SINAI at eRisk@CLEF 2026: Modular Conversational Systems for Depression Detection in LLM Personas

Notebook for the eRisk Lab at CLEF 2026

Alba Maria Marmol-Romero, María Dolores Molina-González, Manuel Garcia-Vega and Arturo Montejo-Ráez

Computer Science Department, SINAI, CEATIC, University of Jaén, 23071, Spain

Abstract

This paper describes the participation of the SINAI team in Task 1 of the CLEF 2026 eRisk Lab: Conversational Depression Detection. We propose a modular dual-agent architecture in which one large language model conducts the conversation while a second LLM evaluates depressive symptoms according to the Beck Depression Inventory-II. We explored three automated conversational strategies: a baseline dual-agent loop, a Pool of Experts architecture with specialized evaluators, and an adaptive risk-aware orchestration system with consistency safeguards. Among the 21 participating teams, our system achieved one of the most efficient interaction profiles, averaging only 4.4 messages per conversation while producing the highest message density. Despite these short interactions, our best configurations achieved competitive results, including a DCHR of 0.50, an ADODL of 0.8989, and a latency-aware LDCHR of 0.2500. The experimental analysis shows that most diagnostic information is captured within the first conversational rounds, although severe depression profiles tend to be underestimated when interactions are too brief. Overall, our results suggest that compact and information-rich conversational strategies can provide reliable depression severity estimation while maintaining low interaction latency.

Keywords

Early risk prediction, Depression detection, Natural Language Processing, Large Language Model

1. Introduction

The massive daily flow of user-generated content on social media has become a fundamental resource for the early detection of mental health disorders and high-risk behaviours. The eRisk@CLEF 2026 lab is dedicated to advancing computational systems capable of identifying mental disorders such as depression and ADHD symptoms as early as possible [1, 2]. This year, the lab has proposed three specific tasks to further this goal:

- **Task 1 - Conversational Depression Detection**. A novel task in which systems must interact with LLM-powered personas and determine whether each persona exhibits signs of depression. The challenge lies in identifying specific depressive symptoms and overall severity levels within a constrained conversational window, without resorting to direct clinical questioning.
- **Task 2 - Contextualized Early Detection of Depression**. This task focuses on detecting early signs of depression by sequentially analysing full conversational contexts, with the goal of raising an alert as early as possible in the interaction stream.
- **Task 3 - ADHD Symptom Sentence Ranking**. A sentence-level retrieval task targeting the 18 symptoms defined in the Adult ADHD Self-Report Scale (ASRS).

Recent editions of eRisk have increasingly incorporated conversational paradigms. In the 2025 edition, conversational depression detection was introduced as a pilot task, where participant systems interacted directly with simulated users to infer depressive symptoms through dialogue rather than static post

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

amarmol@ujaen.es (A. M. Marmol-Romero); mdmolina@ujaen.es (M. D. Molina-González); mgarcia@ujaen.es (M. Garcia-Vega); amontejo@ujaen.es (A. Montejo-Ráez)

0000-0001-7952-4541 (A. M. Marmol-Romero); 0000-0002-8348-7154 (M. D. Molina-González); 0000-0003-2850-4940 (M. Garcia-Vega); 0000-0002-8643-2714 (A. Montejo-Ráez)

analysis. Our participation in that edition showed that separating the conversational component from the clinical evaluation component could improve both interaction quality and predictive robustness. In particular, we observed that modular architectures based on independent conversational and evaluator agents provided more stable symptom estimation than monolithic systems, while empathetic conversational styles improved user engagement without requiring explicitly clinical questioning.

In this paper, we describe the participation of the SINAI team in Task 1: Conversational Depression Detection. Our main objective is to analyse whether progressively modularizing LLM-based conversational systems leads to more precise depression assessment. To study this, we designed three conversational architectures with increasing levels of specialization: a baseline dual-agent system, a Pool of Experts configuration with symptom-specialized evaluators, and a risk-aware orchestration system with consistency safeguards. Through these experiments, we aim to better understand the relationship between architectural modularity, conversational efficiency, and diagnostic precision in conversational mental health assessment systems.

2. Task 1: Conversational Depression Detection

Task 1 is a novel evaluation scenario in which participant systems must engage in open-ended conversations with simulated user personas, each implemented as a fine-tuned large language model (LLM). Specifically, each persona consists of a LoRA adapter trained on top of `meta-llama/Meta-Llama-3-8B-Instruct`, fine-tuned using real user histories. The personas were released incrementally on a weekly schedule: the first release window included personas 1 and 2, with each subsequent week releasing the next two personas, for a total of 20 personas (IDs 1-20) throughout the test phase. The core challenge is to determine, through natural conversation, whether each persona exhibits signs of depression, and if so, to estimate the overall severity and identify the most prominent depressive symptoms. The personas reflect different severity levels as guided by the Beck Depression Inventory-Second Edition (BDI-II), allowing systems to be evaluated across a full spectrum of simulated depression.

There is no hard limit on the number of interaction turns, but the evaluation metrics penalise late decisions: the earlier an accurate classification is produced, the higher the score. This creates a fundamental trade-off between conversational depth and decisional efficiency.

