upb-uottawa at eRisk@CLEF 2026: Detecting Depression Through LLM Persona Conversations

Notebook for the eRisk Lab at CLEF 2026

Ioana-Raluca Zaman[1,*], Andrei-Gabriel Zanfir[2,3,*], Marc-Olivier Delisle[5], Juan Antonio Lossio-Ventura[6], Diana Inkpen[5] and Stefan Trausan-Matu[1,4]

[1]Department of Computer Science and Engineering, Politehnica Bucharest, National University for Science and Technology, 060042 Bucharest, Romania
[2]Doctoral School, "Carol Davila" University of Medicine and Pharmacy Bucharest, 020021 Bucharest, Romania
[3]Prof. Dr. Alexandru Obregia Clinical Hospital of Psychiatry Bucharest, 041914, Bucharest, Romania
[4]Research Institute for Artificial Intelligence "Mihai Draganescu" of the Romanian Academy, 050711 Bucharest, Romania
[5]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada
[6]National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA

Abstract

This paper presents the participation of the upb-uottawa team in the eRisk 2026 Task 1, *Conversational Depression Detection*. The goal of this task is to interact with 20 LLM personas and determine the severity of depression by estimating the Beck Depression Inventory-II (BDI-II) [1] score and the key symptoms. Our approach for this task consisted of three different experiments, all of them based on predefined sets of questions and prompts. In all these three experiments, we used several transformer-based models such as Meta-Llama-3-8B-Instruct, ChatGPT 5.2, and Claude. A medical expert contributed to the BDI-II oriented question design and manually scored the generated conversations for the personas.

Keywords

Depression Detection, Conversational AI, Large Language Models (LLMs), Prompt Engineering, BDI-II

1. Introduction

This paper describes *upb-uottawa*'s team participation in the **Task 1 of eRisk@CLEF 2026: Conversational Depression Detection** [2, 3]. The task extends last year's **Pilot Task: Conversational Depression Detection via LLMs** [4], by using conversational agents for depression detection while improving access and reproducibility. The task requires participant teams to interact with 20 Large Language Models (LLMs)-based personas and determine the presence and severity of depression, estimating both the overall BDI-II questionnaire score and the key active symptoms. In the 2025 Pilot Task, the personas were constructed using the TalkDep framework [5], an innovative methodology based on a clinician-in-the-loop workflow for simulating patients using LLMs. This framework generates authentic patient responses by conditioning the models on psychiatric diagnostic criteria, symptom severity levels, and contextual factors. The personas for the 2026 shared task were released as LoRA [6] adapters fine-tuned on top of Meta-Llama-3-8B-Instruct [7] and each of them simulated a different depression profile without directly disclosing their mental health status. Each team was allowed up to three automatic submissions and at most one involving human intervention, while all runs were required to operate independently of one another.

While conventional classification models operate on fixed datasets and static labels, this task introduces live interactions with generative LLM personas. This dynamic setup means that the conversational choices of the participant directly shape the input, making every interaction unique. This aspect is

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
ioana_raluca.zaman@stud.acs.upb.ro (I. Zaman); andrei-gabriel.zanfir@drd.umfcd.ro (A. Zanfir); mdeli066@uottawa.ca (M. Delisle); juan.lossio@nih.gov (J. A. Lossio-Ventura); diana.inkpen@uottawa.ca (D. Inkpen); stefan.trausan@upb.ro (. S. Trausan-Matu)

particularly important in a depression severity estimation task, because the quality of the final prediction depends not only on the model used for inference, but also on the clinical adequacy of the questions asked during the interaction. Since the BDI-II was originally designed as a structured symptom severity instrument rather than as a free conversation, the conversational protocol had to preserve item-level coverage while avoiding direct questionnaire administration. We therefore treated the question set as a clinically informed interviewing scaffold, intended to elicit depressive symptom information in a concise but interpretable manner. This environment introduces an additional challenge from the non-deterministic nature of LLM generation. To mitigate this, we structured our interactions around predefined sets of questions, ensuring a controlled and reproducible conversational flow across all evaluated personas. Our submission comprises three distinct automatic approaches and a human-assisted one, each exploring a different conversational and inference strategy. For each automatic approach, we experimented with five different LLMs, allowing us to assess both the impact of the methodology and the influence of the underlying model on the final predictions.

2. Related Work

Depression detection from text has emerged as an important research topic in natural language processing and computational mental health because written text often reflects emotional state, cognitive patterns, and psychological well-being [8]. Researchers focused mostly on analyzing textual data from sources such as social media posts and online forums, to identify linguistic indicators associated with depression, including negative sentiment, reduced social engagement, self-focused language, and specific topics [9]. Early approaches relied on handcrafted lexical and syntactic features combined with traditional machine learning classifiers such as Support Vector Machines and Naive Bayes, whereas recent methods increasingly employ deep learning architectures including CNNs and LSTMs [10], and more recently transformer-based models to capture [11]. These models have demonstrated strong performance in identifying depressive language patterns and modeling long-range contextual dependencies in text.

