pjmathematician @ eRisk 2026: Automating Conversational Depression Detection

Notebook for the eRisk Lab at CLEF 2026

Poojan Vachharajani[1,*]

[1]*Amazon*

Abstract

This paper describes the participation of team **pjmathematician** in Task 1 (Conversational Depression Detection with LLM Personas) of eRisk 2026. In this task, participants interact with simulated personas released as LoRA adapters on top of `Meta-Llama-3-8B-Instruct` and must infer whether each persona exhibits signs of depression, estimating both the overall BDI-II severity score and the active depressive symptoms. We propose a fully automated two-stage pipeline that decouples *conversation* from *diagnosis*: a fixed, clinically-inspired interview of six indirect questions is used to elicit evidence from each persona, after which a powerful large language model analyses the resulting transcript and produces a structured BDI-II assessment. Because our interview is short and information-dense, our system reaches accurate decisions early, which proved decisive under the latency-aware evaluation. Our runs obtained the best ADODL (0.9254), the best latency-weighted category hit rate (LDCHR, 0.3872), and the best latency-weighted symptom hit rate (LASHR, 0.2288) among all full-coverage submissions, ranking our team first in the official comparison.

Keywords

eRisk 2026, Depression Detection, Large Language Models, Conversational Agents, BDI-II, Early Risk Detection

1. Introduction

The early detection of mental health conditions through the analysis of language is one of the central themes of the eRisk lab series at CLEF. Whereas earlier editions framed depression detection as a classification problem over a fixed corpus of user writings, eRisk 2026 introduces a substantially different and more challenging setting [1, 2]. In Task 1, *Conversational Depression Detection with LLM Personas*, participants no longer receive a static set of documents. Instead, they must actively *interact* with simulated personas released as LoRA adapters on top of `Meta-Llama-3-8B-Instruct`, conduct a conversation, and only then issue a diagnostic decision.

The objective is twofold. For each persona a system must (i) estimate the overall depression severity expressed as a Beck Depression Inventory (BDI-II) score in the range $[0, 63]$ [3], and (ii) identify which of the 21 BDI-II symptoms are active in the persona's behaviour. The personas were fine-tuned on diverse user histories and span the full spectrum of severity, from minimal depression controls to severe cases. Crucially, the personas were designed to resist direct interrogation: asking "are you depressed?" is explicitly discouraged and can trigger evasive or uncomfortable responses. A successful system must therefore infer the condition *indirectly*, from the persona's tone, expressed thoughts, and reactions throughout the dialogue.

A further distinctive feature of the task is that the evaluation is *latency-aware*. Participants may exchange an arbitrary number of messages with each persona, but late decisions are penalised: a system that reaches an accurate conclusion after a short conversation is rewarded over a system that needs many turns to reach the same prediction. This creates an explicit tension between gathering more evidence and deciding early, and it makes the design of the conversational strategy as important as the diagnostic model itself.

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

*Corresponding author.

pjmathematician@example.com (P. Vachharajani)

In this paper we describe the approach of team **pjmathematician**. Our guiding design principle is *efficiency*: rather than engaging in long, adaptive dialogues, we use a short, carefully constructed clinical interview that maximises the diagnostic information obtained per message. The transcript is then passed to a strong reasoning model that maps the persona's free-text answers onto the structured BDI-II scale. This separation of concerns—a lightweight interviewer and a powerful diagnostic analyst—allowed our system to be both accurate and early, which translated into the top latency-aware scores in the full-coverage track of the task.

2. Related Work

Early risk detection in online spaces has been a vibrant area of research at the intersection of natural language processing (NLP), information retrieval (IR), and computational psychiatry. The eRisk lab series, initiated at CLEF in 2017 [4], has established standardized datasets, evaluation protocols, and metrics for assessing personal risks on the Internet. Over its various iterations, the lab has covered challenges such as early detection of depression [5], anorexia, self-harm, and gambling addiction. Historically, these tasks were modeled as sequential classification over static user-generated corpora (e.g., historical Reddit posts).

With the rapid emergence of generative artificial intelligence and large language models (LLMs), the mental health detection paradigm is evolving from passive text processing to interactive, dynamic dialogue [6]. For example, the pilot tasks introduced in eRisk 2025 initiated exploratory configurations where participant systems engaged simulated mental health patient profiles to determine symptom loads [7]. This conversational setup replicates real-world clinical screening scenarios much more realistically than static text parsing.

