Overview of eRisk 2026 Early Risk Prediction on the Internet: Symptom Ranking and Conversational Approaches for Depression and ADHD (Extended Overview)*

eRisk Extended Overview at CLEF 2026

Anxo Perez[1], Javier Parapar[1], Xi Wang[2] and Fabio Crestani[3]

[1]*Information Retrieval Lab (IRLab), Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña, A Coruña, Spain*

[2]*University of Sheffield, Sheffield, England, United Kingdom*

[3]*Faculty of Informatics, Universitá della Svizzera italiana (USI), Switzerland*

Abstract

This paper presents a condensed overview of eRisk 2026, the tenth edition of the CLEF lab on early risk detection. Since its foundation, eRisk has provided a shared evaluation framework for developing and assessing methods, metrics, and resources for the early identification of personal risks, with a particular focus on mental health and safety-related scenarios. The 2026 edition continues the evolution of the lab toward more contextual, conversational, and interpretable settings. It includes three tasks: i) an official edition of Conversational Depression Detection via LLMs, where participants interact with finetuned LLM personas to infer depression severity and depressive symptoms. ii) The second edition of Contextualized Early Detection of Depression task, where systems process full conversational threads rather than isolated user posts. Finally, iii) a new ADHD Symptom Sentence Ranking task, which extends symptom-oriented retrieval to the 18 symptoms of the Adult ADHD Self-Report Scale. Together, these tasks aim to advance research on early risk detection by promoting methods that can identify clinically meaningful evidence, exploit conversational context, and support symptom-aware analysis across different mental health conditions.

Keywords

Early risk detection, Depression, Conversational analysis, LLMs, Large Language Models, eRisk

1. Introduction

The eRisk lab was created as a benchmark environment to explore evaluation methodologies, new models, and resources for the timely detection of personal health risk situations. Early alert technologies are increasingly relevant in domains focused on safety and personal health, where even small gains in anticipation can support earlier and more effective interventions. Typical use cases range from monitoring mental-health signals to flagging harmful online behaviours, including self-harm risk, eating disorders, pathological gambling, or other psychological conditions. eRisk has traditionally focused on psychological and mental-health risks, where language provides subtle but valuable evidence about a user's state. Since the relationship between linguistic expression and mental condition is complex, robust automatic screening and retrieval tools remain necessary, especially those that can provide interpretable evidence rather than isolated predictions.

The inaugural eRisk 2017 task explored early depression detection with a novel evaluation framework and dataset [1, 2]. In 2018, the scope expanded to eating disorders [3, 4]. The 2019 campaign consolidated

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

*Corresponding author.

†These authors contributed equally.

anxo.pvila@udc.es (A. Perez); javierparapar@udc.es (J. Parapar); xi.wang@sheffield.ac.uk (X. Wang); fabio.crestani@usi.ch (F. Crestani)

0000-0002-0480-006X (A. Perez); 0000-0002-5997-8252 (J. Parapar); 0000-0001-5936-9919 (X. Wang); 0000-0001-8672-0700 (F. Crestani)

work on anorexia detection, introduced self-harm prediction, and proposed a novel task that inferred questionnaire responses on depression severity from social-media activity [5, 6, 7]. The 2020 edition deepened work on self-harm while adding a depression severity estimation task [8, 9, 10]. A year later, in 2021, attention turned to pathological gambling and self-harm, alongside an updated severity estimation task [11, 12, 13]. In 2022, the lab revisited pathological gambling and depression and introduced a new severity estimation challenge for eating-disorder content [14, 15, 16]. The 2023 campaign shifted emphasis toward a sentence-ranking task for individual depression symptoms, complemented by early gambling risk detection and eating-disorder severity estimation [17, 18, 19]. The 2024 lab continued the sentence-ranking challenge covering the 21 BDI-II symptoms, together with an anorexia early detection task and eating-disorder severity estimation track [20, 21, 22]. Finally, eRisk 2025 marked an important transition toward richer contextual and conversational scenarios, introducing contextualized early detection of depression from full conversational threads and a pilot task based on interacting with LLM personas [23, 24].

One of the strengths of eRisk is its interdisciplinary nature, bringing together researchers from information retrieval, natural language processing, machine learning, psychology, and medicine. This combination of expertise has enabled the design of tasks and evaluation protocols that are both technically challenging and clinically meaningful. Across editions, eRisk has fostered the development of computational approaches for early-warning scenarios, where systems aim to identify linguistic signals associated with mental-health deterioration in online environments. Previous work based on eRisk data and tasks has explored a wide range of methods, from traditional machine-learning and feature-based approaches to neural architectures, temporal models, and transformer-based systems [25, 26, 27, 28, 29, 30]. In addition, eRisk has also served as the basis for the creation of relevant resources and datasets for mental-health NLP, including collections and annotations aimed at depression severity estimation, symptom-level analysis, and early risk detection research [31, 32, 33].

The present edition, eRisk 2026 [34, 35], continues this evolution and consolidates the lab's focus on contextual, conversational, and symptom-aware evaluation. This year, the lab included three tasks. Task 1, *Conversational Depression Detection via LLMs*, became an official task after its 2025 pilot edition. Participants interacted with fine-tuned LLM personas and had to infer depression severity and active depressive symptoms from the resulting conversations. Task 2, *Contextualized Early Detection of Depression*, continued for a second edition, requiring systems to process full Reddit discussion threads rather than isolated user writings. Task 3, *ADHD Symptom Sentence Ranking*, introduced a new symptom-oriented retrieval setting focused on the 18 symptoms of the Adult ADHD Self-Report Scale (ASRS-v1.1), extending the sentence-ranking paradigm beyond depression to another mental health condition.

Together, the tasks proposed in eRisk 2026 reflect three central directions for the future of early risk detection. First, the lab moves beyond isolated posts toward conversational and contextual evidence. Second, the evaluation emphasizes interpretable symptom-level analysis, where systems must identify the linguistic evidence associated with specific clinical constructs. Third, this edition explores new evaluation scenarios involving LLM-based simulated personas, enabling controlled interaction while raising new challenges related to latency, questioning strategies, and clinically meaningful inference. The following sections describe the 2026 tasks, datasets, assessment procedures, participating teams, and official results.

2. Task 1: Conversational Depression Detection via LLMs

Task 1 builds on the pilot task introduced in eRisk 2025, where participants interacted with LLM-based personas to detect signs of depression through conversation. In 2026, this setting was consolidated as the first official edition of the task. The main change with respect to the pilot edition was the release of the simulated personas as downloadable models, rather than as hosted conversational agents. This design aimed to improve accessibility, reproducibility, and comparability across participating systems. The task asked participants to determine whether each LLM persona exhibited signs of depression and, if so, to estimate both the overall depression severity and the active depressive symptoms expressed

during the conversation. The diagnostic target was framed in terms of the Beck Depression Inventory-II (BDI-II) [36], as in previous eRisk tasks involving depression severity estimation. The BDI-II is a 21-item self-report questionnaire widely used in clinical psychology. Each item corresponds to a concrete symptom, such as *Sadness*, *Loss of Energy*, or *Indecisiveness*. The 21 symptoms considered in the task are listed in Table 1.

Table 1
The 21 BDI-II depression symptoms.

21 Depression Symptoms		
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

Each BDI-II symptom is scored from 0 to 3 according to severity. The sum of the 21 items yields a global score in the range 0–63. Following the usual BDI-II interpretation, scores were grouped into four severity categories: *minimal depression* (0–9), *mild depression* (10–18), *moderate depression* (19–29), and *severe depression* (30–63). The 2026 collection was balanced across these categories, with five personas assigned to each severity level. Therefore, participating systems were evaluated at two complementary levels: the estimation of the overall BDI-II depression level and the identification of up to four active depressive symptoms expressed by each persona.

As training material, we released the resources associated with the ChatGPT-based personas used in the 2025 pilot task through the TalkDep repository[1] [37]. These resources provided participants with examples of the previous prompted-persona setting and allowed them to study the type of conversational behaviour expected in the task. However, the 2026 test personas were not accessed through ChatGPT. Instead, the official personas were distributed as LoRA adapters on top of `meta-llama/Meta-Llama-3-8B-Instruct` through the eRisk 2026 Hugging Face collection[2]. Participants were required to download the corresponding adapters and run the models locally. The 20 personas were released gradually during the evaluation period, in batches of two personas per weekly window.

Unlike traditional classification tasks based on a fixed collection of user writings, this task required participants to actively interact with the persona before making a decision. Participants could conduct multiple conversational turns with each model and design their own questioning strategies. However, the task encouraged systems to reach accurate conclusions as early as possible. In the 2026 edition, this aspect was explicitly incorporated through latency-aware metrics, which are described in the evaluation subsection. The task was not intended to reward direct diagnostic questioning. Personas were designed to avoid explicit self-diagnosis and could react uncomfortably to overly direct questions about depression or mental health. Instead, participants had to infer the possible condition by analysing the persona's language, tone, expressed thoughts, and reactions throughout the interaction.

Each team could submit up to three runs. At most one run could involve manual intervention or human assistance, while the remaining runs had to be fully automatic. Moreover, each submitted run had to operate independently, without sharing information across methods. For each persona, participants submitted both the conversation log and a structured prediction file. The log preserved the interaction with the persona, while the prediction file contained the estimated BDI-II score and the selected BDI-II symptoms that best explained the inferred depression profile.

[1] https://github.com/Anxo06/TalkDep/
[2] https://huggingface.co/collections/irlab-udc/erisk2026

2.1. Task 1: LLM Personas Design and Construction

The 2026 task included 20 synthetic LLM personas. The persona collection was designed using the same clinician-in-the-loop principles established in the 2025 pilot edition, following the TalkDep framework [37]. However, the implementation was substantially revised. In the pilot task, personas were instantiated as prompted LLM agents: the model was instructed with a profile describing the persona, its background, and its depressive symptomatology. In 2026, by contrast, each persona was released as a fine-tuned model, implemented as a LoRA adapter over `Meta-Llama-3-8B-Instruct`. The personas were created in collaboration with clinicians, and based on synthetic profiles and conversations. Each profile included biographical information, contextual background, communicative style, an overall BDI-II score, and up to four key depressive symptoms. These profiles were then used to generate synthetic training data representing how the persona would interact in a therapeutic or interview-like conversation. To construct these data, we used `Gemini 3.0 Pro` as an external LLM, conditioning it on each persona profile and prompting it to generate dialogue examples consistent with the corresponding background, communicative style, BDI-II severity level, and key depressive symptoms. Importantly, `Gemini 3.0 Pro` was used only to generate synthetic fine-tuning conversations; the released personas themselves were LoRA adapters over `Meta-Llama-3-8B-Instruct`.

Clinicians participated throughout the construction and validation process. The 20 persona profiles were validated by clinical experts, and the generated synthetic conversations were also reviewed in successive rounds. This validation step was introduced to ensure that the generated training samples were clinically meaningful, coherent with the intended BDI-II profile, and suitable for fine-tuning. Feedback from these review rounds was used to refine the generation pipeline and to accept only samples considered appropriate for representing each persona.

The final collection contained five personas in each BDI-II severity category: minimal, mild, moderate, and severe. This balanced design made it possible to evaluate systems across a broad spectrum of simulated depression profiles, including both control-like personas and personas expressing clinically relevant depressive symptoms. The final ground truth for each persona consisted of its BDI-II score, the corresponding severity category, and the set of active BDI-II symptoms.

2.2. Task 1: Participant teams

Table 2 shows the participating teams and several statistics about their interactions with the LLM personas, including the number of submitted runs, the number of personas covered in each run, whether any run involved manual intervention, the mean number of messages per conversation, and the mean number of characters per message. Overall, the numbers reveal a diverse set of interaction strategies, ranging from short and direct conversations to longer multi-turn dialogues. In this table, messages correspond only to those sent by the participant's system to the LLM persona.

A total of 21 teams participated in this first official edition of the task. Most teams submitted multiple runs, with 16 teams submitting three runs. Several teams covered the complete set of 20 personas in all their submitted runs, namely Awakened, BUAP-CRC, DS-GT, SINAI, RecEarlyMind, lexxy, pj-mathematician, and upb-uottawa. Other teams submitted almost complete runs, such as INSA-Lyon, INSALyon-2, and PMH-PsychAI-Lab, with 19 personas per run, or erisk-cedri, with 18 personas per run. By contrast, teams such as SVNITDS100, SymptomAI, Techbbas, The-Interrogators, VANGUARD, Xtra, and NoirByte submitted partial runs, suggesting either exploratory participation, incomplete processing of the released personas, or selective evaluation strategies.