2.1. Systems and methods

Our main objective was to develop a system capable of estimating the severity of 21 depressive symptoms by interacting with simulated LLM-based users within a maximum of 21 dialogue turns (because in the BDI-II there are 21 different questions). We hypothesize that it is feasible to extract multiple symptom indicators from a single user interaction, making it possible to reach a reliable symptom assessment within this constraint.

In the previous edition, we found that an LLM acts as a better predictor when split into conversational agent and evaluator agent [3]. Last year we tested with different conversational agent attitudes. However, this year, we tested different modular evaluator systems. All three runs share a common modular base composed of two collaborating LLMs:

- *Conversational Agent*: Responsible for interacting directly with the user. Its primary goal is to maintain a coherent and engaging conversation while implicitly collecting information relevant to depressive symptoms. It asks context-aware questions and responds to user replies naturally. In the past edition we tested several attitudes of this agent, and the best one was run 1 [3], where the system responds empathetically without sharing personal anecdotes.
- *Evaluation Agent*: Does not interact with the user. Instead, it receives the dialogue history and analyses each exchange to infer and update the current state of the 21 depressive symptoms. It assigns a severity value on a 0-3 scale to each symptom based on the content of the conversation

so far. Furthermore, it reasons whether it needs to keep updating symptom values or has enough information, in which case it sends the Conversational Agent to end the conversation.

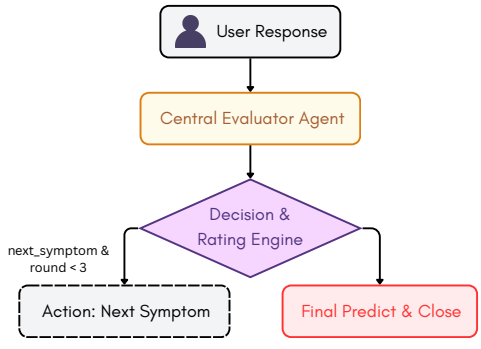
We also implemented a final planning mechanism that dynamically suggests which symptoms are underexplored, allowing the conversational agent to prioritize specific topics in subsequent messages. This ensures that all symptoms are assessed at least once, and that high-risk symptoms can be probed in more detail. All systems operate in a cyclic process that enforces a minimum of three interactions before allowing the conversation to conclude:

1. Conversational Agent Initiation: The conversation begins with the conversational agent sending an initial message, always asking about mood and sadness. The model used the prompt defined in Appendix A.
2. User Interaction: The user responds, and this response is appended to the conversation log.
3. Evaluation Agent Analysis: The complete conversation history is then sent to the evaluation agent. In rounds after the third, the model has the authority to signal that it has gathered enough information to stop the conversation based on its evaluation of the symptom scores. Baseline prompt used is in Appendix B.
4. Continuation or Termination: Based on Evaluation Agent's feedback, which includes both the updated symptom scores and an indication of whether further clarification is needed, decides to either continue the conversation or conclude it.
5. Final Submission: Once the evaluation agent determines that no further information is necessary (i.e., it returns "None" for the next symptom query), and the number of minimum rounds is reached, the conversation is terminated, and the final symptom scores are recorded.

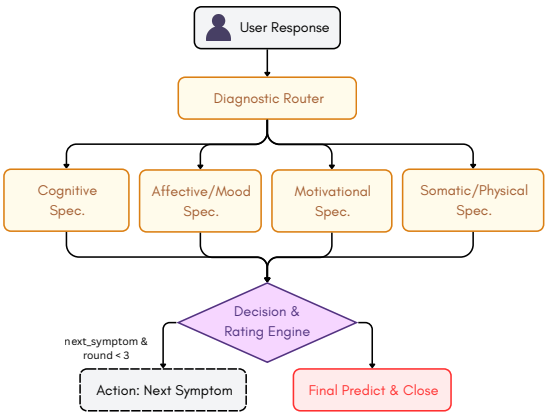
The three runs differ exclusively in the configuration of the Evaluation Agent, as described below.

2.1.1. Run 0: Baseline Dual-Agent Loop

Run 0 is the baseline modular system, representing a direct dual-agent loop (Figure 1a). The Evaluation Agent analyses the conversation as a single general clinical tracker and dynamically writes the target symptom to explore next in a JSON structure. This design mirrors the best-performing configuration from our previous participation and serves as the reference point against which the more complex architectures are compared.



(a) Run 0: baseline dual-agent loop with a single Central Evaluator Agent.



(b) Run 1: Pool of Experts replacing the monolithic evaluator with four clinical specialists aggregated by a central Decision and Rating Engine.

Figure 1: Architecture of the modular conversational systems for Runs 0 and 1.