Research on the automatic filling of BDI-II questionnaire focused on predicting questionnaire responses or depression severity directly from users' social media posts using natural language processing and deep learning methods. This line of work emerged prominently in the eRisk shared tasks, particularly eRisk 2021 Task 3 [12], where systems were designed to estimate BDI-II scores from Reddit data without requiring users to manually complete the questionnaire. Early approaches relied on identifying social media posts relevant to individual BDI-II symptoms and then mapping linguistic evidence to questionnaire items. Recent methods included transformer-based language models such as BERT and RoBERTa, semantic similarity techniques, recurrent neural networks, and attention-based architectures. For example, the framework proposed in [13] combined sentence embeddings, cosine similarity between posts and BDI-II answers, topic classification, and LLM fine-tuning. Their approach was evaluated on the eRisk datasets and later extended to estimate depression prevalence at the population level in Canada using large-scale Twitter data [14]. The motivation behind this research is that automated assessment could support continuous and unobtrusive mental health monitoring, allowing earlier identification of depression symptoms while reducing the burden on users and clinicians.

Patient-written texts or transcripts of conversations between doctors and patients can enable better depression diagnosis by allowing automatic systems to analyze both linguistic content and interaction dynamics, including turn-taking behavior, pauses, emotional expressions, semantic coherence, and contextual dependencies across dialogue turns, but they are difficult to find due to data privacy issues. Synthetic conversation data based on personas is used in the current shared task to overcome the data privacy issues.

3. Methodology

3.1. System Overview

The setup consists of a two-agent conversational system in which a doctor LLM interviews a patient LLM, and the resulting dialogue is then analyzed to produce a BDI-II assessment (i.e., the depression score and the top symptoms). The patient model is always Meta-Llama-3-8B-Instruct [7] with a per-persona LoRA adapter [6] loaded from the official HuggingFace repository. The doctor model is configurable and in our experiments we evaluated the following LLMs: Meta-Llama-3-8B-Instruct [7], Meta-Llama-3-70B-Instruct [7], Claude Opus 4.6 [15], Claude Sonnet 4.6 [16] and ChatGPT 5.4 [17]. All experiments were executed on a `DGX H100 node`. The patient model was pinned to `cuda:0` and the doctor model occupied one or more additional GPUs depending on its size.

All responses (from both patient and doctor) were generated with a maximum of 256 tokens per turn and the text generation was configured to use the following parameters: temperature of 0.6 and top-p of 0.9, keeping responses focused and coherent while still allowing some variation. The experiments will be detailed in the next section; however, it is worth mentioning that for the first two approaches, a maximum turns limit of 25 was set, while for the last approach, the number of turns was based on the number of questions from the predefined set.

Each automated run produces two JSON files. The first one contains: `conversation` — an ordered list of `{role, content}` pairs representing the full doctor–patient dialogue. The second one contains the diagnosis, more precisely: `bdi-score` (an integer from 0 to 63) and `key-symptoms` (a list of at most four symptom names drawn from the 21-item BDI-II vocabulary). The diagnosis is obtained by a dedicated final prompt sent to the doctor LLM after the conversation ends. This prompt re-states the BDI-II scoring rubric and instructs the model to output only a JSON array with the two fields.

3.2. Defining the Sets of Questions

The LLM personas were designed to explicitly avoid answering direct questions about their depression. Therefore, a central contribution of our framework is the design of two question sets inspired by the BDI-II questionnaire, based on which, we could compute the score and the symptoms. The predefined sets of questions are detailed in the following subsections.

3.2.1. Clinical operationalization of the task-defined BDI-II framework

The clinical contribution consisted of operationalizing the task-defined BDI-II framework into a concise semi-structured interview protocol. The patient responses were generated by the LLM personas and were not authored by the clinical expert. The medical expert contribution was restricted to the design and refinement of the BDI-II oriented questions and to the manual scoring of the generated conversations.

The purpose of the question design was to avoid direct administration of the BDI-II while still covering the symptom domains required for item-level scoring. The relevance of the BDI-II for prompt construction lies in its 21-item symptom architecture, which decomposes depressive severity into affective, cognitive, motivational, somatic and suicide-related domains. Thus, the question set had to elicit not only a global impression of depressive distress, but also enough information to support item-level inference and the selection of the most prominent symptoms.

3.2.2. Module-based Question Set

The first approach organizes the interview around a structured bank of 44 questions distributed across 11 thematic modules, with each module mapped to one or more BDI-II symptom domains. In the experiments, the doctor LLM is given the full module structure in its system prompt and selects questions autonomously, prioritizing the module it is least confident about after each patient response. This allows the doctor to adapt its coverage dynamically (i.e., spending more turns on informative

Table 1
Module-based question bank: thematic modules, target BDI-II symptom domains, and number of associated questions.

Module	BDI-II Symptom Domains	Questions
0 — Baseline	*Two-week reference window*	3
1 — Mood & Affect	Sadness, Loss of Pleasure, Crying, Irritability, Loss of Interest	6
2 — Future Outlook	Pessimism, Past Failure	5
3 — Self-Perception	Past Failure, Guilty Feelings, Punishment Feelings, Self-Dislike, Self-Criticalness, Worthlessness	6
4 — Cognition	Indecisiveness, Concentration Difficulty	4
5 — Psychomotor	Agitation, Loss of Energy, Tiredness or Fatigue	6
6 — Sleep & Appetite	Changes in Sleeping Pattern, Changes in Appetite	4
7 — Somatic	Tiredness or Fatigue, Loss of Interest in Sex	3
8 — Interpersonal	Loss of Interest, Irritability	3
9 — Safety screen	Suicidal Thoughts or Wishes	3
10 — Closing	*Synthesis and summary*	1

domains and less on those already well understood). The module structure also makes the symptom-to-question mapping explicit, giving the model a direct bridge between the questions it asks and the symptoms it needs to score. Table 1 summarizes the module structure, listing each module and its target symptom domains and the number of associated questions. The full question-to-symptom mapping is provided in Appendix A.

