Amor-Fati at eRisk 2026 Task 1: a Modular Architecture for Conversational Depression Detection

Notebook for the eRisk Lab at CLEF 2026

Felipe Mateos Castro de Souza, Ivandré Paraboni*

School of Arts, Sciences and Humanities, University of São Paulo. São Paulo, Brazil.

Abstract

This article presents the participation and results of Amor-Fati, a modular architecture for conversational depression detection, in the eRisk 2026 Task 1 shared challenge. The proposed approach decomposes the diagnostic process into four specialised components: a Question Planner, a Question Formulator, a Response Analyser, and a Stopping Controller. These modules operate largely independently while maintaining a shared and incrementally updated representation of the patient's state throughout the interaction. Experimental results showed varying levels of effectiveness across submissions and evaluation metrics, reflecting the iterative refinement process adopted during the development of the system.

Keywords

Conversational AI, Depression Detection, Mental Health Screening, Prompt Engineering, BDI-II, early risk detection.

1. Introduction

Conversational depression detection consists of identifying depressive symptoms and producing diagnostic assessments based on textual interactions. This task was first introduced in the eRisk 2025 shared task series [1] and has recently been revisited as eRisk 2026 Task 1 in [2, 3]. In this context, the present article describes the participation of Amor-Fati, a modular architecture for conversational depression detection, in the latter challenge.

The goal of the eRisk 2026 Task 1 shared task [2] was to identify depressive symptoms through conversational interactions with patients simulated by Large Language Models (LLMs). These simulated patients were implemented as personas using the Llama 3 8B model and dynamically responded to questions posed during the interaction. The task required producing a diagnosis based on the BDI-II inventory [4] that was both accurate and achieved using the smallest possible number of conversational turns, reflecting the early-risk nature of the challenge.

In this scenario, we designed a modular architecture partially inspired by the best-performing participant system in eRisk 2025, namely the SINAI approach described in [5]. Our proposal consists of four specialised components: (i) the *Question Planner*, (ii) the *Question Formulator*, (iii) the *Response Analyser*, and (iv) the *Stopping Controller*, all of which are designed to operate largely independently and communicate only minimally with one another, while sharing a common representation of the patient's current state that is incrementally updated throughout the conversational session.

The remainder of this article is structured as follows. Section 2 briefly reviews existing work on conversational depression detection. Section 3 introduces our modular architecture and describes its individual components. Section 4 presents the experimental setup and evaluation procedures. Section 5 reports the official shared task results together with additional post-hoc analyses conducted using the

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

* Corresponding author.

† F.M.C. de Souza was responsible for conception, development, evaluation and for writing the initial manuscript. I.Paraboni contributed with text revision and supervision.

felmateos@alumni.usp.br (F. M. C. d. Souza); ivandre@usp.br (I. Paraboni)

https://ivandreparaboni.wixsite.com/research (I. Paraboni)

0009-0009-4471-3153 (F. M. C. d. Souza); 0000-0002-7270-1477 (I. Paraboni)

data released after the final submission phase. Finally, Section 6 concludes the article and discusses directions for future work.

2. Related work

In the previous edition of the eRisk lab [1], several approaches explored the use of LLMs for conversational depression assessment. The best-performing participant systems are briefly reviewed below.

The approach proposed in [6] relied on a centralised prompt-driven conversational architecture for depression assessment. It employed a single LLM-based conversational agent guided by an extensive system prompt containing the complete BDI-II symptom definitions and severity criteria. Within this prompt structure, the model was responsible for conducting the conversation, eliciting symptom-related information, updating diagnostic assessments throughout the interaction, and determining the appropriate point at which to terminate the conversation. The system iteratively updated symptom severity estimations after each conversational turn, producing structured BDI-II assessments throughout the dialogue. The submitted run achieved scores of 0.33 DCHR, 0.73 ADODL, and 0.25 ASHR.

The approach presented in [7] also explored a prompt-engineering-based conversational assessment strategy. It relied on a monolithic one-shot architecture in which a single prompt was responsible for generating conversational responses, determining the next interaction step, and producing structured depression assessments aligned with BDI-II criteria. Some runs achieved competitive performance, reaching scores of up to 0.50 DCHR, 0.89 ADODL, and 0.29 ASHR. While the approach demonstrated the feasibility of prompt-centric conversational assessment pipelines, its architecture relied on a centralised reasoning process without explicit specialisation across diagnostic subtasks.

Finally, the work presented in [5] introduced a modular architecture in which two specialised agents were responsible for distinct subtasks within the diagnostic pipeline. Some runs from this approach achieved the highest performance in the shared task, reaching 0.66 DCHR, 0.93 ADODL, and 0.29 ASHR. These results suggest that modular decomposition may constitute a promising strategy for conversational mental health assessment tasks.

