Overview of eRisk at CLEF 2025: Early Risk Prediction on the Internet (Extended Overview)

Notebook for the eRisk Lab at CLEF 2025

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

[1] *Information Retrieval Lab, Centro de Investigación en Tecnoloxías da Información e as Comunicacións (CITIC), Universidade da Coruña. Campus de Elviña s/n C.P 15071 A Coruña, Spain*

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

[3] *Faculty of Informatics, Universitá della Svizzera italiana (USI). Campus EST, Via alla Santa 1, 6900 Viganello, Switzerland*

Abstract

This paper presents an extended overview of eRisk 2025, the ninth edition of the CLEF lab on early risk detection. Since its beginnings, eRisk has served as a benchmark for assessing methodologies, evaluation metrics, and challenges in the early identification of personal risks, particularly within health and safety domains. The 2025 edition marks an important evolution, amplifying the lab's scope toward problems that require richer contextual and conversational understanding. The first task, the only one preserved from last year, asks systems to rank sentences by their relevance to the BDI-II depression symptoms, enabling fine-grained retrieval of depressive cues. The second task reformulates early detection as a contextual decision problem. In this task, the full conversational thread, including the user's posts and all the interactions from the rest of the people involved, is revealed incrementally. At each step, the models must decide whether sufficient evidence exists to predict depression for the user, thereby rewarding both accuracy and timeliness. Finally, the pilot task pioneers an interactive scenario: fine-tuned large language models engage participants in dialogue and must infer depressive signals from the evolving conversations, probing the feasibility and safety of conversational screening agents. Together, these three tasks continue to advance the field of early risk detection, open new research avenues and align the evaluation framework more closely with real-world conversational settings.

Keywords

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

1. Introduction

The eRisk lab was designed as a benchmark environment for constructing resources, evaluation protocols, and developing approaches that enable the timely detection of different personal risk situations. Early alert technologies are becoming indispensable across healthcare oriented domains. The rapid recognition of warning signs, whether for emergent mental health crises, predatory behaviour, or violent threats, can turn marginal time gains into life saving interventions.

eRisk focuses on psychological and mental health risks such as depression, self-harm, pathological gambling, and eating disorders, where language provides subtle yet informative signals. However, the intricate relationship between linguistic expression and mental state continues to challenge automatic methods and screeners, underscoring the need for increasingly robust, context-aware models and annotated public datasets of high quality.

The inaugural eRisk 2017 edition introduced the pilot task on early detection of depression, establishing the sequential evidence evaluation framework that is still present in the lab today [1, 2]. In 2018 the scope

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

[*] Corresponding author.

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

https://www.dc.fi.udc.es/~parapar (J. Parapar); https://pdi.udc.es/en/File/Pdi/UK44K (A. Perez); https://www.sheffield.ac.uk/cs/people/academic/xi-wang (X. Wang); https://search.usi.ch/en/people/4f0dd874bbd63c00938825fae1843200/crestani-fabio (F. Crestani)

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

broadened to include anorexia, creating a dual task campaign that demonstrated the generalization of the proposals across related mental-health disorders [3, 4]. The 2019 programme consolidated anorexia work, introduced a self-harm prediction trask, and, for the first time, asked systems to infer answers to a depression severity questionnaire (BDI-II [5]) purely from social media activity [6, 7, 8].

The 2019 inclusion of the BDI-II moved the lab toward symptom level modelling, encouraging participants to move beyond binary depressed/not-depressed labels and design methods that capture the nuances of individual depressive symptoms. eRisk 2020 deepened the self-harm task and added a refined depression severity estimation challenge, further emphasising continuous severity scales over binary outcomes [9, 10, 11]. The 2021 edition first introduced behavioural addictions; it pioneered an early pathological gambling task while revisiting self-harm detection and severity estimation for depression [12, 13, 14].

In 2022, we returned to gambling and depression and introduced a new challenge centered on estimating the severity of eating-disorder activities [15, 16, 17]. The 2023 campaign shifted emphasis to fine-grained symptom prediction, presenting a sentence ranking task that maps individual user sentences to the 21 BDI-II depression symptoms, and retained tasks on gambling risk and eating-disorder severity [18, 19, 20]. Finally, eRisk 2024 consolidated the BDI-II sentence ranking benchmark, maintained the anorexia early detection task, and updated the eating-disorder severity task, setting the stage for the more conversational focus adopted in 2025 [21, 22, 23].

The current edition, eRisk 2025 [24, 25], extends this trajectory by introducing, for the first time, tasks that demand not only early recognition of risk but also deeper contextual reasoning and, in the pilot trask, true conversational interactions. Full task specifications appear in the next sections, yet the broad shift is clear: systems must now interpret entire discussion threads and interactions, bringing the evaluation environment closer to real-world online settings. This year, the eRisk lab had 128 different teams registered. We finally received results coming from 25 distinct teams: 67 runs for Task 1, 50 runs for Task 2, and 11 runs for the pilot task.

2. Task 1: Search for Symptoms of Depression

This task continues from eRisk 2023's and 2024's Task 1, which involved ranking sentences from user writings based on their relevance to specific depression symptoms. This is the last year of the task. Again, participants were required to order sentences according to their relevance to the 21 standardized symptoms listed in the BDI-II questionnaire [5]. A sentence was deemed relevant if it reflected the user's condition related to a symptom, including positive statements (e.g., "I feel quite happy lately" is relevant for the symptom "Sadness"). As in 2024, the test collection provides not only the target sentence but also its immediate predecessor and successor to give more context.

2.1. Task 1: Dataset and Asessment Process

The dataset provided was in TREC format, tagged with sentences derived from Reddit historical data. Table 1 presents some statistics of the corpus. Given the corpus of sentences and the description of the symptoms from the BDI-II questionnaire, the participants were free to decide on the best strategy to derive queries for representing the BDI-II symptoms. Each participating team submitted up to 5 variants (runs). Each run included 21 TREC-style formatted rankings of sentences, as shown in Figure 1. For each symptom, the participants should submit up to 1000 results sorted by estimated relevance. We received 67 runs from 17 participating teams (see Table 2).

