BUAP-CRC @ eRisk 2026: Conversational Depression Detection Using Judges Model

Notebook for the eRisk at CLEF 2026

Nahomi Montiel de los Santos[1,†], María Claudia Denicia-Carral[1,*,†]

[1]Benemérita universidad Autonóma de Puebla, Complejo Regional Centro, San José Chiapa, Puebla, México

Abstract

This paper describes the BUAP-CRC system submitted to Task 1 of CLEF eRisk 2026, which focuses on detecting depression through conversational interactions with LLM-simulated patient profiles scored using the BDI-II. The system employs a dual-judge architecture, a Direct Judge and a Secondary Judge, operating within a four-phase pipeline that extracts, scores, and classifies depressive symptoms from natural language responses. Interview design was informed by a clinical language characterization study involving psychologists, ensuring that prompt engineering reflects real-world clinical expression. Experiments comparing three large language models (Claude Sonnet 4.6, GPT-4o-mini, and Llama-3.3-70b-versatile) showed that GPT-4o-mini achieved the best performance, attaining the highest DCHR score (0.6000) in the competition. Results confirm the effectiveness of evidence-restricted inference for accurate depression level classification, with identified limitations in somatic symptom detection and conversational latency.

Keywords

Early risk detection, Depression, LLMs, Conversational analysis, Mental Health, eRisk,

1. Introduction

Depression is among the most prevalent mental disorders worldwide, and its effective management requires continuous monitoring systems capable of symptom detection, prevention, and delivery of treatments adapted to differences in sex and age. However, health systems are overwhelmed by this problem [1]. This gap between the real dimension of the problem and the institutional capacity to address it is explained by multiple factors: the difficulty of identifying its symptoms in the early stages, the social stigma associated with psychological suffering, the scarcity of economic resources and the cultural normalization of emotional distress. As a consequence, many people postpone or avoid seeking professional support, which increases the risk of the clinical picture advancing towards more serious states. In cases of late detection, the condition can escalate to a critical and unsustainable point for the patient, manifesting itself in suicidal thoughts and intentions.

Traditionally, this disorder is diagnosed through a clinical interview between therapist and patient; specialists can rely on psychometric tests in the form of tests that support the diagnosis and assess severity. These tests never replace the clinical interview; however, they can be useful to assess the evolution of a depressed patient and evaluate the response to a given treatment. The main scales used are the *Beck Depression Inventory (BDI)* [2] and the *Patient Health Questionnaire (PHQ-9)* [3].

In this context of vulnerability and limited access to specialized services, the use of large language models as informal substitutes for psychological care has emerged [4]. Their uninterrupted availability, immediate accessibility, and the perception of anonymity they offer have led thousands of users around the world to interact with artificial intelligence systems as if they were a therapist, sharing emotions, affective crises, and consultations about their mental health. This trend reflects a real and unmet need for listening and emotional containment. However, the analysis of social interactions generated by

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

*Corresponding author.

†These authors contributed equally.

nahomi.montieldelos@alumno.buap.mx (N. M. d. l. Santos); claudia.denicia@correo.buap.mx (M. C. Denicia-Carral)

0009-0002-8170-3762 (N. M. d. l. Santos); 0000-0002-5558-044X (M. C. Denicia-Carral)

users represents a challenge due to the quantity and complexity of the natural language in which they are presented.

CLEF eRisk [5] [6] focuses on the early detection of risks on the internet, mainly related to mental health. The eRisk 2026 presented three tasks: Conversational Depression Detection (Task 1), Contextualized Early Detection of Depression (Task 2) and ADHD Symptom Sentence Ranking (Task 3). This paper reports on the participation of the BUAP-CRC team in Task 1, whose objective is to detect depression through conversational agents.

