# Contextualized Early Detection of Depression - Hybrid and Time-Aware Approaches: HU at eRisk Task 2 2025

Notebook for the eRisk Lab at CLEF 2025

Muhammad Saad[1,*,†], Asad Ullah Chaudhry[1,†], Meesum Abbas[1,†], Faisal Alvi[1] and Abdul Samad[1]

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

**Abstract**

The early detection of depression in online conversational threads remains a pivotal challenge in computational mental health, particularly under the real-time, context-aware requirements of CLEF eRisk 2025 Task 2. We propose a multifaceted study evaluating five innovative approaches, including transformer-based models (e.g., ModernBERT with time-aware loss), Classification with Partial Information and Decision-Making Component (CPI+DMC) frameworks enhanced by Llama 3.1, data augmentation strategies, simple threshold policies, and zero-shot learning with Llama-4-Scout-17B. Leveraging the eRisk dataset, our methodologies integrate recent advances in time-aware training and hybrid ensembles, addressing the trade-offs between classification accuracy, earliness, and computational efficiency. Our results demonstrate that the CPI+DMC approach achieves a best $F_1$ score of 0.75, securing a 3rd place ranking among 12 teams, with a competitive NDCG@100 of 0.62 for early detection and an $ERDE_{50}$ of 0.05, highlighting its effectiveness in balancing accuracy and latency. These findings offer valuable insights into real-time mental health monitoring and underscore the potential for future research to refine decision policies and enhance long-term ranking stability.

**Keywords**

CLEF eRisk, Early Depression Detection, Conversational Context Analysis, Transformer-based Models

## 1. Introduction

The detection of depression from textual data is a critical area in computational mental health, emphasizing early detection for timely intervention. The CLEF eRisk challenges have advanced this field with real-time, sequential classification tasks, requiring models to analyze user-generated content chronologically [1]. The eRisk 2025 Task 2 focuses on early depression detection via contextualized analysis of Reddit conversations, using full threads to evaluate accuracy and latency with metrics like ERDE.

We propose five innovative approaches integrating deep learning, time-aware modeling, and decision policies. Our contributions include:

- Combining Llama 3.1 summarization with BERT classification (CPI+DMC) to address domain mismatch.
- A zero-shot approach using Llama-4-Scout-17B for efficient detection.
- ModernBERT with time-aware loss and class weighting for enhanced earliness.
- ModernBERT with data augmentation to tackle class imbalance.
- Simple threshold and hybrid decision rules for optimized alerts.

Our CPI+DMC approach achieves an F1 score of 0.75, ranking 3rd among 12 teams, offering insights into real-time mental health monitoring.

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

*Corresponding author.

†These authors contributed equally.

ms08063@st.habib.edu.pk (M. Saad); ac07408@st.habib.edu.pk (A. U. Chaudhry); ma08056@st.habib.edu.pk (M. Abbas); faisal.alvi@sse.habib.edu.pk (F. Alvi); abdul.samad@sse.habib.edu.pk (A. Samad)

0009-0008-7039-7994 (M. Saad); 0009-0000-8626-2797 (A. U. Chaudhry); 0009-0002-7682-1617 (M. Abbas); 0000-0003-3827-7710 (F. Alvi); 0009-0009-5166-6412 (A. Samad)

## 2. Literature Review

Early depression detection from text initially used recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Trotzek et al. [2] combined CNNs with linguistic metadata, like sentiment and syntactic complexity, to improve classification by capturing semantic patterns and psychological insights. Data sparsity and class imbalance necessitated ensemble techniques. Transformer-based models, particularly BERT, advanced the field. Martínez-Castaño et al. [3] showed fine-tuned BERT outperformed RNNs and CNNs in early detection, while Devaguptam et al. [4] noted DeBERTa's slight edge due to better contextual embeddings. These models require threshold tuning to balance sensitivity and false positives, with computational costs limiting real-time use.

Recent work explored large language models (LLMs) like GPT-3.5 and LLaMA2. Munir et al. [5] achieved near-perfect accuracy with fine-tuned LLMs on social media text, surpassing BERT and DeBERTa. Hadzic et al. [6] reported GPT-4's 81% precision and 71% F1 score on conversational data, outperforming BERT. LLMs' computational demands and lack of early decision mechanisms require external policies. Zhang & Poellabauer [7] introduced contextual position encoding (CoPE) for multimodal detection, effective for clinical data but challenging for social media.

