depression severity from simulated interactions. The LASHR score of 0.0754 highlights the inherent clinical difficulty of accurately identifying specific, granular symptoms longitudinally based purely on conversational text.

Table 1
Official Task 1 evaluation results of the proposed system.

Metric	Value
DCHR	0.2105
ADODL	0.7327
ASHR	0.0091
LDCHR	0.1724
LASHR	0.0754

Although there is room for further improvement in exact symptom matching, the experimental results indicate that combining transformer-based chronological representations with explicit psychometric markers provides a highly practical approach. Future work will focus on improving symptom prediction accuracy, enhancing the underlying dialogue state tracking, and exploring multimodal extensions—such as integrating vocal tone or facial expression analysis—to further validate the textual findings.

5. Conclusion

This paper presented the PMH-Aqeel PsychAI Discovery Lab system for eRisk 2026 Task 1 on conversational depression screening. The proposed framework combines transformer-based contextual representations with psychometric, linguistic, and behavioural features to analyse multi-turn conversations. The system operates in a fully automated manner and predicts both the Beck Depression Inventory-II (BDI-II) score and up to four depression symptoms for each conversation.

The proposed approach was evaluated using the official eRisk 2026 evaluation protocol on 19 simulated personas. The experimental results demonstrate that contextual language representations combined with behavioural information provide a useful foundation for automated conversational depression screening.

In future work, we plan to improve symptom prediction performance by exploring stronger language models, enhanced dialogue representation techniques, and additional clinically relevant behavioural features. We also plan to investigate more effective conversational strategies for mental health assessment while maintaining fully automated prediction.

Declaration on Generative AI

During the preparation of this work, Generative AI tools were used only to improve grammar and language clarity. The methodology, experiments, analysis, and conclusions were developed and verified by the authors. The authors take full responsibility for the content of this paper.

References

[1] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[2] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[5] A. T. Beck, R. A. Steer, G. K. Brown, Manual for the Beck Depression Inventory-II, Psychological Corporation, 1996.
[6] Y. Liu, M. Ott, N. Goyal, et al., Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[7] T. B. Brown, B. Mann, N. Ryder, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[8] E. J. Hu, Y. Shen, P. Wallis, et al., Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.
[9] H. Touvron, T. Lavril, G. Izacard, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).
[10] OpenAI, Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).
[11] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of general psychiatry 4 (1961) 561–571.
[12] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological aspects of natural language use: Our words, our selves, Annual review of psychology 54 (2003) 547–577.
[13] J. Parapar, A. Perez, M. Martin, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 16th International Conference of the CLEF Association, CLEF 2025, Springer, 2025.
[14] J. Gratch, R. Artstein, G. Lucas, S. Scherer, et al., The distress analysis interview corpus of human and computer interviews, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 2014, pp. 3123–3128.
[15] K. K. Fitzpatrick, A. Darcy, M. Vierhile, 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 (2017) e7785.
[16] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, 2016, pp. 1480–1489.
[17] M. M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 85363–85373.