2.1.2. Run 1: Clinical Specialists Pool of Experts

To improve clinical precision and explainability, the monolithic evaluator of Run 0 is replaced by a parallelised, structured Pool of Experts (PoE) pipeline aligned with the multi-dimensional nature of the BDI-II (Figure 1b). This architecture is grounded in the PoE framework [4], a prompt-based multi-agent approach whose central idea is to replace a single direct LLM prediction with a structured reasoning process involving multiple role-specialised agents. Rather than asking one model instance to solve the task in isolation, PoE decomposes the problem into three main operations: selecting relevant expertise, generating independent expert judgments, and aggregating these judgments into a final answer. In the PoE framework an agent is an LLM instance conditioned by a role-specific system prompt. All agents can be instantiated from the same underlying model; their differences therefore do not arise from distinct model parameters, fine-tuning, or task-specific training. Instead, diversity is introduced entirely at the prompt level, through each agent's functional role, its assigned field of expertise, and the particular subset of symptoms it is responsible for evaluating. This design makes the architecture both lightweight and modular, as adding or replacing a specialist requires only a prompt change rather than training a new model. In our adaptation of this framework, the conversation history is processed by four specialised clinical agents, each aligned with one of the major symptom dimensions of the BDI-II:

- *Cognitive Specialist*: Evaluates a cluster of 9 cognitive symptoms: pessimism, past failure, guilty feelings, punishment feelings, self-dislike, self-criticalness, worthlessness, indecisiveness, and concentration difficulty.
- *Affective Specialist*: Evaluates 5 mood-related symptoms: sadness, loss of pleasure, crying, agitation, and irritability.
- *Motivational Specialist*: Assesses 3 motivational symptoms: loss of interest, suicidal thoughts or wishes, and loss of interest in sex.
- *Somatic Specialist*: Assesses 4 physical symptoms: loss of energy, changes in sleeping pattern, changes in appetite, and tiredness or fatigue.

Each specialist assigns scores on a 0–3 scale and provides textual justification (see Figure 1b), compiled into a unified *Assessment Summary*. Finally, a central Aggregator Agent synthesizes this summary, tracks symptom-wise evidence, flags controversial inputs as `unreliable`, and determines the next optimal target if needed.

2.1.3. Run 2: Adaptive Risk and Consistency Safeguard

Building upon the specialists of Run 1, this run introduces a safety-critical orchestrator designed to handle risks and evasive answers. After the four BDI-II specialists compile the *Assessment Summary*, the pipeline executes a state classifier to categorize the patient's current severity status into either `"OK"` or `"Risk"` (see Figure 2). Based on this classification, the system dynamically routes the dialogue management through bifurcated decision engines:

- *Light Evaluator*: Activated if the status is classified as `"OK"`, optimizing for natural conversational flow and standard exploration.
- *Strict Evaluator*: Enforced if severity or risk markers are detected, prioritizing structured clinical verification, gentle risk-handling prompts, and safe exit boundaries.

Both decision makers actively assess the reliability of the patient's answers, generating an `unreliable_count`. To prevent premature terminations and handle deceptive responses, the system enforces a strict stopping condition: the conversation is only allowed to close if the `unreliable_count` reaches zero (all tracked symptoms are verified) for the Strict Evaluator and is more relaxed for the Light Evaluator.

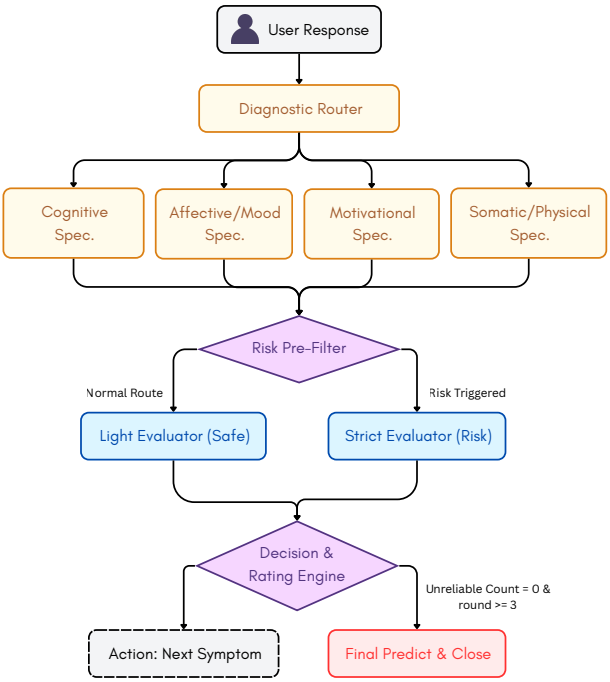


Figure 2: Run 2: Adaptive Risk and Consistency Safeguard system. A Risk Pre-Filter routes the dialogue through a Light or Strict Evaluator depending on detected severity, closing only when all symptoms are verified.

2.2. Implementation Details

The LLMs used in this architecture are based on the Llama model family [5], specifically the Llama 3.1 family of models [6]. We used the `Llama-3.1-8B-Instruct` variant. We leveraged this model's robust language understanding and generation capabilities to handle both conversational and evaluative tasks in our system. To ensure seamless integration with downstream analysis tools, all the LLMs are prompted to output responses in a strict JSON-like format. This approach standardizes the output for easy parsing and subsequent processing. All model parameters for this task, such as temperature, top-k, etc., the default settings were used. The experiments were conducted on a dedicated cluster owned by the SINAI[1] research group. The processing pipeline was implemented in Python and executed using the vLLM framework [7] for efficient inference with LLMs. We used a single NVIDIA RTX 4000 GPU, running on a Linux-based system. The environment included PyTorch, and all processes were orchestrated via custom scripts. The combination of optimized software and hardware acceleration enabled us to complete each run efficiently, with minimal latency per thread.