Thompson & Errecalde [8] used time-aware training with timestamp indicators and an ERDE loss function, achieving top $\text{ERDE}_{50}$ scores. Loyola et al. [9] proposed a Classification-Prediction & Decision (CPI+DMC) approach, with a transformer outputting probabilities and a separate alert module, reducing false positives via adaptive policies. The CPI+DMC approach treats early risk detection as a multi-objective problem, balancing precision through a classification component (CPI) and speed via a decision-making component (DMC) that determines the optimal moment for issuing alerts based on prediction history. This method has shown robustness across eRisk challenges by allowing independent optimization of classification accuracy and timely decision-making. Gui et al. [10] applied reinforcement learning to select relevant text, improving precision by 14.6%, though requiring extensive data.

Evaluation metrics like ERDE [11] penalize late detections but face interpretability issues. Sadeque et al. [12] proposed a latency-weighted F1 score ($\text{F}_{\text{latency}}$) for clarity. Ranking metrics like Precision@10 and NDCG assess real-time prioritization. The UNSL team at eRisk 2024 [8] combined a BERT-based CPI+DMC model with a time-aware transformer for top performance. Future work may integrate RL, time-aware LLMs, and multimodal data. Transformers have replaced CNNs and RNNs, but computational costs, data imbalance, and interpretability remain challenges, with eRisk 2025 emphasizing conversational context for hybrid model advances.

## 3. Methods

For eRisk 2025 Task 2, we developed five approaches[1] to detect depression early in Reddit conversational threads, analyzing posts and comments sequentially [1]. Common preprocessing includes lowercase conversion, URL replacement with "[URL]", HTML/Unicode artifact removal, and truncation to 2048 tokens (except Run 4) to balance context and efficiency [2, 3]. A hashmap tracks user IDs across rounds, and tiered thresholds optimize decision-making [9]. To visualize the pipeline of our five approaches, including preprocessing steps and methodological components, refer to Figure 1. Below, we detail each approach (Runs 0–4) with enumerated steps and key methodological insights.

### 3.1. ModernBERT with Time-Aware Loss and Class Weighting (Run 0)

ModernBERT's 8192-token capacity enables processing of long conversational sequences, unlike BERT's 512-token limit, making it ideal for capturing Reddit thread context. However, due to hardware constraints, we truncate to 2048-token input. In addition, we incorporate a time-aware loss to prioritize early detection, using $\text{ERDE}_{50}$ to penalize late predictions [8].

[1] Source code available at https://github.com/meesuma5/erisk_task2.