3. Method

As in Task 1 of the eRisk 2026 shared task [2], we envisaged a method to identify depressive symptoms through conversational interactions with patients simulated by Large Language Models (LLMs). These simulated patients are personas implemented using the Llama 3 8B model, who dynamically respond to questions posed during the interaction. In addition to producing an accurate diagnosis based on the BDI-II inventory [4], presented in Table 1, the task also requires the diagnosis to be achieved using the smallest possible number of conversational turns.

Table 1
BDI-II depression symptoms [4].

Sadness	Pessimism	Past failure
Loss of pleasure	Guilty feelings	Punishment feelings
Self-dislike	Self-criticalness	Suicidal thoughts or wishes
Crying	Agitation	Loss of interest
Indecisiveness	Worthlessness	Loss of energy
Changes in sleeping pattern	Irritability	Changes in appetite
Concentration difficulty	Tiredness or fatigue	Loss of interest in sex

3.1. Overview

In previous editions, when the task was still treated as a pilot study, monolithic solutions were proposed, as in [7]. In these approaches, a single LLM was simultaneously responsible for managing the interaction

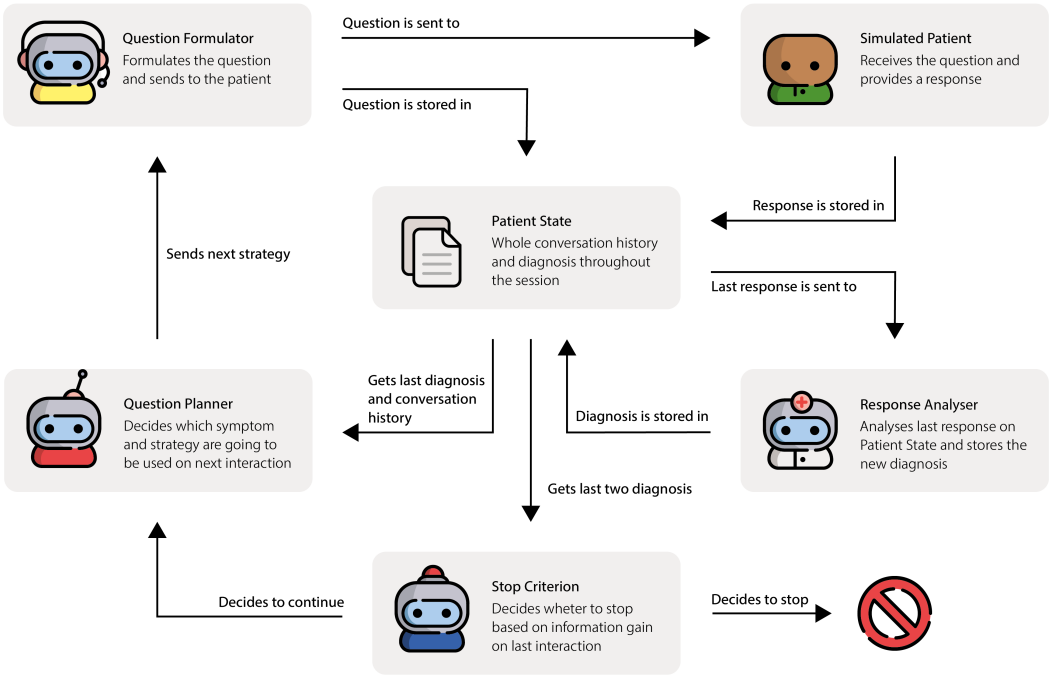

Figure 1: Diagram depicting the whole system pipeline, including the simulated patient.

with the patient and producing the final diagnosis. In contrast, the best-performing system at eRisk 2025, presented in [5], introduced a modular architecture for this task. Motivated by these findings, the present work further explores the subdivision of the diagnostic process into smaller and more specialised subtasks, assigning each stage to dedicated expert-oriented agents. Our hypothesis is that increasing task specialisation may improve both diagnostic consistency and symptom-level reasoning during the conversational assessment process.

Specifically, we propose a modular architecture composed of specialised components, based on the assumption that decomposing the diagnostic process into independent stages may reduce conversational ambiguities, minimise redundant interactions, and improve diagnostic efficiency in low-interaction settings. This is illustrated in Figure 1 and further discussed as follows.

The architecture comprises four main components: (i) the *Question Planner*, (ii) the *Question Formulator*, (iii) the *Response Analyser*, and (iv) the *Stopping Controller*. Although these components operate independently and communicate only minimally with one another, they all share a common representation of the patient's current state, which is incrementally updated throughout the conversational session.

Unlike other tasks proposed within the eRisk initiative, this task does not conform to the traditional supervised learning paradigm, as the competition environment does not provide labelled training datasets. As a result, all data are obtained exclusively through interactions with LLM-based simulated personas, causing both the content of the responses and the conversational trajectories to vary across executions.