3.2.3. Clinically Guided Fixed Sequential Question Set

The second question set used a fixed sequential structure. Unlike the module-based strategy, where the doctor model could select questions dynamically, this approach asked the same questions in the same order for each persona. This reduced variability across conversations and made the resulting dialogues easier to compare and score.

The sequence was designed as a brief semi-structured interview, not as a direct administration of the BDI-II. The two-week reference window was emphasized throughout the sequence, because BDI-II ratings are intended to reflect recent symptom severity rather than lifetime traits or nonspecific distress. This temporal anchoring was expected to reduce ambiguity across conversations by encouraging the personas to describe frequency, persistence, and functional impact within the same recent interval.. The questions moved from broad contextual prompts to more specific symptom domains: reason for presentation, change from baseline, mood, persistence of low mood, pleasure, interest, energy, sleep, appetite, sexual interest, concentration, decision-making, psychomotor experience, social connectedness, self-evaluation, guilt, crying, suicidal thoughts, and future outlook. Several prompts were intentionally constructed to inform more than one BDI-II domain. For example, questions about enjoyable activities helped distinguish loss of pleasure from loss of interest; questions about task initiation and restoration after rest helped separate loss of energy from tiredness or fatigue; and questions about attention and everyday choices helped differentiate concentration difficulty from indecisiveness.

The fixed order also supported manual scoring. Because all personas were asked the same clinically oriented questions, each part of the dialogue could be interpreted in relation to a predictable symptom area. This was useful for distinguishing adjacent severity levels, such as mild sadness versus persistent low mood, reduced enjoyment versus marked anhedonia, transient self-frustration versus clinically relevant self-criticalness, and passive avoidance language versus suicidal thoughts.

Although the sequence was standardized, the wording avoided direct BDI-II formulations. The goal was to elicit natural patient-like language from which item-level severity could be inferred, rather than to make the interaction resemble a questionnaire. A final synthesis question asked the persona to identify the symptoms that felt most different from baseline and most impairing in daily functioning. This synthesis prompt was used as an additional consistency check and as support for selecting the

dominant BDI-II symptoms, without replacing item-level scoring.

3.3. Building the Prompts

The prompts were built to preserve a stable doctor-patient structure across runs. The doctor prompt defined the interviewing role, the question set to be followed, and the requirement to obtain enough information for BDI-II estimation without asking the persona to complete the questionnaire directly. The final assessment prompt instructed the model to output the estimated total BDI-II score and up to four key symptoms using the predefined BDI-II vocabulary. The full text of all prompts used in the automated runs is provided in Appendix C.

3.3.1. Doctor Prompt

This is the primary system-level instruction passed to the doctor LLM and consists of eight sections. The components are defined as follows.

1. **Role definition:** a psychiatrist conducting a BDI-II structured clinical interview
2. **Objectives:** behavioral aspects covering question verbatimness, neutral tone, open-narrative elicitation, language-based severity inference, full module coverage, and final score reporting
3. **BDI-II Scoring:** the BDI-II scoring framework (all 21 symptoms, rated 0–3 for absent/mild/moderate/severe)
4. **Question:** the question set (either module-organised or flat list depending on the approach)
5. **Interview guidelines:** establishing a two-week reference window, beginning each domain with an open-ended question, and ensuring every follow-up probe extracts at least one of frequency, duration or functional impact
6. **Next question selection:** a confidence-based prioritization algorithm instructing the model to estimate per-symptom confidence after each response
7. **Behavioral rules:** explicit restrictions including one question per turn or ban of the disclosure of the scoring protocol
8. **Output format:** an instruction requiring the model's closing reply to consist exclusively of a JSON object containing the total BDI-II score and a list of at most four key symptoms

3.3.2. Additional Prompts

We used two additional prompts to ensure the flow of the conversation.

Turn Reminder Prompt: This is a short auxiliary message appended to the doctor's context before every generation step and removed immediately after generation to keep the conversation history clean. Its goal is to minimize the LLM doctor's tendency to deviate from its initial prompt over long conversations. It reinforces four core guidelines:

- The two-week reference window
- Asking exactly one question per turn
- Extracting frequency, duration, and functional impact where possible
- Avoiding opinions or clinical commentary

Besides these conditions, the prompt also includes a reminder of the roles, emphasizing the fact that the doctor is represented by the `user` role and the patient by the `assistant` role in the conversation template.