The interaction statistics show substantial variation in the number of messages exchanged before producing a decision. The most extensive conversational strategies were followed by lexxy and upb-uottawa, with averages of 29.4 and 25.6 participant messages per conversation, respectively. INSALyon-2 also relied on relatively long dialogues, with 23.4 messages on average, while VANGUARD used 15.0 messages per conversation. At the other end of the spectrum, SINAI, NoirByte, SVNITDS100, erisk-cedri, and Awakened used shorter interactions, all below five and a half participant messages per conversation on average. These differences suggest that participants explored substantially different

Table 2
Task 1: participating teams, number of runs, number of personas submitted per run, presence of manual intervention, mean number of messages per conversation, and mean number of characters per message.

Team	#Runs	Personas/run	Manual	Interaction statistics	
			run(s)	msgs./conv.	chars./msg.
Amor-Fati	3	20; 16; 13	#1	9.3	108.76
Awakened	2	20; 20	–	5.3	84.37
BUAP-CRC	3	20; 20; 20	#3	13.2	140.82
DS-GT	3	20; 20; 20	–	11.5	210.15
INSA-Lyon	3	19; 19; 19	–	6.2	156.96
INSALyon-2	3	19; 19; 19	–	23.4	80.41
NoirByte	1	12	–	4.6	97.71
PMH-PsychAI-Lab	1	19	–	8.0	131.16
RecEarlyMind	1	20	–	7.5	93.23
SINAI	3	20; 20; 20	–	4.4	242.61
SVNITDS100	3	6; 6; 6	#1	4.9	114.51
SymptomAI	3	2; 2; 8	–	–	119.41
Techbbas	3	6; 6; 6	–	8.0	186.49
The-Interrogators	3	6; 6; 6	#1	9.3	206.94
UNED-GELP	3	20; 20; 18	–	10.7	114.47
VANGUARD	3	2; 2; 18	–	15.0	80.62
Xtra	1	2	–	6.0	192.46
erisk-cedri	3	18; 18; 18	–	4.8	82.79
lexxy	3	20; 20; 20	–	29.4	93.84
pjmathematician	3	20; 20; 20	–	6.0	108.88
upb-uottawa	3	20; 20; 20	#1	25.6	119.85

trade-offs between early decision-making and evidence gathering.
Finally, several teams included manual or human-assisted components in one of their runs, as allowed by the task rules. This was the case for Amor-Fati, BUAP-CRC, SVNITDS100, The-Interrogators, and upb-uottawa. The inclusion of both automated and manual-in-the-loop submissions makes it possible to compare fully automatic conversational strategies against approaches where human judgement guided part of the interaction or decision process.

2.3. Evaluation Metrics

Based on evaluation metrics that have been developed from eRisk 2019 [5], which involved the use of BDI-II questionnaires and scores, we extend and develop the evaluation approaches as follows:

- **Depression Category Hit Rate (DCHR)**: Based on the four depression level categories that we have discussed, from minimal depression to severe depression, this effectiveness measure examines the fraction of cases where the BDI-II scores describing simulated personas estimated by the participants lie in the correct depression category.
- **Average DODL (ADODL)**: For this pilot task, we reuse the Average Difference between Overall Depression Levels (ADODL), which measures the closeness between the actual and estimated depression level for effectiveness measurement. The ADODL is calculated by following: $CR = (MAD - |ADL - EDL|)/MAD$, where $|ADL - EDL|$ calculates the absolute value between the Actual Depression Level (ADL) and the Estimated Depression Level (EDL). Then divided by Maximum Absolute Difference (i.e., 63) to obtain a normalised evaluation score in [0,1]. For example, if a simulated persona has a minor depression severity (depression level 5) and a participant estimates the depression level is 9, the DODL is calculated as $(63 - |9 - 5|)/63 = 0.9365$.

- **Average Symptom Hit Rate (ASHR)**: For the last effectiveness measure, aside from estimating the depression level of simulated personas as per BDI-II scores, this pilot task also involves the identification of major depression symptoms of simulated personas. Hence, SHR calculates the ratio of cases where the participants can correctly identify the major symptoms of the simulated personas. For example, each simulated persona has four major symptoms. If a participant accurately identifies two of them, then the SHR equals 2/4 = 0.5.

2.3.1. Latency-aware Evaluation

In addition to the standard effectiveness metrics, we introduce two latency-aware variants that penalize late decisions. In this task, participants could interact with each LLM persona for an arbitrary number of messages before submitting their final prediction. Therefore, systems that reach accurate conclusions after fewer user messages should be rewarded over systems that require long conversations to obtain the same result.

For each persona $u \in U$, let k_u be the number of messages sent by the participant's system to the LLM persona before producing the final prediction. Following the latency penalty used in early risk detection, we define:

$$penalty(k_u) \quad = \quad -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}}, \tag{1}$$

where p controls how quickly the penalty increases with the number of messages. A decision made after the first user message receives no penalty, i.e., $penalty(1) = 0$. The corresponding speed factor is:

$$speed(k_u) \quad = \quad 1 - penalty(k_u). \tag{2}$$

Thus, $speed(k_u)$ is close to 1 for early decisions and decreases as the number of messages increases. In this first edition, we set p so that the speed factor is equal to 0.5 after $K_0 = 10$ user messages. This value was chosen based on the observed interaction statistics of the submitted runs: many systems used approximately 10–30 messages per conversation, while a smaller number of systems relied on substantially longer interactions. Therefore, $K_0 = 10$ provides a balanced threshold that rewards early decisions while avoiding an excessively harsh penalty for systems that conduct moderately long exploratory conversations.

More formally, we set:

$$p \quad = \quad \frac{\ln(3)}{K_0 - 1} = \frac{\ln(3)}{30 - 1} \simeq 0.0379. \tag{3}$$

With this setting, a decision made after the first user message receives the full effectiveness score ($speed(1) = 1$), while a decision made after 30 user messages receives half of the original score ($speed(30) = 0.5$). Decisions made after longer conversations receive progressively lower speed factors.

Let $DCHR_u$ be the persona-level Depression Category Hit Rate score, where $DCHR_u = 1$ if the estimated BDI-II score falls into the correct depression category and $DCHR_u = 0$ otherwise. We define the latency-weighted category hit rate as:

$$LDCHR \quad = \quad \frac{1}{|U|} \sum_{u \in U} DCHR_u \cdot speed(k_u). \tag{4}$$

Similarly, let SHR_u be the persona-level Symptom Hit Rate, computed as the proportion of key symptoms correctly identified for persona u. We define the latency-weighted average symptom hit rate as:

$$LASHR = \frac{1}{|U|} \sum_{u \in U} SHR_u \cdot speed(k_u). \quad (5)$$

These two metrics preserve the interpretation of the original measures while incorporating the cost of delayed interaction. A system that correctly identifies the depression category or symptoms after a short conversation will obtain latency-aware scores close to its original DCHR or ASHR. Conversely, a system that reaches the same prediction only after many messages will receive lower latency-aware scores. In this way, $LDCHR$ and $LASHR$ evaluate not only the correctness of the final prediction, but also the efficiency of the conversational strategy used to obtain it.

2.4. Task 1: Results

The results are reported separately for complete and incomplete submissions. Table 3 includes only those runs that submitted predictions for all 20 LLM personas, while Table 4 reports runs that covered fewer than 20 personas. This distinction is important because the metrics are computed over the personas submitted by each run. Therefore, scores obtained from a very small subset of personas are not directly comparable to those obtained over the full evaluation set. The complete-submission table provides the most reliable basis for comparing systems under the full task setting, whereas the incomplete-submission table is useful for documenting partial participation and exploratory runs. In addition to the original effectiveness metrics, both tables include the latency-aware variants LDCHR and LASHR, which penalize correct decisions that are reached only after longer conversations. Each metric value is also accompanied by its rank within the corresponding table, allowing the reader to identify not only the best overall run but also the relative position of each run for each individual metric.

For complete submissions, the best ADODL score was achieved by PJMATHEMATICIAN in run #3, with an ADODL of 0.9254. The strongest DCHR result was obtained by BUAP-CRC in run #2, with a score of 0.6000, while the best ASHR was achieved by UPB-UOTTAWA in run #3, with 0.3375. When considering the more restrictive latency-aware metrics, with $K_0 = 10$, PJMATHEMATICIAN remained the strongest team: runs #1 and #3 obtained the highest LDCHR value, 0.3872, and run #2 achieved the highest LASHR value, 0.2288. This indicates that PJMATHEMATICIAN combined strong prediction quality with comparatively efficient conversational strategies. However, the latency-aware rankings also show that some systems with good original scores were penalized substantially when they required longer interactions. For example, LEXXY and UPB-UOTTAWA achieved competitive ASHR values, but their LASHR ranks dropped due to longer conversations. Overall, PJMATHEMATICIAN, DS-GT, and BUAP-CRC obtained the highest ADODL values among full-coverage submissions, while the LDCHR and LASHR columns highlight the additional importance of reaching accurate decisions early.

For incomplete submissions, some runs obtained very high values, such as SYMPTOMAI run #2 with an ADODL of 0.9683 and VANGUARD run #2 with a DCHR of 1.0000. However, both results were computed over only 2 out of the 20 personas, and should therefore be interpreted with caution. The same caveat applies to the latency-aware metrics: although XTRA run #1 obtained the highest LASHR value, 0.1760, and VANGUARD run #2 obtained a strong LDCHR value, both runs covered only two personas. Among the partial submissions with broader coverage, UNED-GELP run #2 covered 18 personas and achieved an ADODL of 0.8836 and the best LDCHR value in the incomplete table, 0.4550. Similarly, ERISK-CEDRI submitted three 18-persona runs, with its run #3 obtaining an ADODL of 0.8686 and an LDCHR of 0.2817. The two INSA-LYON variants covered 19 personas per run, with ADODL values mostly in the 0.83–0.85 range, although their latency-aware scores varied depending on the length of the conversations. These results show that several incomplete submissions were close to the full-coverage setting, whereas others correspond to very small subsets and mainly serve as partial evidence of system behaviour.

Table 3
Evaluation for Task 1 considering only runs that submitted predictions for all 20 LLM personas. Teams are ordered by their best ADODL run. Ranks for each metric are shown in parentheses. '*' indicates manual or human-assisted runs.

Team	Run	DCHR	ADODL	ASHR	LDCHR	LASHR
pjmathematician	#1	0.5500 (2)	0.9071 (2)	0.3125 (3)	**0.3872 (1)**	0.2200 (2)
	#2	0.4500 (5)	0.9040 (5)	0.3250 (2)	0.3168 (3)	**0.2288 (1)**
	#3	0.5500 (2)	**0.9254 (1)**	0.3000 (4)	**0.3872 (1)**	0.2112 (3)
DS-GT	#1	0.4500 (5)	0.8841 (11)	0.2500 (8)	0.2091 (10)	0.1138 (8)
	#2	0.4000 (6)	0.8968 (8)	0.2500 (8)	0.1696 (15)	0.1142 (7)
	#3	0.5500 (2)	0.9063 (3)	0.1875 (11)	0.2493 (6)	0.0777 (14)
BUAP-CRC	#1	0.5000 (3)	0.9056 (4)	0.1000 (16)	0.2028 (11)	0.0310 (19)
	#2	**0.6000 (1)**	0.9024 (6)	0.1500 (15)	0.2409 (7)	0.0575 (15)
	#3*	0.4500 (5)	0.8929 (9)	0.1750 (12)	0.1761 (14)	0.0548 (16)
SINAI	#1	0.5000 (3)	0.8905 (10)	0.1875 (11)	0.2500 (5)	0.0938 (11)
	#2	0.4000 (6)	0.8738 (13)	0.2250 (10)	0.2000 (13)	0.1125 (9)
	#3	0.4737 (4)	0.8989 (7)	0.1711 (13)	0.2368 (8)	0.0855 (12)
UNED-GELP	#0	0.4000 (6)	0.8040 (22)	0.1875 (11)	0.0572 (16)	0.0268 (20)
	#1	0.4500 (5)	0.8968 (8)	0.1625 (14)	0.3168 (3)	0.1144 (6)
Awakened	#1	0.2500 (9)	0.8325 (17)	0.1000 (16)	0.2167 (9)	0.0807 (13)
	#2	0.5000 (3)	0.8802 (12)	0.2625 (7)	0.3374 (2)	0.1709 (4)
lexxy	#1	0.3000 (8)	0.8556 (14)	0.1875 (11)	0.0137 (22)	0.0122 (24)
	#2	0.3000 (8)	0.8492 (15)	0.2500 (8)	0.0144 (21)	0.0162 (23)
	#3	0.3000 (8)	0.8413 (16)	0.2875 (5)	0.0186 (20)	0.0196 (22)
Amor-Fati	#1*	0.4500 (5)	0.8254 (18)	0.2375 (9)	0.2514 (4)	0.1410 (5)
upb-uottawa	#1	0.3500 (7)	0.8222 (20)	0.2750 (6)	0.0376 (17)	0.0315 (18)
	#2	0.3000 (8)	0.8230 (19)	0.2625 (7)	0.0304 (18)	0.0355 (17)
	#3	0.3000 (8)	0.7849 (23)	**0.3375 (1)**	0.0214 (19)	0.0241 (21)
RecEarlyMind	#1	0.3000 (8)	0.8071 (21)	0.1625 (14)	0.2013 (12)	0.0977 (10)

Table 4
Evaluation for Task 1 considering runs that submitted predictions for fewer than 20 LLM personas. Teams are ordered by their best ADODL run. Ranks for each metric are shown in parentheses. '*' indicates manual or human-assisted runs.

