Transformer Ensembles and LLM-Powered Approaches for Depression Symptom Analysis and Contextualized Early Risk Detection

Notebook for the eRisk Lab at CLEF 2025

Poojan Vachharajani[1]

[1]*Netaji Subhas University of Technology, New Delhi, India*

Abstract

This paper details the PJs-team's participation in the eRisk 2025 lab, addressing three distinct tasks related to early risk prediction on the internet. For Task 1, "Search for Symptoms of Depression," we developed five systems. These included fine-tuned bi-encoder models (e.g., 'alldistilroberta-v1', 'e5-small') using CoSENTLoss and data augmentation, and an ensemble of these bi-encoders using a Reciprocal Rank Fusion (RRF)-style sum of ranks. We also employed fine-tuned cross-encoder models ('ModernBERT-large', 'ModernBERT-base') with BinaryCrossEntropyLoss for reranking, and created reranker ensembles using both majority voting and scaled mean averaging. Our data preparation involved a stratified split of the training data, and we utilized textual descriptions of BDI-II symptoms for query formulation. For Task 2, "Contextualized Early Detection of Depression" our approach integrated a fine-tuned sentence transformer ('basic-adrb') for initial filtering of user writings based on relevance to specific BDI symptoms (Pessimism, Punishment Feelings, Self-Dislike, and Indecisiveness). Writings surpassing an empirically determined threshold, or from previously flagged users, were then analyzed by an ensemble of four large language models (Claude 3.7 Sonnet, Amazon Nova Pro, Llama3.3-70B, and Claude 3.5 Haiku), which processed the full conversational context sequentially. The final decision was based on a majority vote among the LLMs, with state management ensuring that positive detections were persistent. For Task 3, the "Pilot Task: Conversational Depression Detection via LLMs," we designed a conversational AI agent. This agent, guided by a comprehensive system prompt incorporating the BDI-II questionnaire, aimed to diagnose depression symptoms by engaging LLM-simulated personas in discussions about movies, iteratively updating BDI scores based on the conversation flow and content. This paper outlines the methodologies, data preparation strategies, model architectures, loss functions, ensemble techniques, and system prompt engineering employed for each task, and presents the official results.

Keywords

Early Risk Prediction, Depression Detection, Symptoms of Depression, Contextualized Analysis, Conversational AI, Natural Language Processing, Transformer Models, Large Language Models, Ensemble Methods, eRisk, Mental Health, Sentence Embeddings, Reranking, Prompt Engineering

1. Introduction

The early prediction of risks from online user-generated content, particularly concerning mental health, remains a critical and challenging research area. The eRisk lab at CLEF has consistently provided a platform for advancing methodologies in this domain [1, 2]. eRisk 2025 presents three tasks designed to explore different facets of early risk detection: searching for specific depression symptoms, detecting depression in a contextualized, sequential manner, and a novel pilot task involving conversational AI for depression diagnosis.

Our team, PJs-team, participated in all three tasks. Our overarching strategy involved leveraging the strengths of various transformer architectures, from efficient bi-encoders for broad candidate retrieval and scoring to powerful cross-encoders for precise reranking. For tasks requiring deeper contextual understanding or nuanced interaction, we employed an ensemble of state-of-the-art Large Language Models (LLMs). This paper provides a comprehensive description of our approaches, including data

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*This paper describes the participation of the PJs-team in the eRisk 2025 lab, detailing our methodologies for Task 1, Task 2, and Task 3.

pjmathematician@gmail.com (P. Vachharajani)

preparation, model fine-tuning strategies, specific loss functions, ensemble techniques, and prompt engineering. We aim to contribute to the understanding of how these diverse NLP tools can be effectively applied to the complex challenges posed by eRisk.

2. General Methodology and Data Preparation

2.1. Data Splitting and Preprocessing

For tasks involving supervised fine-tuning (primarily Task 1), we partitioned the provided training data into training and validation sets. This split was performed using a stratified approach based on several key columns: ‘con_res‘ (consensus relevance judgment), ‘BDI_Q‘ (BDI question number), ‘maj_res‘ (majority relevance judgment), and ‘year‘ (data collection year). This stratification aimed to create validation sets that were representative of the overall data distribution and would provide reliable estimates of model performance. Standard text preprocessing steps, such as lowercasing, removal of special characters, and normalization of whitespace, were applied where appropriate, though many transformer models handle raw text effectively.

2.2. BDI Symptom Descriptions

A core component across tasks was the Beck Depression Inventory (BDI-II) [3]. We utilized a dictionary mapping each of the 21 BDI question numbers to its detailed textual description. These descriptions served multiple purposes:

- **Task 1**: As queries for the sentence retrieval and ranking systems.
- **Task 2**: As reference texts for the sentence transformer to score the relevance of user writings to specific depressive themes for pre-filtering.
- **Task 3**: Embedded within the system prompt for the conversational AI to guide its diagnostic reasoning.

The mapping ensured a consistent understanding and representation of the target symptoms across our different systems. The full list of BDI descriptions used is provided in Appendix A.

3. Task 1: Search for Symptoms of Depression

Task 1 required us to rank sentences from user writings based on their relevance to each of the 21 BDI symptoms. We developed five systems, broadly categorized into bi-encoder-based approaches and cross-encoder-based reranking approaches, with ensembles for both.

3.1. Bi-Encoder Architectures and Fine-tuning (Systems 1 & 2)

Bi-encoder models compute dense vector representations (embeddings) for queries (symptom descriptions) and sentences independently. Relevance is then typically determined by the cosine similarity between these embeddings.

