SymptomAI @ eRisk 2026: Context-Aware Conversational Depression Detection*

Notebook for the eRisk Lab at CLEF 2026

Syeda Samah Daniyal[1,*,†], Fakeha Faisal[1,†], Yusra Faisal[1,†], Faisal Alvi[1] and Abdul Samad[1]

[1]Dhanani School of Science and Engineering, Habib University, Karachi, Pakistan

Abstract

The eRisk Lab explores methodologies, evaluation metrics, and real-world applications for early risk detection on the Internet, particularly in health and safety-related domains. In CLEF 2026, eRisk features Task 1, which focuses on conversational depression detection via interactions with LLM personas calibrated according to BDI-II severity levels.

We develop a connected dual-agent system in which a GPT-4 conversational interviewer and a Gemini silent evaluator operate in real time, with evaluation steering conversation toward unresolved symptoms every two turns. The system incorporates a composite confidence scoring mechanism, a multiplicative function of symptom coverage and BDI stability, that drives an adaptive early stopping gate and informs the interviewer's questioning strategy. Additional components include explicit tracking of confirmed-absent symptoms, a clinical priority override for suicidal ideation, and an ensemble final evaluation across three Gemini runs to reduce stochastic variance in the submitted BDI score. These contributions advance symptom-aware, conversation-centric, and temporally grounded early risk detection.

Keywords

CLEF eRisk, Early Risk Detection, Transformer-based Models, Dialogue Context Analysis, Confidence Scoring, Depression

1. Introduction

Early risk detection on the Internet is about spotting signs of mental health issues as they start to appear in online conversations. The eRisk Lab was one of the first to focus on this area. The lab introduced evaluation metrics that don't just measure accuracy, but also consider time and actual world limits. Unlike normal text classification tasks, eRisk is designed to feel more like real life monitoring, where systems must make decisions step by step and within limited time.

The 2026 eRisk edition [1, 2] includes Task 1: Conversational Depression Detection. In this task, systems talk to AI personas (LLMs) that are trained on different user histories. These personas are adjusted based on the Beck Depression Inventory-II (BDI-II). Within a limited number of messages, the system has to decide:

- Whether the persona shows signs of depression.
- Which symptoms are active.
- How severe the depression is overall.

Our contributions emphasise three themes: (i) real-time integration of evaluation feedback into conversational strategy, (ii) principled uncertainty quantification to support adaptive stopping, and (iii) layered system design that combines semantic precision with clinical grounding. The remainder of this paper details our system design, empirical results, and planned extensions.

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

*Corresponding author.

†These authors contributed equally.

samah.daniyal@gmail.com (S. S. Daniyal); fakehaafaisal@gmail.com (F. Faisal); yf08967@st.habib.edu.pk (Y. Faisal); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0004-4765-5271 (S. S. Daniyal); 0009-0000-6147-8567 (F. Faisal); 0009-0009-5181-0349 (Y. Faisal); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

2. Literature Review

2.1. Task 1: Conversational Depression Detection

Detecting depression through conversational agents has changed quickly with increased use of Large Language Models. Recent work in the eRisk 2025 Lab represented a big step forward from looking at single posts to modeling whole conversations [3]. This shift highlights how evidence builds up over time by framing early depression detection as a series of decisions, where each new interaction can change how previous messages are understood. This requires creating clinically grounded simulated patient personas and developing diagnostic agents capable of reliably assessing them within limited conversation windows.

Wang et al. [4] introduced TalkDep, a clinician-in-the-loop framework that used GPT-4o. It generated 12 unique personas ranging from minimal to severe depression, each coming from the Beck Depression Inventory-II (BDI-II). Backstories, adverse life events, and specific linguistic patterns were used in personas to add clinical coherence. A third-party LLM judge, specifically DeepSeek R1-14 B, was used for evaluation and it correctly identified the more severe persona in 86.36% of pairwise comparisons, and certified psychologists rated persona humanness and symptom realism at an average of 3.92 out of 5.

Two main architectures were highlighted for conducting diagnostic interviews. Single-agent systems used a heavily prompted LLM to converse and simultaneously track symptoms. Vachharajani [5] employed Claude 3.7 Sonnet with a system prompt embedding the full BDI-II, instructing the agent to update severity scores across all 21 symptoms each time. Varela et al. [6] used standalone GPT-4 and Falcon to conduct interviews and output symptom assessments in JSON. On the other hand, Mármol-Romero et al. [7] made a dual-LLM architecture with Llama-3.1-8B-Instruct, which used a Conversational Agent and an Evaluation Agent that tracked symptoms and planned follow-up questions. It was found that single-agent approaches struggled to balance coherent dialogue with strict symptom tracking.