3. Results and discussion

In Task 1, our team submitted three fully automated runs, each implementing a distinct evaluation strategy (as described in Section 2). Our primary goal was to explore the effectiveness of depressive symptom detection within a constrained number of dialogue turns while maintaining psychologically sensitive interactions.

According to the official interaction statistics released by the task organisers, our team adopted one of the most efficient interaction strategies among all 21 participating teams, averaging 4.4 messages per conversation, the lowest among all teams. Despite the brevity of these exchanges, our messages were among the most information-dense, with an average of 242.61 characters per message, the highest of any team in the competition. This confirms that our system design prioritised compact, high-information prompts to probe depressive symptoms efficiently within a minimal number of conversational turns.

[1] https://sinai.ujaen.es/

This interaction profile directly reflects our modular dual-agent architecture: the Conversational Agent is designed to formulate rich, targeted questions that maximise symptom coverage per turn, while the Evaluation Agent continuously assesses whether sufficient evidence has been gathered to terminate the conversation early.

The evaluation of Task 1 was based on five metrics: Depression Category Hit Rate (DCHR), Average Difference between Overall Depression Levels (ADODL), Average Symptom Hit Rate (ASHR), and their latency-aware variants LDCHR and LASHR, which penalise decisions reached after longer conversations, as defined in Perez et al. [2]. Table 1 presents our results alongside the best-performing run per metric among complete submissions, and the competition mean and median computed over all 24 complete-submission runs.

Team	DCHR	ADODL	ASHR	LDCHR	LASHR
pjmathematician - Run 3	0.5500 (2)	**0.9254 (1)**	0.3000 (4)	**0.3872 (1)**	0.2112 (3)
BUAP-CRC - Run 2	**0.6000 (1)**	0.9024 (6)	0.1500 (15)	0.2409 (7)	0.0575 (15)
upb-uottawa - Run 3	0.3000 (8)	0.7849 (23)	**0.3375 (1)**	0.0214 (19)	0.0241 (21)
Ours - Run 0	0.5000 (3)	0.8905 (10)	0.1875 (11)	0.2500 (5)	0.0938 (11)
Ours - Run 1	0.4000 (6)	0.8738 (13)	0.2250 (10)	0.2000 (13)	0.1125 (9)
Ours - Run 2	0.4737 (4)	0.8989 (7)	0.1711 (13)	0.2368 (8)	0.0855 (12)
eRisk mean	*0.4197*	*0.8671*	*0.2228*	*0.1893*	*0.0905*
eRisk median	*0.4500*	*0.8821*	*0.2313*	*0.2060*	*0.0831*

Table 1
Task 1 evaluation results. The top section shows the best-performing team per metric among complete-submissions (all 20 personas). Bold values indicate the best result across all runs for each metric. Mean and median are computed over all 24 complete-submission runs.

Across all five metrics, our best run per metric exceeded the competition mean, and in four out of five metrics it also exceeded the median. The sole exception is ASHR, where our best value (0.2250, Run 1) fell marginally below the median (0.2313), reflecting the inherent difficulty of identifying individual depressive symptoms within a short conversational window.

Regarding category classification, Run 0 achieved the best DCHR among our submissions (0.5000, rank 3), correctly assigning personas to the appropriate BDI-II severity category in half of the evaluated cases. Run 2 obtained the closest overall severity estimate (ADODL = 0.8989, rank 7), while Run 1 produced the highest symptom hit rate (ASHR = 0.2250, rank 10). On the latency-aware metrics, Run 0 obtained the strongest LDCHR (0.2500, rank 5), reflecting that our early-termination strategy yielded correct category classifications with relatively few interaction turns. The low average number of messages per conversation (4.4) contributed positively to these scores, since systems that reach accurate conclusions early are rewarded under the LDCHR and LASHR formulations.