Team	Run	Personas	DCHR	ADODL	ASHR	LDCHR	LASHR
SymptomAI	#1	2/20	0.0000 (18)	0.8810 (6)	0.1250 (13)	0.0000 (26)	0.0625 (18)
	#2	2/20	0.5000 (5)	**0.9683 (1)**	0.2500 (4)	0.2500 (7)	0.1250 (6)
	#3	8/20	0.3750 (8)	0.8294 (22)	0.2500 (4)	0.1875 (13)	0.1250 (6)
VANGUARD	#1	2/20	0.5000 (5)	0.8810 (6)	0.1250 (13)	0.1533 (17)	0.0383 (23)
	#2	2/20	**1.0000 (1)**	0.9286 (2)	0.0000 (19)	0.3066 (3)	0.0000 (27)
	#3	18/20	0.2222 (13)	0.7319 (29)	0.1806 (7)	0.0681 (21)	0.0554 (19)
SVNITDS100	#1	6/20	0.2857 (11)	0.8594 (10)	0.1429 (12)	0.1969 (12)	0.0714 (17)
	#2	6/20	0.5000 (5)	0.9259 (3)	0.0417 (17)	0.2500 (7)	0.0208 (26)
	#3*	5/20	0.2000 (15)	0.9111 (4)	0.1500 (10)	0.1521 (18)	0.1199 (8)
UNED-GELP	#2	18/20	0.5556 (4)	0.8836 (5)	0.1667 (8)	**0.4550 (1)**	0.1365 (5)
Techbbas	#1	6/20	0.3333 (9)	0.8280 (23)	0.1667 (8)	0.1990 (11)	0.0995 (12)
	#2	6/20	0.1667 (16)	0.7646 (27)	0.0000 (19)	0.0995 (19)	0.0000 (27)
	#4	6/20	0.1667 (16)	0.8730 (7)	0.1250 (13)	0.0995 (19)	0.0746 (16)
The-Interrogators	#1*	6/20	0.6000 (3)	0.8667 (9)	0.0500 (16)	0.3000 (4)	0.0250 (24)
	#2	6/20	0.6000 (3)	0.8413 (18)	0.0500 (16)	0.3000 (4)	0.0250 (24)
	#3	6/20	0.8000 (2)	0.8730 (7)	0.0500 (16)	0.4000 (2)	0.0250 (24)
erisk-cedri	#1	18/20	0.1111 (17)	0.8325 (20)	0.1667 (8)	0.0878 (20)	0.1200 (7)
	#2	18/20	0.2222 (13)	0.8474 (14)	0.1806 (7)	0.1724 (16)	0.1496 (3)
	#3	18/20	0.3889 (6)	0.8686 (8)	0.0972 (15)	0.2817 (5)	0.0758 (14)
NoirByte	#1	12/20	0.3333 (9)	0.8585 (11)	0.1042 (14)	0.2730 (6)	0.0841 (13)
Amor-Fati	#2	16/20	0.3750 (8)	0.8552 (12)	0.2656 (3)	0.2219 (9)	0.1385 (4)
	#3	13/20	0.3846 (7)	0.8462 (17)	0.1923 (6)	0.2040 (10)	0.1096 (10)
INSALyon-2	#1	19/20	0.2105 (14)	0.8145 (24)	0.2763 (2)	0.0203 (24)	0.0542 (20)
	#2	19/20	0.3158 (10)	0.8463 (16)	0.2500 (4)	0.0679 (22)	0.0513 (21)
	#3	19/20	0.3158 (10)	0.8496 (13)	0.2237 (5)	0.0486 (23)	0.0495 (22)
INSA-Lyon	#1	19/20	0.3158 (10)	0.8388 (19)	0.1579 (9)	0.2285 (8)	0.1142 (9)
	#2	19/20	0.2632 (12)	0.8304 (21)	0.1447 (11)	0.1767 (15)	0.1013 (11)
	#3	19/20	0.2632 (12)	0.8471 (15)	0.2237 (5)	0.1826 (14)	0.1525 (2)
Xtra	#1	2/20	0.0000 (18)	0.8016 (25)	0.2500 (4)	0.0000 (26)	**0.1760 (1)**
upb-uottawa	#1*	18/20	0.2222 (13)	0.7989 (26)	**0.3472 (1)**	0.0159 (25)	0.0248 (25)
PMH-PsychAI-Lab	#1	19/20	0.2105 (14)	0.7327 (28)	0.0091 (18)	0.1724 (16)	0.0754 (15)

3. Task 2: Contextualized Early Detection of Depression

In 2026, eRisk hosted the second edition of the Contextualized Early Detection of Depression task. This task introduces a more realistic scenario for depression detection by incorporating full conversational contexts. Whereas earlier eRisk editions released isolated writings authored by a target user, this task provides the *entire Reddit discussion thread* in which the target user participated. Consequently, both in the training and test phases, systems had access not only to the messages written by the target user, but also to the contributions of other users and to the interaction structure linking the messages.
This design was motivated by the observation that the clinical relevance of a message often becomes apparent only when it is interpreted together with the surrounding conversation. For example, a seemingly neutral reply may become informative when it appears as a response to criticism, emotional disclosure, or a request for help. The task therefore aims to simulate scenarios in which depression detection may require analysing exchanges among multiple participants, rather than treating each user post as an isolated piece of evidence. This setting introduces additional challenges, since systems must model not only the textual content of individual messages, but also the conversational context in which they occur and the way this context affects the interpretation of possible depressive signals.
The task was divided into two phases:

- During the training phase, participants were provided with the test dataset from the 2025 edition of this task, consisting of contextualized writings from depressed and control users. This dataset allowed teams to develop and tune their approaches under the same contextualized setting used in the official evaluation.
- The test phase was carried out interactively. For each target user, the server released a sequence of Reddit discussion threads over time. Each thread constituted a *submission round*. After processing the newly released thread for each target user, together with all information available up to that point, teams had to return a depression label (*positive* or *negative*) and a confidence score before the next thread became available.

3.1. Task 2: Evaluation Metrics

3.1.1. Decision-based Evaluation

This evaluation approach uses the binary decisions made by the participating systems for each user. In addition to standard classification measures such as Precision, Recall, and F1 score (computed with respect to the positive class), we also calculate ERDE (Early Risk Detection Error), used in previous editions of the lab. A detailed description of ERDE was presented by Losada and Crestani in [38]. ERDE is an error measure that incorporates a penalty for delayed correct alerts (true positives). The penalty increases with the delay in issuing the alert, measured by the number of user posts processed before making the alert.
Since 2019, we complemented the evaluation report with additional decision-based metrics that try to capture additional aspects of the problem. These metrics try to overcome some limitations of $ERDE$, namely:

- the penalty associated to true positives goes quickly to 1. This is due to the functional form of the cost function (sigmoid).
- a perfect system, which detects the true positive case right after the first round of messages (first chunk), does not get error equal to 0.
- with a method based on releasing data in a chunk-based way (as it was done in 2017 and 2018) the contribution of each user to the performance evaluation has a large variance (different for users with few writings per chunk vs users with many writings per chunk).
- $ERDE$ is not interpretable.

Some research teams have analysed these issues and proposed alternative ways for evaluation. Trotzek and colleagues [39] proposed $ERDE_o^{\%}$. This is a variant of ERDE that does not depend on the number of user writings seen before the alert but, instead, it depends on the *percentage* of user writings seen before the alert. In this way, user's contributions to the evaluation are normalized (currently, all users weight the same). However, there is an important limitation of $ERDE_o^{\%}$. In real life applications, the overall number of user writings is not known in advance. Social Media users post contents online and screening tools have to make predictions with the evidence seen. In practice, you do not know when (and if) a user's thread of messages is exhausted. Thus, the performance metric should not depend on knowledge about the total number of user writings.

Another proposal of an alternative evaluation metric for early risk prediction was done by Sadeque and colleagues [40]. They proposed $F_{latency}$, which fits better with our purposes. This measure is described next.

Imagine a user $u \in U$ and an early risk detection system that iteratively analyzes u's writings (e.g. in chronological order, as they appear in Social Media) and, after analyzing k_u user writings ($k_u \geq 1$), takes a binary decision $d_u \in \{0, 1\}$, which represents the decision of the system about the user being a risk case. By $g_u \in \{0, 1\}$, we refer to the user's golden truth label. A key component of an early risk evaluation should be the delay on detecting true positives (we do not want systems to detect these cases too late). Therefore, a first and intuitive measure of delay can be defined as follows[3]:

$$\text{latency}_{TP} = \text{median}\{k_u : u \in U, d_u = g_u = 1\} \tag{6}$$

This measure of latency is calculated over the true positives detected by the system and assesses the system's delay based on the median number of writings that the system had to process to detect such positive cases. This measure can be included in the experimental report together with standard measures such as Precision (P), Recall (R) and the F-measure (F):

$$P = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : d_u = 1|} \tag{7}$$

$$R = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : g_u = 1|} \tag{8}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{9}$$

Furthermore, Sadeque et al. proposed a measure, $F_{latency}$, which combines the effectiveness of the decision (estimated with the F measure) and the delay[4] in the decision. This is calculated by multiplying F by a penalty factor based on the median delay. More specifically, each individual (true positive) decision, taken after reading k_u writings, is assigned the following penalty:

$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}} \tag{10}$$

where p is a parameter that determines how quickly the penalty should increase. In [40], p was set such that the penalty equals 0.5 at the median number of posts of a user[5]. Observe that a decision right after the first writing has no penalty (i.e. $penalty(1) = 0$). Figure 1 plots how the latency penalty increases with the number of observed writings.

The system's overall speed factor is computed as:

[3] Observe that Sadeque et al (see [40], pg 497) computed the latency for all users such that $g_u = 1$. We argue that latency should be computed only for the true positives. The false negatives ($g_u = 1$, $d_u = 0$) are not detected by the system and, therefore, they would not generate an alert.

[4] Again, we adopt Sadeque et al.'s proposal but we estimate latency only over the true positives.

[5] In the evaluation we set p to 0.0078, a setting obtained from the eRisk 2017 collection.

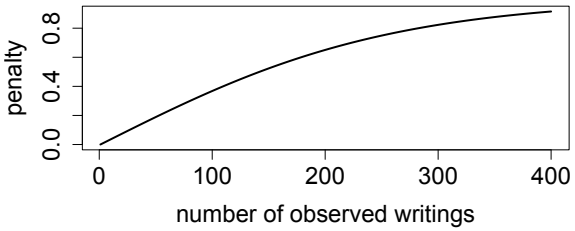

Figure 1: Latency penalty increases with the number of observed writings (k_u)

$$speed \quad = \quad (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\}) \tag{11}$$

where speed equals 1 for a system whose true positives are detected right at the first writing. A slow system, which detects true positives after hundreds of writings, will be assigned a speed score near 0. Finally, the *latency-weighted* F score is simply:

$$F_{latency} \quad = \quad F \cdot speed \tag{12}$$

Since 2019 user's data were processed by the participants in a post by post basis (i.e. we avoided a chunk-based release of data). Under these conditions, the evaluation approach has the following properties:

- smooth grow of penalties;
- a perfect system gets $F_{latency} = 1$;
- for each user u the system can opt to stop at any point k_u and, therefore, now we do not have the effect of an imbalanced importance of users;
- $F_{latency}$ is more interpretable than $ERDE$.

3.1.2. Ranking-based Evaluation

In addition to the evaluation discussed above, we employed an alternative form of evaluation to further assess the systems. After each data release (new user writing, that is post or comment), participants were required to provide the following information for each user in the collection:

- A decision for the user (alert or no alert), which was used to calculate the decision-based metrics discussed previously.
- A score representing the user's level of risk, estimated based on the evidence observed thus far.

The scores were used to create a ranking of users in descending order of estimated risk. For each participating system, a ranking was generated at each data release point, simulating a continuous re-ranking approach based on the observed evidence. In a real-life scenario, this ranking would be presented to an expert user who could make decisions based on the rankings (e.g., by inspecting the top of the rankings). Each ranking can be evaluated using standard ranking metrics such as P@10 or NDCG. Therefore, we report the performance of the systems based on the rankings after observing different numbers of writings.

Table 5
Task 2: participating teams, number of runs, number of threads processed by the team, and lapse of time taken for the entire process.