Elicitation strategies were different for each of them. Vachharajani [5] used movie discussions as a proxy for emotional states, allowing the agent to assess sadness or anhedonia without clinical interrogation. Mármol-Romero et al. [7] tested empathy with self-disclosure, empathy alone, and direct questioning, finding that empathy without personal anecdotes got the most accurate severity estimates. Varela et al. [6] found that grouping related symptoms, such as energy, sleep, fatigue, into single natural-sounding prompts outperformed sequential single-symptom questioning in both efficiency and diagnostic accuracy.

Dual-agent system Mármol-Romero et al. [7] achieved an Average Difference of Overall Depression Levels (ADODL) of 0.93 and a Depression Category Hit Rate (DCHR) of 0.66, suggesting that separating conversational generation from clinical evaluation is better. Varela et al. [6] observed that proprietary models like GPT-4 outperformed open-source alternatives such as Falcon, which suffered from repetition and over-attribution of sadness. It was also important to note that properly prompted LLMs matched the diagnostic accuracy of a non-expert human using the BDI-II framework, suggesting that LLM-based screening tools can also be used for evaluation.

3. Approach

3.1. Task 1: Conversational Depression Detection

3.1.1. Dataset

The dataset is provided by the IRLab research group via the Hugging Face collection `irlab-udc/erisk2026` [8]. It consists of 20 simulated patient personas, constructed as LoRA adapters fine-tuned on Meta-Llama-3-8B-Instruct, each representing a distinct BDI-II severity level. Personas are released at a rate of two per week during the evaluation period. Each completed conversation must

yield three outputs: presence of depression, a set of active depressive symptoms, and the overall severity level.

Figure 1 illustrates the BDI-II clinical severity scale, persona distribution across severity bands, and the conversation output format.

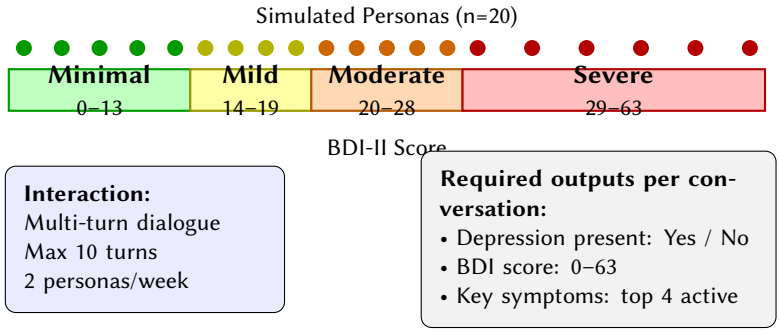

Figure 1: Task 1 dataset overview. The BDI-II severity scale is divided into four clinical bands. Twenty personas are distributed across bands. Each conversation must produce a binary depression judgement, a BDI score estimate, and four key symptoms.

3.1.2. System Design

All Gemini calls in the system were made through the `gemini-flash-latest` alias endpoint; during the experiment period the model was Gemini 3 Flash (`gemini-3-flash`).

We developed three progressively more sophisticated architectures, each addressing a specific limitation of its predecessor.

System 1: Fixed Questions, Separate Evaluator. The interviewing agent posed the same eight predefined questions to every persona regardless of their responses. Each question was designed to probe multiple BDI-II symptoms simultaneously, following the symptom-grouping approach of Varela et al. [6]. A separate Gemini evaluator analysed the completed transcript post-hoc to predict the BDI score and four key symptoms. The two agents operated independently with no information exchange during the conversation.

System 2: Dynamic Questions, Separate Evaluator. The fixed question set was replaced with GPT-4 generating each question dynamically, using the full conversation history and a running list of uncovered BDI-II symptoms as context. This allowed the interviewer to adapt to what the persona had already disclosed, producing more naturalistic, symptom-targeted dialogue. The evaluation agent remained post-hoc.

System 3: Connected Dual-Agent System with Confidence-Driven Early Stopping. This system implements the dual-agent architecture of Mármol-Romero et al. [7], extended with confidence scoring, adaptive early stopping, and ensemble evaluation. The system architecture is illustrated in Figure 2.