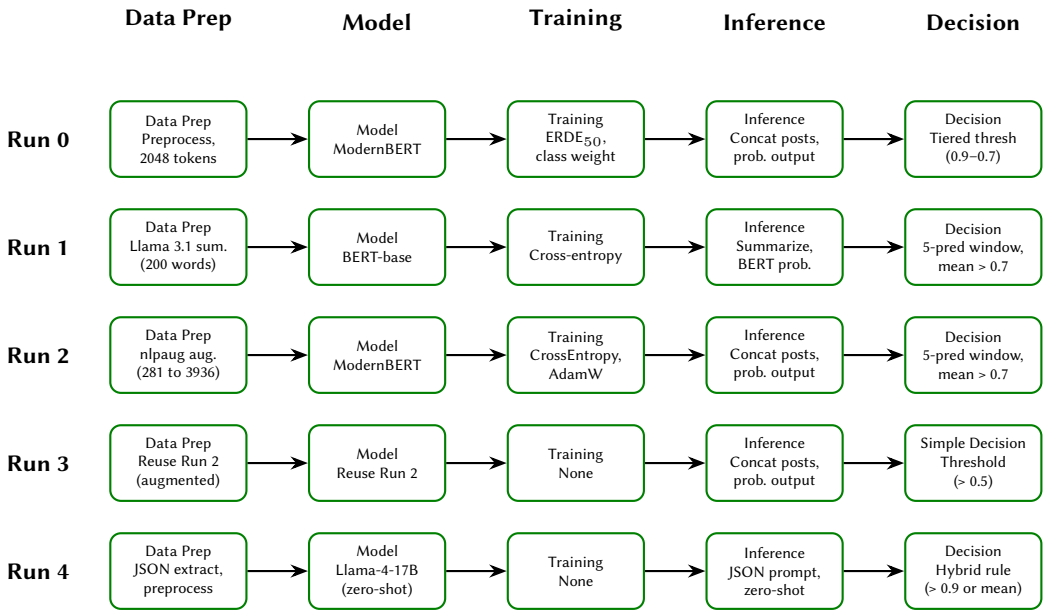


**Figure 1:** Pipeline summary of the five approaches for eRisk 2025 Task 2.

1. **Data Preparation**: Extract <TEXT> or <TITLE> from each user's posts in the training dataset, concatenate posts as `[CLS] post1 [SEP] post2 [SEP] ...`, preprocess with lowercase and URL replacement.
2. **Model**: Fine-tune ModernBERT-base with CrossEntropy and $\text{ERDE}_{50}$ loss, applying class weighting to address 10% positive class imbalance.
3. **Training**: Split data 80:20 (train:validation), train for 5 epochs, learning rate 2e-5, achieving validation F1-macro of 0.66.
4. **Inference**: Process posts with context as (`[CLS] post [CONTEXT] comment1 [SEP] ...`), or if the target user is the author of a comment process it as (`[CLS] comment of target user (`$comment_1$`) [CONTEXT] parent of` $comment_1$ `(`$comment_2$`) [SEP] parent of` $comment_2$ `(`$comment_3$`) [SEP] ...`), and then output depression probability.
5. **Decision**: Apply tiered thresholds on mean scores (0.9 for 1 post, 0.85 for 2, 0.8 for 3, 0.75 for 4 and 0.7 for 5+ posts) using hashmap for user tracking. Tiered thresholds were used to balance precision and speed by requiring higher confidence for early predictions with limited posts, while allowing lower thresholds as more posts provide greater context, thus optimizing early risk detection performance.

### 3.2. Summarization-Enhanced CPI+DMC with BERT (Run 1)

This approach addresses the domain mismatch between training (isolated writings) and test (conversational threads) data by using Llama 3.1 to summarize texts, preserving emotional cues. The CPI+DMC framework balances classification accuracy with timely decisions [9].

1. **Data Preparation**: For each user in the training set, all available posts are concatenated to form a composite document. This aggregated text is then summarized to a maximum of 200 words using Llama 3.1 with prompts designed to retain emotional depth and salient personal context. The prompts used for summarization are as follows:
   *System Prompt:*

   > You are a focused, analytical summarizer. Your role is to extract and condense content into a concise summary that captures the emotional state, notable life events, and communication style expressed in the input text. Your output must be a standalone

summary of no more than 200 words, with no additional commentary or introductory phrases.

*User Prompt:*

> You have been provided with a text for analysis. Summarize the text into a concise summary that focuses on the emotional state and notable life events of the person in it (if any). Include any signs of sadness, depression, or concerning words or phrases you find in the text. Ensure the summary is no longer than 200 words and contains only the summary content. Always answer starting with "The user..."
> Text: {text}

Token length of model output was capped to 500 tokens in line with BERT's input token limit of 512 tokens. [3]

2. **Model**: Fine-tune BERT-base-uncased (512-token input, 12 layers, 110M parameters) with linear layer (768-to-1) and sigmoid for probability [3]. BERT was chosen in line with UNSL 2024's choices and their obtained results [8].
3. **Training**: Use 80:20 split, 3 epochs, Adam (lr 2e-5), binary cross-entropy; validation precision 0.977 (non-depressed), 0.418 (depressed).
4. **Inference**: Summarize threads with Llama 3.1, classify with BERT to output probability.
5. **Decision**: DMC uses 5-prediction sliding window, triggers alert if mean score > 0.7 after 10 predictions.

### 3.3. ModernBERT with Data Augmentation and DMC (Run 2)

To mitigate class imbalance (312 depressed vs. 2824 non-depressed users), we augment positive samples using multiple techniques, enhancing ModernBERT's robustness. The DMC ensures controlled alert triggering [9].

1. **Data Preparation**: Parse XML, concatenate writings, augment 281 positive samples with nlpaug (BackTranslation: English→French/German→English, SynonymAug, ContextualWordEmbsAug) to 3936 samples [1].
2. **Model**: Fine-tune ModernBERT-base (2048-token input) for sequence classification.
3. **Training**: Use 90:10 split, 5 epochs, AdamW (lr 2e-5), CrossEntropy loss; validation F1-macro 0.771.
4. **Inference**: Concatenate post title/body/comments, preprocess, tokenize, output probability.
5. **Decision**: DMC with 5-prediction window, alert if mean score > 0.7 after 5 writings.