Team	#Runs	#User threads	Lapse of time (from 1st to last response)
Awakened	1	118	0 days 23:12
erisk-cedri	5	500	0 days 01:47
VANGUARD	2	129	0 days 12:42
HUTECH-NLP	5	500	2 days 03:02
Lotu-ixa	5	500	0 days 04:24
UET-PsyWar	5	500	0 days 06:57
DeepCare	5	500	0 days 09:14
UNED-GELP	5	500	0 days 10:27
MacEarlyDetect	5	121	0 days 21:51
INSA-Lyon	5	500	6 days 01:45
NoirByte	5	289	53 days 18:49
NYCU-NLP	5	247	11 days 08:35
HUGETIME	1	500	2 days 20:14
LCDAA	5	500	3 days 12:07
MindAILab	1	302	0 days 01:15
Snugi-AI-v2	5	500	0 days 01:26
TeamZed	5	4	0 days 00:53

3.2. Task 2: Participant teams

Table 5 reports the teams that participated in Task 2, together with the number of submitted runs, the number of user threads processed, and the lapse of time between each team's first and last response. As in the previous edition, this lapse of time provides an approximate indication of the level of automation and computational efficiency of the submitted systems.

A total of 17 teams participated in this second edition of the contextualized early detection task. Most teams submitted the maximum number of runs allowed, with 12 teams submitting five runs. In terms of coverage, 10 teams processed the complete set of 500 user threads: erisk-cedri, HUTECH-NLP, Lotu-ixa, UET-PsyWar, DeepCare, UNED-GELP, INSA-Lyon, HUGETIME, LCDAA, and Snugi-AI-v2. The remaining teams processed a partial subset of the test collection, ranging from 4 to 302 user threads. The fastest complete submissions were produced by Snugi-AI-v2 and erisk-cedri, both completing the task in under two hours, followed by Lotu-ixa, UET-PsyWar, and DeepCare, which completed the full set in under ten hours. These results suggest that, despite the additional complexity introduced by full conversational contexts, efficient end-to-end processing remains feasible. At the other end of the spectrum, teams such as INSA-Lyon, LCDAA, NYCU-NLP, and especially NoirByte required substantially longer processing times, which may reflect more computationally demanding pipelines, manual intervention, or interruptions during the interactive evaluation process. Finally, TeamZed submitted five runs but processed only four user threads, and should therefore be interpreted as a very limited partial submission.

3.3. Task 2: Results

Table 6 show the decision-based results of task 2, ordered in terms of best F1. Table 7 shows the ranking-based results, ordered by $NDCG@100$.

Overall, the results show that Task 2 remains challenging, particularly when systems must combine accurate classification with early decisions. In the decision-based evaluation, HUGETIME obtained the best F_1 score, closely followed by UNED-GELP, NoirByte, MindAILab, and INSA-Lyon. The results reveal different trade-offs: some runs achieved high recall with lower precision, while others were more conservative and obtained higher precision but missed more positive cases. The latency-aware scores further show that timely decisions are essential, with HUGETIME and UNED-GELP combining strong

Table 6
Task 2 decision-based evaluation ordered in terms of best F_1.

Team	Run	P	R	F_1	$ERDE_5$	$ERDE_{50}$	latency$_{TP}$	$speed$	$F_{latency}$
HUGETIME	0	0.80	0.87	**0.83**	0.15	**0.05**	9.00	0.97	**0.81**
UNED-GELP	0	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00
	1	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00
	2	0.79	0.85	0.82	0.11	0.06	4.00	0.99	**0.81**
	3	0.78	0.71	0.75	0.10	0.08	2.00	**1.00**	0.74
	4	0.81	0.69	0.75	0.11	0.08	3.00	0.99	0.74
NoirByte	0	0.88	0.76	0.82	0.15	0.09	20.00	0.93	0.76
	1	0.93	0.70	0.80	0.16	0.10	24.00	0.91	0.73
	2	0.93	0.62	0.74	0.16	0.11	28.50	0.89	0.66
	3	0.94	0.54	0.69	0.17	0.12	27.00	0.90	0.62
	4	0.93	0.42	0.58	0.17	0.13	35.50	0.87	0.50
MindAILab	0	0.84	0.80	0.82	0.16	0.08	16.00	0.94	0.77
INSA-Lyon	0	0.76	0.85	0.80	0.16	0.07	13.00	0.95	0.76
	1	0.69	0.89	0.78	0.15	**0.05**	9.00	0.97	0.75
	2	0.46	0.95	0.62	0.14	0.08	7.50	0.97	0.60
	3	0.63	0.92	0.75	0.16	0.06	11.50	0.96	0.72
	4	0.75	0.85	0.80	0.16	0.07	14.00	0.95	0.76
UET-PsyWar	0	0.62	0.87	0.72	0.15	0.08	15.00	0.95	0.68
	1	0.57	0.88	0.69	0.16	0.08	15.00	0.95	0.65
	2	0.73	0.82	0.77	0.15	0.07	15.00	0.95	0.73
	3	0.76	0.80	0.78	0.14	0.08	14.00	0.95	0.74
	4	0.60	0.88	0.71	0.15	0.07	15.00	0.95	0.68
LCDAA	0	0.66	0.81	0.73	0.11	0.07	3.00	0.99	0.72
	1	0.69	0.81	0.74	0.11	0.07	4.00	0.99	0.74
	2	0.78	0.78	0.78	0.13	0.08	10.00	0.96	0.75
	3	0.70	0.70	0.70	0.09	0.07	**1.00**	**1.00**	0.70
	4	0.72	0.74	0.73	0.11	0.07	2.00	**1.00**	0.73
Lotu-ixa	0	0.37	0.96	0.54	0.13	0.07	3.00	0.99	0.53
	1	0.31	0.97	0.47	0.14	0.08	3.00	0.99	0.46
	2	0.38	0.93	0.54	0.14	0.08	6.00	0.98	0.53
	3	0.63	0.86	0.73	0.14	0.07	12.00	0.96	0.70
	4	0.75	0.77	0.76	0.12	0.07	6.50	0.98	0.74
erisk-cedri	0	0.73	0.77	0.75	0.10	0.07	1.50	**1.00**	0.75
	1	0.78	0.51	0.61	0.13	0.10	4.50	0.99	0.60
	2	0.85	0.44	0.58	0.14	0.11	5.50	0.98	0.57
	3	0.84	0.34	0.48	0.16	0.12	9.00	0.97	0.47
	4	**1.00**	0.11	0.20	0.17	0.15	9.00	0.97	0.19
VANGUARD	0	0.62	0.90	0.74	0.14	**0.05**	6.50	0.98	0.72
	1	0.95	0.45	0.61	0.17	0.11	16.00	0.94	0.58
HUTECH-NLP	0	0.31	0.99	0.48	0.09	0.07	**1.00**	**1.00**	0.48
	1	0.48	0.96	0.64	0.09	**0.05**	3.00	0.99	0.63
	2	0.77	0.70	0.74	0.11	0.06	4.00	0.99	0.73
	3	0.40	0.93	0.56	**0.08**	0.06	**1.00**	**1.00**	0.56
	4	0.43	0.92	0.59	0.09	**0.05**	3.00	0.99	0.58
Snugi-AI-v2	0	0.72	0.71	0.72	0.18	0.08	8.00	0.97	0.70
	1	0.77	0.69	0.73	0.17	0.09	14.00	0.95	0.69
	2	0.71	0.66	0.69	0.18	0.07	8.00	0.97	0.67
	3	0.80	0.65	0.72	0.17	0.07	8.00	0.97	0.70
	4	0.66	0.65	0.65	0.18	0.08	8.00	0.97	0.63
DeepCare	0	0.92	0.51	0.65	0.11	0.09	**1.00**	**1.00**	0.65
	1	0.54	0.89	0.67	0.15	0.06	7.00	0.98	0.66
	2	0.96	0.25	0.40	0.16	0.13	8.00	0.97	0.39
	3	0.92	0.53	0.67	0.16	0.10	10.00	0.96	0.65
	4	0.95	0.57	0.71	0.13	0.09	5.00	0.98	0.70
NYCU-NLP	0	0.43	0.79	0.55	0.16	0.09	11.00	0.96	0.53
	1	0.47	0.60	0.53	0.17	0.11	15.00	0.95	0.50
	2	0.49	0.44	0.46	0.18	0.12	13.50	0.95	0.44
	3	0.49	0.64	0.56	0.17	0.11	15.00	0.95	0.52
	4	0.49	0.64	0.56	0.17	0.11	15.00	0.95	0.52
Awakened	0	0.33	0.68	0.45	0.11	0.10	**1.00**	**1.00**	0.45
MacEarlyDetect	0	0.21	0.98	0.35	0.16	0.11	2.00	**1.00**	0.35
	1	0.21	0.99	0.35	0.16	0.12	2.00	**1.00**	0.34
	2	0.22	0.98	0.36	0.17	0.11	3.00	0.99	0.36
	3	0.19	**1.00**	0.33	0.13	0.13	**1.00**	**1.00**	0.33
	4	0.19	**1.00**	0.33	0.13	0.13	**1.00**	**1.00**	0.33
TeamZed	0	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00
	1	0.42	0.12	0.19	0.16	0.16	3.00	0.99	0.19
	2	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00
	3	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00
	4	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00

classification performance with early detection. In the ranking-based evaluation, DeepCare achieved the best results across the different writing cutoffs, while MindAILab, NoirByte, INSA-Lyon, LCDAA,

Table 7
Task 2 ranking-based evaluation ordered by best $NDCG@100$ at 500 writings.