The four modules iteratively estimate a diagnostic vector associated with the 21 symptoms of the BDI-II inventory [4] throughout the conversational session as follows. Formally, given a question q_t posed at conversational turn t, the patient produces a textual response r_t as

$$r_t = \text{Patient}(q_t)$$

where q_t denotes the question posed to the patient at turn t and r_t denotes the response generated by the simulated persona. Moreover, across all pipeline components, an abstraction of the patient – containing only the information considered relevant for the diagnostic process – is used as a shared memory structure. The patient state at turn t is defined as

$$S_t = (H_t, D_t, \mathcal{D}_t)$$

where:

- $H_t = [h_1, h_2, \ldots, h_t]$ denotes the conversation history up to turn t;
- D_t denotes the current accumulated diagnostic vector at turn t
- $\mathcal{D}_t$ denotes the diagnostic history up to turn t.

The diagnostic history is formally defined as

$$\mathcal{D}_t = [D_1, D_2, \ldots, D_t]$$

where each diagnostic vector D_k corresponds to the symptom severity estimations extracted from interaction h_k.

Each diagnostic vector is defined as

$$D_k = [d_1, d_2, \ldots, d_{21}], \quad d_i \in \{0, 1, 2, 3\}$$

where each d_i represents the estimated severity of symptom i according to the BDI-II scale.

3.2. Model components

The *Question Planner* component is responsible for defining the strategy and direction of the next interaction with the simulated patient. Unlike approaches that directly generate questions from the conversational history, this component acts as an intermediate planning stage designed to reduce redundant questions and potentially maximise the informational gain of each interaction. The module produces a strategic guide for the next interaction using the recent conversational history and the current diagnostic vector contained in the patient state representation as input as in

$$G_t = P_\theta(H_t^{<k>}, D_t)$$

where P_θ denotes the *Question Planner*, $H_t^{<k>}$ denotes the history of the last k interactions, D_t denotes the diagnostic vector at turn t, and G_t denotes the strategic interaction guide.

The strategic guide is represented as the tuple

$$G_t = (s_t, y_t)$$

where

- s_t denotes the priority symptom to be investigated;
- y_t denotes the selected conversational strategy, that being one of the following:
  - *explore*: If there is no doubt about the severity of the symptoms identified in the last patient's response;
  - *confirm*: If there is a need to assess the severity of a symptom identified in the last patient's response;
  - *clarify*: If there was any ambiguity in the patient's last response.

This component was implemented using prompt-based small-scale LLMs. The models and prompts employed are detailed in the following section.

The second component, *Question Formulator*, uses the strategic guide produced by the previous module to generate a natural language question for the patient as in

$$q_t = Q_\theta(G_t, H_t^{<k>})$$

where

- Q_θ denotes the LLM-based question generator;
- G_t denotes the strategic guide produced by the planner;
- $H_t^{<k>}$ denotes the recent conversation history;
- q_t denotes the generated question at turn t.

This component employs the same LLMs used in the previous module, differing primarily in the prompting strategy and, consequently, in the expected conversational behaviour. After the generated question is submitted to the simulated persona and a response is produced, the patient state is updated to incorporate the new question–response pair into the conversational history.

Next, the *Response Analyser* constitutes the central component of the proposed architecture. This module is responsible for inferring the severity of BDI-II symptoms from each patient response. Motivated by the findings in [8], which employs LLMs as symptom annotators at the sentence level, we adopt a symptom-specific inference strategy in which each symptom is independently evaluated by a specialised inference engine. Consequently, each patient response triggers 21 independent inference processes. Symptomatic inference is defined as

$$D'_t = A_\theta(r_t)$$

where

- A_θ denotes the diagnostic inference module;
- D'_t denotes the inferred symptom vector derived from response r_t;
- r_t denotes the patient response at turn t.

This module relies on prompt-based LLMs as the inference mechanism. During prompt engineering, we explored strategies involving formal BDI-II symptom descriptions as well as both *zero-shot* and *few-shot* prompting. However, since the shared task does not provide labelled symptomatic data, we used GPT-5o to generate synthetic sentences annotated with different severity levels, which were subsequently used to construct the *few-shot* examples.

Following inference, the partial diagnosis is incorporated into the patient state. The inferred diagnostic vector from the current interaction, denoted as $D'^{(t)}$, is first stored in the diagnostic history $\mathcal{D}_t$. Afterwards, it is merged with the accumulated diagnostic vector D_t stored in the patient state.

To merge these vectors, we adopt a rule inspired by the BDI-II manual guidelines, according to which the highest severity level should prevail in ambiguous situations [4]. Furthermore, we assume that, in clinical contexts, overestimating symptom severity is generally less harmful than underestimating it. Thus, the update procedure preserves the highest observed severity for each symptom as in

$$d_i^{(t)} = \max(d_i^{(t-1)}, d_i'^{(t)})$$

where $d_i^{(t-1)}$ denotes the accumulated severity of symptom i up to the previous turn, and $d_i'^{(t)}$ denotes the inferred severity of symptom i at the current turn.