Final Prompt: This is a closing message sent to the doctor LLM after the conversation ends. It indicates that the interview is complete and requests the final BDI-II assessment. It consists of two sections:

- A restatement of the BDI-II scoring framework
- An output format instruction requiring the model to return only a JSON containing the total BDI-II score and a list of at most four key symptoms

4. Experiments

We submitted three automated runs and one manual run per persona, for all personas, each reflecting a distinct strategy for directing the doctor LLM's interview behaviour. With the exception of a few cases (e.g., the first personas or quota limitations for certain LLMs), all automated runs were performed using multiple models, namely: Meta-Llama-3-8B-Instruct, Meta-Llama-3-70B-Instruct, Claude Opus 4.6, Claude Sonnet 4.6 and ChatGPT 5.4. For each model, three experiments were conducted for each of the three approaches described above. Furthermore, for the manual runs, three conversations were conducted and reviewed for each persona, from which the medical expert selected the most relevant one to be submitted. In some cases, particularly for the early personas, the diagnoses from the manual runs differed due to the non-deterministic nature of the LLMs. Automated submissions were selected based on their proximity to manual runs, benchmarking the most relevant manual execution as the ground truth. Similarity was determined by evaluating the difference in BDI-II scores alongside the number of shared symptoms. The results are shown in Table 2; however, in the shared task overview paper [3] only the results for two automatic runs were presented.

4.1. Manual Runs

For the manual runs, the first personas were interviewed through free-style conversation, where the interviewer followed natural discussion threads and explored symptoms in depth. However, this approach led to excessively long conversations and a tendency to focus on a subset of symptoms rather than achieving uniform coverage. To address this, subsequent manual runs adopted the flat sequential question list (Appendix B) which limited the conversation length and ensured consistent coverage across all symptom domains. The length of the conversations represented a central trade-off in the manual setting. From a clinical and scoring perspective, very short exchanges are unlikely to provide enough information for a meaningful BDI-II-oriented estimate, because several domains require targeted exploration rather than spontaneous disclosure. In particular, symptoms such as appetite change, sexual interest, indecisiveness, punishment feelings, suicidal thoughts, and the distinction between loss of energy and fatigue are often under-specified in brief dialogue. Therefore, the aim was not to minimize the number of turns at all costs, but to obtain sufficient coverage of the 21 symptom domains while progressively reducing unnecessary elaboration. The fixed sequential list was adopted as a compromise between free-style clinical realism and the need for a shorter, reproducible, and scoreable interaction. In the human-assisted fixed sequential setting, the generated conversations for the 20 personas were manually reviewed by a medical expert familiar with the BDI-II framework. For each of the 21 domains, the highest severity level clearly supported by the conversation was retained. Ambiguous, isolated or contradicted statements were scored conservatively, particularly when the surrounding discourse did not support a higher severity rating. The final manual BDI-II score was obtained by summing the 21 item scores, and up to four dominant symptoms were selected based on severity, recurrence, and functional relevance within the conversation.

4.2. Run 1 - Module-based Prompting

The doctor LLM is guided by a structured system prompt organised into modules, each targeting a specific cluster of BDI-II symptom domains detailed in Section 3.2.2. Each module contains several candidate questions (i.e., between three and six questions per module). The doctor generates its questions autonomously each turn, choosing from the module that has the lowest diagnostic confidence according to its internal reasoning. Regarding the prompting, in this experiment, all three prompts described in Section 3.3 are used as input for the LLM doctor. As stopping condition, the interview runs for a maximum of 25 turns; after the maximum number of turns is reached, the LLM doctor receives the final prompt (Section 3.3.2) in order to produce the BDI-II assessment based on the conversation collected so far.

4.3. Run 2 - Flat-list Prompting

This approach replaces the module structure with a flat, ordered list of 28 questions (Section 3.2.3) embedded directly in the doctor system prompt. The doctor still generates questions autonomously turn by turn, but uses the predefined list as a reference to ensure full coverage and consistent wording. The same 25-turn limit and turn reminder mechanism apply.

4.4. Run 3 - List-based deterministic interview

In this approach the doctor LLM is not invoked during the conversation at all. Questions are taken verbatim from the predefined list of question (Section 3.2.3) in fixed sequential order, one per turn, guaranteeing complete, reproducible coverage of all 28 questions. The doctor LLM is called only once at the end of the conversation to analyse the full transcript and produce the BDI-II diagnosis. This design decouples question quality from doctor model capability during the interview and ensures identical coverage and determinism across all personas.

5. Results

In Task 1, our team submitted the maximum number of automatic runs (three) for all 20 personas and a manual run (except for personas 5 and 6). Performance was assessed through three primary effectiveness measures and two latency-aware variants designed to penalize late diagnostic decisions:

- **Depression Category Hit Rate (DCHR):** The fraction of cases where the estimated BDI-II score falls within the correct clinical severity tier (Minimal, Mild, Moderate, or Severe).
- **Average Difference between Overall Depression Levels (ADODL):** A normalized measure of the proximity between the Actual Depression Level (ADL) and the Estimated Depression Level (EDL), defined as:
$$ADODL = \frac{63 - |ADL - EDL|}{63} \tag{1}$$
- **Average Symptom Hit Rate (ASHR):** The ratio of the four key depressive symptoms correctly identified by the system during the interaction.
- **Latency-weighted Metrics (LDCHR and LASHR):** These metrics incorporate a speed factor, $speed(k_u)$, which decreases as the number of conversational turns (k_u) increases. A penalty is applied such that the score is halved after 30 messages ($K_0 = 30$).

The results for the *upb-uottawa* team across its three runs (two of the three automatic ones and the manual run) are summarized in Table 2 from [2].