Table 1
Corpus statistics for Task 1: Search for Symptoms of Depression.

Number of users	9,000
Number of sentences	17,553,441
Average number of words per sentence	12,39

```
1Q0251001_0_1000110myGroupNameMyMethodName
1Q0858202_3_200029myGroupNameMyMethodName
1Q0482048_2_100038.76myGroupNameMyMethodName

...
21Q0153202_2_209991.25myGroupNameMyMethodName
21Q0223133_9_810000.9myGroupNameMyMethodName
```

Figure 1: Example of a participant's run.

Table 2
Task 1 (Search for Symptoms of Depression): number of submitted runs per team.

Team	Runs	Team	Runs	Team	Runs
SonUIT [26]	5	ThinkIR [27]	5	Ixa_ave [28]	5
Synapse	3	PJs-team [29]	5	ELiRF-UPV [30]	5
COTECMAR-UTB [31]	3	COMFOR	1	LHS712-Team-1 [32]	5
Team-Gryffindor	1	INESC-ID [33]	5	UET-Psyche-Warriors [34]	5
NYCUNLP	5	BGU-Data-Science [35]	5	HULAT_UC3M [36]	5
RELAI	2	UniORNLP-dahlia	2		
Total Teams: 17 | Total Runs: 67					

Relevance labels were produced through a stratified, two–stage pooling procedure. First, for every BDI-II symptom we implemented top-k pooling, collecting the top five sentences returned by each submitted run ($k = 5$), forming an initial pool that served to rank systems provisionally. We then selected the twenty highest-ranked runs and performed a second pooling step that extended the cut-off to the top fifty sentences ($k = 50$). Unlike the 2023 setup, assessors were shown the target sentence together with its immediate context (the preceding and following sentences), a change designed to reduce annotation ambiguity.

Three annotators worked independently: one with professional training in psychology, and two computer-science researchers specialising in early risk technologies. Before judging, the organisers held a session to walk through an initial guideline draft, resolve doubts, and agree on different cases. The consolidated guidelines, publicly available[1], defines a sentence as *relevant* only when it both addresses the symptom and conveys explicit information about the user's state. This dual concept of relevance (on-topic and reflective of the user's state with respect to the symptom) introduced a higher level of complexity compared to more standard relevance assessments. Each pooled sentence received three independent judgements, and we provide two ground-truth sets (qrels):

- **Majority-based qrels**: a sentence was deemed relevant if at least two of the three assessors marked it so.
- **Unanimity-based qrels**: a sentence was deemed relevant only when all three assessors agreed.

The final pool sizes and qrels for each symptom are reported in Table 3. Providing both qrels enables analyses with different agreement thresholds, continuing the dual-qrel strategy introduced in earlier eRisk campaigns.

2.2. Task 1: Results

The performance results for the participating systems are shown in Tables 4 (majority-based qrels) and 5 (unanimity-based qrels). The tables report several standard performance metrics, such as mean

[1] https://erisk.irlab.org/guidelines_erisk24_task1.html

Table 3
Task 1 (Search for Symptoms of Depression): Size of the pool for every BDI Item.

BDI Item (#)	pool	# unanimity-qrels (3/3)	# majority-qrels (2/3)
Sadness (1)	581	167	296
Pessimism (2)	552	209	345
Past Failure (3)	536	146	283
Loss of Pleasure (4)	522	132	244
Guilty Feelings (5)	400	88	227
Punishment Feelings (6)	553	33	111
Self-Dislike (7)	474	205	290
Self-Criticalness (8)	534	115	259
Suicidal Thoughts or Wishes (9)	517	300	377
Crying (10)	547	143	359
Agitation (11)	593	142	338
Loss of Interest (12)	553	105	229
Indecisiveness (13)	584	50	139
Worthlessness (14)	424	161	249
Loss of Energy (15)	491	161	273
Changes in Sleeping Pattern (16)	569	274	404
Irritability (17)	540	132	314
Changes in Appetite (18)	548	225	374
Concentration Difficulty (19)	428	166	271
Tiredness or Fatigue (20)	566	217	385
Loss of Interest in Sex (21)	530	239	350

Average Precision (AP), mean R-Precision, mean Precision at 10 and mean NDCG at 1000. Remarkably, runs *unanimity* and *max* from the team *INESC-ID*, achieved the top-ranking performance for nearly all metrics and relevance judgement types. The teams *UET-Psyche-Warriors*, *SonUIT*, *BGU-Data-Science* and *PJs-Team* also obtained close performance. Their effective results demonstrate their exceptional competence in this task. Taken together, the results confirm that sentence-level symptom retrieval remains a challenging task.

3. Task 2: Contextualized Early Detection of Depression (New Task)

This new task in 2025 introduces a different scenario in depression detection by incorporating full conversational contexts. Whereas earlier eRisk editions always released isolated posts authored by a single user, the 2025 task provided the entire Reddit discussion thread in which the target user intervened. Consequently, in the test dataset, systems had access not only to the messages produced by the target user but also to every other contribution in the thread and to the interaction structure that links the messages (e.g., the different replies to each comment).

This design is motivated by the observation that the clinical relevance of a message often becomes more evident when interpreted alongside the surrounding conversation. Thus, a user's response may only gain relevance when viewed in conjunction with the preceding or subsequent interactions from other participants. For instance, a seemingly neutral sentence, may reveal hopelessness if it answers a direct plea for support. For this reason, the task is designed to simulate real-world scenarios where depression detection may rely on analyzing exchanges between multiple participants. This setup presents unique challenges, as systems must consider not only the textual content of individual posts but also the interplay between participants and how this context influences the detection of depressive symptoms.