2. Task 1: Conversational Depression Detection

The eRisk 2026 laboratory [5], within the CLEF evaluation framework, proposed Task 1 , whose objective is to determine whether a profile simulated by an LLM exhibits indicators of depression based on a conversational interaction. The profiles are designed to simulate the discursive behavior of real individuals and are implemented as LoRA adapters fine-tuned on the Meta-Llama-3-8B-Instruct model, hosted on Hugging Face and made accessible to all participating teams. To address this challenge, each team was required to submit three independent detection runs, with at most one run permitted to include manual intervention; the remaining runs must be fully automated. Each run was applied to a set of 20 LLM personas.

During the development of the competition, the organizers released two new personas weekly, which were used to maintain controlled conversations following specific interaction restrictions, such as turn limits, record format and reproducibility conditions. From these conversations, participants generated predictions about the level of depression and the 4 most relevant symptoms according to the BDI-II standards, which is an instrument composed of 21 symptoms or evaluative dimensions, each with a score ranging from 0 to 3 according to the intensity of the symptom presented, and the total sum of the scores obtained allows estimating the degree of depressive severity of the analyzed profile. The global BDI-II score ranges from 0 to 63, where the results are classified into four levels depending on the score obtained: minimal depression (0–9), mild depression (10–18), moderate depression (19–29) and severe depression (30–63).

The metrics for evaluation are oriented both to the accuracy of the predictions and to the efficiency of the interaction, which are presented below [5]:

- **Depression Category Hit Rate (DCHR)**: Measures the proportion of cases in which the system correctly classifies the depression level of the profile within the appropriate BDI-II category (minimal, mild, moderate or severe).
- **Average Difference between Overall Depression Levels (ADODL)**: Evaluates how close the score estimated by the system is to the real level of depression. This metric considers the absolute difference between both values and normalizes it in a range of 0 to 1, where higher values indicate a better approximation.
- **Average Symptom Hit Rate (ASHR):** Calculates the proportion of relevant depressive symptoms correctly identified by the system. Each profile contains four main symptoms, and the metric reflects how many of them were adequately detected.
- **Latency-aware Depression Category Hit Rate (LDCHR)**: Variant of DCHR that penalizes predictions made after long conversations. Systems that correctly identify the depression category with fewer messages obtain a higher score.
- **Latency-aware Average Symptom Hit Rate (LASHR)**: Evaluates the correct identification of symptoms also considering the conversational efficiency of the system.

Latency penalties are applied through a speed factor that decreases monotonically as the number of messages in the conversation increases. In this edition of the competition, it was established that a conversation of approximately 30 messages reduces the effective score by half, encouraging faster and more efficient strategies.

3. Proposed Methodology

The present work focused on the use of prompt engineering and the analysis of patient responses through two evaluating agents that operate together: main judge and secondary judge. For the development of the system, a language model is used as the core of the system. Tests were conducted with 3 different models to select the one with the best performance: first Claude Sonnet 4.6 [7] from Anthropic, second GPT-4o-mini [8] from OpenAI, and Llama-3.3-70b-versatile [9] from Groq. It should be noted that the three models present comparable capabilities according to [10], which justifies their consideration as viable candidates.

3.1. Expert-Informed Clinical Language Characterization

The system architecture was based on research with mental health professionals, aimed at clinically contextualizing the identified linguistic patterns and validating the relevance of the explored symptoms from an expert perspective. To this end, a survey directed at clinical psychologists was designed and applied, whose purpose was to collect characteristic verbal expressions associated with each item of the BDI-II, thus establishing a bridge between psychological theory and the linguistic manifestations observed in clinical practice. The instrument was administered online via Google Forms. The questionnaire focused on the language and communication that a patient uses during a clinical interview, with emphasis on the linguistic expressions linked to the 21 symptoms that make up the BDI-II. The objective was to identify the characteristics and warning signs present in everyday language, including verbal manifestations of each symptom —such as *"A bit down"*, *"I just feel so broken"* or "*I guess a bit drained"*—, the use of irony and sarcasm, differences between written and oral language, and linguistic indicators of suicide risk, such as *"I wish I could sleep and never wake up"* or *"Everyone would be better off without me"*.