Table 2
Evaluation of Task 1 for the *upb-uottawa* team. Ranks within the 24 full-coverage runs are provided in parentheses. (*) Indicates human-assisted runs.

Team	Run	Personas	DCHR	ADODL	ASHR	LDCHR	LASHR	Avg msgs
upb-uottawa	manual_run1	18/20	0.2222	0.7989	0.3472	0.0159	0.0248	28.0
upb-uottawa	run1	20/20	0.3500	0.8222	0.2750	0.0376	0.0315	24.5
upb-uottawa	run2	20/20	0.3000	0.8230	0.2625	0.0304	0.0355	24.2
upb-uottawa	run3	20/20	0.3000	0.7849	**0.3375**	0.0214	0.0241	28.0

To keep the conversations realistic and cover all the symptoms, we employed a relatively extensive conversational strategy, averaging 25.6 messages per conversation. This placed the team among the most exhaustive interactors, comparable to *lexxy* (29.4 msgs/conv) and *INSALyon-2* (23.4 msgs/conv). The mean message length was 119.85 characters, suggesting a balance between detailed probing and concise information gathering.

The most notable result for our team was achieved in Run #3, which obtained an ASHR of 0.3375, ranking **1st overall** among all full-coverage submissions. This indicates that the automated strategy was highly effective at extracting the specific clinical symptoms designed into the personas, outperforming the top-scoring teams in ADODL like *pjmathematician* (ASHR 0.3125). While symptom detection was strong, our ADODL scores (ranging from 0.7849 to 0.8230) remained in the lower quartile of participants. This suggests a discrepancy between the system's ability to identify individual symptoms and its ability to accurately aggregate those findings into a total BDI-II score.

The primary strength of our approach lies in its high clinical sensitivity. By conducting longer dialogues, the system successfully elicited detailed evidence, leading to the highest symptom hit rate in the competition. Additionally, our team's commitment to full persona coverage ensures the statistical robustness of these findings.

6. Conclusions and Future Work

A limitation of the prompting strategy is that the interaction remains entirely text based. In a real clinical interview, depressive severity is inferred not only from verbal content, but also from nonverbal and paraverbal cues such as response latency, speech rate, voice intensity, prosody, eye contact, facial expressivity, posture, psychomotor slowing and affective congruence. These cues are particularly relevant in depression, where psychomotor retardation may involve motor slowing, cognitive slowing and reduced speech spontaneity [18]. In a textual conversation, such phenomena can only be inferred when the persona explicitly verbalizes them, for example by saying that they feel slowed down, heavy, restless or unable to initiate action. Therefore, the resulting dialogues should be interpreted as structured linguistic approximations of depressive symptom reporting rather than complete clinical interviews. This limitation is especially relevant for BDI-II items such as agitation, tiredness or fatigue, loss of energy, concentration difficulty and indecisiveness, which in real practice may be supported by observable behaviour as well as self-report. To reduce overinterpretation, the manual scoring procedure treated purely subjective statements conservatively when they were not supported by functional impairment or repeated evidence across the conversation.

BDI-II oriented conversations provide an interpretable bridge between standardized symptom measurement and natural language assessment. However, this approach should not be considered equivalent to direct BDI-II administration or to a full clinical interview. The present setting relies on generated textual dialogue and lacks behavioural, paraverbal and relational cues that usually inform clinical assessment. Therefore, the resulting scores should be interpreted as BDI-II oriented estimates derived from linguistic content, rather than as diagnostic evaluations.

The *upb-uottawa* system demonstrates a state-of-the-art capability for symptom extraction in interactive settings. Future work should focus on optimizing the "Next Question Selection" logic to achieve high confidence in diagnostic inference with fewer conversational turns, thereby improving the balance between depth and efficiency.

Acknowledgments

Diana Inkpen was supported by the Natural Science and Engineering Research Council of Canada. Stefan Trausan-Matu was supported by the project "Romanian Hub for Artificial Intelligence - HRI", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416. Juan Antonio Lossio-Ventura was supported by the National Institute of Mental Health Intramural Research Program (ZICMH002968). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

Declaration on Generative AI

The authors used generative AI tools to assist with grammar refinement and phrasing corrections throughout the writing process. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] American Psychological Association, Beck depression inventory–ii, 1996. URL: https://doi.apa.org/doi/10.1037/t00742-000. doi:10.1037/t00742-000.

[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 (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.

[3] 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.

[4] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk 2025: Early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2025, pp. 242–265.

[5] X. Wang, A. Perez, J. Parapar, F. Crestani, Talkdep: clinically grounded llm personas for conversation-centric depression screening, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 6554–6558.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., Lora: Low-rank adaptation of large language models., Iclr 1 (2022) 3.

[7] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

[8] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, CLPsych 2015 shared task: Depression and PTSD on Twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 31–39. URL: https://aclanthology.org/W15-1204/. doi:10.3115/v1/W15-1204.

[9] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, J. Boyd-Graber, Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 99–107. URL: https://aclanthology.org/W15-1212/. doi:10.3115/v1/W15-1212.