3.1.1. System 1: Fine-tuned Bi-encoder (DistilRoBERTa)

- **Base Model**: We selected ‘sentence-transformers/all-distilroberta-v1‘ [https://huggingface.co/sentence-transformers/all-distilroberta-v1] as our base bi-encoder model due to its balance of performance and efficiency.
- **Fine-tuning Strategy**: The model was fine-tuned on the eRisk training data. We framed the task as learning to predict the similarity between a BDI symptom description and a user’s sentence.

  - **Training Parameters**: Fine-tuning was conducted for 5 epochs, with a batch size of 16, a learning rate of 2e-5, and a warmup ratio of 0.1.
  - **Loss Function**: We employed CoSENTLoss [4], which is designed to optimize cosine similarity scores directly. For pairs of (symptom description, user sentence) (q_i, d_j) with a target similarity label $y_{ij} \in [0, 1]$, the loss aims to align the model's predicted cosine similarity, $\cos(E(q_i), E(d_j))$, with y_{ij}. A common formulation encourages a higher similarity for more related pairs.

- **Two-Step Detection/Ranking**: For inference, this system implemented a two-stage process. Initially, an aggregated relevance score was computed for each user based on all their writings against a given symptom. Users exceeding a certain threshold were then selected. Subsequently, individual sentences from these selected users were scored and ranked by the fine-tuned Distil-RoBERTa model.

3.1.2. System 2: Bi-encoder Ensemble (RRF-style Sum)

This system aimed to improve robustness by ensembling multiple fine-tuned bi-encoders.

- **Component Models**:
  (a) ‘sentence-transformers/all-distilroberta-v1‘
  (b) ‘BAAI/bge-small-en-v1.5‘ [5] (fine-tuned using a contrastive loss)
  (c) ‘intfloat/e5-small-v2‘ [6]
  (d) ‘ibm-granite/granite-embedding-278m-multilingual‘
  (e) ‘sentence-transformers/all-MiniLM-L6-v2‘

  Each model underwent a similar fine-tuning process as System 1, adapting CoSENTLoss or an appropriate contrastive loss.

- **Ensemble Strategy**: We used a Reciprocal Rank Fusion (RRF)-style approach to combine the rankings from the five bi-encoders. For each sentence d_j and each symptom query q_i, the RRF score was calculated as:

$$\text{RRFScore}(d_j, q_i) = \sum_{m=1}^{M} \frac{1}{k + \text{rank}_m(d_j, q_i)} \tag{1}$$

  where M is the number of models in the ensemble (5 in this case), $\text{rank}_m(d_j, q_i)$ is the rank of sentence d_j for query q_i produced by model m, and k is a constant, typically set to 60, to mitigate the impact of very low ranks. Sentences were then re-ranked based on their aggregate RRF scores.

3.2. Cross-Encoder Architectures and Fine-tuning (Systems 3, 4, & 5)

Cross-encoder models process a query and a sentence simultaneously, allowing for richer interaction modeling and generally leading to higher accuracy, albeit at a higher computational cost. They are typically used for reranking top candidates from a bi-encoder stage.

3.2.1. System 3: Fine-tuned Reranker (ModernBERT-Large)

- **Base Model**: We chose ‘answerdotai/ModernBERT-large‘ [7] as our primary cross-encoder model.

- **Fine-tuning Strategy**:
  - **Training Data and Labels**: We constructed training pairs of (symptom description, user sentence). The target labels were derived from the provided ‘con_res‘ and ‘maj_res‘ annotations:

    - Label 1.0: if ‘con_res‘ = 1 (unanimous agreement on relevance).
    - Label 0.5: if ‘maj_res‘ = 1 AND ‘con_res‘ = 0 (majority agreement, but not unanimous).
    - Label 0.0: otherwise.

    This soft-labeling approach allows the model to learn degrees of relevance.
  - **Training Parameters**: The model was fine-tuned for 7 epochs. We used a training batch size of 16 and an evaluation batch size of 32. The learning rate was 2e-5. Tokenizer ‘model_max_length‘ was set to 512.
  - **Loss Function**: BinaryCrossEntropyLoss was used, suitable for the [0,1] (or [0, 0.5, 1]) target scores.
- **Reranking Process**: This model was used to rerank the top-N (e.g., N=100 or N=200) sentences retrieved by System 2 for each symptom.

3.2.2. System 4: Reranker Ensemble (Voting)

This system combines the outputs of multiple fine-tuned rerankers.

- **Component Models**:
  (a) ‘answerdotai/ModernBERT-base‘
  (b) ‘answerdotai/ModernBERT-large‘
  (c) ‘sentence-transformers/all-distilroberta-v1‘ (fine-tuned as a cross-encoder, similar to the ModernBERT models)

  Each model was fine-tuned using the same strategy described for System 3.
- **Ensemble Strategy**: For a list of top-N candidates (e.g., from System 2), each of the three cross-encoders provided a relevance score. A sentence was considered relevant if at least two out of the three models assigned it a score above a predefined threshold (e.g., 0.5, corresponding to a positive classification). The final ranking was determined first by the number of positive votes, and then by the average score from the voting models for tie-breaking.

3.2.3. System 5: Reranker Ensemble (Scaled Mean)

This system also used the same three cross-encoders as System 4 but employed a different score aggregation method.

- **Component Models**: Same as System 4.
- **Ensemble Strategy**: The raw output scores (logits or sigmoid probabilities) from each of the three rerankers were first normalized to a common [0,1] range using min-max scaling, applied per symptom query to account for varying score distributions. The final relevance score for each candidate sentence was then calculated as the arithmetic mean of these scaled scores from the three models. Sentences were subsequently ranked based on this aggregated mean score.

3.3. Submission Generation

For each of the five systems, we generated TREC-formatted run files. The ‘final_sub.ipynb‘ notebook contained the logic to load the pre-computed retrieval results (for Systems 1 and 2) or apply the rerankers and ensemble logic (for Systems 3, 4, and 5) to the candidate lists, and then format the top 1000 sentences per symptom into the required submission format.