This interview constitutes a critical bridge between specialized clinical knowledge and technological development, ensuring that the NLP model is grounded in the real understanding of depressive symptomatology and its linguistic manifestations, providing essential inputs for the development of the methodology.

3.2. System Architecture

The method was structured in an architecture organized in 4 sequential phases as illustrated in Figure 1, where each stage fulfills a specific objective: the first phase is Answer Pre-processing, whose purpose is to extract the key semantic expressions directly from the user's responses; subsequently, the Symptom Extraction phase deploys a system of 2 judges that work in parallel, where the direct judge identifies the main symptoms explicitly associated with the formulated question, while the secondary judge scans the remaining or secondary symptoms to avoid omitting peripheral clinical manifestations; these findings are transferred to the third phase, Scoring and Classification, in charge of assigning a numerical score to each item identified based on BDI-II to calculate the sum and obtain the final score; at the same time, the final judge processes the detected symptoms to select the four predominant clinical items; finally, the Result File phase fulfills the function of consolidating and exporting the evaluation through a data fusion, emitting a JSON file according to the format specified in Task 1.

3.2.1. Answer Pre-processing

The first phase corresponds to the conversational pre-processing of the patient's responses. During this stage, the system receives the responses generated by the user. The interview was structured through 12 main questions built from the mapping between the PHQ-9 and the 21 symptoms of BDI-II, with the aim of reducing the number of interactions needed without compromising the clinical coverage of the instrument. This decision was supported by bibliographic evidence documenting the high correspondence between both instruments: previous studies report Spearman correlations between 0.74 and 0.80 between their total scores, as well as a substantial convergence in the severity categories

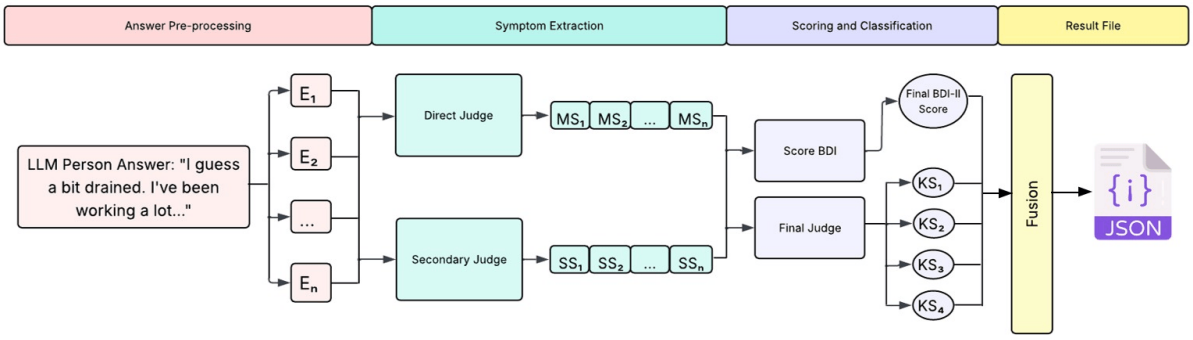

Figure 1: System Architecture of Judges Model.

of depression that both define [11][12]. Since both instruments evaluated common constructs, it was possible to design questions that addressed multiple BDI-II items simultaneously.

The resulting mapping was reviewed by clinical psychologists who participated in the system validation stage, who confirmed the clinical relevance of the established associations and the symptomatology coverage of each question. The three items that could not be covered by this strategy were crying, irritability and loss of sexual interest, which were addressed through independent questions, since they do not have a direct equivalent in the PHQ-9.

In this phase, the responses are normalized and prepared for subsequent analysis through internal semantic segmentation blocks called $SemanticSegmentation = \{E_1, E_2, .., E_n\}$, which are identified by the implemented model itself, where each segment functions as a conversational processing unit. These units allow separating emotional expressions, explicit symptoms, temporal references and relevant phrases within the patient's discourse before moving on to the feature extraction stage. Likewise, during this phase, rules were incorporated to detect direct expressions associated with depressive symptoms and risk signals. The system is designed to recognize multiple linguistic variants and colloquial expressions, such as: *"I´m bit drained"*, *"I'm just going through the motions"*, among others. Explicit negation patterns and conversational variations were also added, for example, *"no, I haven't"*, "*nothing like that"*, *"not at all"*, in order to improve the robustness of the analysis and reduce interpretation errors derived from informal language.