Final System Enhancements. The System 3 architecture was further refined with several key improvements made in System 4 and System 5:

Confidence Scoring. After each Gemini evaluation, a composite confidence score $C \in [0, 1]$ is computed as:

$$C = \underbrace{\frac{N_{\text{total}} - N_{\text{unclear}}}{N_{\text{total}}}}_{\text{coverage}} \times \underbrace{\max\left(0,\ 1 - \frac{|\Delta\text{BDI}|}{63}\right)}_{\text{stability}} \tag{1}$$

where $N_{\text{total}} = 21$ is the number of BDI-II symptoms, N_{unclear} is the count of symptoms Gemini cannot yet classify, and $|\Delta\text{BDI}|$ is the absolute change in interim BDI score between consecutive evaluations. The multiplicative structure ensures that a high coverage with an unstable BDI estimate, or a stable

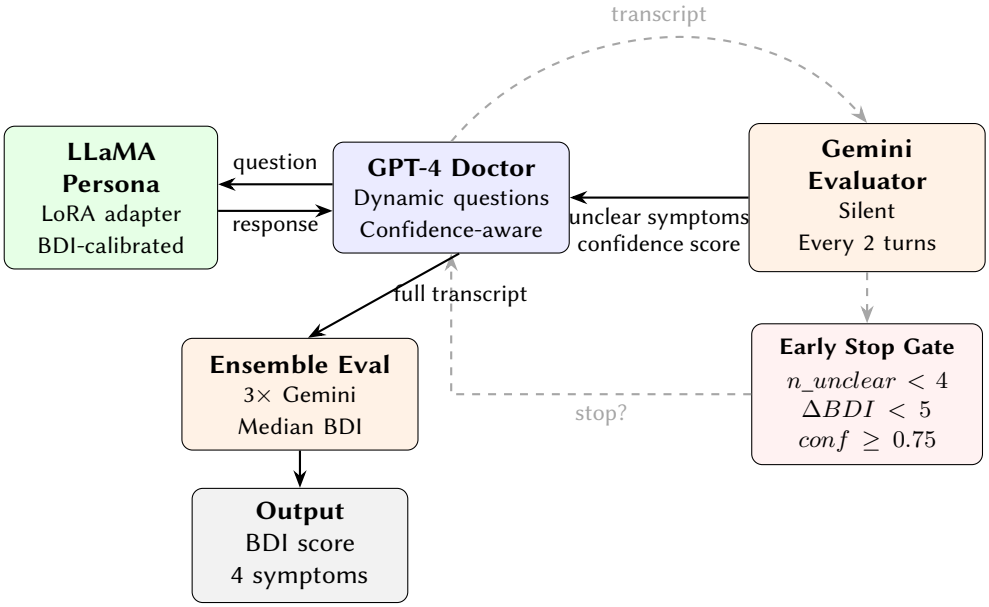

Figure 2: System 3 dual-agent architecture. GPT-4 interviews the LLaMA persona using BDI-II clinical definitions to generate targeted questions. Gemini evaluates silently every two turns using full 0–3 scale criteria, returning a confidence score and a list of unclear symptoms to GPT-4. The early stopping gate halts conversation when three composite conditions are simultaneously satisfied after a minimum of six turns. Final output is produced by an ensemble of three Gemini evaluations.

BDI with many unresolved symptoms, both produce a low composite score. The first evaluation always yields $C = 0$ since no prior BDI reference exists, preventing premature stopping from a single data point.

A critical implementation note: the original Gemini evaluation prompt instructed the model to report the four most important unclear symptoms. This artificially fixed $N_{\text{unclear}} = 4$ regardless of actual conversational coverage, permanently capping coverage at $17/21 \approx 0.81$. The prompt was corrected to report all genuinely unclear symptoms without a count ceiling.

Absent Symptom Tracking. Earlier versions discarded Gemini's ABSENT classifications, treating confirmed-absent symptoms identically to unprobed ones. The updated system explicitly captures a SYMPTOMS_ABSENT list. Coverage in Equation 1 is computed over all resolved symptoms (observed OR absent), giving a more accurate reflection of how thoroughly the BDI-II domain has been covered.