### 3.4. ModernBERT with Simple Decision Threshold (Run 3)

This approach simplifies decision-making to evaluate minimal DMC impact, reusing Run 2's model and data for efficiency [9].

1. **Data Preparation**: Reuse Run 2's augmented dataset (3936 samples) and preprocessing pipeline.
2. **Model**: Reuse Run 2's fine-tuned ModernBERT-base.
3. **Training**: No additional training; leverage Run 2's model.
4. **Inference**: Same as Run 2. Concatenate post title/body/comments, preprocess, tokenize, output probability.
5. **Decision**: Trigger alert if current probability > 0.5, no sliding window.

### 3.5. Zero-Shot with Hybrid Decision Rule (Run 4)

Using Cloudflare Workers AI Llama-4-Scout-17B (17B parameters, 16 experts), this approach avoids training, relying on zero-shot inference for low-latency detection. A hybrid decision rule balances immediate and trend-based alerts [5].

1. **Data Preparation**: Extract user contributions from JSON (title/body for posts, comment/parent for comments), and preprocess.
2. **Model**: Use Llama-4-Scout-17B for zero-shot probability estimation. The prompts used are as follows:
   *System Prompt*

   > Respond with a valid JSON object containing a 'score' field (float, 0-1). Do not include any additional text.

   *User Prompt*

   > You are a mental health expert specializing in the detection of depression through text analysis. Your task is to evaluate the following text written by {`target_subject`} for signs of depression. Look for indicators such as expressions of sadness, hopelessness, worthlessness, guilt, fatigue, loss of interest in activities, social withdrawal, or suicidal thoughts. Consider the tone, word choice, and emotional context of the text. Assign a depression likelihood score as a float between 0 and 1, where 0 indicates no signs of depression and 1 indicates a very high likelihood of depression. Ensure your assessment is balanced and avoids overgeneralization.
   >
   > Text to analyze is {`target_text`}
   >
   > Return a JSON object with a single field `"score"` representing the likelihood of depression as a float between 0 and 1. Do not include any additional text or explanations outside the JSON object. For example: {"score": 0.5}
3. **Training**: None; rely on pretrained model.
4. **Inference**: Query model with JSON prompt for score (0–1), default to 0.5 if invalid.
5. **Decision**: Use 5-score window; alert if current score > 0.9 or mean > 0.7.

## 4. Results and Discussion

We submitted five distinct runs to address Task 2 of the eRisk 2025 challenge, focusing on the early detection of depression through contextualized analysis of Reddit discussion threads. Each run corresponds to a specific methodological approach designed to balance classification accuracy and decision latency, ranging from transformer-based models with time-aware training to zero-shot learning with large language models. Table 1 summarizes the mapping of our approaches to the official run IDs, providing a comprehensive overview of the evaluated strategies. Based on the results released [13] [14], we achieved a commendable 3rd place ranking out of 12 participating teams, underscoring the effectiveness of our hybrid methodologies in this competitive setting.

**Table 1**
Mapping of Approaches to Official Runs for our Team

| Run | Approach Description |
|---|---|
| 0 | ModernBERT with Time-Aware Loss and Class Weighting |
| 1 | Classification with Partial Information and Decision-Making Component |
| 2 | ModernBERT with Data Augmentation and Decision-Making Component |
| 3 | ModernBERT with Simple Decision Threshold |
| 4 | Llama-4-Scout-17B with Hybrid Decision Rule |

### 4.1. Decision-Based Evaluation

The decision-based evaluation metrics, derived from the eRisk 2025 results [13] [14], quantify the performance of our runs in terms of classification accuracy (Precision, Recall, $F_1$), earliness ($ERDE_5$, $ERDE_{50}$, $latency_{TP}$), speed, and latency-weighted performance ($F_{latency}$). Table 2 presents these metrics