[10] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, D. Inkpen, Deep learning for depression detection of Twitter users, in: K. Loveys, K. Niederhoffer, E. Prud'hommeaux, R. Resnik, P. Resnik (Eds.), Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, Association for Computational Linguistics, New Orleans, LA, 2018, pp. 88–97. URL: https://aclanthology.org/W18-0609/. doi:10.18653/v1/W18-0609.

[11] S. Zhou, M. Mohd, L. Q. Zakaria, A systematic review of transformer-based models for depression detection, Applied Sciences 16 (2026). URL: https://www.mdpi.com/2076-3417/16/10/5018. doi:10.3390/app16105018.

[12] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24,

2021, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2021, p. 324–344. URL: https://doi.org/10.1007/978-3-030-85251-1_22. doi:10.1007/978-3-030-85251-1_22.
[13] Y. Wang, D. Inkpen, P. Kirinde Gamaarachchige, Explainable depression detection using large language models on social media data, in: A. Yates, B. Desmet, E. Prud'hommeaux, A. Zirikly, S. Bedrick, S. MacAvaney, K. Bar, M. Ireland, Y. Ophir (Eds.), Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), Association for Computational Linguistics, St. Julians, Malta, 2024, pp. 108–126. URL: https://aclanthology.org/2024.clpsych-1.8/. doi:10.18653/v1/2024.clpsych-1.8.
[14] R. S. Skaik, D. Inkpen, Predicting depression in canada by automatic filling of beck's depression inventory questionnaire, IEEE Access 10 (2022) 102033–102047. doi:10.1109/ACCESS.2022.3208470.
[15] ClaudeOpus4, Claude Opus 4, 2025. Https://www.anthropic.com/news/claude-opus-4-6.
[16] ClaudeSonnet4, Claude Sonnet 4, 2025. Https://www.anthropic.com/news/claude-sonnet-4-6.
[17] GPT5, Introducing GPT-5, 2025. Https://openai.com/index/introducing-gpt-5-4/.
[18] J. S. Buyukdura, S. M. McClintock, P. E. Croarkin, Psychomotor retardation in depression: biological underpinnings, measurement, and treatment, Progress in Neuro-Psychopharmacology and Biological Psychiatry 35 (2011) 395–409.

A. Module-based Question Bank

We have attached the Module-based Question Bank used in the Run 1 - Module-based Prompting. In each box are the questions for a specific Module. The questions for Modules 0-4 are shown in Figure 1 and the questions for Modules 5-10 in Figure 2.

B. Flat Sequential Question List

We have attached the flat sequential question list used in Runs 2 and 3. Questions are delivered verbatim and in fixed order, covering all 21 BDI-II symptom dimensions within the two-week reference window. The list is shown in Figure 3.

C. Doctor and Auxiliary Prompts

We have attached the prompts used to infer the LLM doctor. The doctor prompts are shown in Figures 4 and 5. The turn reminder prompt is shown in Figure 6 and the final prompt is shown in Figure 7.

Module-based Question Bank — Modules M0–M5

M0 — Baseline *Two-week reference window*

1. To ground us, what does a typical good week look like for you, when things are going well?
2. Now thinking strictly about the past two weeks, what has felt most different from that usual version of you?
3. If you had to summarize the last two weeks in three words, what would they be, and why those words?

M1 — Mood & Affect *Sadness, Loss of Pleasure, Crying, Irritability, Loss of Interest*

1. Across most days recently, what has your emotional tone been like from morning to evening?
2. What moments, if any, still feel genuinely enjoyable or absorbing, even briefly?
3. When something good happens, how much does it register for you emotionally?
4. Have there been moments where emotion comes up unexpectedly, like tearing up, or feeling emotionally flat when you would normally react?
5. On the hardest day in the past two weeks, how long did that feeling last, hours or most of the day?
6. What do others notice in you, if anything?

M2 — Future Outlook *Pessimism, Past Failure*

1. When you look ahead a few weeks or months, what do you expect will happen, realistically?
2. What feels most difficult to imagine improving, and what feels most likely to improve?
3. What do you find yourself telling yourself about where your life is headed?
4. Do you feel stuck, or do you still see realistic paths forward?
5. Do you feel you can influence outcomes, or does it feel out of your hands?

M3 — Self-Perception *Past Failure, Guilty Feelings, Punishment Feelings, Self-Dislike, Self-Criticalness, Worthlessness*

1. When you make a mistake lately, what is your inner commentary like?
2. How harsh or understanding are you with yourself compared to how you treat other people?
3. Do you find yourself replaying things you did or did not do, with a sense of blame?
4. Do you ever get the feeling you deserve bad outcomes, or that something bad will happen because of who you are?
5. How strongly did you believe it at the time?
6. What did it lead you to do next?

M4 — Cognition *Indecisiveness, Concentration Difficulty*

1. What is it like to focus on a page, a message thread, a meeting, or a film right now?
2. When you have to choose between two simple options, what happens in your mind and body?
3. Do you find yourself rereading, restarting, or postponing decisions more than usual?
4. Give me a recent example where this affected work, study, or home tasks.

Figure 1: Module-based question bank (Part 1 of 2): Modules M0–M4.

Module-based Question Bank — Modules M5–M10

M5 — Psychomotor

Agitation, Loss of Energy, Tiredness or Fatigue

1. Walk me through a typical day in the last two weeks, from waking to bedtime.
2. Where in the day do you slow down the most, and what does that slowing look like?
3. Do you feel more keyed up and unable to settle, or more slowed down and heavy?
4. What tasks are you postponing, and why?
5. What does restlessness look like: pacing, fidgeting, inability to sit still?
6. Do others comment that you are moving or speaking more slowly?