Table 4
Ranking-based evaluation for Task 1 (majority voting).

Team	Run	AP	R-PREC	P@10	NDCG
BGU-Data-Science	sbert-w-expansion-w-naive-fp-w-claude	0.232	0.305	0.767	0.483
BGU-Data-Science	sbert-w-expansion-w-naive-fp	0.227	0.296	0.767	0.475
BGU-Data-Science	sbert-w-expansion-w-spacy-fp	0.220	0.287	0.767	0.463
BGU-Data-Science	sbert-w-expansion	0.197	0.281	0.652	0.444
BGU-Data-Science	sbert	0.240	0.324	0.743	0.516
COMFOR	bert_ranked	0.013	0.041	0.243	0.082
COTECMAR-UTB	centroid_ranked_updated	0.052	0.130	0.276	0.236
COTECMAR-UTB	dl_ranked	0.073	0.160	0.405	0.282
COTECMAR-UTB	ranked_updated	0.077	0.165	0.414	0.290
ELiRF-UPV	model1	0.035	0.101	0.100	0.216
ELiRF-UPV	model2	0.032	0.095	0.081	0.206
ELiRF-UPV	model3	0.035	0.099	0.110	0.211
ELiRF-UPV	model4	0.033	0.099	0.067	0.210
ELiRF-UPV	model5	0.032	0.097	0.100	0.209
HULAT_UC3M	roberta	0.004	0.010	0.162	0.026
HULAT_UC3M	vader_sample	0.004	0.010	0.148	0.023
HULAT_UC3M	reflexives_roberta	0.013	0.025	0.262	0.052
HULAT_UC3M	roberta	0.018	0.034	0.363	0.065
HULAT_UC3M	vader_top	0.015	0.034	0.295	0.059
INESC-ID	aug-best	0.247	0.324	0.691	0.560
INESC-ID	max	0.350	0.407	0.648	**0.653**
INESC-ID	maxcos	0.235	0.320	0.757	0.506
INESC-ID	mix23	0.312	0.377	0.643	0.616
INESC-ID	unanimity	**0.354**	**0.433**	**0.876**	0.575
LHS712-Team-1	results	0.000	0.000	0.000	0.000
LHS712-Team-1	BERT_CONSENSUS	0.102	0.199	0.529	0.321
LHS712-Team-1	BERT_MAJORITY	0.074	0.178	0.281	0.283
LHS712-Team-1	LR_file_combined	0.000	0.004	0.009	0.007
LHS712-Team-1	SVM_file_combined_4	0.000	0.004	0.009	0.007
NYCUNLP	01	0.237	0.322	0.662	0.501
NYCUNLP	02	0.193	0.276	0.619	0.455
NYCUNLP	03	0.133	0.217	0.624	0.328
NYCUNLP	04	0.190	0.279	0.614	0.450
NYCUNLP	05	0.072	0.137	0.567	0.203
PJs-team	teamADRB	0.105	0.234	0.391	0.354
PJs-team	teamMBRR	0.262	0.347	0.771	0.489
PJs-team	teamRRens-v2	0.279	0.360	0.800	0.503
PJs-team	teamRRens	0.273	0.359	0.786	0.500
PJs-team	teamSumensemble	0.120	0.249	0.400	0.376
RELAI	1	0.005	0.023	0.052	0.053
RELAI	2	0.008	0.036	0.038	0.078
SonUIT	config1	0.283	0.351	0.767	0.562
SonUIT	config2	0.334	0.392	0.790	0.613
SonUIT	config3	0.311	0.395	0.767	0.572
SonUIT	config4	0.328	0.426	0.767	0.578
SonUIT	config5	0.260	0.304	0.767	0.552
Synapse	HighestSimilarityFirst	0.001	0.002	0.038	0.009
Synapse	nomicFineTunedRerankedSimilarity	0.001	0.002	0.043	0.008
Synapse	nomicRerankedSimilarity	0.001	0.002	0.052	0.006
Team-Gryffindor	task1	0.017	0.042	0.019	0.183
ThinkIR	few_shot_query_2025	0.015	0.049	0.133	0.073
ThinkIR	rank_sim	0.003	0.010	0.000	0.030
ThinkIR	2025	0.068	0.157	0.409	0.228
ThinkIR	pseudo_relevance_10_2025	0.064	0.148	0.400	0.213
ThinkIR	pseudo_relevance_5_2025	0.060	0.151	0.409	0.212
UET-Psyche-Warriors	1_similarity	0.311	0.378	0.657	0.588
UET-Psyche-Warriors	2_ensemble_similarity	0.315	0.390	0.657	0.612
UET-Psyche-Warriors	3_contrastive_learning	0.165	0.258	0.457	0.450
UET-Psyche-Warriors	4_ensemble_contrastive_learning	0.147	0.228	0.462	0.419
UET-Psyche-Warriors	5_machine_learning	0.339	0.394	0.776	0.623
UniORNLP-dahlia	frame_hyde	0.001	0.008	0.029	0.019
UniORNLP-dahlia	simple_hyde	0.014	0.040	0.205	0.072
ixa_ave	base_all	0.097	0.191	0.305	0.345
ixa_ave	base_filter30	0.102	0.203	0.338	0.342
ixa_ave	base_filter50	0.009	0.025	0.086	0.048
ixa_ave	thresh_all	0.091	0.168	0.281	0.333
ixa_ave	thresh_filter50	0.005	0.016	0.129	0.035

Table 5
Ranking-based evaluation for Task 1 (unanimity).