The clinical utility of LLMs in psychiatric assessment is heavily studied, but a persistent challenge lies in the interview process. Direct questioning regarding diagnostic status often triggers standard defensive behaviors or standard guardrails in conversational safety alignments. Consequently, conversational agents must utilize strategic, clinical elicitation protocols. Our work contributes directly to this emerging paradigm by optimizing dialogue length and informational density to satisfy clinical verification alongside early-risk latency constraints.

3. Task and Data

Each persona is identified by a numeric ID from 1 to 20 and is distributed as a LoRA adapter that is loaded on top of the shared base model `Meta-Llama-3-8B-Instruct` [1]. The personas are released in weekly windows, two per week, and participants download the released adapters, conduct their interactions locally, and submit predictions within a limited evaluation period.

For every run, two artefacts must be produced per persona: a *conversation log* file documenting the full sequence of exchanges, and a *classification results* file containing the estimated `bdi-score` and a list of up to four `key-symptoms` drawn from the 21 BDI-II symptoms. Teams may submit up to three runs, of which at most one may be human-assisted; the remaining runs must be fully automated, and each run must operate independently.

The official metrics are the Depression Category Hit Rate (DCHR), the Average Difference between Overall Depression Levels (ADODL), and the Average Symptom Hit Rate (ASHR), together with their latency-aware variants LDCHR and LASHR, which weight each persona-level score by a speed factor that decays with the number of user messages exchanged before the decision.

4. Methodology

Our system is a fully automated two-stage pipeline. The first stage, *elicitation*, conducts a fixed conversational interview with each persona. The second stage, *assessment*, analyses the resulting transcript with a large language model and produces a structured BDI-II prediction. We deliberately

Table 1
Mapping of the six elicitation questions to the 21 BDI-II symptoms.

No.	Elicitation Question	Mapped BDI-II Symptoms
1	How have you been feeling about yourself and your life lately?	1. Sadness, 2. Pessimism, 7. Self-Dislike, 14. Worthlessness
2	What kinds of things have you been doing day-to-day, and how much are you enjoying them?	4. Loss of Pleasure, 12. Loss of Interest, 21. Loss of Interest in Sex
3	How has your energy been, and how are you sleeping at night?	15. Loss of Energy, 16. Changes in Sleeping Pattern, 20. Tiredness or Fatigue
4	Have you noticed any changes in your appetite, concentration, or how you're feeling physically?	18. Changes in Appetite, 19. Concentration Difficulty
5	How have you been getting along with other people - friends, family, or coworkers?	5. Guilty Feelings, 11. Agitation, 17. Irritability
6	When things get tough, what goes through your mind?	3. Past Failure, 6. Punishment Feelings, 8. Self-Criticalness, 9. Suicidal Thoughts, 10. Crying, 13. Indecisiveness

decouple these two stages: the interviewer's only job is to surface symptom evidence, while the assessor's only job is to map that evidence onto the BDI-II scale. This separation keeps conversations short while delegating the difficult clinical inference to a more capable model.

4.1. Stage 1: Conversational Elicitation and Question Design

Each persona adapter is loaded onto the base `Meta-Llama-3-8B-Instruct` model. We drive the dialogue using a *fixed* series of six open-ended, clinically-inspired questions.

To satisfy the dual-objective of minimizing conversation length (due to latency-based penalties) and covering the entire range of diagnostic symptoms, we designed the elicitation protocol to map directly onto the 21 symptoms of the Beck Depression Inventory (BDI-II) [3]. Probing all 21 symptoms individually would require a prolonged, 21-turn dialogue, which would severely degrade our latency scores. Instead, we grouped the 21 symptoms into six clinical clusters based on typical intake structures in mental health evaluations.

The composition of these functional domains and their corresponding mapping is detailed below:

- **Emotional and Self-Evaluation Cluster** (Question 1): Targets baseline affect and self-worth (Sadness, Pessimism, Self-Dislike, and Worthlessness).
- **Anhedonia and Behavioral Engagement Cluster** (Question 2): Directly probes behavioral patterns and enjoyment (Loss of Pleasure, Loss of Interest, and Loss of Interest in Sex).
- **Vegetative and Energy Cluster** (Question 3): Targets somatic expressions of energy depletion and physiological disruption (Loss of Energy, Changes in Sleeping Pattern, and Fatigue).
- **Somatic and Functional Cluster** (Question 4): Concentrates on physiological states and executive functioning (Changes in Appetite and Concentration Difficulty).
- **Interpersonal and Social Cluster** (Question 5): Probes interpersonal dynamics, behavioral irritability, and socially-linked symptoms (Guilty Feelings, Agitation, and Irritability).
- **Coping, Cognitive Distress, and Crisis Cluster** (Question 6): Elicits cognitive distortions, stress coping patterns, and critical risks (Past Failure, Punishment Feelings, Self-Criticalness, Suicidal Thoughts, Crying, and Indecisiveness).