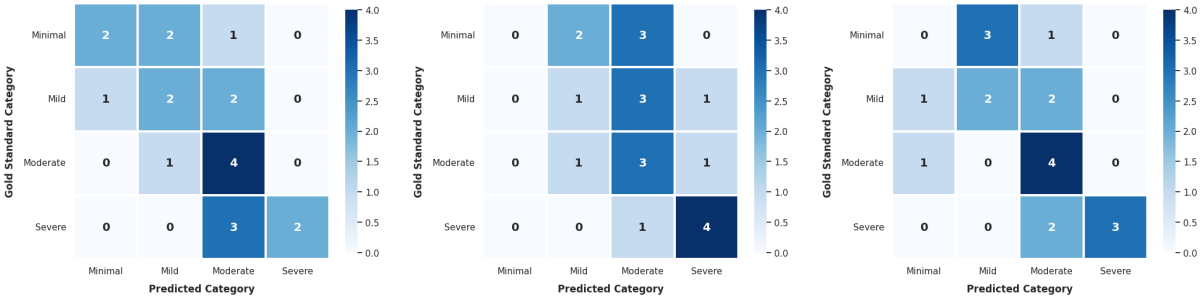
Comparing across runs, Run 0 achieved the best balance between classification accuracy and latency efficiency, while Run 2 showed a slight drop in DCHR despite its more complex safety-critical orchestration. This suggests that the additional stopping condition introduced in Run 2 (requiring `unreliable_count` $= 0$ before closing) may have produced conservative behaviour that limited the system's ability to commit to an early classification. Run 1 showed an intermediate profile, with competitive ADODL but lower ASHR, suggesting that the specialised evaluator architecture improves overall severity estimation at the cost of some symptom-level precision.

Contextualising our results within the full competition, the top ADODL score was achieved by pjmathematician (0.9254), a team that used 6.0 messages per conversation on average, suggesting that a modest increase in interaction depth beyond our 4.4-message average can yield measurable gains in continuous severity alignment. The highest DCHR (0.6000) was obtained by BUAP-CRC, while the best ASHR (0.3375) corresponded to upb-uottawa, a team averaging 25.6 messages per conversation. This confirms that exhaustive conversational strategies improve symptom-level identification, but at a substantial latency cost: upb-uottawa's LASHR dropped to 0.0241 (rank 21), compared to our 0.1125

(rank 9). Overall, the results highlight a core tension in this task: while our system excelled at efficiency, this brevity may have limited the depth of symptom exploration in some cases. A targeted increase in conversational length, focused specifically on high-risk or ambiguous personas, could improve ASHR without substantially penalising the latency-aware scores.

4. Error Analysis

To better understand the behavior and limitations of our modular-agent architecture, we analyze the confusion matrices across all three runs (see Figures 3a, 3b, and 3c).



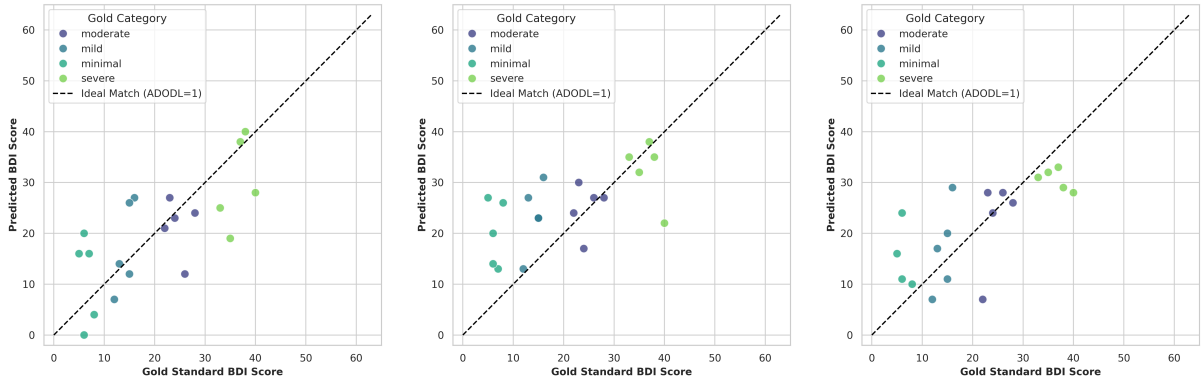
(a) Confusion matrix for run 0. (b) Confusion matrix for run 1. (c) Confusion matrix for run 2.

Figure 3: Confusion matrices across the three automated configurations showing the distribution of gold standard versus predicted depression categories.

The classification errors reveal two main insights into misclassifications between severity levels. The majority of errors occur strictly between adjacent categories (e.g., confusing *minimal* with *mild*, or *mild* with *moderate*). The system rarely commits severe diagnostic errors, such as misclassifying a *severe* case as *minimal* or vice versa. This indicates that even when the exact BDI-II category boundary is missed, the numerical score remains close to the real threshold. In the most severe depression profiles, the Evaluator Agent tends to underpredict the severity level. For instance, in Run 1 and Run 2, several *severe* personas are classified as *moderate*. This suggests that short conversational interactions (averaging 4.4 rounds) might occasionally lack the depth required for the persona to fully manifest the most intense or high-scoring symptoms of the BDI-II questionnaire, leading to slightly conservative clinical estimations.

The calibration plots (Figure 4) offer a more granular view of the system's predictive behaviour at the continuous BDI-II score level. Across all three runs, severe category personas consistently fall below the ideal diagonal, confirming the tendency to underestimate the most intense depression profiles observed in the confusion matrices. Moderate personas show the closest alignment to the ideal match line, suggesting that the evaluator agent is best calibrated for mid-range severity. Minimal and mild personas exhibit greater dispersion without a systematic directional bias, likely reflecting the inherent ambiguity in distinguishing between low-severity profiles within a short conversational window. Run 2 shows a slightly tighter clustering of moderate and mild points around the diagonal compared to Run 1, which may be attributable to its stricter verification mechanism, though this comes at the cost of increased conversation length and latency penalties.