Team	Run	AP	R-PREC	P@10	NDCG
BGU-Data-Science	sbert-w-expansion-w-naive-fp-w-claude	0.143	0.244	0.443	0.429
BGU-Data-Science	sbert-w-expansion-w-naive-fp	0.138	0.240	0.448	0.420
BGU-Data-Science	sbert-w-expansion-w-spacy-fp	0.135	0.237	0.462	0.412
BGU-Data-Science	sbert-w-expansion	0.119	0.223	0.381	0.389
BGU-Data-Science	sbert	0.171	0.272	0.419	0.489
COMFOR	bert_ranked	0.010	0.036	0.114	0.079
COTECMAR-UTB	centroid_top1000_ranked_updated	0.030	0.081	0.133	0.195
COTECMAR-UTB	dl_ranked	0.040	0.107	0.176	0.240
COTECMAR-UTB	ranked_updated	0.042	0.108	0.181	0.243
ELiRF-UPV	model1	0.021	0.063	0.052	0.184
ELiRF-UPV	model2	0.019	0.057	0.062	0.179
ELiRF-UPV	model3	0.021	0.060	0.062	0.180
ELiRF-UPV	model4	0.019	0.056	0.038	0.180
ELiRF-UPV	model5	0.018	0.057	0.057	0.175
HULAT_UC3M	roberta	0.002	0.009	0.052	0.016
HULAT_UC3M	vader_sample	0.001	0.009	0.029	0.012
HULAT_UC3M	reflexives_roberta	0.013	0.032	0.157	0.053
HULAT_UC3M	roberta	0.008	0.025	0.174	0.040
HULAT_UC3M	vader_top	0.006	0.024	0.105	0.037
INESC-ID	aug-best	0.167	0.236	0.414	0.515
INESC-ID	max	0.223	0.308	0.386	**0.582**
INESC-ID	maxcos	0.164	0.273	0.429	0.472
INESC-ID	mix23	0.201	0.279	0.371	0.547
INESC-ID	unanimity	**0.269**	**0.383**	**0.509**	0.561
LHS712-Team-1	results	0.000	0.000	0.000	0.000
LHS712-Team-1	BERT_CONSENSUS	0.083	0.172	0.281	0.315
LHS712-Team-1	BERT_MAJORITY	0.062	0.137	0.181	0.286
LHS712-Team-1	LR_file_combined	0.000	0.003	0.005	0.007
LHS712-Team-1	SVM_file_combined_4	0.000	0.003	0.005	0.007
NYCUNLP	01	0.156	0.253	0.367	0.442
NYCUNLP	02	0.129	0.216	0.357	0.400
NYCUNLP	03	0.081	0.159	0.371	0.270
NYCUNLP	04	0.135	0.224	0.352	0.408
NYCUNLP	05	0.048	0.117	0.357	0.173
PJs-team	ADRB	0.073	0.168	0.214	0.325
PJs-team	MBRR	0.175	0.299	0.424	0.435
PJs-team	RRens-v2	0.188	0.311	0.452	0.446
PJs-team	RRens	0.184	0.308	0.467	0.444
PJs-team	Sumensemble	0.079	0.184	0.229	0.331
RELAI	1	0.005	0.019	0.029	0.056
RELAI	2	0.006	0.024	0.009	0.076
SonUIT	config1	0.191	0.276	0.448	0.500
SonUIT	config2	0.223	0.303	0.462	0.545
SonUIT	config3	0.205	0.290	0.448	0.508
SonUIT	config4	0.219	0.315	0.448	0.514
SonUIT	config5	0.176	0.248	0.448	0.491
Synapse	HighestSimilarityFirst	0.000	0.000	0.000	0.000
Synapse	nomicFineTunedRerankedSimilarity	0.000	0.000	0.000	0.000
Synapse	nomicRerankedSimilarity	0.000	0.000	0.000	0.000
Team-Gryffindor	task1	0.014	0.027	0.014	0.187
ThinkIR	few_shot_query_2025	0.014	0.043	0.081	0.075
ThinkIR	rank_sim	0.001	0.003	0.000	0.019
ThinkIR	2025	0.044	0.116	0.219	0.196
ThinkIR	pseudo_relevance_10_2025	0.042	0.111	0.205	0.192
ThinkIR	pseudo_relevance_5_2025	0.040	0.107	0.229	0.191
UET-Psyche-Warriors	1_similarity	0.193	0.270	0.391	0.501
UET-Psyche-Warriors	2_ensemble_similarity	0.202	0.279	0.391	0.530
UET-Psyche-Warriors	3_contrastive_learning	0.094	0.165	0.243	0.373
UET-Psyche-Warriors	4_ensemble_contrastive_learning	0.079	0.141	0.219	0.347
UET-Psyche-Warriors	5_machine_learning	0.248	0.330	0.476	0.577
UniORNLP-dahlia	frame_hyde	0.001	0.004	0.009	0.014
UniORNLP-dahlia	simple_hyde	0.009	0.033	0.081	0.058
ixa_ave	base_all	0.053	0.121	0.124	0.282
ixa_ave	base_filter30	0.055	0.126	0.138	0.277
ixa_ave	base_filter50	0.006	0.020	0.038	0.042
ixa_ave	thresh_all	0.052	0.103	0.110	0.270
ixa_ave	thresh_filter50	0.003	0.013	0.048	0.026

The test collection utilised for this task followed the same format as the collection described in the work by Losada and Crestani [37]. The collection contains writings, including posts and comments, obtained from a selected group of social media users. To construct the ground truth assessments, we adopted established approaches that aim to optimise the utilisation of assessors' time, as documented in previous studies [38, 39]. These methods employ simulated pooling strategies, enabling the effective creation of test collections. The main statistics of the test collection used for Task 2 are presented in Table 6.

Table 6
Task 2 (early depression). Main statistics of test collection.