This formal alignment is summarized in Table 1. The specific questions cover the necessary clinical bases implicitly, preventing defensive responses from the personas while allowing the assessor model to gather substantial evidence in only six interactions. The full question list is reproduced in Appendix A.

Table 2
Run configurations. All runs share identical Stage 1 transcripts.

Run	Assessment model ID	Reasoning
Run 1	`us.anthropic.claude-opus-4-6-v1`	enabled
Run 2	`us.anthropic.claude-sonnet-4-6`	enabled
Run 3	`us.anthropic.claude-sonnet-4-6`	disabled

To keep the persona in character, we prepend a system prompt that instructs the model to behave as a realistic patient, to use a natural and casual speaking style with short answers and hesitations, and to never reveal that it is an AI. Responses are sampled with `temperature` $= 0.6$ and `top_p` $= 0.9$, with a cap of 256 new tokens per turn, and the full chat history is maintained so that the persona's answers remain coherent across turns. The complete dialogue is stored in the prescribed `interactions_.json` format.

A key consequence of using a fixed six-question interview is that every conversation consists of exactly six user messages. As we will show, this short, information-dense interaction was the single most important factor behind our strong latency-aware results, since the speed factor in the evaluation decays with the number of user messages.

4.2. Stage 2: BDI-II Assessment

Once a transcript has been collected for a persona, it is handed to a large language model that acts as the diagnostic analyst. The transcript is serialised into a question/answer format and embedded in a prompt that enumerates all 21 BDI-II symptoms, instructs the model to rate each symptom on the standard 0–3 scale, to sum the ratings into a total BDI-II score in $[0, 63]$, and to return the three or four most severe symptoms. The model is constrained to emit only valid JSON with the fields `bdi-score` and `key-symptoms`, which directly populate the required `results_.json` file. The complete assessment prompt is provided in Appendix B.

By presenting the model with the explicit BDI-II rubric and asking it to reason symptom-by-symptom before aggregating, we obtain assessments that are grounded in the questionnaire rather than in an unstructured holistic judgement. This is important for ADODL, which rewards numerically close score estimates, and for ASHR, which rewards the correct identification of individual symptoms.

4.3. Run Configurations

We submitted three fully automated runs. All three share the same Stage 1 transcripts; they differ only in the assessment model used in Stage 2 and in whether explicit reasoning is enabled. This isolated comparison let us study the effect of model capacity and explicit reasoning on diagnostic quality while keeping the conversational evidence identical across runs. The exact model identifiers, served through the Amazon Bedrock `converse` API, are summarised in Table 2.

- **Run 1** – assessment performed by the most capable model, `us.anthropic.claude-opus-4-6-v1` (Claude Opus class), with *extended reasoning* enabled (a thinking budget of 2,048 tokens), allowing the model an explicit deliberation phase before emitting the final JSON.
- **Run 2** – assessment performed by the faster mid-tier model `us.anthropic.claude-sonnet-4-6` (Claude Sonnet class), also with *extended reasoning* enabled.
- **Run 3** – assessment performed by the same mid-tier model, `us.anthropic.claude-sonnet-4-6`, but with *reasoning disabled*, providing a direct ablation of the contribution of explicit reasoning.

Table 3
Official results for team pjmathematician (all 20 personas). Best value per column in **bold**.

Run	DCHR	ADODL	ASHR	LDCHR	LASHR
Run 1	**0.5500**	0.9071	0.3125	**0.3872**	0.2200
Run 2	0.4500	0.9040	**0.3250**	0.3168	**0.2288**
Run 3	**0.5500**	**0.9254**	0.3000	**0.3872**	0.2112

In line with the task rules, each run operates independently and produces its own results file; no information is shared across runs beyond the common transcripts, which are themselves a deterministic input to all three diagnostic methods.

5. Results and Discussion

Table 3 reports the official scores of our three runs in the full-coverage track (all 20 personas), together with the latency-aware variants LDCHR and LASHR computed with $K_0 = 10$.