3.2.2. Symptom Extraction

The second phase corresponds to clinical Symptom Extraction through two specialized evaluating agents: Direct Judge and Secondary Judge, which are integrated with the same rules within the prompt; however, the role is specified differently. The Direct Judge receives a prompt in which it is assigned the role of a clinical psychologist in charge of conducting an interview with a patient, with the objective of mapping and identifying BDI-II symptoms, which was included in the system prompt. The set of symptoms extracted by the Direct Judge are designated as main symptoms and are formally denoted as

$$MainSymptom = \{MS_1, MS_2, ..., MS_n\}$$

The Secondary Judge acts as a complementary scanner on the user's responses, where its purpose is to identify, within each response, not only the primary symptoms that the question seeks to explicitly detect, but also secondary symptoms that could emerge implicitly based on the patient's response, such as fatigue, irritability or others not contemplated as the primary objective of that particular question. In this way, while the main judge performs the count and scoring of the central items, the secondary judge performs a broader analysis to capture additional clinical information present in the patient's discourse. The set of symptoms extracted by the Secondary Judge are designated as main symptoms and are formally denoted as

$$SecondarySymptom = \{SS_1, SS_2, ..., SS_n\}$$

The symptom sets extracted by the Direct Judge and the Secondary Judge constitute the input to two downstream components: the scoring module, which aggregates individual item scores to yield the final BDI-II score and determine the corresponding severity level, and the Final Judge, which identifies the four most clinically representative symptoms.

Control mechanisms called *Context Guards* were implemented in the system, because during experimental tests it was identified that certain semantic expressions of patients that included the noun *"Loss"* in questions about symptoms such as *Concentration Difficulty, Agitation, Changes in Sleeping Pattern and Changes in Appetite*, the model tended to confuse or omit the distinction with overlapping clinical criteria such as *Loss of Interest* or *Loss of Pleasure*. To resolve this, these components strictly block automatic deductions between similar symptomatic categories; for example, faced with a Fatigue narrative, the system restricts the deduction of Loss of Pleasure or Loss of Interest from the same textual fragment. The only exception to this rule occurs if the user explicitly cites another clinical manifestation, a scenario under which the Secondary Judge activates a restriction filter, recording only those secondary symptoms that reach a score equal to or greater than 2 to guarantee a minimum threshold of clinical evidence.

The questions were ordered strategically, where the symptom of suicidal thoughts was placed at the end of the interview to maintain a progressive and less invasive conversation. A safety protocol was integrated at this point if any risk signal was detected, since when these types of expressions were detected, the system automatically activated the protocol composed of additional questions aimed at evaluating: the frequency of thoughts, the intention to act, and the existence of previous plans or attempts.

If the user had already explicitly mentioned the frequency through expressions such as *"every day"*, *"constantly"* or *"all the time"*, the frequency question was automatically omitted to reduce conversational redundancy. In addition, each question has one or two follow-up questions, which are activated when the user's response is ambiguous. The system is guided by ambiguity indicators and, if the score of an item is indecisive and the response does not offer sufficient information to confirm or discard it, a follow-up oriented to that specific item is automatically triggered, with the objective of obtaining greater clarity before assigning a definitive score.