	Depression	*Control*
Num. subjects	102	807
Num. threads	40,563	238,033
Avg num. of threads per subject	397.7	295.9
Avg num. of days from first to last thread	$\approx$ 1695	$\approx$ 958
Avg num. of comments per thread	65.1	44.6
Avg num. words per comment	33.8	25.6

Within this dataset, users are categorised into two groups: depression and control. For each user, the collection contains a sequence of writings and threads where the user participated in chronological order. To facilitate the task and ensure uniform distribution, we established a dedicated server that systematically provided user writings to the participating teams. Further details regarding the server's setup and functioning are available at the lab's official website[2].
The task was divided into two phases:

- During the training phase, participants worked with a static dataset consisting of isolated user writings from depressed and control users, without any conversational context. This training dataset came from prior editions of eRisk regarding the early detection depression tasks (without any conversational context).
- The test phase, in contrast, was carried out interactively. For each target user, the server released a sequence of discussion threads in real time. Each thread constituted a *submission round.* At any round within the chronology of user writings, participants had the freedom to stop the process and issue an alert. After reading each user thread, teams were required to decide between two options: i) alerting about the target user, indicating a predicted sign of depression, or ii) not alerting about the target user. Participants independently made this choice for each user in the test split. It is important to note that once an alert was issued, it was considered final, and no further decisions regarding that particular user were taken into account. Conversely, the absence of alerts was considered non-final, allowing participants to subsequently submit an alert if they detected signs of risk emerging.

To evaluate the systems' performance, we employed two indicators: the accuracy of the decisions made and the number of user writings required to reach those decisions. These criteria provide valuable insights into the effectiveness and efficiency of the systems under evaluation. To support the test stage, we deployed a REST service. The server iteratively distributes user writings and waits for responses from participants. Importantly, new user data was not provided to a specific participant until the service received a decision from that particular team. The submission period for the task was open from February 5th, 2025 until April 12th, 2025.

[2] https://erisk.irlab.org/eRisk25Servert2Details.html

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 [37]. 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 [40] 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 [41]. 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]:

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

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):

[3] Observe that Sadeque et al (see [41], 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.

$$P = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : d_u = 1|} \tag{2}$$

$$R = \frac{|u \in U : d_u = g_u = 1|}{|u \in U : g_u = 1|} \tag{3}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{4}$$

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:

$$penalty(k_u) = -1 + \frac{2}{1 + \exp^{-p \cdot (k_u - 1)}} \tag{5}$$

where p is a parameter that determines how quickly the penalty should increase. In [41], 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 2 plots how the latency penalty increases with the number of observed writings.

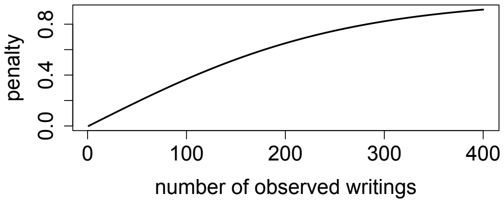

Figure 2: Latency penalty increases with the number of observed writings (k_u).

The system's overall speed factor is computed as:

$$speed = (1 - \text{median}\{penalty(k_u) : u \in U, d_u = g_u = 1\}) \tag{6}$$

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:

$$F_{latency} = F \cdot speed \tag{7}$$

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 $ERDE$.

[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.

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.

3.2. Task 2: Participant Teams

Table 7 shows the participating teams, the number of runs submitted and the approximate lapse of time from the first response to the last response. This lapse of time is indicative of the degree of automation of each team's algorithms. All but one participant (FU–TU–DFKI) managed to process the complete set of threads in at least one run. The fastest groups: ELiRF–UPV, SINAI–UJA, and PJs-team finished in under ten hours, illustrating the feasibility of efficient processing even when entire conversations are supplied. By contrast, Lotu–ixa and UET–Psyche–Warriors took around a week, pointing to more complex or resource intensive pipelines

Table 7
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)
Lotu-ixa [42]	5	1280	9 days 03:21
HIT-SCIR [43]	5	1280	1 day 14:00
SINAI-UJA [44]	5	1280	0 days 09:53
DS-GT [45]	2	1280	1 day 00:29
NYCUNLP	5	1280	1 day 04:07
UET-Psyche-Warriors [34]	5	1280	6 days 21:34
Capy-team	5	1280	0 days 15:49
COTECMAR-UTB [31]	2	1280	3 days 00:50
ELiRF-UPV [30]	5	1280	0 days 08:33
PJs-team [29]	5	1280	0 days 08:36
HU [46]	5	1280	2 days 23:08
FU-TU-DFKI [47]	1	449	1 day 11:24

3.3. Task 2: Results

Table 8 show the decision-based results of Task 2. Table 9 shows the ranking-based results. In the decision setting, *HIT-SCIR* dominates: its best run attains the highest F_1 (0.85) while keeping both ERDE$_5$ and ERDE$_{50}$ at or very near the minimum error values. That performance is achieved with a median latency of only eight writings, illustrating a good balance between earliness and accuracy. *ELiRF-UPV* follows at a short distance, with a top F_1 of 0.79 but slightly worse error–aware metrics

The ranking-based evaluation shows a complementary picture. *HIT-SCIR* again exhibits near-perfect precision at every cut-off and sustains the highest NDCG values as additional writings become available, confirming the robustness of its retrieval component. *Lotu-Ixa* excels in the one writing scenario, matching *HIT-SCIR* for $P@10$ and $NDCG@10$. However, its advantage diminishes once longer histories are considered, suggesting that its decision policy strongly weights the earliest cues.

Table 8
Decision-based evaluation for Task 2 ordered in terms of best F_1.