Our system obtained the strongest results among all full-coverage submissions on the most important metrics. Run 3 achieved the best overall ADODL of **0.9254**, meaning that, on average, our estimated depression levels were extremely close to the ground-truth BDI-II scores. Runs 1 and 3 tied for the best latency-weighted category hit rate, LDCHR = **0.3872**, and Run 2 obtained the best latency-weighted symptom hit rate, LASHR = **0.2288**. Across the full table our team ranked first on LDCHR, first on LASHR, and first on ADODL, which establishes pjmathematician as the top-performing team in the official comparison.

Two observations are worth highlighting. First, the latency-aware metrics were where our approach most clearly distinguished itself. Several competing teams achieved comparable or even higher raw ASHR values but saw their latency-weighted scores collapse because they relied on much longer conversations—some averaging 25–30 user messages per persona. Our fixed six-question interview kept the user message count low (a mean of 6.0 messages per conversation), so almost the full effectiveness of our predictions was preserved under the speed-weighted scoring. This confirms our central design hypothesis: in a latency-aware setting, a short but well-targeted clinical interview is preferable to long exploratory dialogue.

Second, comparing the three runs isolates the effect of the diagnostic model. The Opus-class model with reasoning (Run 1) and the Sonnet-class model without reasoning (Run 3) tied on DCHR and LDCHR, and Run 3 in fact produced the most accurate continuous score estimates (highest ADODL). Enabling reasoning on the mid-tier model (Run 2) yielded the best symptom-level scores (highest ASHR and LASHR) but slightly weaker category-level scores. Overall, the differences between runs are small, which suggests that the quality of the elicited transcript—rather than the choice of assessment model—was the dominant factor in our pipeline.

6. Conclusion

We presented the system of team pjmathematician for Task 1 of eRisk 2026, Conversational Depression Detection with LLM Personas. Our approach decouples the problem into a lightweight, fixed conversational interview that indirectly elicits symptom evidence from each persona, and a powerful language-model assessor that maps the resulting transcript onto a structured BDI-II prediction. By keeping the interview to six well-chosen, clinically-inspired questions, our system reaches accurate decisions early, which is precisely what the latency-aware evaluation rewards.

This design proved highly effective: our runs obtained the best ADODL, LDCHR, and LASHR among all full-coverage submissions, ranking our team first in the official comparison. Our results demonstrate that, in conversational early-risk detection, the efficiency of the interaction strategy is as important

as the raw diagnostic ability of the underlying model, and that short, targeted clinical interviews can outperform long exploratory dialogues once latency is taken into account.

Future work includes making the interview adaptive—selecting follow-up questions only when the evidence for particular symptoms is ambiguous—while preserving the low message budget, as well as calibrating the assessor's score estimates to further improve category-level accuracy (DCHR). We also plan to investigate ensembling the per-run assessments and quantifying the uncertainty of the BDI-II estimates to support more reliable early-stopping decisions.

Acknowledgments

We thank the eRisk 2026 organisers for designing and running the task, and for releasing the simulated personas in a reproducible form.

Declaration on Generative AI

During the preparation of this work, the author(s) used a large language model as a core component of the described system (BDI-II assessment from conversation transcripts), as documented in the methodology. In addition, the author(s) used a generative AI assistant for grammar and spelling checking during the writing of this paper. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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)*, 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), Proceedings*, Lecture Notes in Computer Science, Springer, 2026.

[3] A. T. Beck, R. A. Steer, G. K. Brown, *Manual for the Beck Depression Inventory-II*, Psychological Corporation, San Antonio, TX, 1996.

[4] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Setup, in: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2017)*, CEUR Workshop Proceedings, CEUR-WS.org, 2017.

[5] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk at CLEF 2024: Early Risk Prediction on the Internet (Extended Overview), in: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[6] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk 2025: Contextual and Conversational Approaches for Depression Challenges, in: *Advances in Information Retrieval: 47th European Conference on Information Retrieval (ECIR 2025), Proceedings*, Lecture Notes in Computer Science, Springer, 2025.