**Table 2**
Decision-based evaluation for Task 2 (Top 3 Teams)

| Team | Run | $P$ | $R$ | $F_1$ | $ERDE_5$ | $ERDE_{50}$ | $latency_{TP}$ | $speed$ | $F_{latency}$ |
|---|---|---|---|---|---|---|---|---|---|
| HIT-SCIR | 0 | 0.72 | 0.96 | 0.82 | 0.06 | **0.03** | 4.00 | 0.99 | 0.81 |
| | 1 | 0.72 | 0.95 | 0.82 | 0.06 | **0.03** | 4.00 | 0.99 | 0.81 |
| | 2 | 0.74 | 0.94 | 0.83 | 0.06 | **0.03** | 4.00 | 0.99 | **0.82** |
| | 3 | 0.73 | 0.94 | 0.82 | 0.08 | **0.03** | 7.00 | 0.98 | 0.80 |
| | 4 | 0.77 | 0.94 | **0.85** | 0.09 | **0.03** | 8.00 | 0.97 | **0.82** |
| ELiRF-UPV | 0 | 0.78 | 0.81 | 0.79 | 0.08 | 0.04 | 7.00 | 0.98 | 0.78 |
| | 1 | 0.37 | 0.62 | 0.46 | 0.07 | 0.06 | **1.00** | **1.00** | 0.46 |
| | 2 | 0.83 | 0.47 | 0.60 | 0.10 | 0.07 | 8.00 | 0.97 | 0.58 |
| | 3 | 0.68 | 0.67 | 0.67 | 0.09 | 0.05 | 7.00 | 0.98 | 0.66 |
| | 4 | 0.68 | 0.67 | 0.67 | 0.09 | 0.05 | 7.00 | 0.98 | 0.66 |
| HU | 0 | 0.61 | 0.77 | 0.68 | 0.09 | 0.05 | 10.00 | 0.96 | 0.66 |
| | 1 | 0.72 | 0.77 | 0.75 | 0.10 | 0.05 | 11.00 | 0.96 | 0.72 |
| | 2 | 0.14 | 0.94 | 0.25 | 0.15 | 0.09 | 6.00 | 0.98 | 0.24 |
| | 3 | 0.11 | 1.00 | 0.20 | 0.11 | 0.10 | **1.00** | **1.00** | 0.20 |
| | 4 | 0.27 | 0.88 | 0.41 | 0.10 | 0.07 | 11.00 | 0.96 | 0.40 |

for the top three teams—HIT-SCIR, ELiRF-UPV, and us—highlighting the competitive landscape and our standing.

Run 1 from our team, integrating Classification with Partial Information (CPI) and a Decision-Making Component (DMC) with Llama 3.1 summarization and BERT-base-uncased classification, achieved an $F_1$ score of 0.75, securing our 3rd place ranking. This score reflects a balanced precision (0.72) and recall (0.77), demonstrating robust classification across diverse conversational contexts. The ERDE metrics ($ERDE_5$ = 0.10, $ERDE_{50}$ = 0.05) indicate competitive earliness, though they are outperformed by HIT-SCIR's best runs ($ERDE_{50}$ = 0.03 across all runs), which also achieved the highest $F_1$ of 0.85 in Run 4. ELiRF-UPV's Run 0, with an $F_1$ of 0.79, shows a strong precision-recall balance (0.78 and 0.81), but its earliness ($ERDE_{50}$ = 0.04) is slightly less optimal than HIT-SCIR's. Run 0 from our team, employing ModernBERT with time-aware loss, achieved an $F_1$ of 0.68 and an $ERDE_{50}$ of 0.05, closely aligning with Run 1's earliness but with reduced accuracy. Runs 2 and 3, despite high recall (0.94 and 1.00), suffer from low precision (0.14 and 0.11), reflecting over-prediction issues, with Run 3's simple threshold yielding the fastest $latency_{TP}$ (1.00) and perfect speed (1.00) at the cost of $F_{latency}$ (0.20). Run 4, leveraging a zero-shot Cloudflare Workers AI model (Llama-4-Scout-17B), achieved a moderate $F_1$ of 0.41 with a recall of 0.88, but its $ERDE_{50}$ of 0.07 suggests reasonable earliness, albeit with lower precision (0.27).

Compared to the top teams, our best performance (Run 1) trails HIT-SCIR's peak $F_1$ (0.85) and ELiRF-UPV's Run 0 (0.79), but our $F_{latency}$ of 0.72 is competitive with HIT-SCIR's 0.82, indicating a strong latency-weighted performance. The superior earliness of HIT-SCIR ($ERDE_{50}$ = 0.03) and ELiRF-UPV's Run 1 speed (1.00) suggest that we could improve by refining decision thresholds or leveraging faster inference mechanisms.