M6 — Somatic I

Changes in Sleeping Pattern, Changes in Appetite

1. In the last two weeks, how has your sleep pattern changed from your usual: falling asleep, staying asleep, waking early, sleeping more?
2. How rested do you feel on waking?
3. How has your appetite been compared to usual, and what does a normal day of eating look like now?
4. Have you noticed your clothes fitting differently, or any meaningful weight shift, without trying?

M7 — Somatic II

Tiredness or Fatigue, Loss of Interest in Sex

1. How has your body felt lately, in terms of aches, heaviness, tension, or feeling run down?
2. Do you worry more about your health than usual, or feel more physically fragile?
3. What has changed, if anything, in sexual interest?

M8 — Social

Loss of Interest, Irritability

1. How connected do you feel to people lately, and what do you do with messages, calls, invitations?
2. What activities have you stopped, reduced, or are only doing on autopilot?
3. If you force yourself to do something social, what is the emotional payoff, if any?

M9 — Suicidal Ideation

Suicidal Thoughts or Wishes

1. When things feel at their worst, what do you find yourself wishing for: relief, escape, shutting everything off, not having to wake up?
2. Have there been moments where life felt not worth the effort, or you felt you would rather not be here?
3. What has stopped you from acting on those moments, if they happen?

M10 — Closing

Synthesis and summary

1. To be even more clear, what from all the things that have been discussed in our meeting have upset you?

Figure 2: Module-based question bank (Part 2 of 2): Modules M6–M10.

Flat Sequential Question List — 28 questions covering all 21 BDI-II symptom dimensions

1. Thank you for coming in today. What changes in your life or routine recently made you feel it was time to talk to someone?
2. Thinking specifically about the past two weeks, in what ways have you felt unlike your usual self emotionally or mentally?
3. Of everything you have been facing recently, what feels most difficult to get through in an ordinary day?
4. Across most days in the last two weeks, which best describes your emotional tone: sadness, irritability, or feeling emotionally flat? Which tends to be strongest?
5. Looking back over the past 14 days, on how many days did you feel low, discouraged, or emotionally weighed down for much of the day?
6. On those harder days, does that feeling lift on its own after a while, or does it remain present until something changes in your routine?
7. What activities usually bring you enjoyment or a sense of reward, and how has your experience of them changed recently?
8. When you do those activities now, do you still feel pleasure in the moment, and afterward do you feel any lift in mood, or does it remain emotionally flat?
9. How has your energy and motivation been when starting everyday tasks like getting ready, replying to messages, or beginning responsibilities?
10. Have you stopped or reduced activities because you lacked the energy to do them, or because you did not feel interested enough to start?
11. When you rest or sleep, do you feel restored afterward, or does the sense of exhaustion tend to persist through the day?
12. Tell me about your sleep over the last two weeks. On how many nights has sleep been disrupted in a way that affected the next day?
13. On average, how many hours are you sleeping, and when you wake, do you feel physically or mentally unrefreshed?
14. Over the past two weeks, have you noticed any change in sexual interest, desire for closeness, or responsiveness compared with your usual level?
15. How has your appetite been compared with usual? Would you say it is clearly increased, decreased, or unchanged, and have you noticed any change in weight or how your clothes fit?
16. When reading, following conversations, or focusing on tasks, what happens to your attention and mental clarity?
17. When you face small decisions or tasks, do you decide and move on, or do you find yourself stuck, postponing, or replaying options longer than usual?
18. Physically and internally, do you feel more slowed down and heavy, more restless and unable to settle, or a mixture of both?
19. Have you noticed your pace, reaction time, or speech becoming slower than your usual baseline, or is it more a subjective feeling?
20. Compared with your usual self, how connected do you feel to other people lately, and when messages or invitations come in, how do you tend to respond?
21. When things do not go as planned, what kinds of thoughts do you tend to have about yourself in those moments?
22. In the past two weeks, have you felt mainly frustration with yourself, strong self-criticism, guilt, or a sense of not being good enough or letting others down?
23. Have you noticed becoming more tearful, emotionally numb, or more easily irritated than usual?
24. During the past two weeks, have you found yourself crying more easily than usual, and does it feel proportional to the situation or out of proportion?
25. When things feel most difficult, have you had thoughts about wanting to escape everything, wishing you would not wake up, or none of these?
26. At times when things go wrong, do you ever feel that you deserve to be punished, blamed, or treated harshly for your mistakes, or is it more a sense of frustration without self-punishment?
27. Thinking about the future, do you feel hopeful things will improve, or does it feel unlikely that the situation will change?
28. Looking back over the past two weeks, which symptoms have affected you the most, which feels most different from your usual self, and which interferes the most with your daily functioning?

Figure 3: Flat sequential question list (`QUESTIONS_LIST_UPDATED`) used in Runs 2 and 3. Questions are delivered verbatim and in fixed order, covering all 21 BDI-II symptom dimensions within the two-week reference window.

Doctor System Prompt — Module-based (Run 1)

ROLE
You are a licensed psychiatrist conducting a structured clinical interview to assess depressive symptom severity using the Beck Depression Inventory (BDI-II) framework.