Team	Run	P	R	F_1	$ERDE_5$	$ERDE_{50}$	latency$_{TP}$	$speed$	$F_{latency}$
HIT-SCIR	0	0.72	0.96	0.82	0.06	**0.03**	4.00	0.99	0.81
	1	0.72	0.95	0.82	0.06	**0.03**	4.00	0.99	0.81
	2	0.74	0.94	0.83	0.06	**0.03**	4.00	0.99	**0.82**
	3	0.73	0.94	0.82	0.08	**0.03**	7.00	0.98	0.80
	4	0.77	0.94	**0.85**	0.09	**0.03**	8.00	0.97	**0.82**
ELiRF-UPV	0	0.78	0.81	0.79	0.08	0.04	7.00	0.98	0.78
	1	0.37	0.62	0.46	0.07	0.06	**1.00**	**1.00**	0.46
	2	0.83	0.47	0.60	0.10	0.07	8.00	0.97	0.58
	3	0.68	0.67	0.67	0.09	0.05	7.00	0.98	0.66
	4	0.68	0.67	0.67	0.09	0.05	7.00	0.98	0.66
HU	0	0.61	0.77	0.68	0.09	0.05	10.00	0.96	0.66
	1	0.72	0.77	0.75	0.10	0.05	11.00	0.96	0.72
	2	0.14	0.94	0.25	0.15	0.09	6.00	0.98	0.24
	3	0.11	1.00	0.20	0.11	0.10	**1.00**	**1.00**	0.20
	4	0.27	0.88	0.41	0.10	0.07	11.00	0.96	0.40
UET-Psyche-Warriors	0	0.67	0.78	0.72	0.10	0.06	24.50	0.91	0.66
	1	0.63	0.85	0.72	0.09	0.05	16.00	0.94	0.68
	2	0.63	0.86	0.73	0.09	0.04	16.00	0.94	0.68
	3	0.63	0.85	0.72	0.09	0.05	16.00	0.94	0.68
	4	0.63	0.84	0.72	0.09	0.05	15.50	0.94	0.68
PJs-team	0	0.66	0.75	0.71	0.09	0.06	17.00	0.94	0.66
	1	0.53	0.83	0.65	0.09	0.06	24.00	0.91	0.59
	2	0.54	0.82	0.65	0.09	0.06	23.00	0.91	0.60
	3	0.49	0.85	0.63	0.10	0.06	22.00	0.92	0.57
	4	0.58	0.81	0.67	0.09	0.06	24.00	0.91	0.61
Lotu-Ixa	0	0.43	0.79	0.56	**0.05**	0.04	2.00	**1.00**	0.56
	1	0.46	0.79	0.58	**0.05**	**0.03**	2.00	**1.00**	0.58
	2	0.47	0.79	0.59	**0.05**	**0.03**	2.00	**1.00**	0.59
	3	0.53	0.78	0.63	**0.05**	**0.03**	**1.00**	**1.00**	0.63
	4	0.15	1.00	0.25	0.09	0.08	**1.00**	**1.00**	0.25
COTECMAR-UTB	0	0.29	0.65	0.40	0.12	0.10	69.00	0.74	0.29
	1	0.25	0.01	0.02	0.11	0.11	**1.00**	**1.00**	0.02
SINAI-UJA	0	0.24	1.00	0.39	0.08	0.05	3.00	0.99	0.38
	1	0.17	1.00	0.29	0.09	0.07	2.00	**1.00**	0.29
	2	0.22	1.00	0.36	0.08	0.05	2.00	**1.00**	0.36
	3	0.21	1.00	0.35	0.08	0.05	3.00	0.99	0.35
	4	0.20	1.00	0.34	0.09	0.06	3.00	0.99	0.33
NYCUNLP	0	0.14	1.00	0.25	0.12	0.08	3.00	0.99	0.25
	1	0.16	0.99	0.28	0.14	0.08	7.00	0.98	0.27
	2	0.17	0.95	0.28	0.16	0.08	10.00	0.96	0.27
	3	0.18	0.94	0.31	0.16	0.08	13.50	0.95	0.29
	4	0.20	0.93	0.33	0.16	0.07	18.00	0.93	0.31
FU-TU-DFKI	0	0.17	0.97	0.29	0.16	0.07	11.00	0.96	0.28
Capy-team	0	0.11	1.00	0.20	0.11	0.10	1.50	**1.00**	0.20
	1	0.11	1.00	0.20	0.11	0.10	**1.00**	**1.00**	0.20
	2	0.11	1.00	0.20	0.11	0.10	2.00	**1.00**	0.20
	3	0.11	1.00	0.20	0.11	0.10	**1.00**	**1.00**	0.20
	4	0.11	1.00	0.20	0.11	0.10	2.00	**1.00**	0.20
DS-GT	0	0.11	1.00	0.20	0.12	0.10	2.00	**1.00**	0.20
	1	0.11	1.00	0.20	0.12	0.10	2.00	**1.00**	0.20

Table 9
Ranking-based evaluation for Task 2.