Clinical Definitions in the Silent Evaluator. The symptom descriptions and 0–3 severity criteria used in both the silent evaluator and doctor agent prompts were taken directly from the official BDI-II manual [9]. The original Gemini interim evaluation prompt listed only BDI-II symptom names without clinical definitions. Without the scoring criteria, Gemini was effectively classifying symptoms by surface-level name matching rather than clinical meaning, a passing mention of fatigue, for example, could incorrectly trigger *Loss of Energy* even though the BDI-II definition requires energy too low to perform even basic tasks. The full 0–3 scale descriptions for all 21 BDI-II symptoms were added to the silent evaluator prompt, giving Gemini the same clinical reference used by trained assessors. This makes interim symptom classifications more precise and reduces incorrect coverage and severity signals being passed to the GPT-4 Doctor.

Clinical Descriptions in the Doctor Agent. Previously, GPT-4 received only a list of symptom names to probe, e.g., *Worthlessness, Pessimism, Concentration Difficulty*, with no definition of what each symptom means clinically. GPT-4 had to infer probing strategy from the name alone, which frequently produced vague questions that failed to surface the intended symptom. Brief clinical descriptions of

the 15 most conversationally accessible BDI-II symptoms were added to GPT-4's system instructions. Knowing, for instance, that *Loss of Pleasure* refers specifically to activities the patient previously enjoyed allows GPT-4 to ask naturally about hobbies or social activities rather than a generic question about mood. This aligns both agents on the same clinical framework, improving coherence across the full pipeline from conversation steering to final assessment.

Early Stopping. The system halts before reaching the maximum turn budget if all three conditions hold simultaneously and turn $\geq$ MIN_TURNS:

$$N_{\text{unclear}} < 4 \quad \text{AND} \quad |\Delta\text{BDI}| < 5 \quad \text{AND} \quad C \geq 0.75 \tag{2}$$

MIN_TURNS is set to 6. Mild and moderate depression can present with subtle, understated signals that only emerge across several exchanges; stopping earlier risks missing key symptoms and producing an inaccurate BDI score. Six turns provides coverage across the main BDI-II symptom groups while still permitting early stopping between turns 6 and 10 when confidence and coverage indicate a clear diagnosis. This turn penalty is an acceptable tradeoff for improved diagnostic accuracy. An AND gate is used rather than an OR gate because clinical reliability requires both sufficient symptom coverage and a stable severity estimate. Each failed condition is logged individually to support diagnostic analysis.

Suicidal Ideation Priority Flag. Suicidal Thoughts is the highest-stakes item on the BDI-II. If it remains unclear after MIN_TURNS, the Gemini feedback includes a priority override that suspends the confidence-based directive and instructs GPT-4 to probe this domain through indirect, empathetic language before addressing any other symptom. This ensures the most clinically consequential item is not crowded out by less critical unresolved symptoms.

Ensemble Final Evaluation. The final Gemini evaluation is run three times on the completed transcript. The submitted BDI score is the median across runs. The four key symptoms are those appearing in at least two of the three evaluations (majority vote), ranked by frequency of appearance then individual score. This reduces the impact of stochastic variation in a single Gemini completion, which can shift the BDI estimate by several points across identical inputs.

3.1.3. GPT-4 Confidence-Aware Questioning

The confidence score is passed to GPT-4 in natural language as a directive within the clinical feedback block. Three regimes govern question strategy:

- $C \geq 0.75$: *High confidence* — GPT-4 is instructed to begin wrapping up, asking one final confirmatory question.
- $0.50 \leq C < 0.75$: *Moderate confidence* — GPT-4 continues targeted probing of unclear symptoms.
- $C < 0.50$: *Low confidence* — GPT-4 is directed to prioritise the unclear symptom list with focused indirect questions.

4. Results and Discussion

4.1. Task 1: Conversational Depression Detection

Validation findings. Table 1 reports System 5 system validation results against three TalkDep-derived personas using BDI-II ground truth scores. Severity category was correctly predicted in all three cases with a Mean Average Error (MAE) of 5.3 BDI points. The ensemble mechanism performed as intended: for Maria (true BDI = 40), three Gemini runs returned [32, 35, 35], producing a median of 35 and absorbing the outlier at 32. For Laura (true BDI = 23), the prediction of 28 overshoots by 5 points, but in the clinically safer direction. A false positive toward Severe is less harmful than a missed detection. Maya (true BDI = 6) returned [0, 0, 2] with median 0, correctly rejecting any depression

classification despite a non-zero ground truth, within acceptable clinical noise at the minimal end of the scale.