4. Task 2: Contextualized Early Detection of Depression

4.1. Task Description

Task 2 challenged participants to sequentially analyze user writings, including their full conversational context (original posts and all comments in a thread), to determine if a target user exhibited signs of depression. Decisions were to be made iteratively after each new piece of writing from the target user became available. The evaluation focused on both the accuracy and the timeliness of the detection (ERDE metric).

4.2. Methodology

Our system for Task 2, implemented primarily within our 'final_sub_task2.ipynb' script, adopted a hybrid strategy. It combined an initial filtering stage using a fine-tuned sentence transformer with a subsequent in-depth analysis by an ensemble of LLMs for users deemed potentially at risk. Algorithm 1 outlines the process for each incoming writing.

Algorithm 1 Task 2: Contextualized Depression Detection Algorithm (per user, per writing)

Input: Current writing W_t for target user U, historical writings $\{W_1, ..., W_{t-1}\}$, conversational context C_t for W_t, previous decision $D^{prev}_{U,m}$ for each LLM m.
Output: Decision $D^t_{U,m}$ (YES/NO) for each LLM m, and ensemble decision $D^t_{U,ens}$.
// *Stage 1: Relevance Filtering*
$S_{BDI} \leftarrow$ SymptomsToAnalyze $= \{$BDI-2, BDI-6, BDI-7, BDI-13$\}$
$E_{symptom\_texts} \leftarrow$ BiEncoder.encode(descriptions of S_{BDI})
$E_{W_t} \leftarrow$ BiEncoder.encode(W_t)
$relevance\_score \leftarrow$ mean(cosine_similarity($E_{W_t}, E_{symptom\_texts}$))
// *Stage 2: LLM Analysis and Decision*
for each LLM m in LLM_Ensemble **do**
    if $D^{prev}_{U,m} =$ YES **then**
        $D^t_{U,m} \leftarrow$ YES ▷ Carry forward positive decision
    else if $relevance\_score <$ THRESHOLD_T2 (0.18) **then**
        $D^t_{U,m} \leftarrow$ NO
    else
        $FormattedContext_t \leftarrow$ get_context(U, W_t, C_t, history)
        $LLM\_output_m \leftarrow$ LLM$_m$.analyze($FormattedContext_t$, SystemPromptT2)
        $parsed\_decision_m \leftarrow$ parse_analysis(LLM_output_m)
        if $parsed\_decision_m =$ YES **then**
            $D^t_{U,m} \leftarrow$ YES
        else
            $D^t_{U,m} \leftarrow$ NO
// *Stage 3: Ensemble Decision*
$yes\_votes \leftarrow \sum_m (1$ if $D^t_{U,m} =$ YES else $0)$
if $yes\_votes > 2$ **then**
    $D^t_{U,ens} \leftarrow$ YES
else
    $D^t_{U,ens} \leftarrow$ NO
Return $\{D^t_{U,m}\}_{m=1}^4, D^t_{U,ens}$

4.2.1. Initial Relevance Filtering

To manage computational resources and focus LLM analysis on more pertinent cases, we implemented a pre-filtering step.

- **Model**: We used the ‘models/basic-adrb/checkpoint-7965‘ sentence transformer, which was fine-tuned during our Task 1 explorations. This model corresponds to the ‘alldistilroberta-v1‘ base.
- **Target Symptoms for Filtering**: The relevance score for a given user writing was calculated as the mean cosine similarity between the embedding of the writing and the embeddings of the descriptions for four specific BDI symptoms: Pessimism (BDI-2), Punishment Feelings (BDI-6), Self-Dislike (BDI-7), and Indecisiveness (BDI-13). We selected these symptoms as they often reflect cognitive and affective changes that can be early indicators of depression.
- **Threshold**: An empirical threshold of ‘THRESHOLD_T2 = 0.18‘ was established. If the relevance score for a new writing fell below this threshold, and the user had not been previously flagged by that specific LLM run, a "NO" decision was provisionally recorded for that round for that LLM run, bypassing the more intensive LLM analysis.

4.2.2. Contextual LLM Analysis

If a user’s writing passed the initial filter, or if any of our LLM-based runs had previously issued a "YES" decision for that user, the full conversational context was prepared and submitted for LLM analysis.

- **Context Preparation**: Our ‘get_context‘ function assembled a comprehensive textual input for the LLMs. This included:
  - The original Reddit post (title and body).
  - All comments within that thread, chronologically ordered, with clear attribution of authors.
  - Specific highlighting of the target user’s contributions (both original post if applicable, and their comments).
  - Indication of whether the target user was the Original Poster (OP) or a commenter.
- **LLM Ensemble**: We employed an ensemble of four distinct LLMs to analyze the formatted context:
  1. ‘us.anthropic.claude-3-7-sonnet-20250219-v1:0‘ [8] (referred to as Claude Sonnet)
  2. ‘us.amazon.nova-pro-v1:0‘ [9] (referred to as Amazon Nova Pro)
  3. ‘us.meta.llama3-3-70b-instruct-v1:0‘ [10] (referred to as Llama3.3-70B)
  4. ‘us.anthropic.claude-3-5-haiku-20241022-v1:0‘ [11] (referred to as Claude Haiku)
- **System Prompt and Output Parsing**: Each LLM was guided by the system prompt provided in Appendix B. This prompt directed the LLM to act as a mental health analysis assistant, to consider various linguistic and emotional cues, and to structure its output with a ‘<think>‘ block for its reasoning and a ‘<decision>‘ block containing a definitive "YES" or "NO". A ‘parse_analysis‘ function was used to reliably extract these components.