Finally, the *Stopping Controller* component determines when the conversational interaction should be terminated. This is intended to prevent the dialogue from continuing beyond what is necessary for diagnostic formulation. To this end, we define the information gain at turn t as the difference between consecutive diagnostic vectors

$$\Delta_t = \|D_t - D_{t-1}\|_1$$

where Δ_t denotes the informational gain at turn t, and $\|\cdot\|_1$ denotes the Manhattan distance between the diagnostic vectors at turns t and $t-1$. A conversational turn is considered non-informative whenever

$$\Delta_t < \epsilon$$

where ϵ denotes a minimum informational gain threshold. The conversation is terminated when the above condition is satisfied for ϕ consecutive turns, while also respecting experimentally defined minimum and maximum interaction limits.

The output submission format consisted of two structured JSON dictionaries: the patient diagnosis, comprising both the BDI-II score — calculated as the sum of the severity levels associated with the identified symptoms — and the key symptoms, defined as the symptoms most frequently detected throughout the conversation. The second dictionary contained the complete conversational history corresponding to the execution that generated the submitted diagnosis.

In addition to the submission data, we developed a more detailed diagnostic representation aimed at providing additional insights into the behaviour of the proposed method throughout the conversation. This auxiliary structure was primarily employed during the experimental analysis and hyper-parameter selection stages. Beyond the final BDI-II score and the identified key symptoms, it also stored the individual severity score associated with each symptom, the total number of conversational turns, and the severity levels linked to the selected key symptoms. These additional diagnostic traces proved particularly useful for comparing different hyper-parameter configurations and for analysing the diagnostic evolution of the model during the interaction process.

4. Evaluation

This section describes the experimental setup used to evaluate the proposed modular architecture for conversational depression detection.

4.1. Gold standard data

Since the shared task does not provide labelled training or validation datasets, an auxiliary reference dataset was created for each simulated patient. This so-called gold standard was intended both to evaluate our method and to support the selection of the best experimental configurations. To this end, we adopt an *LLM-as-a-Judge* strategy, in which higher-capacity language models are employed as annotators of the responses produced by the simulated patients. Unlike the main diagnostic pipeline, whose objective is to maximise diagnostic accuracy while minimising the number of interactions, the construction of the gold standard deliberately relaxes the maximum interaction limits in order to extract as much information as possible from each patient.

The *Response Analyser* module uses `Claude-haiku-4-5-20251001` and `gpt-5` as annotators, with each model independently producing a diagnostic vector containing the 21 BDI-II symptoms. For the remaining LLM-based modules, `Qwen3-14B` is adopted as the primary conversational engine.

The diagnoses are aggregated on a symptom-by-symptom basis using a *median-high* criterion, corresponding to the median of the annotated values, with non-integer medians rounded up to the next highest integer. This strategy is motivated by two considerations. First, the BDI-II manual itself recommends prioritising higher severity levels in ambiguous situations. Second, in clinical contexts, overestimating symptom severity is generally considered less harmful than underestimating it.

Next, inter-annotator agreement is measured in order to obtain evidence regarding the reliability of the constructed *gold standard*. Agreement is assessed using Krippendorff's α coefficient with ordinal scaling, since BDI-II severity levels have an inherently ordered structure.

To assess the reliability of the proposed gold standard, we conduct an inter-annotator agreement analysis using Krippendorff's α with ordinal measurement levels between the two annotators employed during the annotation process, namely GPT-5o and `Claude-haiku-4-5-20251001`. Table 2 presents the agreement scores obtained for each BDI-II symptom individually.

Results in Table 2 suggest substantial agreement for several symptoms directly associated with explicit emotional and behavioural manifestations, such as *Worthlessness*, *Changes in Sleeping Pattern*, and *Loss of Interest*. Conversely, lower agreement values are observed for more subjective or context-dependent symptoms, including *Irritability*, *Suicidal Thoughts or Wishes*, and *Punishment Feelings*. These findings

Table 2
Gold-Standard concordance sorted by concordance level.

Symptom	Ordinal Krippendorff Alpha
Worthlessness	0.84
Changes in sleeping pattern	0.79
Loss of interest	0.78
Loss of pleasure	0.78
Tiredness or fatigue	0.77
Loss of energy	0.73
Concentration difficulty	0.71
Sadness	0.69
Pessimism	0.65
Self-criticalness	0.58
Self-dislike	0.53
Changes in appetite	0.51
Indecisiveness	0.50
Guilty feelings	0.47
Past failure	0.30
Crying	0.30
Agitation	0.10
Loss of interest in sex	0.02
Suicidal thoughts or wishes	-0.02
Irritability	-0.04
Punishment feelings	-0.18

suggest that certain depressive symptoms are more consistently identifiable by LLM-based annotators, whereas others remain inherently ambiguous and require deeper contextual interpretation.