## 4.2. Ranking-Based Evaluation

The ranking-based evaluation metrics assess the ability to prioritize at-risk users using Precision@10 (P@10), NDCG@10, and NDCG@100 across varying numbers of writings (1, 100, 500, 1000). Table 3 presents these metrics for the top three teams, providing insights into early detection capabilities.

Run 1 from our team exhibits exceptional early detection capabilities, achieving perfect P@10 (1.00) and NDCG@10 (1.00) scores after processing a single writing, with an NDCG@100 of 0.62. This performance underscores the efficacy of the CPI+DMC approach, enhanced by Llama 3.1 summarization, in prioritizing at-risk users from minimal data. Run 0 follows with strong early metrics (P@10 = 0.90, NDCG@100 = 0.53), maintaining consistency up to 1000 writings (NDCG@100 = 0.49). HIT-SCIR

**Table 3**
Ranking-based evaluation for Task 2 (Top 3 Teams)

| Team | Run | 1 writing | | | 100 writings | | | 500 writings | | | 1000 writings | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *P*@10 | *NDCG*@10 | *NDCG*@100 | *P*@10 | *NDCG*@10 | *NDCG*@100 | *P*@10 | *NDCG*@10 | *NDCG*@100 | *P*@10 | *NDCG*@10 | *NDCG*@100 |
| HIT-SCIR | 0 | **1.00** | **1.00** | 0.58 | **1.00** | **1.00** | **0.84** | **1.00** | **1.00** | **0.89** | **1.00** | **1.00** | **0.90** |
| | 1 | **1.00** | **1.00** | 0.58 | **1.00** | **1.00** | **0.84** | **1.00** | **1.00** | **0.89** | **1.00** | **1.00** | **0.90** |
| | 2 | **1.00** | **1.00** | 0.58 | **1.00** | **1.00** | **0.84** | **1.00** | **1.00** | **0.89** | **1.00** | **1.00** | **0.90** |
| | 3 | **1.00** | **1.00** | 0.58 | **1.00** | **1.00** | **0.84** | **1.00** | **1.00** | **0.89** | **1.00** | **1.00** | **0.90** |
| | 4 | **1.00** | **1.00** | 0.58 | **1.00** | **1.00** | 0.83 | **1.00** | **1.00** | **0.89** | **1.00** | **1.00** | **0.90** |
| ELiRF-UPV | 0 | 0.90 | 0.88 | 0.36 | **1.00** | **1.00** | 0.69 | 0.90 | 0.94 | 0.74 | 0.90 | 0.81 | 0.74 |
| | 1 | 0.30 | 0.25 | 0.32 | **1.00** | **1.00** | 0.45 | **1.00** | **1.00** | 0.44 | **1.00** | **1.00** | 0.46 |
| | 2 | 0.20 | 0.31 | 0.14 | **1.00** | **1.00** | 0.45 | **1.00** | **1.00** | 0.44 | **1.00** | **1.00** | 0.46 |
| | 3 | 0.90 | 0.94 | 0.35 | **1.00** | **1.00** | 0.68 | 0.60 | 0.46 | 0.60 | 0.70 | 0.63 | 0.63 |
| | 4 | 0.60 | 0.75 | 0.27 | **1.00** | **1.00** | 0.68 | 0.60 | 0.46 | 0.60 | 0.70 | 0.63 | 0.63 |
| HU | 0 | 0.90 | 0.81 | 0.53 | 0.80 | 0.87 | 0.49 | 0.70 | 0.68 | 0.48 | 0.70 | 0.66 | 0.49 |
| | 1 | **1.00** | **1.00** | **0.62** | 0.90 | 0.88 | 0.57 | 0.60 | 0.71 | 0.35 | 0.40 | 0.60 | 0.26 |
| | 2 | 0.30 | 0.21 | 0.11 | 0.20 | 0.16 | 0.12 | 0.00 | 0.00 | 0.11 | 0.40 | 0.60 | 0.26 |
| | 3 | 0.30 | 0.21 | 0.11 | 0.20 | 0.16 | 0.12 | 0.00 | 0.00 | 0.11 | 0.40 | 0.60 | 0.26 |
| | 4 | 0.60 | 0.53 | 0.33 | 0.40 | 0.58 | 0.36 | 0.30 | 0.37 | 0.24 | 0.40 | 0.60 | 0.26 |