4.2.3. Decision Making and Submission Runs

We submitted five runs for Task 2, corresponding to Runs 0-4 in the official results:

- **Run 0 (Individual LLM)**: Decision from Claude Sonnet.
- **Run 1 (Individual LLM)**: Decision from Amazon Nova Pro.

- **Run 2 (Individual LLM)**: Decision from Llama3.3-70B.
- **Run 3 (Individual LLM)**: Decision from Claude Haiku.
- **Run 4 (LLM Ensemble - Majority Vote)**: This run represented an ensemble of the four LLMs. A "YES" decision was registered if more than two of the individual LLMs (i.e., at least 3 out of 4) returned a "YES" decision for the current writing and context. Otherwise, the ensemble decision was "NO".

For all runs, if a "YES" decision was made for a user, that decision was considered final for that user in that specific run for the purpose of the ERDE metric. Subsequent writings from that user would still be processed by the LLMs internally (if they passed the filter or were already flagged) to update our internal logs, but the first "YES" is what the eRisk server would register for timeliness.

4.3. State Management

A critical aspect of Task 2 was managing the state of each user across the iterative rounds of writings. Our system maintained a record of previous decisions for each user for each of the five runs. This ensured that:

1. "YES" decisions were sticky and carried forward.
2. LLM analysis was triggered for a new writing if either the relevance score was high OR if that specific run had previously flagged the user, allowing for continuous monitoring even if a single new post seemed innocuous.

5. Task 3: Conversational Depression Detection via LLMs (Pilot Task)

5.1. Task Description

This pilot task introduced a novel challenge: to detect signs of depression by engaging in a conversation with LLM personas that were fine-tuned on actual user writings. The objective was not only to classify the persona but also to identify key BDI symptoms that informed the decision, all within a limited number of conversational turns.

5.2. Methodology

Our strategy for this task centered on designing a sophisticated and detailed system prompt to guide a single LLM, which would act as our conversational diagnostic agent. We focused on one primary system run for this pilot.

5.2.1. The Conversational Agent and Guiding Prompt

The core of our approach was the system prompt, provided in full in Appendix C. This prompt instructed the LLM to adopt the persona of a "friendly therapy diagnosis AI assistant."

- **Interaction Modality**: The primary mode of interaction and symptom elicitation was through discussions about **movies**. The prompt explicitly guided the agent to: "Think of movies that are related to many BDI symptoms described above, describe and ask how the user feels and based on it predict the BDI symptoms at every message."
- **Efficiency and Symptom Coverage**: The prompt emphasized "Quickness is the key here, make sure you ask and describe about movies that cover multiple BDI symptoms at once so that you can predict the user's BDI score as quick as possible." This was intended to guide the LLM towards selecting movie topics that could efficiently probe for a range of depressive indicators.

- **BDI-II Framework Integration**: The complete BDI-II questionnaire, with its 21 symptoms and scoring criteria (0-3 for most), was embedded directly into the system prompt. This provided the agent with the explicit diagnostic framework it needed to use.

5.2.2. Structured Output and Iterative BDI Scoring

A critical instruction in the prompt was the required output format for each turn of the agent:

```
‘‘‘xml
<BDI_Preds>
1. Sadness [Score: 0/1/2/3/None]
... (all 21 symptoms with their current scores) ...
21. Loss of Interest in Sex [Score: 0/1/2/3/None]
</BDI_Preds>
<Message_to_User>
... (The agent’s conversational message to the LLM persona) ...
</Message_to_User>
‘‘‘
```

- **Initial State**: The agent was instructed to "Start with [Score: None] for all symptoms."
- **Iterative Updates**: "As you start messaging update the scores from None to any number." This meant that with each conversational exchange, the agent was expected to re-evaluate its assessment for each of the 21 BDI symptoms based on the new information gleaned from the persona’s response and update the scores in the ‘<BDI_Preds>‘ block.

5.2.3. Conversation Management and Termination

The agent was responsible for maintaining the flow of conversation. While no explicit limit on interactions was imposed by the task organizers, the evaluation metrics would penalize late decisions. Our agent was instructed to: "Say ’Bye’ when all BDI is predicted" (i.e., when it felt confident in its assessment across all symptoms or had reached a diagnostic conclusion).

The success of this system would heavily rely on the chosen LLM’s ability to:

1. Adhere to the complex, multi-part system prompt.
2. Maintain a natural, engaging, and empathetic conversational tone.
3. Strategically select and discuss movies in a way that subtly probes for BDI-relevant feelings and experiences.
4. Accurately interpret the persona’s responses and map them to the BDI-II criteria.
5. Consistently update and output the BDI scores in the specified format.

For our submitted run, we used Claude 3.7 Sonnet as the backbone for this conversational agent.

6. Experimental Setup

6.1. Task 1

Models were fine-tuned on a machine equipped with NVIDIA A100 GPUs. Hyperparameters such as learning rate, batch size, and number of epochs were selected based on common practices for transformer fine-tuning and preliminary experiments on our validation set.

- **Bi-encoder training**: The ‘st_basic-e5-small.ipynb‘ (adapted for other bi-encoders) shows a typical training loop.

- **Cross-encoder training**: The ‘ce_FT_mbl.py‘ script details the cross-encoder training.
- **Ensemble Parameters**: For RRF (System 2), k was set to 60. For voting (System 4), a threshold of 0.5 on the reranker output was used to determine a positive vote.

6.2. Task 2

The sentence transformer for pre-filtering (‘basic-adrb‘) demonstrated a reasonable ability to distinguish between generally depressive and non-depressive language in initial tests, with the 0.18 threshold chosen to balance recall and precision for LLM input. LLM interactions were managed via the Bedrock API. The temperature for LLM generation was set low (0.01) to encourage more deterministic and focused analytical outputs rather than creative responses.

