erisk_cedri at eRisk 2026 Task 1: Conversational Depression Detection Using LoRA-Adapted Symptom Classifiers

Filipa Afonso[1,*], Rui Pedro Lopes[1]

[1]CeDRI – Research Centre in Digitalization and Intelligent Robotics, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Abstract

This paper describes the participation of team erisk_cedri in Task 1 of eRisk 2026, which focused on conversational depression detection via LLM-based personas. The proposed approach used 21 specialized LoRA-adapted DeBERTa-v3-Large classifiers to analyze persona responses and estimate BDI-II depression scores. The system employed three different conversational strategies across three runs, with an early stopping mechanism to minimize the number of conversational turns while maintaining detection accuracy. Predictions were submitted for 18 out of 20 personas. The best run achieved a Depression Category Hit Rate (DCHR) of 0.3889, placing 6th among incomplete submissions. The latency-aware symptom hit rate (LASHR) of 0.1496 in Run 2 ranked 3rd, demonstrating effective symptom identification with low conversational overhead. This paper discusses the challenges of indirect depression inference through conversation and presents lessons learned for future iterations.

Keywords

Depression detection, Conversational AI, BDI-II, LoRA fine-tuning, eRisk 2026

1. Introduction

Depression is a prevalent mental health condition affecting millions worldwide, and its early detection through digital traces has become an active research area. The eRisk Lab [1] has organized shared tasks on early risk prediction since 2017, focusing on identifying signs of depression, self-harm, and other mental health conditions from social media data.

Task 1 of eRisk 2026 introduced a novel conversational paradigm for depression detection [2, 3]. Unlike previous editions that analyzed static user-generated content, this task required participants to actively interact with LLM-based personas to infer their depression status. The personas were implemented as LoRA adapters on Meta-Llama-3-8B-Instruct and were designed to simulate individuals with varying levels of depression severity according to the Beck Depression Inventory-II (BDI-II).

The main objectives were twofold: (i) estimate the overall depression severity level (BDI-II score ranging from 0 to 63), and (ii) identify which of the 21 BDI-II symptoms were present in each persona. A key challenge was that personas were explicitly designed to avoid answering direct questions about their mental health, requiring systems to infer depression indicators indirectly through conversational analysis.

Team erisk_cedri participated with three fully automated runs, employing different conversational strategies while using the same underlying symptom detection architecture. This paper describes the approach and presents the results from this first official edition of the conversational depression detection task.

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

*Corresponding author.

a61183@alunos.ipb.pt (F. Afonso); rlopes@ipb.pt (R. P. Lopes)

0000-0000-0000-0000 (F. Afonso); 0000-0000-0000-0000 (R. P. Lopes)

2. Related Work

Depression detection from text has been extensively studied using social media data [4]. Traditional approaches relied on linguistic features, sentiment analysis, and machine learning classifiers trained on user-generated content. Recent work has leveraged transformer-based models such as BERT and RoBERTa for improved performance [5].

The eRisk Lab has been instrumental in advancing this field. Task 1 in previous editions focused on sentence-level ranking for symptom detection, where systems ranked sentences by their relevance to specific BDI-II symptoms. Top-performing systems in eRisk 2024 and 2025 combined semantic embeddings, contrastive learning, and LLM-generated synthetic data.

The conversational paradigm introduced in 2025 as a pilot and consolidated in 2026 represents a significant shift. Rather than analyzing existing content, systems must actively engage in dialogue to elicit relevant information. This setting mirrors clinical interviews where practitioners ask open-ended questions to assess mental health. Similar conversational approaches have been explored in mental health chatbots [6], though the evaluation of automated depression inference through conversation remains nascent.

Parameter-efficient fine-tuning techniques such as LoRA (Low-Rank Adaptation) [7] have enabled the adaptation of large language models with reduced computational requirements. The proposed approach builds on this technique to create specialized symptom classifiers from a shared base model.

3. System Description

The system consisted of two main components: (i) a conversational agent that generated questions and interacted with the personas, and (ii) a symptom detection module that analyzed responses to estimate BDI-II scores.

3.1. Conversational Strategy

Three distinct conversational strategies were designed for the three runs, each with a different number and style of questions:

Run 1 – Comprehensive approach: Up to 10 questions covering all major symptom clusters, starting with general wellbeing and progressively exploring sleep, energy, activities, enjoyment, appetite, mood, future outlook, and self-perception.