A critical caveat: these results validate the Gemini evaluator component only, tested on static TalkDep personas rather than the live LoRA personas from the eRisk 2026 dataset. Full end-to-end evaluation against the competition personas is pending completion of the weekly release schedule.

Table 1
System 5 Gemini evaluator validation results. Pred BDI is the median of three ensemble runs. Severity category is correct in all three cases.

Persona	True BDI	Pred BDI	All Runs	MAE	True Category	Pred Category
Maria	40	35	[32, 35, 35]	5	Severe	Severe ✓
Laura	23	28	[26, 28, 31]	5	Moderate	Moderate ✓
Maya	6	0	[0, 0, 2]	6	Minimal	Minimal ✓
Mean MAE				5.3	3/3 correct	

Error direction analysis. Figure 3 visualises the BDI prediction error for each persona relative to clinical severity boundaries. Laura's over-prediction is conservative toward the Severe boundary; Maya's under-prediction approaches zero but remains within the Minimal band. Neither error crosses a severity boundary.

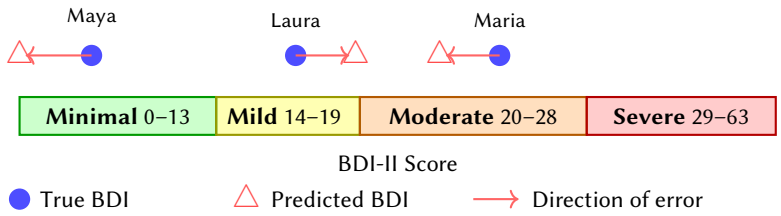

Figure 3: BDI prediction error relative to clinical severity bands. All errors remain within the correct severity category. Laura's over-prediction is conservative toward Severe; Maya's under-prediction remains within Minimal.

Confidence trajectory findings. Two live persona runs produced confidence and BDI trajectory data that directly stress-tested the early stopping logic. Persona 15 (moderate depression, final BDI = 20) stabilised quickly: BDI delta between turns 4 and 6 was only 1 point, and confidence reached 0.797 at turn 6, above the 0.75 threshold. The gate correctly permitted early stopping at turn 6, saving four turns from the maximum budget. Persona 14 (severe depression, final BDI = 31) showed BDI deltas of 6, 10, and 5 across consecutive evaluations, keeping confidence at 0.745 at turn 8. The gate correctly blocked stopping throughout. This distinction, mild and quickly characterised versus severe and evolving, is precisely what the composite gate was designed to handle. The AND logic prevents premature termination when either coverage or stability is insufficient, not only when both are poor. Trajectories are shown in Figures 4 and 5.

Threshold sensitivity. Figure 6 shows that for Persona 15, early stopping at turn 6 is achievable across any threshold below 0.797. Beyond that value confidence never meets the bar and the conversation runs to full length. For Persona 14, no threshold setting produces early stopping because BDI instability blocks the stability condition regardless of the confidence bar. The gate's discriminative behaviour is therefore driven by clinical signal quality, not threshold calibration.

What the system progression reveals. Table 2 summarises the failure mode each system transition exposes. The five-system progression is not iterative engineering alone as each transition surfaces a

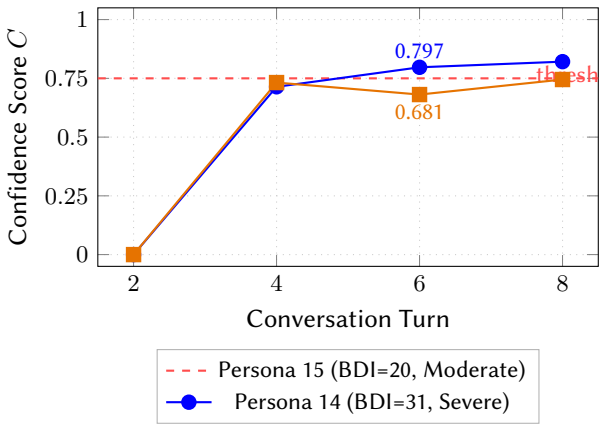

Figure 4: Confidence score trajectory over conversation turns. Turn 2 always yields $C = 0$ (no prior BDI reference). Persona 15 crosses the stopping threshold at turn 6; Persona 14 remains below it at turns 6 and 8 due to BDI instability.