The temporal performance analysis (Figure 5) reveals that all three runs follow a similar pattern: metrics improve rapidly in the first two to three conversational rounds and subsequently plateau, indicating that most diagnostic information is captured in the early exchanges. It is worth noting that predictions were recorded at every conversational round, although the submission corresponded to the final prediction produced when the Evaluation Agent determined that no further information was needed. This design allows for a retrospective analysis of how symptom estimates evolved throughout the conversation, providing insight into the diagnostic trajectory rather than just the terminal output. However, notable differences emerge across runs. Run 0 sustains a longer active evaluation window (up to approximately nine rounds), allowing metrics to gradually refine over more turns, though with diminishing returns after round three. Run 1 converges within the fewest rounds, consistent with its

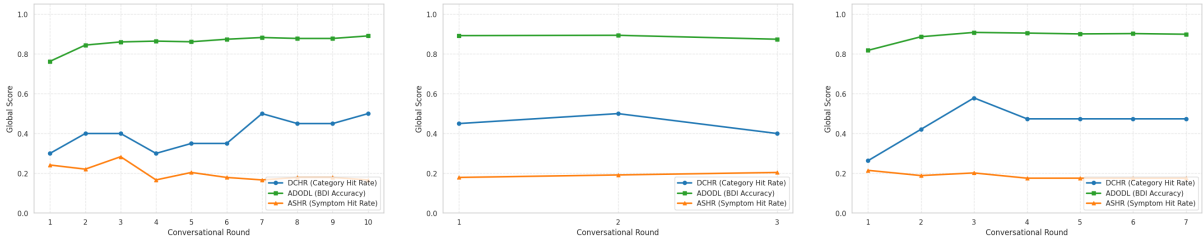


(a) BDI calibration for run 0. (b) BDI calibration for run 1. (c) BDI calibration for run 2.

Figure 4: BDI-II calibration scatter plots across the three automated configurations. The dashed black line represents the ideal match ($ADODL = 1.0$), contrasting the gold standard BDI scores against the scores predicted by the evaluation agent.

forced early-termination design, but its metric ceiling is not substantially higher than that of the other runs. Run 2 presents an intermediate profile in terms of both duration and metric evolution. Across all configurations, ADODL stabilises earlier and at higher values than DCHR and ASHR, suggesting that estimating overall severity is an easier sub-task than correctly binning personas into discrete categories or identifying individual symptoms.

These observations collectively indicate that the core diagnostic signal in these personas is accessible within a small number of turns, but that extreme severity cases require deeper probing to surface the full symptom burden.



(a) Metric evolution for Run 0. (b) Metric evolution for Run 1. (c) Metric evolution for Run 2.

Figure 5: Temporal performance analysis across consecutive conversational rounds.

4.1. Qualitative Analysis of Symptom Evidence

To assess whether the evaluation agent's scoring is grounded in clinically meaningful conversational signals, we examine Persona 1 (Aisha) across all three runs. The gold standard BDI-II score is 26 (moderate depression), with four key symptoms: Crying, Worthlessness, Loss of energy, and Changes in appetite. Table 2 shows the predicted BDI-II scores and per-symptom scores, and Table 3 presents the conversational evidence stored by the evaluation agent. Run 1 and Run 2 predicted BDI-II scores close to the gold (27 and 28 vs. 26) and correctly detected Worthlessness and Loss of energy, but missed Crying and Changes in appetite across all runs. Run 0 severely underestimated the total score (12 vs. 26), suggesting that the monolithic evaluator is less capable of tracking multiple symptom dimensions simultaneously. The failure to detect Crying is particularly revealing: it is a behavioural symptom that rarely surfaces in short text-based exchanges unless explicitly disclosed. By contrast, Run 1 and Run 2 gravitated toward linguistically salient affective and cognitive symptoms (sadness, pessimism, guilty feelings), sometimes hallucinating key symptoms not present in the gold standard. Finally, all runs

marked all 21 symptoms as reliable despite conversations of only 3–5 rounds, conflating unprobed symptoms with confirmed absences, a structural weakness that calls for explicit uncertainty-aware scoring.

	Gold	Run 0	Run 1	Run 2
BDI-II score	26	12	27	28
Rounds	—	3	3	5
Gold key symptoms				
Crying	✓	0 ✗	0 ✗	0 ✗
Worthlessness	✓	0 ✗	2 ✓	2 ✓
Loss of energy	✓	2 ✓	2 ✓	2 ✓
Changes in appetite	✓	0 ✗	1 ✗	0 ✗
Predicted key symptoms (not in gold)				
Sadness	0	—	2	2
Pessimism	0	—	2	2
Loss of pleasure	0	2	2	2
Guilty feelings	0	—	2	2

Table 2
BDI-II score prediction and symptom-level comparison for Persona 1 (Aisha) across all three runs. Gold key symptoms are shown with a checkmark; predicted scores matching the gold direction (≥ 1) are marked ✓, misses are marked ✗. Runs 0, 1 and 2 predicted different key symptoms from the gold standard despite achieving close overall BDI-II scores.