demonstrates superior long-term ranking performance, with all runs achieving an NDCG@100 of 0.90 across 1000 writings, reflecting a robust ability to sustain prioritization over time. ELiRF-UPV's Run 0 excels early with an NDCG@100 of 0.36 after one writing, but its performance improves to 0.74 by 1000 writings, indicating a more stable ranking capability compared to our runs. Runs 2 and 3 from our team show poor early performance (NDCG@10 = 0.21) and a complete drop by 500 writings, while Run 4 achieves moderate early scores (NDCG@100 = 0.33) but converges to 0.26 by 1000 writings, aligning with Runs 1, 2 and 3.

Against the top teams, our runs perform well early but fall short as writings increase (except for run 0), suggesting that our approaches require enhancement for long-term stability. The convergence of our runs to similar NDCG@100 scores (0.26) by 1000 writings also indicates a common challenge in maintaining ranking quality with increased data.

## 4.3. Discussion

The results affirm that Run 1 from our team, with an $F_1$ of 0.75 and $F_{\text{latency}}$ of 0.72, represents our strongest approach, securing a 3rd place ranking among 12 teams. The integration of Llama 3.1 for summarization effectively captures contextual nuances, while the DMC component balances accuracy and earliness, as evidenced by its competitive $ERDE_{50}$ (0.05) and perfect early ranking scores (P@10 = 1.00, NDCG@10 = 1.00). Run 0's solid performance ($F_1$ = 0.68, $ERDE_{50}$ = 0.05) highlights the potential of time-aware training, suggesting a promising direction for refinement. However, Runs 2 and 3's low precision (0.14 and 0.11) despite high recall (0.94 and 1.00) indicates over-prediction, with Run 3's speed (1.00) and $\text{latency}_{TP}$ (1.00) compromised by an $F_{\text{latency}}$ of 0.20. Run 4's zero-shot approach with Cloudflare Workers AI model (Llama-4-Scout-17B) offers moderate performance ($F_1$ = 0.41, $ERDE_{50}$ = 0.07), but its reliance on pretrained knowledge limits precision (0.27), underscoring the need for task-specific adaptation.

Compared to HIT-SCIR, whose Run 4 achieves the highest $F_1$ (0.85) and $ERDE_{50}$ (0.03), and ELiRF-UPV's Run 0 ($F_1$ = 0.79), our best runs demonstrate competitive accuracy but lag in earliness and long-term ranking stability (NDCG@100 = 0.90 for HIT-SCIR vs. 0.26 for us at 1000 writings). The

superior earliness of HIT-SCIR and ELiRF-UPV's Run 1 speed (1.00) suggest that we could improve by refining decision thresholds or leveraging faster inference mechanisms. The degradation in ranking metrics over time across all our runs points to the need for dynamic models, potentially incorporating temporal embeddings or ensemble techniques combining Run 1's accuracy with Run 3's speed.

## 5. Conclusion

Our exploration of hybrid and time-aware approaches for eRisk 2025 Task 2 has yielded meaningful insights into the challenge of detecting depression early within Reddit's dynamic conversational threads. Achieving an F1 score of 0.75 with our CPI+DMC approach, securing a 3rd place ranking among 12 teams, underscores the potential of integrating Llama 3.1 summarization with BERT-based classification to navigate complex social media contexts. The competitive $ERDE_{50}$ of 0.05 and strong early ranking metrics (P@10 = 1.00, NDCG@10 = 1.00) reflect our success in balancing timely detection with accuracy, a critical step toward supporting timely mental health interventions. While our approaches faced challenges in long-term ranking stability, these findings highlight opportunities to refine decision-making policies and model architectures. This work not only contributes to the growing field of computational mental health, but also reinforces the importance of context-aware, real-time systems in addressing global mental health challenges. We are optimistic that continued research will build on these foundations, paving the way for more effective tools to identify and support individuals at risk.