Run 2 – Focused approach: Up to 6 questions with more direct probing of mood, enjoyment, sleep/energy, and self-image.

Run 3 – Deep exploration: Up to 8 questions emphasizing emotional states, daily experiences, coping mechanisms, and social connection.

All questions were designed to be indirect and conversational, avoiding explicit mental health terminology. For example, rather than asking "Do you feel depressed?", the system asked "How would you describe your mood lately?" or "Is there anything that brings you joy these days?"

The official system prompt provided by the organizers was used verbatim, as required by the task protocol.

3.2. Symptom Detection Module

The symptom detection component analyzed each persona response using 21 specialized binary classifiers, one for each BDI-II symptom. The architecture was based on DeBERTa-v3-Large [8] fine-tuned with LoRA adapters.

Base model training: A multilabel classifier was first trained on the eRisk 2025 Task 1 training data, which contained approximately 17.5 million sentences from Reddit users with expert annotations for symptom relevance. Training used BCEWithLogitsLoss, batch size of 4 with gradient accumulation of 4, learning rate of 2×10^{-5}, and 3 epochs.

LoRA specialization: Subsequently, 21 individual LoRA adapters were trained on the multilabel base model, each specializing in detecting a single symptom. LoRA configuration used rank $r = 16$, scaling factor $\alpha = 32$, dropout of 0.1, and targeted the query and value projection layers. This reduced trainable parameters from 435 million to approximately 1.5 million per adapter.

Data augmentation: To address data scarcity for certain symptoms (particularly S6 Punishment Feelings, S9 Suicidal Thoughts, and S11 Agitation), 1,350 synthetic training sentences were generated using Claude 3.5 Sonnet. These were designed to capture diverse linguistic expressions of each symptom in conversational contexts.

3.3. BDI-II Score Estimation

For each response, all 21 symptom classifiers produced probability scores. Cumulative maximum probabilities were maintained across the conversation and converted to a BDI-II score using threshold-based mapping:

$$\text{BDI} = \sum_{i=1}^{21} \text{score}_i, \quad \text{score}_i = \begin{cases} 3 & \text{if } p_i \geq \tau_h \\ 2 & \text{if } \tau_m \leq p_i < \tau_h \\ 1 & \text{if } \tau_l \leq p_i < \tau_m \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where p_i is the maximum probability observed for symptom i, and thresholds τ_h, τ_m, τ_l were set to 0.7, 0.5, and 0.3 for Runs 1 and 3, and slightly lower (0.65, 0.45, 0.25) for Run 2.

3.4. Early Stopping Mechanism

To minimize conversational turns while maintaining detection accuracy, an early stopping mechanism was implemented based on multiple criteria:

- Severe depression (BDI $\geq$ 25): stop after 3 turns
- Moderate depression (BDI $\geq$ 20): stop after 4 turns
- Mild depression (BDI $\geq$ 16): stop after 5 turns
- Strong symptom signal (5+ symptoms with probability $\geq$ 0.7): stop after 3 turns
- Core depression cluster active (Sadness, Loss of Pleasure, Loss of Interest, Loss of Energy): stop after 4 turns

This approach aimed to balance early detection with avoiding false positives for control personas.

4. Results

Predictions were submitted for 18 out of 20 personas due to time constraints during the final submission weeks. Table 1 presents the official results.

Table 1
Official results for erisk_cedri (18/20 personas). DCHR = Depression Category Hit Rate, ADODL = Average Difference between Overall Depression Levels, ASHR = Average Symptom Hit Rate. L-prefixed metrics are latency-weighted versions.

Run	DCHR	Rank	ADODL	Rank	ASHR	Rank
#1	0.1111	17	0.8325	20	0.1667	7
#2	0.2222	13	0.8474	14	0.1806	6
#3	0.3889	6	0.8686	8	0.0972	14

Table 2
Latency-weighted results for erisk_cedri.

Run	LDCHR	Rank	LASHR	Rank
#1	0.0878	20	0.1200	7
#2	0.1724	16	0.1496	3
#3	0.2817	5	0.0758	14

Run 3 achieved the best overall performance with DCHR = 0.3889 (6th among incomplete submissions) and ADODL = 0.8686 (8th). The latency-weighted LDCHR of 0.2817 ranked 5th, indicating efficient conversational strategies.