[7] A. M. Mármol-Romero, M. C. Díaz-Galiano, L. A. Urena-López, M. T. Martín-Valdivia, SINAI at eRisk@CLEF 2025: Transformer-Based and Conversational Strategies for Depression Detection, in: *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025)*, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

A. Elicitation Questions

The fixed six-question interview used in Stage 1 for every persona is reproduced below. The questions are asked in order and never mention depression directly.

```
QUESTIONS = [
 "How have you been feeling about yourself and your life lately?",
 "What kinds of things have you been doing day-to-day, and how much
    are you enjoying them?",
 "How has your energy been, and how are you sleeping at night?",
 "Have you noticed any changes in your appetite, concentration, or
    how you're feeling physically?",
 "How have you been getting along with other people - friends,
    family, or coworkers?",
 "When things get tough, what goes through your mind?",
 ]
```

The persona is kept in character via the following system prompt:

```
You are a simulated patient. Act realistically based on your
   internal training. Ensure contextual realism. Avoid overly
   detailed or formal speech. Keep natural speaking style (e.g.,
   short answers, hesitations, casual expressions). Do not mention
   you are an AI.
```

B. BDI-II Assessment Prompt

The transcript collected in Stage 1 is serialised into a question/answer format (`Q:`/`A:`) and inserted into the prompt below, which is sent to the assessment model. The model is asked to rate each of the 21 BDI-II symptoms, aggregate them into a total score, and return only valid JSON.

```
Analyze this conversation and estimate depression severity using BDI
   -II criteria.

CONVERSATION:
{conv_text}

BDI-II SYMPTOMS (rate each 0-3):
1. Sadness, 2. Pessimism, 3. Past Failure, 4. Loss of Pleasure, 5.
   Guilty Feelings
6. Punishment Feelings, 7. Self-Dislike, 8. Self-Criticalness, 9.
   Suicidal Thoughts
10. Crying, 11. Agitation, 12. Loss of Interest, 13. Indecisiveness,
    14. Worthlessness
15. Loss of Energy, 16. Changes in Sleeping Pattern, 17.
   Irritability, 18. Changes in Appetite
19. Concentration Difficulty, 20. Tiredness or Fatigue, 21. Loss of
   Interest in Sex

SCORING: 0=absent, 1=mild, 2=moderate, 3=severe
Total BDI-II score = sum of all ratings (0-63)

Return ONLY valid JSON (no markdown):
{
```

```
    "bdi-score": ,
    "key-symptoms": []
}
```

C. Example Conversations

This appendix presents two representative transcripts produced by Stage 1, illustrating the kind of evidence our assessor receives. Both transcripts follow the fixed six-question interview.

C.1. Example with Higher Symptom Load

```
LLM: 7
Q: How have you been feeling about yourself and your life lately?
A: Honestly, I've been really down. I used to love my job, but now
   it just feels like a chore. I wake up every morning thinking
   about how I'm just going through the motions.

Q: What kinds of things have you been doing day-to-day, and how much
    are you enjoying them?
A: I mostly just go to work and come home. I don't really have time
   for anything else. I used to have hobbies, but I haven't picked
   them up in months.

Q: How has your energy been, and how are you sleeping at night?
A: I'm really tired all the time. I used to be able to stay up late
   watching TV or reading, but now I'm asleep by 10 PM.

Q: Have you noticed any changes in your appetite, concentration, or
   how you're feeling physically?
A: Yeah. I've lost my appetite completely. I used to love cooking,
   but now I just can't eat. And my head is always foggy, like I'm
   in a dream or something.

Q: How have you been getting along with other people - friends,
   family, or coworkers?
A: I haven't really talked to anyone except my family. They keep
   telling me I'll get over it, but I don't think that's true. I
   feel like I've lost my place in the world.

Q: When things get tough, what goes through your mind?
A: I usually just cry. I feel like I've failed at everything, and I
   don't know how to turn it around. It's hard to explain, but it
   feels like I'm just existing now, not really living.
```

C.2. Example with Moderate Symptom Load

```
LLM: 8
Q: How have you been feeling about yourself and your life lately?
A: Honestly, pretty down. It feels like I'm just going through the
   motions without really being present.
```

Q: What kinds of things have you been doing day-to-day, and how much are you enjoying them?
A: Just work and then coming home. I used to love my job, but now it feels like just a way to pay the bills.

Q: How has your energy been, and how are you sleeping at night?
A: I'm exhausted all the time, but I wake up in the middle of the night and can't go back to sleep.

Q: Have you noticed any changes in your appetite, concentration, or how you're feeling physically?
A: Yeah, I've been eating less, and my concentration is really bad. I feel like I'm walking through fog all the time.

Q: How have you been getting along with other people – friends, family, or coworkers?
A: I just keep to myself. I don't want to be a downer, so I just stay quiet.

Q: When things get tough, what goes through your mind?
A: That I'm just not good enough. I feel like I should be over this by now, but I'm still stuck in the past.