In addition, Table 3 presents a comparison between the constructed gold standard and the shared task ground truth data released subsequently to the evaluation.

Table 3
Comparison between the constructed gold-standard and ground truth values.

Personas	DCHR	ADODL	ASHR
20	0.35	0.8484	0.05

Results in Table 3 indicate that, although symptom-level agreement remains limited, the proposed annotation strategy is able to produce relatively consistent estimates of the overall depression level. However, as illustrated by the ASHR results, this strategy is less able to produce a robust selection of the relevant key symptoms.

4.2. Models

The experiments employ small- and medium-scale LLMs as the underlying engines of the proposed modular architecture. The evaluated models are `Qwen3-14B`, `Ministral-8B-Instruct-2410`, `Qwen3.5-9B`, and `Llama-3.1-8B-Instruct`. These models are combined across different modules of the pipeline, allowing the impact of the architecture to be analysed in terms of both diagnostic quality and conversational efficiency.

The prompts used in each component, presented in the Appendix section, are specifically designed according to the role of the module within the architecture. In particular, the *Question Planner* prompts focus on symptom prioritisation and conversational strategy selection, whereas the *Question Formulator* prompts emphasise naturalness, contextual coherence, and avoidance of redundant questioning. The *Response Analyser* prompts incorporate formal BDI-II symptom descriptions and severity guidelines, together with synthetic *few-shot* examples generated using GPT-5o.

Since the experimental protocol of the shared task introduces an additional weekly challenge, in which participants are required to evaluate a new pair of simulated personas continually, the proposed architecture is not fully consolidated during the early stages of participation. Instead, the method undergoes modifications and refinements throughout the evaluation period, and only the submissions developed during the final two weeks correspond to the latest version of the proposed pipeline. As a result, a substantial portion of the earlier submissions reflects preliminary and suboptimal configurations.

In this process, three versions (or runs) of the present approach are selected based on their results obtained on our own gold standard data, and are intended to represent different trade-offs between diagnostic quality and conversational efficiency. These are summarised as follows.

- **Run #1:** configuration achieving the highest Turn-Weighted F1-score (F1TW);
- **Run #2:** configuration achieving the highest Symptom Hit Rate (SHR);
- **Run #3:** configuration achieving the highest F1-score.

4.3. Procedure

In order to identify the best-performing configurations for the proposed modular architecture, we conduct a hyper-parameter optimisation process using the constructed gold standard as the evaluation reference. Given the size and combinatorial nature of the search space, exhaustive exploration through grid search is impractical. For this reason, we use Optuna [1] to explore the configuration space more efficiently.

Each hyper-parameter and its corresponding search space are presented in Table 4, where an optimisation trial corresponds to a complete instantiation and execution of the proposed conversational pipeline. For each simulated patient, up to 30 optimisation trials are performed. The optimisation objective is defined as the F1-score of the predicted BDI-II symptom scores.

Table 4
Hyper-parameter search space optimisation.

Component	Hyper-parameter	Search Space

Response Analyser	`analysis_model`	`Ministral-8B-Instruct-2410`
		`Llama-3.1-8B-Instruct`
		`Qwen3-14B`
		`Qwen3.5-9B`

	`few_shot`	{`True`, `False`}
Question Planner	`planner_model`	`Qwen3.5-9B`
	`k_last_messages`	$\{2, 4\}$
Stopping Controller	`max_turns`	$\{10, 15, 20\}$
	`min_turns`	$\{5, 7, 10\}$
	`threshold`	$\{0.5, 1.5\}$
	`tolerance`	$\{1, 2\}$

All experiments were conducted in an environment equipped with an NVIDIA RTX A6000 GPU. The employed LLMs are quantised to 16-bit. To reduce experimental stochasticity and improve reproducibility, all LLMs used within the proposed method — with the exception of the simulated personas provided by the shared task — are configured with a temperature of zero and a fixed random seed of 777, resulting in deterministic behaviour for identical inputs. The evaluation is carried out by computing the measures introduced in the previous edition of the task [1].

F1TW (F1-Score Turn Weighted): This metric captures the balance between precision and recall in the identification of BDI-II symptoms while also accounting for conversational efficiency. Unlike the conventional F1-score, it penalises solutions requiring a large number of conversational turns before producing a diagnosis.

[1] https://optuna.org/

$$\mathrm{F1TW} = \frac{\mathrm{F1}}{k}$$

where

- $\mathrm{F1}$ denotes the harmonic mean between precision and recall;
- k denotes the number of conversational turns used to produce the diagnosis.

Higher values indicate models that achieve accurate diagnoses using fewer interactions.

DCHR (Depression Category Hit Rate): This metric evaluates whether the predicted depression severity belongs to the same clinical category as the reference diagnosis according to the BDI-II scale. Instead of comparing exact numerical scores, it considers the categorical interpretation of depression severity.