		1 writing			100 writings			250 writings			500 writings		
Team	Run	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
DeepCare	0	**1.00**	**1.00**	**0.74**	**1.00**	**1.00**	**0.88**	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	**0.91**
	1	0.90	0.93	0.65	**1.00**	**1.00**	0.79	**1.00**	**1.00**	0.83	**1.00**	**1.00**	0.85
	2	0.90	0.93	0.65	**1.00**	**1.00**	0.79	**1.00**	**1.00**	0.83	**1.00**	**1.00**	0.85
	3	0.90	0.93	0.67	**1.00**	**1.00**	0.79	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.85
	4	**1.00**	**1.00**	0.65	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.83	**1.00**	**1.00**	0.85
MindAILab	0	**1.00**	**1.00**	0.69	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.89	**1.00**	**1.00**	0.90
NoirByte	0	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.88	**1.00**	**1.00**	0.89
	1	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.88	**1.00**	**1.00**	0.89
	2	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.88	**1.00**	**1.00**	0.89
	3	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.88	**1.00**	**1.00**	0.89
	4	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.86	**1.00**	**1.00**	0.88	**1.00**	**1.00**	0.89
INSA-Lyon	0	0.80	0.87	0.64	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.87	**1.00**	**1.00**	0.89
	1	0.80	0.87	0.64	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.87	**1.00**	**1.00**	0.89
	2	0.90	0.81	0.55	**1.00**	**1.00**	0.80	**1.00**	**1.00**	0.83	**1.00**	**1.00**	0.85
	3	0.80	0.87	0.64	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.87	**1.00**	**1.00**	0.89
	4	0.80	0.87	0.63	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.87	**1.00**	**1.00**	0.89
LCDAA	0	0.90	0.92	0.61	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.87	0.90	0.94	0.88
	1	0.90	0.92	0.61	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.87	0.90	0.94	0.88
	2	0.90	0.94	0.63	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.86	0.90	0.94	0.88
	3	0.90	0.92	0.71	0.90	0.93	0.79	0.90	0.94	0.79	0.90	0.88	0.84
	4	0.90	0.93	0.73	0.90	0.93	0.78	0.90	0.92	0.79	0.90	0.86	0.81
HUTECH-NLP	0	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.85
	1	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.85
	2	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.85
	3	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.85
	4	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.84	**1.00**	**1.00**	0.85
Snugi-AI-v2	0	0.20	0.22	0.20	**1.00**	**1.00**	0.76	**1.00**	**1.00**	0.78	**1.00**	**1.00**	0.79
	1	0.20	0.22	0.20	**1.00**	**1.00**	0.76	**1.00**	**1.00**	0.78	**1.00**	**1.00**	0.79
	2	0.20	0.22	0.20	**1.00**	**1.00**	0.76	**1.00**	**1.00**	0.78	**1.00**	**1.00**	0.79
	3	0.20	0.22	0.20	**1.00**	**1.00**	0.80	**1.00**	**1.00**	0.82	**1.00**	**1.00**	0.85
	4	0.20	0.22	0.20	**1.00**	**1.00**	0.73	0.90	0.92	0.74	0.90	0.88	0.72
UNED-GELP	0	0.00	0.00	0.11	0.20	0.19	0.11	0.10	0.06	0.10	0.10	0.06	0.12
	1	0.30	0.37	0.36	0.30	0.36	0.42	0.30	0.31	0.38	0.60	0.59	0.42
	2	0.90	0.94	0.68	1.00	1.00	0.87	1.00	1.00	0.82	1.00	1.00	0.83
	3	0.90	0.94	0.69	1.00	1.00	0.88	1.00	1.00	0.85	1.00	1.00	0.84
	4	0.90	0.93	0.72	1.00	1.00	0.87	1.00	1.00	0.83	1.00	1.00	0.83
erisk-cedri	0	0.80	0.86	0.65	0.90	0.94	0.81	**1.00**	**1.00**	0.84	0.90	0.94	0.84
	1	0.70	0.72	0.60	0.90	0.93	0.77	0.90	0.94	0.77	0.90	0.94	0.76
	2	0.80	0.87	0.63	**1.00**	**1.00**	0.77	**1.00**	**1.00**	0.76	**1.00**	**1.00**	0.77
	3	0.80	0.87	0.63	**1.00**	**1.00**	0.75	**1.00**	**1.00**	0.73	**1.00**	**1.00**	0.75
	4	0.80	0.87	0.63	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.71	**1.00**	**1.00**	0.70
VANGUARD	0	1.00	1.00	0.71	0.90	0.94	0.75	0.90	0.94	0.75	0.90	0.94	0.75
	1	1.00	1.00	0.67	0.90	0.94	0.80	0.90	0.94	0.80	0.90	0.94	0.80
HUGETIME	0	0.70	0.81	0.32	0.60	0.44	0.73	0.60	0.44	0.79	0.60	0.44	0.79
Awakened	0	1.00	1.00	0.54	0.90	0.94	0.63	0.90	0.94	0.63	0.90	0.94	0.63
NYCU-NLP	0	0.60	0.62	0.46	0.70	0.79	0.46	0.80	0.76	0.47	0.80	0.76	0.51
	1	0.60	0.62	0.46	0.70	0.79	0.46	0.80	0.76	0.47	0.70	0.76	0.49
	2	0.60	0.62	0.46	0.70	0.79	0.46	0.80	0.76	0.47	0.70	0.76	0.49
	3	0.60	0.62	0.48	0.70	0.75	0.46	0.70	0.75	0.48	0.70	0.75	0.51
	4	0.60	0.62	0.48	0.70	0.75	0.46	0.70	0.75	0.48	0.70	0.75	0.51
Lotu-ixa	0	0.90	0.88	0.63	0.60	0.68	0.56	0.80	0.82	0.41	0.00	0.00	0.23
	1	0.90	0.94	0.60	0.80	0.86	0.59	0.80	0.87	0.42	0.00	0.00	0.25
	2	0.90	0.94	0.60	0.80	0.86	0.59	0.80	0.87	0.42	0.00	0.00	0.25
	3	0.90	0.94	0.60	0.80	0.86	0.59	0.80	0.87	0.42	0.00	0.00	0.25
	4	**1.00**	**1.00**	0.67	0.90	0.88	0.58	0.90	0.88	0.49	0.80	0.85	0.38
MacEarlyDetect	0	0.20	0.16	0.27	0.20	0.18	0.39	0.00	0.00	0.23	0.00	0.00	0.23
	1	0.20	0.16	0.27	0.20	0.18	0.39	0.00	0.00	0.23	0.00	0.00	0.23
	2	0.20	0.16	0.27	0.20	0.18	0.39	0.00	0.00	0.23	0.00	0.00	0.23
	3	0.50	0.58	0.37	0.90	0.94	0.55	0.00	0.00	0.25	0.00	0.00	0.25
	4	0.50	0.58	0.37	0.90	0.94	0.55	0.00	0.00	0.25	0.00	0.00	0.25
UET-PsyWar	0	0.60	0.71	0.32	0.50	0.64	0.30	0.00	0.00	0.23	0.00	0.00	0.23
	1	0.60	0.61	0.30	0.50	0.64	0.30	0.00	0.00	0.23	0.00	0.00	0.23
	2	0.60	0.60	0.30	0.50	0.64	0.30	0.00	0.00	0.23	0.00	0.00	0.23
	3	0.60	0.60	0.30	0.50	0.64	0.30	0.00	0.00	0.23	0.00	0.00	0.23
	4	0.60	0.61	0.30	0.50	0.64	0.30	0.00	0.00	0.23	0.00	0.00	0.23
TeamZed	0	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20
	1	0.50	0.57	0.30	0.50	0.57	0.30	0.50	0.57	0.30	0.50	0.57	0.30
	2	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20
	3	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20
	4	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20	0.20	0.22	0.20

and HUTECH-NLP also obtained strong rankings. These results suggest that conversational context is useful for prioritizing at-risk users, although producing accurate, calibrated, and timely binary decisions remains difficult.

4. Task 3: ADHD Symptom Sentence Ranking

In 2026, we introduced a new sentence-level ranking task focused on the retrieval of evidence related to ADHD symptoms. This task extends the symptom-oriented retrieval setting explored in previous eRisk editions. In this case, we target the ADHD disorder, and the 18 symptoms defined in the Adult ADHD Self-Report Scale (ASRS-v1.1). Participants were asked to rank candidate sentences according to their relevance to each ADHD symptom. A sentence was considered relevant when it conveyed information about the user's state with respect to the target symptom, independently of polarity or stance. Therefore, both sentences expressing the presence of a symptom and sentences providing evidence about its absence could be relevant, as long as they were informative for the corresponding symptom.

The task was designed to encourage systems to identify clinically meaningful evidence at sentence granularity, rather than relying only on surface keyword matching. This is especially important in ADHD-related language, where symptoms such as inattention, impulsivity, disorganization, restlessness, or difficulty sustaining effort may be expressed indirectly, through everyday experiences and self-descriptions. By formulating the task as a ranking problem, participants were required to prioritize the most informative candidate sentences for each symptom.

As this was the first edition of the ADHD symptom ranking task, no annotated training data was provided. Participants received the test inputs only, following the formatting conventions of recent eRisk ranking tasks. For each of the 18 ASRS-v1.1 symptoms, teams submitted a ranking of up to 1000 candidate sentences, ordered by decreasing estimated relevance. Relevance assessments were then produced using top-k pooling and expert annotation, and systems were evaluated using standard information retrieval metrics such as MAP@1000, nDCG@1000, and P@10.

4.1. Task 3: Dataset

The dataset released for this task consisted of sentence-tagged documents derived from publicly available social media writings collected to contain ADHD-related expressions. The corpus was provided in a format compatible with previous eRisk ranking tasks, where each sentence receives a unique identifier that participants use in their submitted rankings. Table 8 presents the main statistics of the corpus.

4.2. Task 3: Assessment Process

Participants were given the corpus of candidate sentences and the description of the 18 symptoms from the ASRS-v1.1 questionnaire. They were free to decide the best strategy to represent each ADHD symptom as a query or retrieval target. Each team could submit up to 5 variants, or runs. Each run included 18 TREC-style formatted rankings of sentences, one ranking per ADHD symptom, as illustrated in Figure 2. For each symptom, participants submitted up to 1000 results sorted by estimated relevance.

```
1   Q0  sent_000251_0    0001    10.0    myGroupNameMyMethodName
1   Q0  sent_085820_3    0002     9.4    myGroupNameMyMethodName
...
18  Q0  sent_015320_2    0998     1.2    myGroupNameMyMethodName
18  Q0  sent_022313_9    1000     0.9    myGroupNameMyMethodName
```

Figure 2: Example of a participant's run for Task 3.

Table 8
Corpus statistics for Task 3: ADHD Symptom Sentence Ranking.

Number of sentences	4 170 875
Average number of words per sentence	13.05
Number of ADHD symptoms	18

Table 9
Task 3: participating teams and number of submitted runs for the ADHD Symptom Sentence Ranking task.

Team	#Runs
AI-MH-Z	4
Awakened	2
DS-GT	5
INSA-Lyon	5
NaiveNeuron	5
NeuroRankUB	4
PMH-PsychAI-Lab	1
Silvermask	1
SymptomAI	4
UET-PsyWar	5
VANGUARD	4
erisk-cedri	5
Total	45

The relevance assessment process followed the same methodology used in recent eRisk sentence-ranking tasks. For each ADHD symptom, a pool of candidate sentences was created from the highest-ranked results submitted by the participating systems. These pooled sentences were then manually assessed by expert annotators. A sentence was judged as relevant when it provided evidence about the user's state with respect to the target ADHD symptom, regardless of whether the sentence indicated the presence, absence, uncertainty, or contextual manifestation of that symptom.
All sentences included in the assessment pool were independently labelled by three expert assessors. From these annotations, we derived two versions of the relevance judgments. In the *majority-vote* version, a sentence was considered relevant if at least two of the three assessors judged it as relevant. In the *unanimity* version, a sentence was considered relevant only if all three assessors agreed on its relevance. Therefore, the unanimity version constitutes a more restrictive evaluation setting, focusing only on sentences with full agreement among assessors.
The resulting relevance judgments were used to evaluate the submitted rankings with standard information retrieval metrics. In particular, we report Mean Average Precision (MAP), normalized Discounted Cumulative Gain (nDCG), including nDCG@1000, and Precision at 10 (P@10). These measures capture complementary aspects of ranking quality: MAP rewards systems that retrieve relevant sentences throughout the ranking, nDCG@1000 emphasizes the ordering of relevant evidence near the top of the ranking, and P@10 reflects the quality of the highest-ranked sentences that would be most immediately inspected by a user or expert. We received 45 runs from 12 participating teams, as shown in Table 9.

4.3. Task 3: Results

The submitted runs were evaluated under the two relevance-judgment settings derived from the expert annotations. Table 10 reports the results obtained with the majority-vote relevance judgments, where a sentence is considered relevant when at least two of the three assessors marked it as relevant. Table 11 reports the results obtained with the unanimity relevance judgments, where only sentences marked as relevant by all three assessors are considered relevant. The unanimity setting is therefore more restrictive and provides a complementary view of system performance on sentences with stronger

Table 10
Task 3 results using majority-vote relevance judgments. A sentence is considered relevant if at least two of the three expert assessors marked it as relevant.

Team	Run	AP	R-PREC	P@10	NDCG@1000
NeuroRankUB	final ADHD ranking	0.248	0.328	0.528	0.547
NeuroRankUB	gte rrf final	0.203	0.264	0.450	0.516
erisk-cedri	EnsembleV3	0.190	0.262	0.400	0.461
erisk-cedri	EnsembleV2	0.186	0.259	0.406	0.459
NeuroRankUB	ADHD submission	0.171	0.242	0.417	0.475
erisk-cedri	EnsembleV1	0.167	0.239	0.372	0.446
INSA-Lyon	Ensemble	0.159	0.206	0.317	0.406
INSA-Lyon	LLM cascade	0.158	0.180	0.344	0.402
INSA-Lyon	HiPerT full	0.137	0.169	0.294	0.388
INSA-Lyon	BiEnc baseline	0.117	0.160	0.311	0.321
erisk-cedri	VecSearch	0.105	0.162	0.278	0.350
erisk-cedri	NLI zero-shot	0.097	0.162	0.267	0.379
NeuroRankUB	final ranking	0.092	0.185	0.283	0.312
DS-GT	asrs run5	0.085	0.115	0.300	0.207
AI-MH-Z	ST equal weights	0.076	0.110	0.239	0.317
AI-MH-Z	ST option weights	0.075	0.110	0.228	0.316
AI-MH-Z	ST variable weights	0.074	0.109	0.228	0.313
UET-PsyWar	run5 RRF robust	0.072	0.134	0.272	0.270
UET-PsyWar	run4 hybrid key bonus	0.069	0.121	0.222	0.267
UET-PsyWar	run3 balanced hybrid	0.068	0.121	0.222	0.264
SymptomAI	run4	0.067	0.096	0.300	0.215
INSA-Lyon	DepTransfer	0.054	0.087	0.100	0.216
SymptomAI	run1	0.052	0.098	0.194	0.267
UET-PsyWar	run2 CE only	0.050	0.107	0.200	0.236
Silvermask	v2 semantic	0.045	0.090	0.167	0.241
NaiveNeuron	ENSEMBLE	0.044	0.094	0.144	0.220
DS-GT	asrs run4	0.043	0.078	0.194	0.146
DS-GT	asrs run2	0.038	0.075	0.189	0.134
NaiveNeuron	RERANK MultiQA	0.037	0.090	0.139	0.204
SymptomAI	run3	0.037	0.056	0.189	0.155
AI-MH-Z	ST by description	0.035	0.062	0.094	0.150
NaiveNeuron	Rerank BioMed	0.034	0.081	0.150	0.152
NaiveNeuron	HYBRID RRF	0.031	0.073	0.128	0.127
PMH-PsychAI-Lab	run1	0.029	0.075	0.117	0.189
UET-PsyWar	run1 sim only	0.029	0.072	0.106	0.172
VANGUARD	run3	0.028	0.065	0.156	0.113
DS-GT	asrs run1	0.026	0.044	0.094	0.109
DS-GT	asrs run3	0.023	0.058	0.122	0.112
VANGUARD	run5	0.023	0.049	0.111	0.106
VANGUARD	run1	0.014	0.033	0.056	0.089
VANGUARD	run4	0.014	0.030	0.083	0.052
SymptomAI	run2	0.012	0.041	0.056	0.094
Awakened	run1	0.006	0.015	0.022	0.057
Awakened	run2	0.006	0.019	0.028	0.058
NaiveNeuron	S-Biomed-Roberta	0.006	0.011	0.022	0.095

assessor agreement.