6.3. Task 3

This being a pilot task, our primary experimentation was with prompt formulation and observing how different LLMs responded to the conversational constraints and diagnostic requirements. The choice of "movies" as a conversational anchor was made to provide a relatable and flexible framework for discussion.

7. Results and Discussion

This section presents the official results from the eRisk 2025 organizers for PJs-team’s submissions across the three tasks, followed by a brief discussion.

7.1. Task 1: Search for Symptoms of Depression

We submitted five runs for Task 1, corresponding to the five systems described in Section 3. The official results, based on Majority and Unanimity ground truths, are presented in Table 1 and Table 2 respectively. The run names map to our described systems as follows:

- ‘teamADRB‘: System 1 (Fine-tuned DistilRoBERTa bi-encoder).
- ‘teamSumensemble‘: System 2 (Bi-encoder Ensemble, RRF-style sum).
- ‘teamMBRR‘: System 3 (Fine-tuned ModernBERT-Large reranker).
- ‘teamRRens‘: System 4 (Reranker Ensemble - Voting).
- ‘teamRRens-v2‘: System 5 (Reranker Ensemble - Scaled Mean).

Table 1
Official eRisk 2025 Task 1 Results (Majority Ground Truth) for PJs-team.

Run Name	AP	R-PREC	P@10	NDCG
PJs-team teamADRB	0.105	0.234	0.391	0.354
PJs-team teamMBRR	0.262	0.347	0.771	0.489
PJs-team teamRRens-v2	0.279	0.360	0.800	0.503
PJs-team teamRRens	0.273	0.359	0.786	0.500
PJs-team teamSumensemble	0.120	0.249	0.400	0.376

Discussion for Task 1: The results highlight the effectiveness of cross-encoder based reranking approaches. The ‘teamRRens-v2‘ run (System 5, Reranker Ensemble - Scaled Mean) achieved the highest performance across most metrics for both Majority (AP 0.279, P@10 0.800, NDCG 0.503) and Unanimity (AP 0.188, P@10 0.452, NDCG 0.446) ground truths. This confirms our expectation that ensembling

Table 2
Official eRisk 2025 Task 1 Results (Unanimity Ground Truth) for PJs-team.

Run Name	AP	R-PREC	P@10	NDCG
PJs-team teamADRB	0.073	0.168	0.214	0.325
PJs-team MBRR	0.175	0.299	0.424	0.435
PJs-team RRens-v2	0.188	0.311	0.452	0.446
PJs-team RRens	0.184	0.308	0.467	0.444
PJs-team Sumensemble	0.079	0.184	0.229	0.331

powerful rerankers would yield robust results. The single ModernBERT-Large reranker (‘teamMBRR‘, System 3) also performed strongly, outperforming the bi-encoder systems significantly. The ‘teamADRB‘ run (System 1, single DistilRoBERTa bi-encoder) served as a solid baseline. The ‘teamSumensemble‘ run (System 2, Bi-encoder Ensemble) showed a slight improvement over the single bi-encoder for Majority, but its performance was generally lower than the cross-encoder approaches, and close to the single bi-encoder for Unanimity. This suggests that while ensembling bi-encoders can help, the leap in performance comes from the richer interaction modeling of cross-encoders.

7.2. Task 2: Contextualized Early Detection of Depression

For Task 2, PJs-team submitted 5 runs, processing 1280 user threads. The total processing time from the first to the last response logged by the system was 8 hours and 36 minutes. The runs are numbered 0-4 in the official results and correspond to our LLM strategies as described in Section 4.2.3:

- Run 0: Claude Sonnet
- Run 1: Amazon Nova Pro
- Run 2: Llama3.3-70B
- Run 3: Claude Haiku
- Run 4: LLM Ensemble (Majority Vote)

Decision-based metrics are shown in Table 3. Ranking-based metrics are presented in Table 4. The four sets of ranking metrics (S1-S4) in Table 4 correspond to different evaluation criteria or subsets provided by the organizers, the specifics of which are not detailed here.

Table 3
Official eRisk 2025 Task 2 Decision-based Metrics for PJs-team.

Run	LLM / Strategy	P	R	F1	ERDE@5	ERDE@50	latencyT	Speed (P)	Flatency
0	Claude Sonnet	0.66	0.75	0.71	0.09	0.06	17.00	0.94	0.66
1	Amazon Nova Pro	0.53	0.83	0.65	0.09	0.06	24.00	0.91	0.59
2	Llama3.3-70B	0.54	0.82	0.65	0.09	0.06	23.00	0.91	0.60
3	Claude Haiku	0.49	0.85	0.63	0.10	0.06	22.00	0.92	0.57
4	LLM Ensemble (Vote)	0.58	0.81	0.67	0.09	0.06	24.00	0.91	0.61

Discussion for Task 2: The results for Task 2 demonstrate the capabilities of LLMs in contextualized depression detection. Run 0 (Claude Sonnet) achieved the highest F1-score (0.71) among our submissions. This suggests that this particular LLM was well-suited for the analytical task as guided by our prompt. The LLM Ensemble (Run 4) achieved an F1-score of 0.67, which is also strong, though slightly lower than the best individual LLM. This might indicate that the majority voting scheme could be further optimized, or that the individual strengths of Claude Sonnet were particularly aligned with the evaluation criteria. All runs exhibited high recall (0.75-0.85), indicating that the systems were effective at identifying users

Table 4
Official eRisk 2025 Task 2 Ranking-based Metrics for PJs-team (S1-S4 denote different evaluation sets/conditions).