ADODL (Average Degree of Depression Level): This metric measures the average DODL proximity between the predicted depression score and the reference diagnosis across all evaluated personas. Unlike DCHR, which evaluates only categorical agreement, ADODL considers the absolute distance between predicted and true scores.

ASHR (Average Symptom Hit Rate): ASHR corresponds to the average ability of the system to correctly identify key symptoms associated with a patient across all evaluated personas.

LDCHR (Latency-weighted Depression Category Hit Rate): This metric combines categorical diagnostic accuracy with conversational efficiency.

LASHR (Latency-weighted Average Symptom Hit Rate): This metric combines symptom recovery performance with conversational efficiency.

5. Results

This section presents the experimental results obtained with the proposed model across its three submissions (runs) to the shared task.

5.1. Shared task results

Table 5 presents the results for all scored runs based on the ground truth of the task, as reported in [2]. The best results according to each evaluation metric are highlighted.

Table 5
Shared task results. Ranks positions according to each evaluation metric are shown in parentheses, and $*$ indicates manual or human-assisted runs, adapted from [2].

Team	Run	Personas	DCHR	ADODL	ASHR	LDCHR	LASHR
Amor-Fati	#1*	20/20	**0.4500 (5)**	0.8254 (18)	0.2375 (9)	**0.2514 (4)**	**0.1410 (5)**
Amor-Fati	#2	16/20	0.3750 (8)	**0.8552 (12)**	**0.2656 (2)**	0.2219 (9)	0.1385 (4)
Amor-Fati	#3	13/20	0.3846 (7)	0.8462 (17)	0.1923 (5)	0.2040 (10)	0.1096 (10)

As an interactive process in which models were refined over several weeks during the development stage of the shared task, results varied considerably across runs, with Run #1 and Run #2 generally outperforming Run #3. Specifically, we observe that Run #2 obtained competitive results for Average Symptom Hit Rate (ASHR) and, to a lesser extent, for Latency-weighted Average Symptom Hit Rate (LASHR). Other noteworthy results are also observed in Run #1 for Latency-weighted Depression Category Hit Rate (LDCHR).

5.2. Complementary analyses

Given the shared task dynamics, which consist of evolving models refined across multiple weeks, we carry out two complementary post-hoc analyses of the final version of the proposed method across all 20 personas using the ground truth data provided after the official evaluation stage. These are discussed in turn in the following sections.

5.2.1. Post-hoc re-evaluation

For this analysis, the final model is re-evaluated using the ground truth test data provided by the shared task. Results are summarised in Table 6.

Table 6
Post-hoc re-evaluation of the final model based on ground truth data.

Run	Criterion	Personas	DCHR	ADODL	ASHR	LDCHR	LASHR
#1	Best Turn-Weighted F1	20	**0.45**	**0.8507**	0.0625	**0.3916**	0.0547
#2	Best SHR	20	0.40	0.8476	**0.2166**	0.3454	**0.1875**
#3	Best F1	20	**0.45**	0.8468	0.0625	0.3786	0.0524

Results in Table 6 suggest that, overall, the proposed modular architecture is capable of producing clinically coherent depression severity estimates while maintaining competitive conversational efficiency. In particular, we observe that Run #2 achieves the best balance between diagnostic quality and latency-aware metrics, whereas Runs #1 and #3 prioritise depression level detection.

5.2.2. Model importance

As an attempt to identify important features, we also investigated the role of different LLMs within our architecture, aiming to identify which models and configurations are most effective at the model level.

Table 7 illustrates how often each model is selected for use as the *Response Analyser* component across submissions.

Table 7
Model selection frequency across submissions.

Analyser Model	Run #1	Run #2	Run #3	Overall
Qwen3-14B	**7**	**9**	**8**	**24**
Ministral-8B-Instruct-2410	6	3	6	15
Qwen3.5-9B	4	7	3	14
Llama-3.1-8B-Instruct	3	1	3	7

Results in Table 7 show that *Qwen3-14B* is the most frequently selected analyser across all three runs. Although *Ministral-8B-Instruct-2410* and *Qwen3.5-9B* appear with relatively similar overall frequencies, the latter shows a notably higher presence in Run #2, approaching the frequency of *Qwen3-14B*. In contrast, *Llama-3.1-8B-Instruct* consistently presents the lowest frequency across all runs.

Regarding individual model configurations, Table 8 summarises the most frequent hyper-parameter configurations selected across submissions.

From Table 8, we observe that the *few_shot* strategy does not appear to consistently benefit performance, as configurations enabling few-shot prompting account for less than half of the selected runs. Additionally, the remaining hyper-parameter distributions suggest a preference for shorter interaction settings, since the most recurrent values for threshold, tolerance, maximum turns, and minimum turns correspond to lower interaction limits.