5. Participant Teams

Table 12 reports the participating teams and the runs that they submitted for each eRisk task. The next paragraphs give a brief summary on the methods implemented by each of them. Further details are available at the CLEF 2026 working notes proceedings for the participants.

SymptomAI [62]. The SymptomAI team, from Habib University, participated in Task 3, submitting four runs based on a progressively enhanced retrieval pipeline for ranking sentences according to the 18 ASRS-

Table 11
Task 3 results using unanimity relevance judgments. A sentence is considered relevant only if all three expert assessors marked it as relevant.

Team	Run	AP	R-PREC	P@10	NDCG@1000
NeuroRankUB	final ADHD ranking	0.207	0.274	0.367	0.488
NeuroRankUB	gte rrf final	0.184	0.233	0.306	0.466
NeuroRankUB	ADHD submission	0.166	0.232	0.311	0.439
erisk-cedri	EnsembleV3	0.138	0.193	0.261	0.387
INSA-Lyon	LLM cascade	0.134	0.167	0.272	0.369
erisk-cedri	EnsembleV2	0.134	0.191	0.278	0.384
INSA-Lyon	Ensemble	0.129	0.160	0.244	0.363
INSA-Lyon	BiEnc baseline	0.122	0.154	0.256	0.313
erisk-cedri	EnsembleV1	0.122	0.190	0.233	0.372
INSA-Lyon	HiPerT full	0.107	0.131	0.172	0.336
erisk-cedri	VecSearch	0.090	0.148	0.200	0.304
DS-GT	asrs run5	0.080	0.121	0.217	0.188
UET-PsyWar	run4 hybrid key bonus	0.079	0.122	0.156	0.276
AI-MH-Z	ST equal weights	0.078	0.108	0.167	0.291
AI-MH-Z	ST option weights	0.078	0.105	0.161	0.289
AI-MH-Z	ST variable weights	0.076	0.102	0.167	0.288
NeuroRankUB	final ranking	0.073	0.117	0.183	0.268
UET-PsyWar	run5 RRF robust	0.072	0.113	0.200	0.262
UET-PsyWar	run3 balanced hybrid	0.071	0.128	0.167	0.268
SymptomAI	run4	0.070	0.093	0.206	0.192
UET-PsyWar	run2 CE only	0.058	0.099	0.161	0.242
erisk-cedri	NLI zero-shot	0.057	0.093	0.117	0.297
INSA-Lyon	DepTransfer	0.054	0.065	0.078	0.202
DS-GT	asrs run4	0.052	0.078	0.133	0.142
NaiveNeuron	ENSEMBLE	0.049	0.093	0.117	0.208
DS-GT	asrs run2	0.048	0.065	0.144	0.135
SymptomAI	run1	0.048	0.072	0.133	0.234
NaiveNeuron	Rerank BioMed	0.043	0.093	0.122	0.159
NaiveNeuron	RERANK MultiQA	0.042	0.082	0.111	0.191
Silvermask	v2 semantic	0.041	0.070	0.094	0.199
DS-GT	asrs run1	0.040	0.057	0.083	0.116
NaiveNeuron	HYBRID RRF	0.037	0.076	0.089	0.133
VANGUARD	run3	0.034	0.071	0.117	0.115
AI-MH-Z	ST by description	0.031	0.051	0.072	0.129
PMH-PsychAI-Lab	run1	0.030	0.066	0.083	0.180
SymptomAI	run3	0.030	0.040	0.128	0.126
VANGUARD	run5	0.029	0.051	0.078	0.110
DS-GT	asrs run3	0.026	0.053	0.083	0.100
UET-PsyWar	run1 sim only	0.020	0.044	0.050	0.149
VANGUARD	run1	0.018	0.031	0.039	0.090
VANGUARD	run4	0.018	0.027	0.044	0.056
Awakened	run1	0.009	0.017	0.017	0.054
NaiveNeuron	S-Biomed-Roberta	0.006	0.010	0.017	0.090
Awakened	run2	0.005	0.011	0.017	0.043
SymptomAI	run2	0.005	0.015	0.028	0.066

v1.1 ADHD symptoms. Run 1 was a cosine-similarity baseline using all-MiniLM-L6-v2 embeddings. Run 2 used a clinical-domain embedding model, S-PubMedBERT-MS-MARCO, together with query expansion based on clinical synonyms and informal paraphrases, and encoded the surrounding context of each sentence. Run 3 added first-person boosting and GPT-4o-mini re-ranking of the top candidates, combining LLM and embedding scores. Run 4 introduced a judge-driven diagnostic cycle, where GPT-4o acted as an independent evaluator to identify failure modes and guide further refinements, including third-person pronoun penalties, targeted negation detection, iterative query rewriting, an expanded re-ranking pool, and per-symptom prompting. This final run achieved the team's best AP and P@10 under both relevance settings. The authors also reported a gap between their internal LLM-as-judge

Table 12
Overview of participating teams and number of submitted runs per eRisk 2026 task.

Team	Task 1 #runs	Task 2 #runs	Task 3 #runs
AI-MH-Z [41]	–	–	4
Amor-Fati [42]	3	–	–
Awakened [43]	2	1	2
BUAP-CRC [44]	3	–	–
DeepCare [45]	–	5	–
DS-GT [46]	3	–	5
erisk-cedri [47]	3	5	5
HUGETIME [48]	–	1	–
HUTECH-NLP [49]	–	5	–
INSA-Lyon [50]	3	5	5
INSALyon-2 [51]	3	–	–
LCDAA [52]	–	5	–
lexxy [53]	3	–	–
Lotu-ixa [54]	–	5	–
MacEarlyDetect	–	5	–
MindAILab [55]	–	1	–
NaiveNeuron	–	–	5
NeuroRankUB [56]	–	–	4
NoirByte	1	5	–
NYCU-NLP	–	5	–
pjmathematician [57]	3	–	–
PMH-PsychAI-Lab [58]	1	–	1
RecEarlyMind [59]	1	–	–
Silvermask	–	–	1
SINAI [60]	3	–	–
Snugi-AI-v2 [61]	–	5	–
SVNITDS100	3	–	–
SymptomAI [62]	3	–	4
TeamZed [63]	–	5	–
Techbbas	3	–	–
The-Interrogators	3	–	–
UET-PsyWar [64]	–	5	5
UNED-GELP [65]	3	5	–
upb-uottawa [66]	3	–	–
VANGUARD [67]	3	2	4
Xtra	1	–	–
Total	54	70	45

precision estimates and the official scores, attributing it to shared model-family bias between the re-ranker and the judge.

Amor-Fati [42]. The Amor-Fati team, from the University of São Paulo, participated in Task 1 with a modular architecture inspired by the 2025 SINAI system. The pipeline decomposes the diagnostic process into four specialised components that share an incrementally updated patient state: a Question Planner, a Question Formulator, a Response Analyser, and a Stopping Controller. The Question Planner selects the next symptom to investigate and the strategy to follow, while the Response Analyser infers the severity of each of the 21 BDI-II symptoms from each response and merges the evidence using a max-severity rule. The Stopping Controller terminates the dialogue based on information gain, measured as the L1 distance between consecutive diagnostic vectors. The conversational modules used Qwen3-14B as the primary engine, with hyper-parameter search conducted through Optuna against an LLM-as-judge gold standard built from GPT-5 and Claude Haiku annotations. The team submitted three

runs exploring different trade-offs between diagnostic quality and conversational efficiency; Run #2 obtained competitive ASHR, while Run #1 performed well on LDCHR.

VANGUARD [67]. The VANGUARD team, from the National University of Science and Technology Politehnica Bucharest, the Academy of Romanian Scientists, and Fraunhofer FOKUS, participated in Task 3 with a zero-shot hybrid information-retrieval pipeline. For each of the 18 ASRS-v1.1 symptoms, the team built a query profile including the official ASRS question, a short label, expanded keywords, and a hand-crafted lexical query string, adopting a polarity-agnostic notion of relevance. The pipeline combined BM25 sparse retrieval with multi-query fusion, dense bi-encoder retrieval using all-mpnet-base-v2, cross-encoder re-ranking with ms-marco-MiniLM-L-6-v2, optional local LLM re-scoring using Llama 3.2 via Ollama and LangChain, and reciprocal rank fusion. Of five notebook runs, four were submitted. On the released majority-vote qrels, the cross-encoder run was the strongest, reaching nDCG@1000 = 0.113 and MAP = 0.028.

Awakened [43]. The Awakened team, from the National University of Science and Technology POLITEHNICA Bucharest and the Academy of Romanian Scientists, participated in Task 3 with a dense-retrieval approach based on sentence embeddings. Each ASRS-v1.1 symptom was expanded into a descriptive natural-language query and matched against the Reddit corpus by cosine similarity in a shared embedding space. The team submitted two runs: Run 1 encoded only the target sentence using all-MiniLM-L6-v2, while Run 2 used BAAI/bge-base-en-v1.5 and enriched each sentence with its preceding and following context. The pipeline emphasised scalability on consumer-grade hardware through GPU-accelerated chunked similarity computation and per-symptom checkpointing. The two runs achieved similar, comparatively low effectiveness, which the authors attribute to the use of a vanilla bi-encoder without domain adaptation, query reformulation, or re-ranking.

TeamZed [63]. TeamZed, from the University of Johannesburg, South Africa, participated in Task 2 on early detection of signs of depression from contextualized Reddit discussions. Their system deployed five parallel runs, contrasting four rule-based strategies—Conservative, Aggressive, Balanced, and Ensemble—with a neural run using MentalBERT for per-user risk scoring. All runs shared a User-State Layer that accumulated posts chronologically and applied semantic guardrails based on minimum post and word-count thresholds, as well as direct risk-cue filters, before issuing alerts. Official results were affected by a user-tracking bug that expanded the tracked population from the target subjects to a much larger set of conversation participants; only the Aggressive run registered non-zero binary decisions on the server. Offline evaluation against the released ground truth indicated that the MentalBERT run achieved the highest precision and ranking quality among the team's configurations, while the Aggressive run obtained the best F1.

PMH-PsychAI-Lab [58]. The PMH-PsychAI-Lab team, from the Department of Clinical Psychology, Shifa Tameer-e-Millat University, Islamabad, Pakistan, participated in Task 3 on ADHD symptom sentence ranking. Their approach used a zero-shot bi-encoder architecture based on bge-large-en-v1.5 to map user sentences and the 18 ASRS-v1.1 diagnostic criteria into a shared 1024-dimensional vector space. Similarity was computed through cosine scoring over L2-normalized vectors, and the top 1000 sentences per symptom were retrieved using FAISS indexing. No task-specific fine-tuning was performed. The team submitted a single run, which provided a competitive zero-shot dense-retrieval baseline.

erisk-cedri [47]. The erisk-cedri team, from CeDRI, Instituto Politécnico de Bragança, Portugal, participated in Task 3 with a zero-shot ensemble retrieval pipeline. The system combined four components: multi-query semantic similarity using all-mpnet-base-v2 with max-pooling over symptom reformulations, NLI-based entailment scoring using DeBERTa-v3-large fine-tuned on multiple NLI datasets, a keyword module with confidence dictionaries derived from the ASRS-v1.1 items, and cross-encoder re-ranking with ms-marco-MiniLM-L-12-v2 applied to the top candidates. The five submitted runs ranged from single-component baselines to weighted ensembles with increasing NLI and cross-encoder contributions. The best configuration, EnsembleV3, obtained strong results under majority-vote relevance judgements and ranked among the top-performing systems in the task.

DS-GT [46]. The DS-GT team, from the Georgia Institute of Technology, participated in Task 3 with a staged retrieval design in which BM25 retrieved candidates at scale and subsequent re-rankers refined

the final rankings. The submitted runs included a BM25 baseline with noise filtering and evidence-aware re-scoring for self-referential symptom reports, two embedding-based re-rankers using MiniLM and BGE-small over BM25 candidates, a hybrid prototype re-ranker using Likert-style query expansions with MiniLM, and an LLM re-ranker using Gemini through chunk-level listwise scoring over BM25 candidates. Under majority-vote judgements, the Gemini-based re-ranker achieved the team's strongest results, while prototype query expansion was the best non-LLM alternative by MAP.