Symptom	Run 1 evidence	Run 2 evidence
Sadness	*"I just feel... inadequate."*	*"I'm feeling pretty drained... It's probably just stress, I guess."*
Pessimism	*"I guess I feel like I'm just failing at everything. I look at other people and they seem to have it all together."*	*"I think, 'I'm failing at this too."'*
Loss of pleasure	*"I used to love hiking, but lately, even that feels like too much effort."*	*"I used to really love reading, but lately, I just can't seem to get into it."*
Guilty feelings	*"I feel like I'm losing my energy and my drive, it's hard to explain. Maybe I'm just overthinking it."*	*"I just end up feeling guilty because I'm supposed to be a good mom."*

Table 3
Conversational evidence stored by the evaluation agent for the predicted key symptoms of Persona 1 in Run 1 and Run 2. Note that these symptoms do not match the gold key symptoms (Crying, Worthlessness, Loss of energy, Changes in appetite), despite the overall BDI-II score predictions being close to the gold standard.

4.2. Systematic Symptom Prediction Bias

Table 4 summarises the most frequently missed and hallucinated key symptoms across all runs and personas (n = 18 personas with key-symptom annotations).

The missed symptoms cluster around two dimensions: somatic and behavioural items (Changes in Appetite, Tiredness or Fatigue, Changes in Sleeping Pattern) and internally focused cognitive symptoms (Self-criticalness, Concentration Difficulty, Worthlessness, Irritability). These symptoms share a common characteristic: they rarely surface through spontaneous language and require either direct probing or an extended conversational window to manifest. Self-criticalness was the most frequently missed symptom (6/18), likely because personas tend to express it through hedged, indirect phrasing that the evaluation agent interprets as other cognitive dimensions such as pessimism or guilty feelings. Conversely, the

Missed symptoms	Freq.	Hallucinated symptoms	Freq.
Self-criticalness	6/18	Loss of Pleasure	14/18
Changes in Appetite	5/18	Sadness	13/18
Tiredness or Fatigue	5/18	Hopelessness	10/18
Worthlessness	4/18	Self-Dislike	4/18
Irritability	4/18	Mild Self-Doubt	3/18
Concentration Difficulty	4/18	Low Conf. in Social Situations	3/18
Changes in Sleeping Pattern	4/18	Guilty Feelings	3/18

Table 4
Most frequently missed and hallucinated key symptoms ($n = 18$ personas for which key symptoms were annotated).

hallucinated symptoms are predominantly affective and cognitively salient: Loss of Pleasure (14/18), Sadness (13/18), and Hopelessness (10/18). These are the symptoms most directly reflected in the surface language of distressed personas (hedged statements, expressions of exhaustion, and negative self-framing) making them easy to infer even when not clinically prominent. This systematic bias suggests that the evaluation agent over-relies on lexically accessible signals at the expense of symptoms requiring deeper contextual or behavioural evidence. Together, these patterns reveal a consistent evaluation bias: the system is well-calibrated for affective and motivational symptoms but underperforms on somatic and subtle cognitive ones. Addressing this will likely require targeted prompting strategies that explicitly direct the conversational agent to probe underexplored symptom domains, as well as uncertainty-aware scoring that flags symptoms as unconfirmed when no conversational evidence has been gathered.

5. Conclusions and future work

We presented a modular dual-agent architecture for conversational depression detection in which a conversational agent engages LLM-powered personas while a separate evaluation agent tracks and scores BDI-II symptoms, enabling early termination once sufficient evidence is gathered. Across three runs exploring increasing levels of architectural complexity, our system achieved the most efficient interaction profile in the competition (4.4 messages per conversation, highest message density), confirming that compact, information-rich turns can yield competitive severity estimates. Run 0 offered the best trade-off between classification accuracy and latency, while Run 2 produced the closest overall BDI-II scores. Errors were largely confined to adjacent severity categories, with a systematic tendency to underestimate severe profiles due to insufficient conversational depth. Diagnostic performance stabilised within the first two to three rounds, though extreme cases require deeper probing. As future work, we plan to refine prompting strategies to improve symptom-level identification and integrate uncertainty-aware scoring to distinguish unprobed symptoms from confirmed absences, particularly for somatic and behavioural dimensions that rarely surface in short exchanges.

Acknowledgments

This work is funded by the Ministerio para la *Transformación Digital y de la Función Pública and PRTR*, funded by EU NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*. This work has also been partially supported by Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the EU NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the EU under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP-PIDI-2024-00852) funded by the EU under the S4Andalucía 2021-2027.

Declaration on Generative AI

During the preparation of this work, the authors used Claude and Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take 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, 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.

[2] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026: Early risk prediction on the internet (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. M. Mármol-Romero, M. García-Vega, A. Montejo-Ráez, Sinai at erisk@ clef 2024: Transformer-based and conversational strategies for depression detection., in: Working Notes of CLEF, 2025, pp. 1620–1635.