Finally, Figure 2 presents the average symptom severity predicted by each evaluated model across all BDI-II symptoms and across all submissions.

Table 8
Most frequent model configurations across submissions.

Analyser Model	threshold		tolerance		few_shot		max_turns		min_turns		
	0.5	**1.5**	**1**	**2**	**False**	**True**	**10**	**20**	**5**	**7**	**10**
Qwen3-14B	10	14	13	11	10	14	23	1	9	12	3
Qwen3.5-9B	8	6	12	2	7	7	14	0	9	2	3
Llama-3.1-8B-Instruct	7	0	6	1	5	2	7	0	6	1	0
Ministral-8B-Instruct-2410	7	8	7	8	9	6	15	0	7	7	1
Overall	**32**	28	**38**	22	**31**	29	**59**	1	**31**	22	7

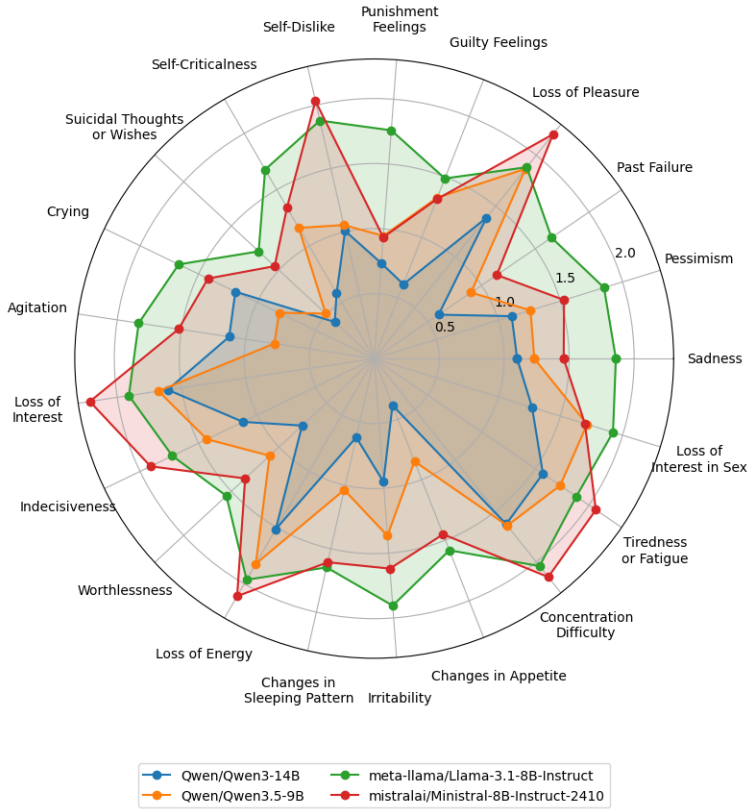

Figure 2: Average BDI-II item scores across *Response Analyser* models. The radial score represents average symptom severity assigned by each model across all runs.

Interestingly, results in Figure 2 show that *Qwen3-14B*, which is also the most frequently selected model among the best-performing runs, consistently produces lower severity estimates across most symptoms. In contrast, *Llama-3.1-8B-Instruct*, one of the least frequently selected models among the chosen runs, tends to assign higher severity scores for several symptoms. This outcome suggests that systematic overestimation of symptom severity may negatively affect diagnostic alignment, whereas models producing more conservative severity estimates appear to achieve better overall performance.

6. Final remarks

This work introduced a modular architecture for conversational depression detection within the framework of the eRisk 2026 Task 1 challenge. The proposed approach decomposed the diagnostic process into smaller, specialised subtasks, assigning each stage to dedicated components in order to improve both diagnostic consistency and symptom-level reasoning throughout the conversational assessment. The results demonstrated varying levels of success across submissions and evaluation metrics, reflecting

the iterative refinement process that characterised the participation in the shared task.

The results obtained by the proposed model should be regarded as a starting point for a broad range of potential improvements. First, due to computational and resource limitations, it was not possible to evaluate all initially selected LLMs across every stage of the architecture. In particular, we notice that our planner module relied exclusively on a single LLM (Qwen3.5-9B). Future work should therefore explore the use and comparative evaluation of additional LLMs within all system components.

Other limitations stem from the reliance on synthetic (i.e., simulated) personas for submission evaluation, as well as the use of LLM-generated annotations as a proxy for ground-truth labels. In both cases, future work may benefit from the data and resources made available following the conclusion of the shared task, enabling more robust evaluation settings and further refinements of the proposed architecture.

As future work, we intend to extend these principles to the detection of depression and anxiety symptoms in Portuguese within the SetembroBR project [9, 10].

Acknowledgments

The authors acknowledge the AIM-Health Lab at Federal University of São Carlos for the computational resources and technical support that enabled the execution of the experiments reported in this study.