HUGETIME [48]. The HUGETIME team, from KAIST, Republic of Korea, participated in Task 2, treating contextualized depression detection as a joint evidence-selection and classification problem. The system constructs target-centred candidate evidence units from full conversational threads, assigns each sample to a writing-count bucket according to the cumulative number of target-authored writings, and applies bucket-specific Top-K evidence pruning. Two complementary pruning branches are used: Risk Centroid Pruning, which aligns candidates with a depression-risk centroid derived from a seed sentence lexicon, and Prototype Query Bank Pruning, which learns multiple prototype directions through a Query Transformer trained with a ranking objective and query-usage regularisation. The pruned evidence is classified by bucket-specific SVM and ProtoTransformer local models with validation-selected ensemble weights. The team submitted a single run, achieving the best decision-based results in the task.

INSA-Lyon [50]. The INSA-Lyon team, from INSA Lyon, CNRS, Université Claude Bernard Lyon 1, and LIRIS, participated in Task 2 with DUET, a Dual-view Evidence Tracking system. DUET builds an incremental per-user representation at each round, combining target-text embeddings, BDI-II symptom activations, lexical ratios, community-channel signals, multi-scale Wasserstein temporal-shift detection over symptom and emotion trajectories, Mahalanobis-based drift scores, a Theory-of-Mind module contrasting self-view and observer-view symptom estimates obtained through LLM calls, and Thompson-sampled per-symptom personalisation. The team submitted five runs spanning XGBoost, MLP, and stacked ensemble classifiers with ERDE-aware adaptive thresholds. Runs 0 and 4 achieved the team's best F1 scores, and several runs obtained strong ranking performance.

SINAI [60]. The SINAI team participated in Task 1 with a modular conversational system based on a dual-agent architecture. One LLM conducts natural, empathetic conversations with simulated users, while a second LLM evaluates depression severity using the 21 BDI-II symptoms. The team explored three increasingly sophisticated variants: a baseline dual-agent loop, a Pool of Experts with specialised evaluators for different symptom dimensions, and a risk-aware orchestration system that adapts behaviour based on detected severity and enforces consistency. The approach emphasises efficiency and early decision-making, showing that short but information-dense conversations can capture relevant diagnostic signals, although very brief interactions may underestimate severe depression cases.

RecEarlyMind [59]. The RecEarlyMind team participated in Task 1 with a transformer-based conversational framework for detecting depression from interactions with LLM personas. Their system focuses on contextual, semantic, and emotional patterns in the generated dialogue. It combines preprocessing, contextual feature extraction, and prompt-based reasoning with transformer models to classify interactions as depression-related or not. The approach emphasises conversational context and emotional cues such as hopelessness or social withdrawal. Unlike modular multi-agent approaches, the system follows a single pipeline centred on representation learning and classification.

HUTECH-NLP [49]. The HUTECH-NLP team participated in Task 2 with a two-stage pipeline combining clinical knowledge and transformer models. First, user posts are filtered using BDI-II symptom similarity via an E5 encoder and then classified with a fine-tuned MentalBERT ensemble incorporating conversational context. Second, depression is detected at the user level by aggregating post-level predictions over the full history through a weighted scoring scheme and a majority-based decision rule designed to reduce false positives. The approach emphasises contextualized analysis, symptom-guided screening, and low-latency early detection, obtaining strong performance while balancing precision, recall, and timeliness.

MindAILab [55]. The MindAILab team, from Northwestern University Qatar, participated in Task 2 with a lightweight sequential modelling framework for contextualized, multi-speaker conversational data. The system extracts and accumulates only the target user's writings across rounds, represents

them through TF-IDF features with unigrams and bigrams, and applies a Logistic Regression classifier with balanced class weights to produce round-by-round probability scores. An evidence-accumulation strategy triggers a final irreversible alert when the probability exceeds a fixed threshold. The team reported strong held-out validation results, suggesting that interpretable lexical models combined with temporal decision strategies remain competitive for early risk detection while requiring substantially lower computational cost than transformer-based systems.

upb-uottawa [66]. The upb-uottawa joint team participated in Task 1 with a two-agent setup in which a configured "doctor" interviewer model interacted with a Llama-3-8B-Instruct "patient" model. The team explored three automated prompting strategies: a module-based prompt for autonomous question routing based on diagnostic confidence, a flat ordered list of clinically guided questions embedded in the system prompt, and a deterministic sequential interview in which the doctor LLM was used only at the end to generate the final diagnosis from the transcript. A medical expert manually designed the underlying interview questions and also conducted human-assisted baseline runs. Although the exhaustive multi-turn strategy did not produce the most accurate BDI-II scores, the deterministic interview framework obtained the best ASHR among complete submissions, showing strong sensitivity for symptom identification.

AI-MH-Z [41]. The AI-MH-Z team participated in Task 3 with an unsupervised dense retrieval pipeline for ranking social-media sentences according to the 18 ASRS-v1.1 ADHD symptoms. Candidate sentences were extracted and normalized from the provided corpus, combining each target sentence with its surrounding context fields. Texts were represented with all-MiniLM-L6-v2 sentence-transformer embeddings and indexed in a Qdrant vector database for approximate nearest-neighbour retrieval. The four submitted runs explored different query formulations and option-level similarity aggregation strategies, including uniform weighting, ordinal weighting, symptom-specific weighting based on the ASRS rubric, and direct holistic symptom-description similarity. The team's results indicate that option-level decomposition was more effective than direct holistic symptom descriptions.

INSALyon-2 [51]. The INSALyon-2 team participated in Task 1 with a safety-first hybrid agent-to-agent pipeline for interviewing fine-tuned patient personas and predicting BDI-II scores and key symptoms. The system separates language processing from deterministic reasoning guardrails across six specialised modules: Risk Router, Prober, Template Evidence, Extractor, Stopper, and Scorer. Language-understanding and extraction tasks were delegated to DeepSeek models, while structured text matching relied on all-MiniLM-L6-v2 sentence-transformer embeddings. The framework includes an acute safety ladder to clarify suicidal ideation and fatalistic cues, structured YAML question banks for balanced symptom coverage, and three operating profiles instantiated through parameter and threshold variation. The submitted runs covered 19 out of 20 personas and obtained strong symptom-identification results among incomplete submissions.

pjmathematician [57]. The pjmathematician team participated in Task 1 with a fully automated two-stage pipeline that decouples dialogue elicitation from clinical diagnosis. In the first stage, the system conducts a fixed, lightweight clinical interview consisting of six open-ended, indirect questions designed to probe the BDI-II dimensions without triggering defensive responses from the personas. In the second stage, the dialogue transcript is passed to an LLM-based diagnostic analyst, which reviews the transcript symptom by symptom against the BDI-II rubric before producing a structured JSON prediction. Across three runs, the team evaluated different model capacities and reasoning settings using Claude Opus and Claude Sonnet through the Amazon Bedrock API. Their short and information-dense strategy achieved the best overall ADODL among complete submissions and the strongest latency-weighted scores.

UET-PsyWar [64]. The UET-PsyWar team, from the VNU University of Engineering and Technology, Vietnam, participated in Task 2 with graph-based frameworks for early depression detection from multi-participant Reddit conversations. The system models conversational threads as tree-structured graphs, prunes them for depressive relevance, and initializes them with a hierarchical Semantic Post Encoder using an evidence-based DepROBERTa encoder. The first approach uses a Graph Attention Network followed by a Time-sensitive LSTM to model longitudinal progression, while the second maintains a persistent memory state per user updated through a GRU across interactions. Both pipelines were augmented with holistic thread-level summaries generated by Phi-4. Across five submitted runs,

the dynamic architecture consistently outperformed the static alternative, with the best configuration achieving competitive F1 and latency-aware performance.

lexxy [53]. The lexxy team participated in Task 1 with a zero-shot two-agent conversational pipeline. A Conversational Agent interviews the LLM patient personas across the 21 BDI-II depression dimensions using one of three strategies: self-disclosure, empathy, or direct questioning. After each turn, an Evaluation Agent re-scores all BDI-II dimensions using a two-pass method that first extracts explicit symptom markers and then infers implicit contextual cues after negation normalization. A Symptom Planner selects the next symptom to ask about based on low-confidence or underexplored dimensions. Final severity is computed by combining BDI scores, first-person negative-affect density, and contextual risk cues. The team submitted three runs, one for each conversational strategy.

BUAP-CRC [44]. The BUAP-CRC team participated in Task 1 with a dual-judge conversational system for depression detection. The system comprises four stages: answer preprocessing, symptom extraction, scoring and classification, and result-file construction. The interview consists of 12 questions developed by mapping PHQ-9 items to the 21 BDI-II symptoms, with input from clinical psychologists to ensure clinically plausible wording. Two LLM-based judges analyse the responses: the Direct Judge scores the symptoms targeted by each question, while the Secondary Judge identifies additional symptoms that may appear implicitly. The system then computes the final BDI-II score and selects the four most relevant symptoms. The team experimented with three LLMs across the submitted runs: Claude Sonnet 4.6, GPT-4o-mini, and Llama-3.3-70B-versatile.

DeepCare [45]. The DeepCare team, from the University of Bucharest, participated in Task 2 with five runs built on a shared contextualized-depression-detection pipeline. The runs included a Mental-Longformer model processing complete discussion threads, three configurations based on a five-fold MentalBERT ensemble with different score-aggregation and alert policies, and a lightweight character n-gram system combining a calibrated linear SVM with gradient-boosted trees. GPT-4o-mini was additionally used to tag writings with topic and distress cues. The team treated continuous risk ranking separately from the latency-aware binary alert policy. Its Mental-Longformer run achieved the team's best ranking results, while the lightweight character n-gram run produced the strongest decision-based performance.

UNED-GELP [65]. The UNED-GELP team participated in Task 1 with three conversational strategies for estimating depression severity from interactions with LLM personas. Run 0 directly administered the complete BDI-II questionnaire to each persona and computed the final score from the item responses. Runs 1 and 2 used shorter open-ended conversational protocols followed by a secondary LLM evaluator acting as a psychologist-like assessor to assign BDI-II item scores from the persona responses. To reduce evaluator variability, the authors repeated the assessment five times and aggregated the results using a trimmed mean. In the official results, Run 1 achieved the team's best ADODL among complete submissions, while Run 2, evaluated on 18 personas, obtained the team's best category and latency-weighted scores.

LCDAA [52]. The LCDAA team participated in Task 2 with two strategies for contextualized early depression detection. The first explored a temporal transformer-based approach, where each target user was represented as a chronological sequence of round-level text embeddings. The second was a symptom-oriented early detection pipeline that identified and accumulated symptom-relevant evidence over time. Their Random Forest configuration within the symptom-oriented pipeline obtained the team's best result, with an F1 score of 0.78.

NeuroRankUB [56]. The NeuroRankUB team participated in Task 3 with several embedding-based retrieval and re-ranking pipelines for ADHD symptom sentence search. The systems first applied preprocessing and filtering steps to reduce the large sentence collection, including language identification, translation of non-English content, duplicate removal, first-person pronoun filtering, length filtering, and noise removal. The team then explored retrieval strategies combining dense sentence embeddings, symptom reformulations, BM25 lexical retrieval, cross-encoder re-ranking, LLM-based relevance scoring, and Reciprocal Rank Fusion. The submitted runs included configurations based on MiniLM with RoBERTa and Qwen re-ranking, GTE-base with RRF and Qwen/Gemini refinement, E5-large-v2 with BM25 and BGE re-ranking, and MPNet with RoBERTa re-ranking. Their strongest

configuration combined bi-encoder retrieval, cross-encoder re-ranking, LLM refinement, and rank fusion.

Snugi-AI-v2 [61]. The Snugi-AI-v2 team, from the Georgia Institute of Technology, participated in Task 2 with a lightweight early depression detection pipeline. Their main contribution was a learned multi-layer perceptron stopping policy trained to balance the asymmetric cost structure of ERDE50, replacing fixed or heuristic thresholds. The system extracts text representations using a frozen MentalRoBERTa-base feature extractor, aggregates features across rounds through an incremental running mean, and projects these vectors with an MLP classifier to produce a continuous depression probability trajectory. The team's best configuration integrated surrounding-context embeddings, achieving competitive decision-based results while completing the full multi-round evaluation in 1 hour and 26 minutes, the fastest complete submission in the task.

Lotu-ixa [54]. The Lotu-ixa team participated in Task 2 with an explainable-AI-guided transfer-learning approach for contextualized early depression detection. The system used historical eRisk datasets from 2017, 2018, and 2022 to train a preliminary DistilBERT classifier and extract token-level SHAP attributions. These attributions were transformed into a Dual Relevance Mask containing both symptom-indicative and protective-factor vocabulary, which supervised the attention mechanism of a hierarchical Transformer architecture. The model processed Reddit threads through separate channels for the thread-start node, target-user messages, and replies from other users, aggregating their risk scores into a final thread-level decision. The team submitted five runs exploring an unguided baseline, several XAI-guided thresholding configurations, and a temporal rolling-mean strategy.