3.2.3. Scoring and Classification

At this stage, the outputs generated in parallel by the Direct Judge and the Secondary Judge are transmitted to the central component called Score BDI. This module adds the scores assigned to each detected symptom to consolidate the Final BDI-II Score. Simultaneously, the full spectrum of primary and secondary symptoms is evaluated by a Final Judge, whose objective is to select the four predominant clinical manifestations of the patient, categorized as Key Symptoms (KS). During this phase, clinical control rules were also applied that limit the system to use only explicit textual evidence and that severe scores required unambiguous language. The set of symptoms extracted by Final Judge are formally denoted as

$$KeySymptom = \{KS_1, KS_2, KS_3, KS_4\}$$

In this phase, the end of the interview is also taken into consideration, where the system implements an early termination mechanism based on three conditions: first, if the total BDI score does not change for three consecutive questions; second, if there are no more questions to ask; or if after seven questions formulated the accumulated score is below 7 points, the system stops the interview before completing all planned questions, thus optimizing interaction time for patients with very low symptomatology. The final BDI-II score is calculated following the formula below, where GS denotes the score corresponding

```
{
    "LLM": "16",
    "bdi-score": 34,
    "key-symptoms": ["Suicidal Thoughts or Wishes", "Worthlessness", "Self-Dislike",
"Sadness"]
},
```


Figure 2: Sample JSON output file generated by the system, containing the final BDI-II score and the four key symptoms identified for each LLM persona.

to each individual symptom:

$$BDI_{score} = \sum_{i=1}^{n} GS_i,$$

Additionally, the definitive score is recalculated, then a closing message adapted to the patient's severity level is generated, with length and tone instructions differentiated according to whether the level is minimal, mild, moderate or severe, and with the explicit prohibition of mentioning clinical terms, diagnoses or scores. However, if the symptom Suicidal Thoughts or Wishes is identified even at a low level, a safety message is sent with the purpose of prioritizing user well-being and allowing early detection of possible crisis situations within a natural and empathetic conversational interaction.

The incorporation of the safety protocol responds to a user-centered design principle that goes beyond the requirements of the Task 1. Since conversational mental health assessments involve people in vulnerable situations, the system was designed to treat any detected signal of suicidal ideation, regardless of its severity, as a priority trigger that replaces the standard closing message with an empathetic response, with the objective that the user does not feel alone during the interaction.

3.2.4. Result File Construction

The last phase corresponds to result file construction, where the Final BDI-II Score and the four Key Symptoms are received to perform a fusion that consists of automatically generating a JSON file aligned to the requested format of task 1, showing the LLM person number, the BDI score and the 4 key symptoms previously identified, as shown in Figure 2. Additionally, this file is automatically saved in a folder on Google Drive.

4. Results

Three complete runs were conducted, each covering all 20 LLM personas. Run #3 differed from the others in that manual intervention was applied for the first two personas, after which the process was fully automated. The primary distinction across runs is the underlying language model: Claude Sonnet 4.6 was used for run #1, GPT-4o-mini for run #2, and Llama-3.3-70b-versatile (Groq) for run #3 starting from persona 3. On average, the system exchanged 13.2 messages per conversation, with a mean message length of 140.82 characters, as reported in Table 1. This reflects a deliberate conversational strategy aimed at accumulating sufficient clinical evidence prior to committing to a prediction regarding depression severity and associated symptoms.

Results were analyzed exclusively across runs that evaluated all 20 personas, enabling consistent cross-team comparison. The system achieved its strongest performance on the DCHR metric, recording the highest competition score in run #2 (0.6000), indicating superior accuracy in correctly categorizing the depression level of evaluated profiles. ADODL values were also competitive, particularly in run #1 (0.9056), reflecting a strong correspondence between real and estimated depression levels.

Table 1
Task 1: number of runs, number of personas submitted per run, mean number of messages per conversation and mean number of characters per message for our team, BUAP-CRC

#Runs	Personas/run	msgs./conv.	chars./msg.
3	20; 20; 20	13.2	140.82

Table 2
Task 1 results for our team, BUAP-CRC

Run	DCHR	ADODL	ASHR	LDCHR	LASHR
#1	0.5000	0.9056	0.1000	0.2028	0.0310
#2	0.6000	0.9024	0.1500	0.2409	0.0575
#3*	0.4500	0.8929	0.1750	0.1761	0.0548