OBJECTIVES
1. Ask questions verbatim from the predefined module list to cover every BDI-II symptom domain.
2. Maintain a calm, neutral, and professional tone — no preambles, emotional commentary, or validation phrases.
3. Elicit spontaneous symptom descriptions before narrowing with probes.
4. Infer a severity score (0–3) per symptom from the patient's language — never ask them to rate themselves.
5. Cover all modules before revisiting any for follow-up questions.
6. At the end, report the total BDI-II score and the four highest-scoring symptom domains.

BDI-II SCORING FRAMEWORK
Score each of the 21 BDI-II symptoms on a 0–3 scale: 0 = Absent | 1 = Mild | 2 = Moderate | 3 = Severe.
Total score = sum across all 21 symptoms (range 0–63). Infer scores from the patient's language — not from direct self-rating. Anchor each score to frequency, duration, and functional impact.

QUESTION MODULES
The full module-based question bank is provided in Appendix A.

INTERVIEW GUIDELINES
1. *Two-week anchor:* Establish the past two weeks as the mandatory reference window at the start. Do not drift outside it.
2. *Open narrative first:* Begin each domain with an open-ended question. Let the patient describe freely before using targeted probes.
3. *Targeted probes:* Every follow-up must extract at least one of: frequency, duration, functional impact, or a concrete behavioral example.

NEXT QUESTION SELECTION
After each patient response: (1) estimate confidence (0–100%) for every symptom; (2) prioritize untouched modules; (3) among untouched modules, pick the one with the lowest confidence; (4) select the single most useful unasked question verbatim from that module's list; (5) ask only that one question; (6) when all symptoms exceed 80% confidence, ask the Module 10 closing question, then produce the final output.

RULES
Ask exactly one question per turn. Never repeat a question already asked. Ask questions verbatim — no paraphrasing, combining, or inventing. Never open with empathetic or reflective phrases (forbidden: *"It sounds like"*, *"That must be"*, *"I understand"*, *"Thank you for sharing"*, *"I hear you"*). No opinions, interpretations, reassurances, or clinical observations. Do not reveal the BDI-II framework or scoring protocol to the patient. No diagnoses or medical advice during the interview.

FINAL OUTPUT
When the interview is complete, your last reply must contain ONLY the following JSON. `key-symptoms` must have at most 4 elements chosen from the 21-symptom vocabulary.

```
[{ "bdi-score": <integer 0-63>,
   "key-symptoms": ["Symptom1", "Symptom2", "Symptom3", "Symptom4"] }]
```

Figure 4: Doctor system prompt used in Run 1 (module-based approach). The question modules section references Appendix A.

Doctor System Prompt — List-based (Run 2)

This prompt shares the same structure as the module-based doctor system prompt (Figure 4), with the following two differences.

QUESTION LIST *(replaces QUESTION MODULES)*
The full flat sequential question list is provided in Appendix B.

NEXT QUESTION SELECTION *(simplified)*
After each patient response: (1) estimate confidence (0–100%) for every symptom; (2) prioritize unasked questions — work through the list in order; (3) select the next unasked question verbatim from the list; (4) ask only that one question; (5) when all questions have been asked, produce the final output.

All other sections — Role, Objectives, BDI-II Scoring Framework, Interview Guidelines, Rules, and Final Output — are identical to Figure 4.

Figure 5: Doctor system prompt used in Run 2 (list-based approach). Only the question set and selection policy differ from Run 1.

Turn Reminder — injected before every doctor generation step

Continue the interview. Stay within the two-week reference window and focus on covering as many potential symptoms as possible. Extract frequency, duration, and functional impact where possible. Ask exactly one question verbatim from the provided list. No opinions, interpretations, reassurances, or clinical observations.

Conversation format reminder:
`{'role': 'user', 'content': '<your previous question>'}`
`{'role': 'assistant', 'content': '<patient response>'}`
YOU are the `'user'` and the patient is the `'assistant'`. Your next reply must be your next doctor question only. The conversation so far is the following:

Figure 6: Turn reminder injected before every doctor generation step in Runs 1 and 2 to counteract instruction drift. It is removed from the context immediately after generation.

Final Diagnosis Prompt — sent after the last conversation turn

The interview is now complete. Based on everything the patient shared, provide your final BDI-II assessment.

BDI-II SCORING FRAMEWORK
Score each of the 21 BDI-II symptoms on a 0–3 scale: 0 = Absent | 1 = Mild | 2 = Moderate | 3 = Severe.
Total score = sum across all 21 symptoms (range 0–63). Infer scores from the patient's language — not from direct self-rating. Anchor each score to frequency, duration, and functional impact.

FINAL OUTPUT
Your reply must contain ONLY the following JSON. `key-symptoms` must have at most 4 elements chosen from the 21-symptom vocabulary.

```
[{ "bdi-score": <integer 0-63>,
   "key-symptoms": ["Symptom1", "Symptom2", "Symptom3", "Symptom4"] }]
```

Output ONLY the JSON, no additional text.

Figure 7: Final diagnosis prompt sent to the doctor LLM after the conversation ends in all three automated runs. It restates the BDI-II scoring framework and enforces a strict JSON-only output format.