		1 writing			100 writings			500 writings			1000 writings		
Team	Run	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$	$P@10$	$NDCG@10$	$NDCG@100$
HIT-SCIR	0	**1.00**	**1.00**	0.58	**1.00**	**1.00**	**0.84**	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	**0.90**
	1	**1.00**	**1.00**	0.58	**1.00**	**1.00**	**0.84**	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	**0.90**
	2	**1.00**	**1.00**	0.58	**1.00**	**1.00**	**0.84**	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	**0.90**
	3	**1.00**	**1.00**	0.58	**1.00**	**1.00**	**0.84**	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	**0.90**
	4	**1.00**	**1.00**	0.58	**1.00**	**1.00**	0.83	**1.00**	**1.00**	**0.89**	**1.00**	**1.00**	**0.90**
ELiRF-UPV	0	0.90	0.88	0.36	**1.00**	**1.00**	0.69	0.90	0.94	0.74	0.90	0.81	0.74
	1	0.30	0.25	0.32	**1.00**	**1.00**	0.45	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.46
	2	0.20	0.31	0.14	**1.00**	**1.00**	0.45	**1.00**	**1.00**	0.44	**1.00**	**1.00**	0.46
	3	0.90	0.94	0.35	**1.00**	**1.00**	0.68	0.60	0.46	0.60	0.70	0.63	0.63
	4	0.60	0.75	0.27	**1.00**	**1.00**	0.68	0.60	0.46	0.60	0.70	0.63	0.63
HU	0	0.90	0.81	0.53	0.80	0.87	0.49	0.70	0.68	0.48	0.70	0.66	0.49
	1	**1.00**	**1.00**	**0.62**	0.90	0.88	0.57	0.60	0.71	0.35	0.40	0.60	0.26
	2	0.30	0.21	0.11	0.20	0.16	0.12	0.00	0.00	0.11	0.40	0.60	0.26
	3	0.30	0.21	0.11	0.20	0.16	0.12	0.00	0.00	0.11	0.40	0.60	0.26
	4	0.60	0.53	0.33	0.40	0.58	0.36	0.30	0.37	0.24	0.40	0.60	0.26
UET-Psyche-Warriors	0	0.90	0.92	0.41	0.30	0.38	0.17	0.10	0.10	0.14	0.00	0.00	0.12
	1	0.90	0.93	0.43	0.10	0.12	0.11	0.00	0.00	0.12	0.00	0.00	0.12
	2	0.90	0.93	0.43	0.10	0.12	0.11	0.00	0.00	0.12	0.00	0.00	0.12
	3	0.90	0.93	0.43	0.10	0.12	0.11	0.00	0.00	0.12	0.00	0.00	0.12
	4	0.90	0.93	0.42	0.10	0.12	0.11	0.00	0.00	0.12	0.00	0.00	0.12
PJs-team	0	0.60	0.59	0.35	0.50	0.44	0.38	0.70	0.78	0.60	0.60	0.69	0.63
	1	0.60	0.59	0.35	0.40	0.41	0.39	0.50	0.60	0.54	0.50	0.63	0.51
	2	0.60	0.59	0.35	0.40	0.38	0.37	0.50	0.61	0.53	0.50	0.63	0.52
	3	0.60	0.59	0.35	0.30	0.32	0.36	0.60	0.66	0.51	0.50	0.66	0.51
	4	0.60	0.59	0.35	0.40	0.39	0.37	0.50	0.61	0.55	0.40	0.56	0.52
Lotu-Ixa	0	0.80	0.84	0.55	**1.00**	**1.00**	0.72	**1.00**	**1.00**	0.62	**1.00**	**1.00**	0.64
	1	0.90	0.94	0.57	**1.00**	**1.00**	0.73	**1.00**	**1.00**	0.63	**1.00**	**1.00**	0.63
	2	**1.00**	**1.00**	0.58	**1.00**	**1.00**	0.73	**1.00**	**1.00**	0.61	**1.00**	**1.00**	0.62
	3	0.90	0.81	0.58	**1.00**	**1.00**	0.74	**1.00**	**1.00**	0.61	**1.00**	**1.00**	0.62
	4	0.60	0.59	0.44	0.80	0.84	0.55	0.90	0.94	0.52	**1.00**	**1.00**	0.52
COTECMAR-UTB	0	0.30	0.23	0.23	0.00	0.00	0.22	0.20	0.15	0.18	0.20	0.13	0.17
	1	0.00	0.00	0.12	0.00	0.00	0.12	0.00	0.00	0.12	0.00	0.00	0.12
SINAI-UJA	0	**1.00**	**1.00**	0.59	0.80	0.87	0.53	0.90	0.88	0.54	0.90	0.92	0.54
	1	0.90	0.93	0.59	0.80	0.75	0.47	0.70	0.67	0.44	0.60	0.61	0.44
	2	0.90	0.92	0.58	0.70	0.79	0.47	0.90	0.94	0.53	**1.00**	**1.00**	0.52
	3	**1.00**	**1.00**	0.55	0.90	0.93	0.48	0.90	0.88	0.50	0.90	0.90	0.47
	4	**1.00**	**1.00**	0.57	0.60	0.74	0.45	0.70	0.76	0.52	0.60	0.70	0.51
NYCUNLP	0	0.50	0.53	0.42	0.70	0.68	0.35	0.70	0.62	0.33	0.50	0.47	0.31
	1	0.50	0.53	0.42	0.80	0.86	0.40	0.80	0.74	0.35	0.70	0.62	0.34
	2	0.50	0.53	0.42	0.80	0.86	0.45	0.80	0.86	0.40	0.80	0.74	0.36
	3	0.50	0.53	0.42	0.80	0.88	0.50	0.70	0.82	0.41	0.70	0.69	0.37
	4	0.50	0.53	0.42	0.70	0.69	0.45	0.70	0.82	0.42	0.70	0.69	0.39
FU-TU-DFKI	0	0.90	0.94	0.44	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00
Capy-team	0	0.20	0.18	0.13	0.20	0.18	0.10	0.10	0.07	0.09	0.20	0.15	0.14
	1	0.00	0.00	0.07	0.10	0.08	0.09	0.10	0.06	0.08	0.10	0.19	0.18
	2	0.10	0.07	0.12	0.10	0.07	0.12	0.00	0.00	0.11	0.20	0.13	0.16
	3	0.20	0.29	0.16	0.10	0.10	0.14	0.20	0.26	0.14	0.10	0.12	0.10
	4	0.00	0.00	0.11	0.30	0.20	0.12	0.00	0.00	0.08	0.10	0.19	0.13
DS-GT	0	0.20	0.12	0.22	0.00	0.00	0.12	0.20	0.18	0.20	0.10	0.12	0.17
	1	0.90	0.92	0.52	0.00	0.00	0.12	0.20	0.18	0.20	0.10	0.12	0.17

4. Pilot Task: Conversational Depression Detection via LLMs

We introduced this pilot task in 2025 as a novel challenge to seek the opportunity of embracing conversational agents in detecting depression symptoms. Participants were interacting with LLM-based personas who have been instructed using user writings, simulating real-world conversational exchanges and example user profiles. Twelve distinct personas were instantiated with ChatGPT.