6. Conclusions and Future Work

This paper presented the extended overview of eRisk 2026, the tenth edition of the lab. The 2026 campaign continued the evolution of eRisk toward more contextual, conversational, and symptom-aware forms of early risk prediction. Across the three tasks, participants addressed complementary challenges: conversational depression assessment with LLM-based personas, contextualized early detection from Reddit threads, and sentence-level retrieval of ADHD symptom evidence.

Task 1 consolidated the conversational depression-detection setting by replacing hosted prompted agents with downloadable LoRA-based personas, improving reproducibility and local experimentation. The results show that estimating the overall BDI-II severity level is feasible, especially for full-coverage systems, but identifying the specific active depressive symptoms remains considerably more difficult. The latency-aware metrics also highlight the importance of reaching reliable decisions with concise conversations. Task 2 confirmed the value of conversational context for early depression detection from Reddit threads. Several systems achieved strong ranking performance, particularly as more user evidence became available. However, the decision-based evaluation shows that timely and accurate binary alerts remain challenging, with systems often trading precision for recall or earlier decisions. Task 3 introduced ADHD symptom sentence ranking, extending the symptom-oriented retrieval setting beyond depression. The results show that systems can retrieve meaningful evidence for ASRS-v1.1 symptoms, although the task remains difficult because many symptoms are expressed indirectly in everyday language. The comparison between majority-vote and unanimity qrels further underlines the relevance of assessor agreement in symptom-level evaluation.

Overall, eRisk 2026 reflects a shift from isolated risk classification toward richer, more interpretable, and more realistic evaluation settings. Future editions should continue improving conversational protocols, latency-aware evaluation, contextual modeling, and symptom-level retrieval resources. As LLMs become increasingly central in this area, robustness, transparency, ethical safeguards, and the limits of simulated data will remain essential research directions.

Acknowledgments

The authors thank the financial support supplied by the grant PID2022-137061OB-C21 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU". The authors also thank the funding supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditations ED431G 2023/01 and ED431C 2025/49) and the European Regional Development Fund, which acknowledges the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

Declaration on Generative AI

Generative AI tools were used only for language editing, proofreading, and improving the clarity of the manuscript. All scientific content, analyses, results, and conclusions were produced and verified by the authors.

References

[1] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations, in: G. J. Jones, S. Lawless, J. Gonzalo, L. Kelly, L. Goeuriot, T. Mandl, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2017, pp. 346–360.

[2] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental foundations, in: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2017, Dublin, Ireland, 2017.

[3] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: Early Risk Prediction on the Internet, in: P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J. Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2018, pp. 343–361.

[4] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2018: Early Risk Prediction on the Internet (extended lab overview), in: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2018, Avignon, France, 2018.

[5] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019: Early risk prediction on the Internet, in: F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, 2019, pp. 340–357.

[6] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk at CLEF 2019: Early risk prediction on the Internet (extended overview), in: CEUR Proceedings of the Conference and Labs of the Evaluation Forum, CLEF 2019, Lugano, Switzerland, 2019.

[7] D. E. Losada, F. Crestani, J. Parapar, Early detection of risks on the internet: An exploratory campaign, in: Advances in Information Retrieval - 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14-18, 2019, Proceedings, Part II, 2019, pp. 259–266.

[8] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2020: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, 2020, pp. 272–287.

[9] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk at CLEF 2020: Early risk prediction on the internet (extended overview), in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, 2020.

[10] D. E. Losada, F. Crestani, J. Parapar, eRisk 2020: Self-harm and depression challenges, in: Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, 2020, pp. 557–563.

[11] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2021: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, 2021, pp. 324–344.

[12] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk at CLEF 2021: Early risk prediction on the internet (extended overview), in: Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, 2021, pp. 864–887.

[13] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2021: Pathological gambling, self-harm and depression challenges, in: Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part II, 2021, pp. 650–656.

[14] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2022: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, 2022, p. 233–256.

[15] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk at CLEF 2022: Early risk prediction on the internet (extended overview), in: Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5–8, 2022, 2022, pp. 821–850.

[16] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2022: Pathological gambling, depression, and eating disorder challenges, in: Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part II, 2022, pp. 436–442.

[17] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2023: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, 2023, p. 233–256.

[18] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk at CLEF 2023: Early risk prediction on the internet (extended overview), in: Proceedings of the Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 18–21, 2023, 2023.

[19] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2023: Depression, pathological gambling, and eating disorder challenges, in: Advances in Information Retrieval - 45th European Conference on IR Research, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III, 2023, p. 585–592.

[20] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, eRisk 2024: Depression, anorexia, and eating disorder challenges, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, volume 14612 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 474–481. URL: https://doi.org/10.1007/978-3-031-56069-9_65. doi:10.1007/978-3-031-56069-9_65.

[21] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2024: Early risk prediction on the internet (extended overview), in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 759–781. URL: https://ceur-ws.org/Vol-3740/paper-72.pdf.

[22] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2024: Early risk prediction on the internet, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio,

L. Soulier, P. Galuscáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II, volume 14959 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 73–92. URL: https://doi.org/10.1007/978-3-031-71908-0_4. doi:10.1007/978-3-031-71908-0_4.

[23] J. Parapar, A. Perez, X. Wang, F. Crestani, eRisk 2025: contextual and conversational approaches for depression challenges, in: European Conference on Information Retrieval, Springer, 2025, pp. 416–424.

[24] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk 2025: Early risk prediction on the internet (extended overview), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[25] E. Bao, A. Perez, D. Otero, J. Parapar, How does depression talk on social media? modeling depression language with relevance-based statistical language models, Online Social Networks and Media 50 (2025) 100339.

[26] M. E. Aragon, A. P. Lopez-Monroy, L. C. González-Gurrola, M. Montes-y Gómez, Detecting mental disorders in social media through emotional patterns-the case of anorexia and depression, IEEE transactions on affective computing 14 (2021) 211–222.

[27] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197.

[28] M. Couto, A. Perez, J. Parapar, D. E. Losada, Temporal word embeddings for early detection of psychological disorders on social media, Journal of Healthcare Informatics Research (2025) 1–30.

[29] 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 and Data Engineering 32 (2018) 588–601.

[30] Z. Zhang, S. Chen, M. Wu, K. Q. Zhu, Psychiatric scale guided risky post screening for early detection of depression, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5220–5226. URL: https://doi.org/10.24963/ijcai.2022/725. doi:10.24963/ijcai.2022/725, aI for Good.

[31] E. Bao, A. Perez, J. Parapar, Redsm5: A reddit dataset for dsm-5 depression detection, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 6323–6327. URL: https://doi.org/10.1145/3746252.3761610. doi:10.1145/3746252.3761610.

[32] 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: Depresym: A depression symptom annotated corpus..., Lang. Resour. Eval. 59 (2025) 2737–2762.

[33] A. Pérez, J. Parapar, A. Barreiro, S. Lopez-Larrosa, Bdi-sen: A sentence dataset for clinical symptoms of depression, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2996–3006.

[34] A. Perez, J. Parapar, X. Wang, F. Crestani, erisk 2026: Tasks on symptoms ranking, contextual and conversational approaches for early mental health detection, in: European Conference on Information Retrieval, Springer, 2026, pp. 233–241.

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

[36] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An Inventory for Measuring Depression, JAMA Psychiatry 4 (1961) 561–571.

[37] X. Wang, A. Perez, J. Parapar, F. Crestani, Talkdep: clinically grounded llm personas for conversation-centric depression screening, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, 2025, pp. 6554–6558.
[38] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Proceedings Conference and Labs of the Evaluation Forum CLEF 2016, Evora, Portugal, 2016.
[39] M. Trotzek, S. Koitka, C. Friedrich, Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences, IEEE Transactions on Knowledge and Data Engineering (2018).
[40] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: WSDM, ACM, 2018, pp. 495–503.
[41] Z. Ramani, H. Diwanji, AI-MH-Z at eRisk@CLEF 2026: Ranking ADHD symptom sentences using weighted semantic embedding strategies, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[42] F. de Souza, I. Paraboni, Amor-Fati at eRisk 2026 task 1: A modular architecture for conversational depression detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[43] C.-G. Becheru, C.-O. Truică, E.-S. Apostol, Awakened at eRisk 2026: Sentence embedding-based ADHD symptom ranking from Reddit posts, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[44] N. M. de Los Santos, M. C. D. Carral, BUAP-CRC @ eRisk 2026: Conversational depression detection using judges model, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[45] G. Nistor, A.-S. Uban, DeepCare at eRisk 2026: Decoupling risk ranking from latency-aware decisions for contextualized early detection of depression, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[46] D. Guecha, DS@GT-ARC at eRisk 2026 task 3: Sparse, semantic, and LLM reranking for ADHD symptom sentences, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[47] F. Afonso, R. P. Lopes, erisk-cedri at eRisk 2026 task 3: A zero-shot ensemble pipeline with multi-query retrieval, NLI and cross-encoder re-ranking for ADHD symptom sentence ranking, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[48] I. Shin, H. Kim, C. Park, HUGETIME at eRisk 2026: Bucket-wise evidence pruning for contextualized early depression detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[49] T. X. Bui, T. T. N. Le, Contextualized early detection of depression on social media using BDI-II guided screening and MentalBERT ensemble, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[50] D. Nurbakova, INSA-Lyon at eRisk 2026 task 2: DUET, dual-view evidence tracking for contextualized early detection of depression, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[51] S. Shinde, S. H. Ong, D. Nurbakova, INSALyon2 at eRisk 2026 task 1: Safety-first hybrid multi-agent pipeline for conversational depression detection at eRisk 2026, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[52] C. A. Higuera-Guti'errez, S. Mendez-Mendoza, J. A. Navarrete-L'opez, J. M. Raygoza-Romero, J. N. Lazo-Vazquez, I. H. L'opez-Nava, M. Montes-Y-G'omez, LCDAA@eRisk 2026: Symptom-level evidence for contextualized early detection of depression, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[53] L. S. Chouhan, R. Kumar, lexxy at eRisk 2026: Conversational depression detection via LLM personas using a multi-strategy two-agent pipeline, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[54] X. Larrayoz, A. P'erez, P. Papapetrou, XAI-guided transfer learning: Supervised attention for contextualized early depression detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[55] E. Shlkamy, W. Zaghouani, Early depression detection from contextualized online conversations using sequential modeling, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[56] A.-M. Popeangă, T. Berca, D. Horga, R. Negoiţă-Creţu, NeuroRankUB at eRisk 2026: Embedding-based retrieval and re-ranking for ADHD symptom search, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[57] P. Vachharajani, pjmathematician @ eRisk 2026: Automating conversational depression detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[58] M. Aqeel, M. F. Khan, M. A. Khan, Psychometrically anchored zero-shot bi-encoder alignment, dense space feature projection, and semantic ranking layouts for deep adult ADHD symptom retrieval, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[59] S. Eugin, B. A, S. V, S. M. P, RecEarlyMind @ eRisk 2026: Conversational depression detection using transformer-based language models, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[60] A.-M. M'armol-Romero, M. D. Molina-Gonz'alez, M. Garc'ia-Vega, A. Montejo-R'aez, SINAI at eRisk@CLEF 2026: Modular conversational systems for depression detection in LLM personas, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[61] Y. Chiu, Snugi-AI-v2 @ eRisk 2026 task 2: Early depression detection via a learned stopping policy with sustained confidence gate, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[62] S. S. Daniyal, Y. Faisal, F. Faisal, F. Alvi, A. Samad, SymptomAI @ eRisk 2026: Layered retrieval for ADHD symptom sentence ranking, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[63] C. Kabuya, Multi-strategy early risk detection with rule-based guardrails and transformer scoring: CLEF eRisk 2026 task 2 working notes, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[64] D.-L. Tran, P.-A. Chu, H.-D. Do, T.-P. Mai, D.-C. Can, H.-Q. Le, UET-PsyWar@eRisk 2026: Localized context encoding and memory-augmented graph networks for early detection of depression, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[65] J. Fernandez-Hernandez, UNED-GELP at eRisk 2026: Conversational LLMs for depression detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[66] I.-R. Zaman, A.-G. Zanfir, M.-O. Delisle, J. A. Lossio-Ventura, D. Inkpen, S. Trausan-Matu, upb-uottawa at eRisk@CLEF 2026: Detecting depression through LLM persona conversations, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.
[67] A. Vişan, C.-O. Truică, A. Paschke, E.-S. Apostol, VANGUARD at eRisk 2026 task 3: ADHD symptom sentence ranking, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), Jena, Germany, September 21–24, 2026, CEUR Workshop Proceedings, CEUR-WS.org, 2026.