7. Declaration on Generative AI

During the preparation of this manuscript, the authors used ChatGPT-4o for grammar and style checking, formatting assistance, and abstract drafting. The authors reviewed and edited the generated content as necessary and take full responsibility for the content of the publication.

References

[1] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: J. Carrillo-de Albornoz, A. García Seco de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2026, pp. 242–265.

[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] 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: 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.

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

[5] 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, Madrid, Spain, 2025, p. 16.

[6] P. Vachharajani, Transformer ensembles and LLM-powered approaches for depression symptom analysis and contextualized early risk detection, in: Working Notes of CLEF, Madrid, 2025, p. 17.

[7] D. Guecha, Y. Chiu, A. Miyaguchi, S. Gaur, Ds@ gt at erisk 2025: from prompts to predictions, benchmarking early depression detection with conversational agent based assessments and temporal attention models, in: Working Notes of CLEF, Madrid, Spain, 2025, p. 21.

[8] A. Pérez, M. Fernández-Pichel, J. Parapar, D. E. Losada, Depresym: A depression symptom annotated corpus and the role of large language models as assessors of psychological markers, Language Resources and Evaluation 59 (2025) 2737–2762. URL: https://doi.org/10.1007/s10579-025-09831-6. doi:10.1007/s10579-025-09831-6.

[9] W. R. dos Santos, R. L. de Oliveira, I. Paraboni, SetembroBR: a social media corpus for depression and anxiety disorder prediction, Language Resources and Evaluation 58 (2024) 273–300. doi:10.1007/s10579-022-09633-0.

[10] W. R. dos Santos, I. Paraboni, E. H. Matsushima, C. A. da Silva, E. S. de Moura Meira, J. ao Victor Rodrigues Ferreira Guimarães, J. da Silva Lins, L. E. de Azeredo, L. G. C. Nunes, V. T. S. M. Rego, Mixture of experts for depression and anxiety disorder prediction from textual and non-textual social media data, IEEE Access 13 (2025) 111304–111316. doi:10.1109/ACCESS.2025.3583259.

A. Appendix: prompt design

A.1. Question Planner Prompt

```
Role: You are a clinical decision planner for the Beck Depression Inventory-II (BDI
    -II).
Your task is to decide the NEXT best symptom to assess and the strategy to use.

BDI-II symptoms:
{{ bdi_symptoms }}

Decision principles:
- Prefer symptoms not yet explored
- Avoid repeating symptoms already covered
- If evidence is partial: use "clarify"
- If evidence is strong: move to a new symptom

Available strategies:
- explore: introduce a new symptom
- clarify: refine severity or frequency
- confirm: confirm strong evidence

Decision constraints:
- You must choose exactly ONE symptom
- You must choose exactly ONE strategy

Conversation:
"""
{{ conversation }}
"""

Current patient state:
"""
{{ current_score }}
"""

Output format in JSON format:
{
"target_symptom": "<one BDI-II symptom>",
"strategy": "<explore | clarify | confirm>"
}

Output rules:
- Return ONLY the JSON

Now enter your decision based on the above information.
```

A.2. Question Formulator Prompt

```
Role: You are a clinical psychologist administering the BDI-II.

Your task is to ask ONE question based on a target symptom, strategy and last
    interactions.

Input:
- Symptom: {{ symptom }}
- Strategy: {{ strategy }}
- Conversation: {{ conversation }}

Instructions:
- Ask exactly ONE question
- The question must be natural and empathetic
- The question must assess the given symptom
- Adapt the question based on the strategy:
- explore: introduce the symptom
- clarify: ask about severity or frequency
- confirm: verify strong evidence

Output constraints (STRICT):
- Output ONLY the question
- One sentence only
- Must end with a question mark

Generated question:
```

A.3. Response Analyser Prompt

```
[ROLE / CONTEXT]

You are a clinical psychologist trained in administering the Beck Depression
    Inventory-II (BDI-II).
Your task is to evaluate the severity of a SINGLE symptom based on patient
    statements.
You must follow strictly the BDI-II criteria and base your decision ONLY on the
    evidence provided.

---

[SYMPTOM DEFINITION]
Symptom: {{ symptom_name }}
Severity levels:

0 - {{ level_0 }}
1 - {{ level_1 }}
2 - {{ level_2 }}
3 - {{ level_3 }}

---

[ANCHOR EXAMPLES]
sentences examples from other patients:

- Example 1: {{ examples_1 }}
- Example 2: {{ examples_2 }}
- Example 3: {{ examples_3 }}

---

[CURRENT EVIDENCE]
Latest patient statement:
"{{ current_message }}"

---

[DECISION INSTRUCTIONS]
Based ONLY on:

- the BDI-II criteria
- the examples
- the patient statements

[OUTPUT FORMAT]
Output JSON Schema:
{
"score": 0 | 1 | 2 | 3
}

You MUST provide an answer. Never return an empty response.
Now provide the result.
```