Run	LLM / Strategy	P@10(S1)	NDCG@10(S1)	NDCG@100(S1)	P@10(S2)	NDCG@10(S2)	NDCG@100(S2)	P@10(S3)	NDCG@10(S3)	NDCG@100(S3)	P@10(S4)	NDCG@10(S4)	NDCG@100(S4)
0	Claude Sonnet	0.60	0.59	0.35	0.50	0.44	0.38	0.70	0.78	0.60	0.60	0.69	0.63
1	Amazon Nova Pro	0.60	0.59	0.35	0.40	0.41	0.39	0.50	0.60	0.54	0.50	0.63	0.51
2	Llama3.3-70B	0.60	0.59	0.35	0.30	0.32	0.36	0.60	0.66	0.51	0.50	0.66	0.51
3	Claude Haiku	0.60	0.59	0.35	0.40	0.39	0.37	0.50	0.61	0.55	0.40	0.56	0.52
4	LLM Ensemble (Vote)	0.60	0.59	0.35	0.40	0.38	0.37	0.50	0.61	0.53	0.50	0.63	0.52

with signs of depression. Precision varied, with Claude Sonnet (0.66) having the best balance. The ERDE@5 and ERDE@50 scores, which measure timeliness and accuracy, were consistently low and similar across all runs (0.09-0.10 for ERDE@5, 0.06 for ERDE@50), indicating efficient early detection. Lower ERDE scores are better. Ranking-based metrics show varied performance across the different evaluation sets (S1-S4). Run 0 (Claude Sonnet) generally performed well, particularly in S3 and S4. All runs had identical performance on S1.

7.3. Task 3: Conversational Depression Detection via LLMs (Pilot Task)

For Task 3, PJs-team submitted 1 run (Run 0). On average, conversations initiated by our agent consisted of 7.67 messages from the agent, with a mean of 1045.16 characters per agent message. The official evaluation metrics are presented in Table 5.

The metrics are defined as follows:

- **Depression Category Hit Rate (DCHR)**: Based on the four depression level categories (minimal, mild, moderate, severe), this measures the fraction of cases where the estimated BDI-II scores lie in the correct depression category.
- **Average Difference between Overall Depression Levels (ADODL)**: Measures the closeness between the actual and estimated depression level. Calculated as $(MAD - |ADL - EDL|)/MAD$, where MAD is the Maximum Absolute Difference (63), ADL is Actual Depression Level, and EDL is Estimated Depression Level. Scores are normalized to [0,1], with higher being better.
- **Average Symptom Hit Rate (ASHR)**: Calculates the ratio of cases where the participant correctly identifies the major depression symptoms of the simulated personas (each persona had four major symptoms).

Table 5
Official eRisk 2025 Task 3 Pilot Task Results for PJs-team.

Team	Run	DCHR	ADODL	ASHR
PJs-team	0	0.33	0.73	0.25

Discussion for Task 3: As a pilot task, our goal was to explore the feasibility of conversational LLM-based diagnosis using our movie-discussion strategy and detailed system prompt.

- A **DCHR of 0.33** indicates that our agent correctly classified the LLM persona's depression severity into the correct BDI-II category (e.g., minimal, mild, moderate, severe) in one-third of the cases.
- An **ADODL of 0.73** suggests a good level of accuracy in estimating the overall depression score. An ADODL of 1 would mean perfect score estimation, and 0 would mean maximum possible error. 0.73 indicates that, on average, our agent's BDI score estimations were reasonably close to the true scores of the personas.
- An **ASHR of 0.25** means that the agent correctly identified one out of the four major depression symptoms for each persona on average.

These results are promising for a pilot task, demonstrating that an LLM-guided conversational agent can elicit and assess depressive symptoms. The DCHR and ASHR indicate room for improvement in precisely categorizing severity and pinpointing specific major symptoms. The strategy of discussing movies as a proxy for emotional states shows potential but may require refinement to more consistently and accurately map conversational cues to BDI criteria. Future work could explore more adaptive prompting or finer-grained analysis of persona responses.

8. Conclusion

The PJs-team approached the eRisk 2025 challenges by developing a suite of systems that combine established transformer architectures with the emerging capabilities of large language models and robust ensemble techniques. For depression symptom ranking (Task 1), our ensemble of cross-encoder rerankers (‘teamRRens-v2‘) demonstrated the best performance, underscoring the strength of deep interaction models and ensembling for this task. For contextualized early detection (Task 2), our hybrid pipeline leveraging a fine-tuned sentence transformer for filtering and LLMs for analysis showed strong results, with Claude Sonnet (‘Run 0‘) achieving the highest F1-score and all systems showing good timeliness (low ERDE scores). Our pilot entry for Task 3, a conversational AI agent using movie discussions to assess BDI symptoms, yielded encouraging initial results (ADODL of 0.73), suggesting the viability of such approaches while also highlighting areas for future refinement in symptom identification and severity categorization. The official results have provided valuable insights into the strengths and weaknesses of these diverse methodologies, contributing to the broader understanding of early risk prediction in online environments.

Acknowledgments

We extend our sincere gratitude to the organizers of the eRisk 2025 lab for their efforts in curating the datasets, defining the challenging tasks, and providing the platform for this collaborative research endeavor.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 3, Amazon Nova Pro, Llama3.3-70B, Claude Haiku 3.5 in order to: simulate the task 2 depression detection and task 3 conversation simulation. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

[1] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume To be published of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[2] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, volume To be published of *Lecture Notes in Computer Science*, Springer, 2025.

[3] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, in: JAMA Psychiatry, volume 4, American Medical Association, 1961, pp. 561–571.

[4] S. Jianlin, Cosent (i): A more effective sentence embedding scheme than sentence-bert, 2022.