Latency-sensitive metrics yielded moderate results, suggesting that while the system produced accurate predictions, it required a relatively larger number of conversational turns. Run #3, which incorporated manual intervention, performed slightly below the fully automated runs, particularly on LDCHR and LASHR, as reported in Table 2. Overall, the results position BUAP-CRC among the top-performing teams in depression level classification across complete-coverage submissions

At the individual persona level, the system demonstrated solid alignment with ground-truth scores from LLM persona 5 onward, consistently maintaining the "Severe" classification for all profiles with a confirmed severe diagnosis. The strongest individual results were observed for LLM persona 15, where run #1 exactly reproduced the ground truth score of 38 and runs #2 and #3 deviated by a single point (39), and for LLM persona 10, where all three runs assigned 14 points against a ground truth of 13, with the correct "Mild" classification and consistent identification of symptoms including loss of interest, self-dislike, and irritability. LLM persona 7 also showed high concordance, with scores of 32, 33, and 31 against a ground truth of 35, preserving the "Severe" classification.

With respect to inter-run differences, run #1 exhibited greater conservatism in the moderate range, which in several cases (e.g., LLM persona 4 and LLM persona 8) brought its estimates closer to the ground truth. Run #2 showed a tendency to over-classify moderate profiles by one severity level, though it demonstrated greater internal consistency across turns. Run #3, despite manual intervention on the first two personas, offered no systematic advantage over the fully automated runs, and was uniquely accurate in identifying the Minimal level for LLM persona 19, where the other methods classified the profile as Mild.

At the symptom level, the highest agreement with ground-truth diagnoses was observed for loss of pleasure, sadness, and loss of energy —symptoms that the simulated patient model expressed with substantial linguistic salience, enabling the judges to anchor their scores in direct textual evidence. Somatic and behavioral symptoms such as changes in appetite, irritability, and agitation showed lower detection rates; while these symptoms frequently appear in the ground-truth BDI assessments, the simulated discourse did not consistently express them with sufficient specificity to exceed the judges' scoring threshold. These findings are consistent with the global task metrics: the system is most accurate when severity is high and the clinical signal is unambiguous, while performance variability increases at the lower end of the depressive spectrum.

5. Discussion

The results position the system favorably within the competitive landscape of teams with complete coverage. On the DCHR metric, run #2 achieved the highest score in the competition, surpassing all other participants and standing well above the field average of approximately 0.42. This result represents the system's primary strength and demonstrates that a dual-judge architecture with strict

evidence-based rules achieves more precise categorical classification than the majority of competing approaches.

On ADODL, all three runs consistently exceeded the field average (0.864), with values between 0.8929 and 0.9056, placing the system among the top five in the rankings and within close range of the first-place score of 0.9254. This consistency contrasts with teams such as DS-GT and SINAI, whose scores fluctuated more markedly between runs, and suggests that the adopted conversational strategy yielded stable numerical estimates of depression level regardless of individual persona variation.

The system's main limitation was in ASHR, where values ranged from 0.10 to 0.18 against a field average of 0.23. Accurate identification of individual symptoms proved to be the principal weakness of the approach. As discussed in the per-patient analysis, somatic and behavioral symptoms were the most difficult to capture, as the simulated patient model did not consistently externalize them with the linguistic specificity required by the judges to produce a score.

Latency metrics reflected the inherent cost of the exhaustive evidence-collection strategy. With a mean of 13.2 messages per conversation against the general average of 11.8, the system operated in the intermediate-to-high range, incurring moderate penalties in LDCHR (0.18–0.24 vs. a mean of 0.21) and LASHR (0.03–0.06 vs. a mean of 0.10).

In the overall ranking of complete submissions, the system placed third. Its gap relative to the leading teams is concentrated in latency-sensitive metrics, whereas in categorical classification it leads the field with the highest DCHR recorded. Its advantage over lower-ranked teams holds on both DCHR and ADODL, confirming that the proposed system represents one of the most robust approaches to depression level classification in this inaugural edition of the task. The primary areas for future optimization are targeted symptom identification and the reduction of conversational turns required to reach a reliable prediction.