## 6. Future Work

The promising results of our study open several avenues for advancing early depression detection in conversational contexts. A primary direction is the development of ensemble methods that integrate the strengths of our five approaches. For instance, combining the time-aware precision of Run 0's ModernBERT with the contextual summarization of Run 1's CPI+DMC framework could yield a model that excels in both earliness and accuracy. Exploring ModernBERT's full 8192-token capacity may further enhance context capture, particularly for complex Reddit threads, though this would require optimizing computational efficiency to ensure scalability in real-time applications. Another promising area is refining data augmentation strategies, as seen in Run 2, to reduce noise and improve generalization across diverse user expressions. For Run 4's zero-shot approach, fine-tuning large language models like Llama-4-Scout-17B on eRisk-specific data could address precision limitations, potentially bridging the gap with supervised methods. Additionally, incorporating temporal embeddings to model user behavior over time could improve long-term ranking stability, addressing the degradation observed in our NDCG@100 scores. Beyond technical enhancements, we aim to explore cross-domain applications, such as adapting our models for other mental health conditions or platforms like X, to broaden the impact of real-time monitoring. These directions collectively aim to create more robust, adaptive systems for early intervention in mental health.

## Acknowledgments

We would like to acknowledge the support provided by the Office of Research (OoR) at Habib University, Karachi, Pakistan for funding this project through the internal research grant IRG-2235.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References


[1] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, A. Giachanou, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, 2023, pp. 294–315. doi:10.1007/978-3-031-42448-9_22.

[2] M. Trotzek, S. Koitka, C. M. Friedrich, Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences , IEEE Transactions on Knowledge & Data Engineering 32 (2020) 588–601. URL: https://doi.ieeecomputersociety.org/10.1109/TKDE.2018.2885515. doi:10.1109/TKDE.2018.2885515.

[3] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, BERT-based transformers for early detection of mental health illnesses, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, 2021, pp. 189–200. doi:10.1007/978-3-030-85251-1_15.

[4] S. Devaguptam, T. Kogatam, N. Kotian, A. Kumar, Early detection of depression using bert and deberta, in: Conference and Labs of the Evaluation Forum, 2022. URL: https://api.semanticscholar.org/CorpusID:251471697.

[5] S. M. Shah, S. A. Gillani, M. S. A. Baig, M. A. Saleem, M. H. Siddiqui, Advancing depression detection on social media platforms through fine-tuned large language models, 2024. URL: http://arxiv.org/abs/2409.14794. doi:10.48550/arXiv.2409.14794. arXiv:2409.14794 [cs].

[6] C. Tank, S. Pol, V. Katoch, S. Mehta, A. Anand, R. R. Shah, Depression detection and analysis using large language models on textual and audio-visual modalities, 2024. URL: http://arxiv.org/abs/2407.06125. doi:10.48550/arXiv.2407.06125. arXiv:2407.06125 [cs].

[7] E. Zhang, C. Poellabauer, Multimodal depression detection with contextual position encoding and latent space regularization, 2025. URL: https://openreview.net/forum?id=miOYgWl60q.

[8] H. Thompson, M. Errecalde, A time-aware approach to early detection of anorexia: UNSL at eRisk 2024, 2024. URL: http://arxiv.org/abs/2410.17963. doi:10.48550/arXiv.2410.17963. arXiv:2410.17963 [cs].

[9] J. M. Loyola, S. Burdisso, H. Thompson, L. C. Cagnina, M. Errecalde, UNSL at eRisk 2021: A comparison of three early alert policies for early risk detection., in: CLEF (working notes), 2021, pp. 992–1021. URL: https://ceur-ws.org/Vol-2936/paper-81.pdf.

[10] T. Gui, Q. Zhang, L. Zhu, X. Zhou, M. Peng, X. Huang, Depression detection on social media with reinforcement learning, in: Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings, Springer-Verlag, 2019, pp. 613–624. URL: https://doi.org/10.1007/978-3-030-32381-3_49. doi:10.1007/978-3-030-32381-3_49.

[11] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 9822, Springer International Publishing, 2016, pp. 28–39. URL: http://link.springer.com/10.1007/978-3-319-44564-9_3. doi:10.1007/978-3-319-44564-9_3, series Title: Lecture Notes in Computer Science.

[12] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media (2018) 495–503. URL: https://dl.acm.org/doi/10.1145/3159652.3159725. doi:10.1145/3159652.3159725, conference Name: WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining ISBN: 9781450355810 Place: Marina Del Rey CA USA Publisher: ACM.

[13] 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.

[14] 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.