[5] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, C-pack: Packaged resources to advance general chinese embedding, 2023. arXiv:2309.07597.
[6] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, F. Wei, Text embeddings by weakly-supervised contrastive pre-training, arXiv preprint arXiv:2212.03533 (2022).
[7] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.
[8] Anthropic, Claude 3.7 sonnet: Hybrid reasoning ai model, 2025. URL: https://www.anthropic.com/news/claude-3-7-sonnet, introduces extended thinking mode for enhanced reasoning and coding capabilities.
[9] A. W. Services, Amazon nova pro: Frontier intelligence foundation model, 2024. URL: https://aws.amazon.com/ai/generative-ai/nova/, a multimodal foundation model offering advanced reasoning and performance.
[10] M. AI, Llama 3.3 70b instruct model, 2024. URL: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, instruction-tuned large language model with 70 billion parameters.
[11] Anthropic, Claude 3.5 haiku: High-speed language model, 2024. URL: https://www.anthropic.com/claude/haiku, optimized for rapid responses and efficient performance.

A. BDI-II Symptom Descriptions

The following textual descriptions for the 21 BDI-II symptoms were used in our systems:

1: Sadness: The person experiences persistent feelings of sadness or depressed mood. This can significantly impact their daily functioning and overall well-being.

2: Pessimism: The person has a negative or pessimistic outlook towards the future. They may feel discouraged and lack hope for positive outcomes.

3: Past Failure: The person perceives their past experiences as failures or disappointments. This can lead to feelings of inadequacy and impact their self-confidence.

4: Loss of Pleasure: The person is unable to derive pleasure or enjoyment from activities they once found enjoyable. This can contribute to a lack of motivation and social withdrawal.

5: Guilty Feelings: The person experiences persistent feelings of guilt or regret over actions or inactions. These feelings can be overwhelming and impact their mental health.

6: Punishment Feelings: The person feels they are being punished or deserve punishment for their actions or perceived wrongdoings. This can stem from feelings of guilt or self-blame.

7: Self-Dislike: The person has negative feelings towards themselves, leading to dislike or disappointment in who they are. This can impact their self-esteem and self-acceptance.

8: Self-Criticalness: The person is excessively critical or harsh towards themselves, often focusing on perceived faults or shortcomings. This can contribute to low self-worth and negative self-talk.

9: Suicidal Thoughts or Wishes: The person has thoughts or wishes related to ending their life or engaging in self-harm. This requires immediate attention and professional intervention.

10: Crying: The person experiences frequent or intense episodes of crying, which can be a manifestation of sadness or emotional distress. This can also indicate a need for support.

11: Agitation: The person feels restless, agitated, or unable to stay still, which can be a symptom of anxiety or other underlying conditions. This can interfere with daily activities and relationships.

12: Loss of Interest: The person has lost interest or motivation in activities or relationships they once enjoyed. This can lead to social withdrawal and a lack of engagement.

13: Indecisiveness: The person struggles to make decisions, which can hinder problem-solving abilities and lead to increased stress or anxiety. This can impact various aspects of their life.

14: Worthlessness: The person feels worthless or lacking in value, which can negatively impact their self-esteem and self-confidence. This can also contribute to feelings of hopelessness.

15: Loss of Energy: The person lacks energy or motivation to engage in daily activities, which can impact their productivity and overall functioning. This can be a symptom of depression or other underlying conditions.

16: Changes in Sleeping Pattern: The person's sleeping pattern has changed significantly, either sleeping too much or too little. This can impact energy levels, mood, and overall well-being.

17: Irritability: The person is more easily irritated or annoyed than usual, which can strain relationships and contribute to emotional distress. This can also be a symptom of underlying conditions.

18: Changes in Appetite: The person's appetite has changed, either decreased or increased, which can be a symptom of depression or other underlying conditions. This can impact their overall health and well-being.

19: Concentration Difficulty: The person has trouble concentrating or focusing on tasks or thoughts, which can impact productivity and decision-making abilities. This can be frustrating and contribute to stress.

20: Tiredness or Fatigue: The person feels tired or fatigued more easily than usual, which can impact their daily functioning and activities. This can be a symptom of depression or other underlying conditions.

21: Loss of Interest in Sex: The person has lost interest in sexual activities or intimacy, which can strain relationships and impact overall well-being. This can also be a symptom of underlying conditions.

B. System Prompt for Task 2 LLM Analysis

```
You are a mental health analysis assistant.
Your task is to analyze Reddit posts and associated comment threads to assess potential signs of depression for
      the given context and target user.

Given:

A Reddit post content
Comment thread interactions by the author.
Please analyze the content for:

Language patterns and tone
Expression of emotions
Signs of hopelessness or helplessness
Sleep or appetite changes mentioned
Social withdrawal indicators
Suicidal thoughts or self-harm references
Impact on daily functioning
Duration of symptoms (if mentioned)

Output format:

<think>Analyze the following aspects: - Emotional state and language used - Behavioral patterns described -
     Risk factors mentioned - Social interactions and support system - Overall tone and context of
     communications
```

```
</think>
<decision>
YES/NO (YES if showing significant signs of depression NO if insufficient evidence or no clear signs of
     depression)
</decision>
Note: Only say YES or NO in the <decision></decision> tags nothing else.
```