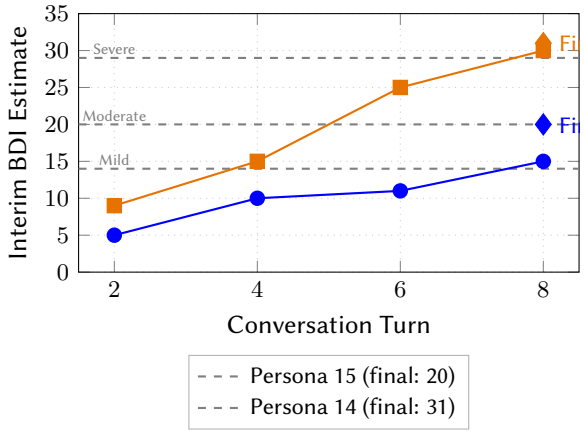

Figure 5: Interim BDI estimation over conversation turns. Dashed lines mark clinical severity boundaries. Persona 15 stabilises in the mild–moderate range by turn 6; Persona 14 escalates through moderate to severe with large inter-turn deltas that correctly block early stopping.

specific diagnostic about what LLM-based clinical interviewing requires. The move from System 1 to System 2 showed that fixed questions leave systematic gaps: a persona may disclose information that makes a pre-planned question redundant or misdirected. The move from System 2 to System 3 showed that post-hoc evaluation misses the dynamic value of symptom feedback. Systems 4 and 5 revealed two failure modes that only surface under real conversational load: Gemini classifying symptoms by surface name rather than clinical criteria (producing spurious coverage signals), and GPT-4 generating vague questions when given symptom names without clinical definitions. Both are failures of shared grounding between agents, not failures of architecture.

Preliminary competition results. Table 3 reports SymptomAI's official preliminary scores from the eRisk 2026 Task 1 evaluation. All three runs appear in the incomplete-submission table. Late onboarding due to course scheduling constraints limited total coverage to 12 of 20 personas. eRisk's three submission slots map to distinct methodological strategies: Run 1 corresponds to our fixed-question baseline, Run 2 to our dynamic symptom-targeted questioning approach, and Run 3 to the connected dual-agent system with progressively added components across the eight-persona evaluation window.

The primary constraint on interpreting these results is sample size. Runs 1 and 2 each cover two personas, making any ranking position statistically fragile; a single well-estimated case is sufficient to produce a high rank at this scale. Run 3, evaluated over eight personas, is the more meaningful basis for assessment, but carries its own problem: those eight personas were processed as the system

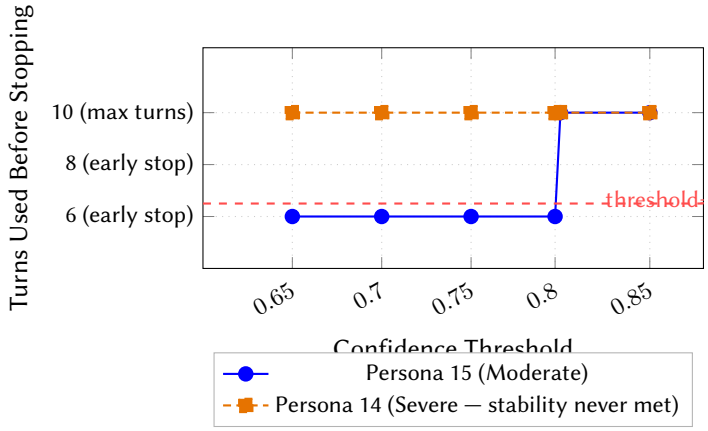

Figure 6: Confidence threshold sensitivity. Persona 15 stops at turn 6 for any threshold ≤ 0.797. Persona 14 always reaches the maximum turn budget because BDI instability blocks the stability condition.

Table 2
Failure mode exposed by each system transition in Task 1.

Transition	Change Introduced	Failure Mode Revealed
System 1 → System 2	Fixed → dynamic questions	Static questions cannot adapt when personas volunteer information early or redirect the dialogue
System 2 → System 3	Post-hoc → real-time evaluation	Post-hoc evaluation discards mid-conversation symptom signals that should redirect subsequent questions
System 3 → System 4	Added confidence scoring and early stopping	No quantified criterion for stopping; conversations ran to full length regardless of diagnostic certainty
System 4 → System 5	Clinical definitions added to both agents	Gemini classified symptoms by name matching, not clinical criteria; GPT-4 generated vague questions from symptom names alone

Table 3
SymptomAI preliminary results (incomplete-submission table). Ranks for each metric are shown in parentheses and are relative to all incomplete runs.