Notably, Run 2 achieved the third-highest LASHR (0.1496) among all submissions in the incomplete category, demonstrating effective symptom identification with low latency. The system averaged 4.8 messages per conversation, among the lowest in the task, reflecting the efficiency of the early stopping mechanism.

4.1. Interaction Statistics

Across all submissions, the system averaged 4.8 messages per conversation (one of the lowest among participants) with an average message length of 82.79 characters (also among the shortest). This concise interaction style was deliberate, prioritizing efficient evidence gathering over exhaustive questioning.

5. Discussion

5.1. Analysis of Results

The significant variation in performance across the three runs (DCHR ranging from 0.11 to 0.39) suggests high sensitivity to the conversational strategy employed. Run 3's deep exploration approach, which emphasized emotional states and coping mechanisms, proved most effective for depression category classification.

The discrepancy between ASHR and DCHR in Run 3 (low symptom identification but relatively good category classification) indicates that the system may have achieved correct severity estimates without precisely identifying the constituent symptoms. This could reflect that certain linguistic patterns correlate with overall depression severity without mapping cleanly to individual BDI-II items.

The strong LASHR performance in Run 2 (3rd place) demonstrates that the early stopping mechanism successfully balanced detection speed with accuracy, at least for symptom identification.

5.2. Limitations

This work has several limitations that should be acknowledged:

Incomplete submission: Only 18 of 20 personas were submitted due to time constraints, limiting direct comparison with teams that achieved full coverage.

Short messages: The concise questioning style (82 characters average) may have limited the depth of responses available for analysis. Teams with longer, more detailed prompts achieved higher performance on several metrics.

Indirect inference challenge: The personas' reluctance to discuss mental health directly made symptom identification particularly difficult. The classifiers, trained on explicit symptom expressions, may not have fully captured the subtle, indirect manifestations in conversational responses.

Gap to top performers: The best-performing team (pjmathematician) achieved ADODL = 0.9254 and LDCHR = 0.3872 with complete coverage, substantially outperforming these results.

5.3. Lessons Learned

Several insights emerged from this participation:

First, conversational strategy matters significantly. The choice of questions, their ordering, and their phrasing substantially impacted detection accuracy.

Second, classifiers trained on static social media text may not transfer directly to conversational contexts. The personas' responses had a different linguistic character than Reddit posts used for training.

Third, early stopping can be effective but requires careful calibration. The conservative thresholds avoided false positives but may have missed subtle depression signals in some personas.

6. Conclusion and Future Work

This paper presented the erisk_cedri system for conversational depression detection at eRisk 2026 Task 1. The approach combined LoRA-adapted symptom classifiers with strategic questioning and early stopping. Despite submitting an incomplete set of predictions, competitive results were achieved on several metrics, particularly the latency-weighted symptom hit rate.

Future work should address several directions: (i) developing classifiers specifically trained on conversational data rather than static posts, (ii) exploring adaptive questioning strategies that adjust based on persona responses, (iii) investigating longer, more exploratory conversations that may elicit richer evidence, and (iv) ensuring complete coverage of all evaluation personas.

Acknowledgments

This work was supported by FCT – Fundação para a Ciência e Tecnologia, I.P. by projects: CeDRI, UID/05757/2025 (DOI: 10.54499/UID/05757/2025) and UID/PRR/05757/2025 (DOI: 10.54499/UID/PRR/05757/2025); SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

Declaration on the Use of Generative AI

Claude (Anthropic) was used for assistance with drafting, paraphrasing, and grammar checking during the preparation of this manuscript. The experimental design, implementation, and analysis were conducted by the authors.

References

[1] Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019 early risk prediction on the Internet. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2019. LNCS, vol. 11696, pp. 340–357. Springer (2019)

[2] Perez, A., Parapar, J., Wang, X., Crestani, F.: 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. CEUR Workshop Proceedings. CEUR-WS.org (2026)

[3] Perez, A., Parapar, J., Wang, X., Crestani, F.: 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. Lecture Notes in Computer Science. Springer (2026)

[4] Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: CLEF 2016. LNCS, vol. 9822, pp. 28–39. Springer (2016)

[5] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT 2019, pp. 4171–4186 (2019)

[6] Abd-Alrazaq, A.A., et al.: An overview of the features of chatbots in mental health: A scoping review. Int. J. Med. Inform. 132, 103978 (2019)
[7] Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR 2022 (2022)
[8] He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: ICLR 2021 (2021)