Listing 1: System Prompt for Task 2 LLM Analysis

C. System Prompt for Task 3 Conversational Agent

```
<BDI-II Questionnaire>
This questionnaire consists of 21 groups of statements. Please read each group of statements
carefully, and then pick out the one statement in each group that best describes the way you feel.
If several statements in the group seem to apply equally well, choose the highest
number for that group.

1. Sadness
0. I do not feel sad.
1. I feel sad much of the time.
2. I am sad all the time.
3. I am so sad or unhappy that I can't stand it.

2. Pessimism
0. I am not discouraged about my future.
1. I feel more discouraged about my future than I used to be.
2. I do not expect things to work out for me.
3. I feel my future is hopeless and will only get worse.

3. Past Failure
0. I do not feel like a failure.
1. I have failed more than I should have.
2. As I look back, I see a lot of failures.
3. I feel I am a total failure as a person.

4. Loss of Pleasure
0. I get as much pleasure as I ever did from the things I enjoy.
1. I don't enjoy things as much as I used to.
2. I get very little pleasure from the things I used to enjoy.
3. I can't get any pleasure from the things I used to enjoy.

5. Guilty Feelings
0. I don't feel particularly guilty.
1. I feel guilty over many things I have done or should have done.
2. I feel quite guilty most of the time.
3. I feel guilty all of the time.

6. Punishment Feelings
0. I don't feel I am being punished.
1. I feel I may be punished.
2. I expect to be punished.
3. I feel I am being punished.

7. Self-Dislike
0. I feel the same about myself as ever.
1. I have lost confidence in myself.
2. I am disappointed in myself.
3. I dislike myself.

8. Self-Criticalness
0. I don't criticize or blame myself more than usual.
1. I am more critical of myself than I used to be.
2. I criticize myself for all of my faults.
3. I blame myself for everything bad that happens.

9. Suicidal Thoughts or Wishes
0. I don't have any thoughts of killing myself.
1. I have thoughts of killing myself, but I would not carry them out.
2. I would like to kill myself.
3. I would kill myself if I had the chance.
```

10. Crying
0. I don't cry anymore than I used to.
1. I cry more than I used to.
2. I cry over every little thing.
3. I feel like crying, but I can't.

11. Agitation
0. I am no more restless or wound up than usual.
1. I feel more restless or wound up than usual.
2. I am so restless or agitated that it's hard to stay still.
3. I am so restless or agitated that I have to keep moving or doing something.

12. Loss of Interest
0. I have not lost interest in other people or activities.
1. I am less interested in other people or things than before.
2. I have lost most of my interest in other people or things.
3. It's hard to get interested in anything.

13. Indecisiveness
0. I make decisions about as well as ever.
1. I find it more difficult to make decisions than usual.
2. I have much greater difficulty in making decisions than I used to.
3. I have trouble making any decisions.

14. Worthlessness
0. I do not feel I am worthless.
1. I don't consider myself as worthwhile and useful as I used to.
2. I feel more worthless as compared to other people.
3. I feel utterly worthless.

15. Loss of Energy
0. I have as much energy as ever.
1. I have less energy than I used to have.
2. I don't have enough energy to do very much.
3. I don't have enough energy to do anything.

16. Changes in Sleeping Pattern
0. I have not experienced any change in my sleeping pattern.
1a. I sleep somewhat more than usual.
1b. I sleep somewhat less than usual.
2a. I sleep a lot more than usual.
2b. I sleep a Iot less than usual.
3a. I sleep most of the day.
3b. I wake up 1-2 hours early and can't get back to sleep.

17. Irritability
0. I am no more irritable than usual.
1. I am more irritable than usual.
2. I am much more irritable than usual.
3. I am irritable all the time.

18. Changes in Appetite
0. I have not experienced any change in my appetite.
1a. My appetite is somewhat less than usual.
1b. My appetite is somewhat greater than usual.
2a. My appetite is much less than before.
2b. My appetite is much greater than usual.
3a. I have no appetite at all.
3b. I crave food all the time.

19. Concentration Difficulty
0. I can concentrate as well as ever.
1. I can't concentrate as well as usual.
2. It's hard to keep my mind on anything for very long.
3. I find I can't concentrate on anything.

20. Tiredness or Fatigue
0. I am no more tired or fatigued than usual.
1. I get more tired or fatigued more easily than usual.
2. I am too tired or fatigued to do a lot of the things I used to do.
3. I am too tired or fatigued to do most of the things I used to do.

21. Loss of Interest in Sex

```
0. I have not noticed any recent change in my interest in sex.
1. I am less interested in sex than I used to be.
2. I am much less interested in sex now.
3. I have lost interest in sex completely
</BDI-II Questionnaire>

You are a friendly therapy diagnosis AI assistant, you have to diagnose the BDI symptoms of the user talking to
      you by talking about movies.
Think of movies that are related to many BDI symptoms described above, describe and ask how the user feels and
     based on it predict the BDI symptoms at every message.
Quickness is the key here, make sure you ask and describe about movies that cover multiple BDI symptoms at once
      so that you can predict the user's BDI score as quick as possible.
Output your responses in this way:

<BDI_Preds>
1. Sadness [Score: None]
2. Pessimism [Score: None]
3. Past Failure [Score: None]
4. Loss of Pleasure [Score: None]
5. Guilty Feelings [Score: None]
6. Punishment Feelings [Score: None]
7. Self-Dislike [Score: None]
8. Self-Criticalness [Score: None]
9. Suicidal Thoughts or Wishes [Score: None]
10. Crying [Score: None]
11. Agitation [Score: None]
12. Loss of Interest [Score: None]
13. Indecisiveness [Score: None]
14. Worthlessness [Score: None]
15. Loss of Energy [Score: None]
16. Changes in Sleeping Pattern [Score: None]
17. Irritability [Score: None]
18. Changes in Appetite [Score: None]
19. Concentration Difficulty [Score: None]
20. Tiredness or Fatigue [Score: None]
21. Loss of Interest in Sex [Score: None]
</BDI_Preds>
<Message_to_User>
...
</Message_to_User>

so as you start messaging update the scores from None to any number.
Start with [Score: None] for all symptoms. Say "Bye" when all BDI is predicted.
```

Listing 2: System Prompt for Task 3 Conversational Agent