Run	Personas	DCHR	ADODL	ASHR	LDCHR	LASHR
Run 1 (baseline)	2/20	0.0000 (18)	0.8810 (6)	0.1250 (12)	0.0000 (25)	0.0625 (18)
Run 2 (dynamic)	2/20	0.5000 (5)	0.9683 (1)	0.2500 (3)	0.2500 (7)	0.1250 (6)
Run 3 (multi-agent)	8/20	0.3750 (8)	0.8294 (22)	0.2500 (3)	0.1875 (13)	0.1250 (6)

evolved across our internal Systems 3 through 5, meaning the submitted results reflect a non-uniform system rather than a single consistent method. This limits what can be concluded about any individual component's contribution.

Within these constraints, two observations hold. First, the step from Run 1 to Run 2, fixed to dynamic questioning, produced a clear improvement across every metric: ADODL rose from 0.8810 to 0.9683, DCHR from 0.0000 to 0.5000, and ASHR from 0.1250 to 0.2500. Run 1's DCHR of 0.0000 alongside a reasonable ADODL indicates the system estimated severity numerically with some accuracy but placed it in the wrong BDI-II category, a boundary sensitivity failure consistent with static questioning that cannot adapt when a persona discloses symptoms out of the expected order. The dynamic approach

in Run 2 directly addresses this. Second, ASHR rank 3 is consistent across Runs 2 and 3 despite the coverage and system differences, suggesting symptom identification is the most robust component of the pipeline. For reference, the best ASHR across all complete twenty-persona submissions was 0.3375, placing our raw score of 0.2500 within a comparable range, though full-coverage evaluation is required before drawing any firm conclusion.

Latency scores were weak across all runs, with Run 1 recording LDCHR = 0 (rank 25). The adaptive stopping mechanism was introduced progressively within Run 3's evaluation window and had not been calibrated at the point of the earlier submissions. Latency performance therefore reflects system maturity rather than an inherent architectural limitation, and remains the clearest target for improvement in any full-coverage submission.

5. Conclusion

This paper presented SymptomAI's approach to eRisk 2026 Task 1. We developed five progressively more sophisticated systems. System 1 and 2 established static and adaptive baselines with post-hoc evaluation. System 3 connected conversation and evaluation in real time, following Mármol-Romero et al. [7]. System 4 extended this with a composite confidence score driving early stopping and modulating GPT-4's questioning posture. System 5 addressed the clinical grounding gap: BDI-II definitions were embedded in both agents, absent symptom tracking corrected coverage calculation, a priority override ensured suicidal ideation is probed, and ensemble final evaluation reduced stochastic variance. Validation results on three personas show a mean MAE of 5.3 BDI points with correct severity category prediction in all cases.

Quantitative evaluation of Task 1 against BDI-II ground truth across all 20 personas remains the primary next step. The confidence trajectory data already collected provides a basis for calibrating the stopping thresholds, in particular, determining whether the current composite conditions produce better BDI accuracy and symptom recall than the full-turn baseline, and whether the three-run ensemble materially reduces submission variance across personas.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checking.

References

[1] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, 21–24 September, 2026, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2026.

[2] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of erisk 2026 early risk prediction on the internet: Symptom ranking and conversational approaches for depression and adhd, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.

[3] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk at CLEF 2025: Early risk prediction on the internet (extended overview), in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_117.pdf.

[4] 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 (CIKM '25), ACM, 2025. URL: https://arxiv.org/abs/2508.04248.

[5] P. Vachharajani, Transformer ensembles and LLM-powered approaches for depression symptom analysis and contextualized early risk detection, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_132.pdf.

[6] A. Varela, M. Oronoz, A. Casillas, A. Perez, Detection of depression with symptom similarity: Data reduction and LLM personas, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_133.pdf.

[7] A. M. Mármol-Romero, M. García-Vega, M. García-Cumbreras, A. Montejo-Ráez, SINAI at eRisk@CLEF 2025: Transformer-based and conversational strategies for depression detection, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4038/paper_130.pdf.

[8] IRLab-UDC, erisk 2026: Early risk prediction on the internet, https://huggingface.co/collections/irlab-udc/erisk2026, 2026. Accessed: 2026-04-17.

[9] A. T. Beck, R. A. Steer, G. K. Brown, BDI-II: Beck Depression Inventory Manual, 2nd ed., The Psychological Corporation, San Antonio, TX, 1996.