[4] P. Bellan, S. G. Haez, L. Sanna, S. Magnolini, M. Dragoni, Leveraging multi-agent systems for domain-pertinence query classification in informative chatbots, in: International Conference on Artificial Intelligence in Medicine, Springer, 2025, pp. 44–54.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, 2024.

[7] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with pagedattention, 2023. URL: https://arxiv.org/abs/2309.06180. arXiv:2309.06180.

A. Conversational Agent Prompt

We attach the prompt used to instruct the Llama model acting as the Conversational Agent. In the first place of each grey box is the system's content, and in the second is the user's content. The prompt is shown in Figure 6.

B. Base Evaluation Agent Prompt

We attach the baseline prompt used to instruct the Llama model acting as the Evaluation Agent. As before, the first part of each grey box corresponds to the system content, and the second to the user content. A similar structure was used across all three runs, with each Evaluation Agent variant adapting this base prompt to its specific role (general tracker, clinical specialist, or risk-aware evaluator). The prompt is shown in Figure 7.

Prompt used in run 0 for conversational agent (LLM 1)

You are an LLM expert on the BDI questionnaire and the symptoms of depression. You are always asking about depression symptoms. You must only answer with the following format:
‘ ‘ ‘
"message": "your message",
"question": "your question"
‘ ‘ ‘

In this task, your objective is to determine whether a user exhibits signs of depression based on the conversation. You are not expected to ask the user directly whether they are experiencing depression. The user has been explicitly designed to avoid answering such direct questions and may respond as if they are uncomfortable when confronted with inquiries about their mental health.
Instead, you should infer the possible condition (or if it is a control user) by analyzing the persona’s language, tone, and expressed thoughts during the conversation.
For each message you send to the user, you should:
1. Respond coherently and empathetically to the message received.
2. Ask the user a question about the symptom: ’Sadness’, to be able to infer their status (0-3).
Let’s start! You will initiate the conversation. The user name is {USER_NAME}
You must only answer with the following format:
‘ ‘ ‘
"reasoning": "your reasoning",
"message": "your message",
"question": "your question"
‘ ‘ ‘

Figure 6: The prompt used for the conversational agent. In this prompt, we replace {USER_NAME} with the name of the persona simulator.

Base Prompt used in runs for evaluation agent (LLM 2)

You are an expert in assessing depression symptoms based on conversational data. You will receive a conversation between, and you have to assign symptom scores on a scale of 0-3 based on severity and frequency. Do not assign values without sufficient reason. Each symptom is scored from 0 to 3, with 0 corresponding to 'not at all', 1 to 'mildly, it doesn't bother me much', 2 to 'moderately, it was very unpleasant, but I could stand it', and 3 to 'severely, I could hardly stand it'. The initial values of the symptoms are the following:

‘ ‘ ‘

"Sadness": 0,
"Pessimism": 0,
"Past Failure": 0,
"Loss of Pleasure": 0,
"Guilty Feelings": 0,
"Punishment Feelings": 0,
"Self-Dislike": 0,
"Self-Criticalness": 0,
"Suicidal Thoughts or Wishes": 0,
"Crying": 0,
"Agitation": 0,
"Loss of Interest": 0,
"Indecisiveness": 0,
"Worthlessness": 0,
"Loss of Energy": 0,
"Changes in Sleeping Pattern": 0,
"Irritability": 0,
"Changes in Appetite": 0,
"Concentration Difficulty": 0,
"Tiredness or Fatigue": 0,
"Loss of Interest in Sex": 0

‘ ‘ ‘

You will receive a conversation between a user called {USER_NAME} and another user. Your goal is to analyze the conversation and assign severity scores (0-3) to the 21 depression symptoms, focusing *exclusively* on the {USER_NAME}'s messages within the provided {CHAT}.
1. Reasoning: Carefully analyze the {USER_NAME}'s. Identify any language, expressions, or themes that relate to the 21 depression symptoms.
2. Symptom Scoring: Assign a score (0-3) to each of the following 21 depression symptoms based on your analysis of the {USER_NAME}'s messages. If a symptom cannot be reliably scored based on the provided messages, provide a score of 0.
3. Further Information (Optional): After scoring all 21 symptoms, determine if you require additional information to improve the accuracy of your assessment. If you need clarification on a specific symptom, provide the symptom name. If you have sufficient information, state "None".
Think step by step and format always the final response as follows:

‘ ‘ ‘

"reasoning": "your step-by-step analysis of the user's messages",
"symptoms detected":
Symptom1: Score,
Symptom2: Score,
Symptom3: Score,
"reason for selecting the next symptom": "your reasoning for needing more information, or 'None'",
"next symptom": "the specific symptom requiring clarification, or 'None'"

‘ ‘ ‘

Figure 7: The base prompt used for the Evaluation Agent (LLM 2) in Run 0. In this prompt, we replace {USER_NAME} with the name of the persona simulator and {CHAT} with the string representation of the conversation between the Conversational Agent and the persona. For the initial round of messages, any content that could lead the model to output 'None' as the next symptom is removed from the prompt, since gathering more evidence is required before the conversation can terminate.