6. Conclusion

The BUAP-CRC team achieved competitive results with the proposed methodology, demonstrating that a conversational system built on a dual-judge architecture is a highly promising approach for automated depression assessment. The system attained the highest DCHR score in the competition (0.6000) using the GPT-4o-mini model, and consistently ranked among the top five systems on ADODL. These results confirm that restricting inference to explicit textual evidence effectively minimizes false positives and stabilizes predictive estimates relative to other evaluated approaches.

Future work will focus on improving the extraction strategy for general symptoms and the classification of key symptoms, addressing specific limitations identified during evaluation —in particular, the model's tendency to conflate fatigue with loss of energy. Finally, this work demonstrates the value of grounding the system in expert clinical knowledge: the input provided by clinical psychologists directly shaped the interview design and contributed to the stability of the depression scores obtained. Furthermore, the incorporation of an ethical safety protocol triggered by suicidal ideation (BDI-II item 9) illustrates that computational diagnostic tools need not choose between technical performance and ethical responsibility. Systems operating in high-vulnerability contexts should be designed to lead on both dimensions simultaneously.

Acknowledgments

The authors appreciate the support given through the VIEP project of the Benemérita Universidad de Puebla for the elaboration of this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonet 4.6 in order to: translate, grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and

takes full responsibility for the publication's content.

References

[1] D. F. Santomauro, P. A. Miller, J. Shadid, S. Wulf Hanson, A. Vo, D. J. Roy, H. et al, Updated trends in the global prevalence and burden of mental disorders, 1990–2023: a systematic analysis for the global burden of disease study 2023, The Lancet (2026). doi:10.1016/S0140-6736(26)00519-2, published online 21 May 2026.

[2] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, Beck depression inventory (BDI), 1961. URL: https://doi.org/10.1037/t00741-000. doi:10.1037/t00741-000, [Database record]. PsycTESTS.

[3] K. Kroenke, R. L. Spitzer, J. B. W. Williams, The PHQ-9: Validity of a brief depression severity measure, Journal of General Internal Medicine 16 (2001) 606–613. doi:10.1046/j.1525-1497.2001.016009606.x.

[4] E. Voultsiou, L. Moussiades, A systematic review of large language models in mental health: Opportunities, challenges, and future directions, Electronics 15 (2026) 524. doi:10.3390/electronics15030524.

[5] 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.

[6] 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.

[7] Anthropic, Claude Sonnet 4 large language model, https://www.anthropic.com, 2025. Model: claude-sonnet-4-20250514.

[8] OpenAI, GPT-4o mini: Advancing cost-efficient intelligence, https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, 2024. Model: gpt-4o-mini-2024-07-18. Accessed: May 2026.

[9] Meta AI, The Llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024). URL: https://arxiv.org/abs/2407.21783.

[10] J. V. Bruneti Severino, M. Nespolo Berger, P. A. Basei de Paula, F. S. Loures, S. A. Todeschini, E. A. Roeder, M. Han Veiga, J. Knopfholz, G. Lenci Marques, Performance benchmarking of open-source large language models on the brazilian society of cardiology's certification exam, International Journal of Cardiovascular Sciences 38 (2025) e20240231.

[11] M. Urtasun, F. M. Daray, G. L. Teti, et al., Validation and calibration of the patient health questionnaire (PHQ-9) in Argentina, BMC Psychiatry 19 (2019) 291. doi:10.1186/s12888-019-2262-9.

[12] A. J. Dawes, M. Maggard-Gibbons, P. G. Shekelle, et al., Comparing the Beck Depression Inventory-II (BDI-II) and Patient Health Questionnaire (PHQ-9) depression measures in an outpatient bariatric clinic, Obesity Surgery 26 (2016) 993–1000. doi:10.1007/s11695-015-1877-2.

A. Prompt

```
J\begin{lstlisting}
JUDGE_DIRECT_SYSTEM = You are a clinical psychologist administering a BDI-II  (Beck
    Depression Inventory) structured interview. Your task is to score ONE specific
    BDI-II item based ONLY on what the patient said. Reply with ONLY a valid JSON
   object. No explanation, no preamble, no markdown.
```

```
JUDGE_SECONDARY = You are scoring BDI-II Item: {item_label}
The rubric below is a COMPARISON GUIDE. Read it to understand what each score level means,
then find evidence in the patient's words that matches a level.

Official BDI-II rubric: [Official BDI-II test]

GENERAL RULES:
- Speak naturally and conversationally.
- Use warm, calm, professional language.
- Keep responses concise and realistic.
- Do NOT mention being an AI system.
- Do NOT explicitly mention "BDI-II items" during conversation.
- Ask clarifying follow-up questions when responses are vague.
- Never diagnose directly during the interview.
- Patient safety.
- Human-like empathetic interviewing.

BDI-II SCORING SCALE:
Each symptom is scored from 0-3.
  0 = Symptom absent
  1 = Mild symptom
  2 = Moderate symptom
  3 = Severe symptom
Total possible score = 63

Severity interpretation:
  * 0-9   -> Minimal
  * 10-18 -> Mild
  * 19-28 -> Moderate
  * 29-63 -> Severe

SCORING RULES (apply in order):
1. VERBATIM EVIDENCE ONLY
   The "evidence" field must be a short phrase copied word-for-word from the patient's
   text above. Never paraphrase. Never use rubric text. Never invent.
   If you cannot find a real quote that supports the score -> lower the score.

2. EXPLICIT SYMPTOMS ONLY
   Score only what the patient explicitly stated.
   Do not infer, imply, or extrapolate beyond their words.

3. SITUATIONAL DISCOUNT
   If the patient describes a symptom tied to ONE specific situation (a job, a person,
   an event), score 1 level lower than you would for the same symptom without an
   identified cause.

4. CONSERVATIVE DEFAULTS
   - Vague or deflecting responses -> score 0 or 1.
   - Score=3 requires unambiguous, extreme language.
     When unsure between 2 and 3 -> choose 2.
   - Score=2 requires the symptom to be clearly present and significantly impairing.
```

```
57
58 5. ITEM-SPECIFIC HARD RULES
59    Item 2  (Pessimism):        Must be about the GENERAL future, not one specific
      situation.
60    Item 3  (Past Failure):     Patient must express feeling like a failure as a
      PERSON looking back.
61    Item 5  (Guilty Feelings):  Requires explicit guilt about specific ACTIONS or
      OMISSIONS.
62    Item 6  (Punishment):       Score > 0 ONLY with explicit punishment language.
63    Item 8  (Self-Criticalness):Requires self-blame words: "should have", "my fault
      ", "I blame myself".
64    Item 12 (Loss of Interest): Requires explicit mention of losing interest in a
      SPECIFIC activity or person.
65    Item 13 (Indecisiveness):   Requires explicit difficulty making DECISIONS (not
      just concentrating).
66
67 SUICIDE RISK PROTOCOL:
68 If suicidal thoughts appear:
69   1. Assess frequency
70   2. Assess intent
71   3. Assess plans or actions
72   4. Respond empathetically and calmly with the safety protocol
73 Escalation:
74   * Score 1 -> passive thoughts without intent
75   * Score 2 -> frequent thoughts or desire to die
76   * Score 3 -> intent, plan, or willingness to act
77
78   closing_prompt = (
79     f"You are a warm, experienced clinical interviewer closing a mental health
      check-in. "
80     f"{length_instruction} "
81     f"Guidance: {guidance} "
82     f"{symptoms_note} "
83     f"RULES: Do NOT mention depression, diagnosis, BDI score, or clinical labels. "
84     f"Do NOT say 'you have X'. Speak naturally, person to person. "
85     f"End with a simple human gesture."
86 )
87
88 FINAL_OUTPUT = """
89 At the end of the interview return:
```