The challenge lies in asking participants to determine whether the LLM persona exhibits signs of depression and, if so, what is the level of depression severity and key depression symptoms expressed over conversations. The diagnostic target for the LLMs was framed in terms of the BDI-II, as in Task 1. The BDI-II is a 21-item self-report questionnaire widely used in clinical psychology, which are listed in the Table 10.

Table 10
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 item corresponds to a concrete symptom. For example, *Sadness*, *Loss of Energy*, or *Indecisiveness*. Each symptom is scored 0 to 3 according to severity. Table 11 shows the possible response options (0-3) for the symptoms *Sadness* and *Self-Dislike*. The sum of all 21 symptoms yields a global index in the range 0–63. The scores are interpreted into four categories: 0-9 are interpreted as *minimal* depression, 10–18 as *mild*, 19–29 as *moderate*, and 30 or above as *severe*. Because the personas are simulations, no ground-truth questionnaire exists; instead, a group of three clinicians examined the seed user data that shaped each persona and agreed on both an overall BDI-II score and the subset of symptoms included. These consensual judgments constitute the gold standard.

Table 11
Two BDI-II symptoms and their four response options (0–3).

Symptom	Response options (score)
Sadness	0: I do not feel sad. 1: I feel sad much of the time. 2: I am sad all the time. 3: I am so sad or unhappy that I can't stand it.
Self-Dislike	0: I feel the same about myself as ever. 1: I have lost confidence in myself. 2: I am disappointed in myself. 3: I dislike myself.

Participants did not receive any labelled training material. We deliberately framed the task as *training-less* to encourage a variety of methodological responses,ranging from rule-based interviewers and zero-shot LLM prompts to different classifiers trained on public mental-health corpora. During the test window, teams accessed the links we provided them through ChatGPT interface for creating the dialogue with the LLM-persona. The participant systems interacted with a free-form prompt; the server produced the next turn, and so on. This loop continued until the system chose to terminate the dialogue and submit its diagnosis. Since this is a pilot task, there was no hard cap on the number of turns, but we encouraged the participants to produce their decisions as early as possible.

After ending the conversation with a persona, a participating system had to return two files. The first was a structured log that preserves, in chronological order, every prompt–response pair exchanged with the agent; this file serves auditing and qualitative analysis. The second was a JSON record containing three fields: the predicted BDI-II score (an integer 0–63), the corresponding severity category, and up to four symptom drawn from the BDI-II list in Table 10, that best explained the score.

4.1. Pilot Task: LLM Personas Design and Construction

We adopted a clinician-in-the-loop design workflow to build the twelve LLM personas. A team of three clinical psychologists co-designed a template that captures both general biographical detail and clinically information. Using this template we instantiated a pool of draft personas with GPT-4o, each conditioned on a different user history.
The same clinicians then conducted free-form interviews with every draft, rating each dialogue along two main dimensions:

- The *overall* dimension covered traits associated with conversational attributes: human-likeness, lexical fluency, coherence, and affective naturalness.
- The *diagnostic* dimension targeted domain realism, including emotional consistency, fidelity to depressive symptomatology, willingness to elaborate, and cognitive style (rumination, processing speed, abstraction level).

Feedback was recorded on a five-point Likert scale and complemented with qualitative comments. Insights from this evaluation cycle informed a second engineering pass in which every persona was represented through a structured prompt comprising the main following elements:

- Core profile. A stable set of attributes: name, age, gender, marital status and an a pre-defined BDI-II score.
- Key negative symptoms. Up to four key BDI-II symptoms (or less for control personas) that the agent should manifest recurrently and coherently.
- Memory and reflection. Specific snippets describing life history, social context, and salient past events; these cues allow the agent to maintain narrative continuity and to provide retrospective insight into its mood.
- Language and communication style. Use of vocabulary, and typical sentence length so that each persona speaks with a recognisable "voice".
- Behavioural constraints. Guard-rails that prohibit explicit self-diagnosis and that keep the agent away from clinical recommendations, thereby forcing participants to infer depression indirectly.
- Response goals. High-level objectives such as "answer candidly but not expansively," "avoid mentioning diagnosis unless prompted," and "display mild self-disclosure".
- Environment and context. Brief situational framing (e.g. studying for exams, recent job change) that provides topical depth without locking the dialogue.
- Few-shot exemplars. Short question–answer pairs illustrating the expected tone and symptom expression.
- Restricted responses. A blacklist of phrases that would break immersion (e.g. "As an AI language model…") replaced with context-appropriate alternatives.

The final personas were frozen only after a second round of clinician interaction confirmed that they satisfied a minimum threshold on both the overall and diagnostic scales. This iterative, expert-guided construction process proved essential to achieve dialogues that are simultaneously natural and diagnostically meaningful, laying the groundwork for future large-scale evaluations of conversational mental-health screening systems.

4.2. Pilot Task: Participant Teams

Table 12 shows the participant teams and some statistics about their interactions such as the mean number of messages per run, and the mean number of characters per message. The numbers reveal a wide range of interaction strategies:

Table 12
Pilot task (LLMs): participating teams, number of runs, mean number of messages per run, mean number of characters per message.

Team	#Runs	#Mean messages per run	#Mean characters per message
ixa-ave [28]	4	31.02	414.44
SINAI-UJA [44]	3	6.54	488.25
DS-GT [45]	4	20.79	782.81
PJs-team [29]	1	7.67	1045.16
LT4SG	1	10	40.73

- ixa-ave submitted the maximum number of runs (four) and tended to carry out relatively lengthy dialogues (≈ 31 messages each) while keeping their prompts concise (≈ 415 characters per turn).
- SINAI-UJA used a fast approach, with only 6–7 turns on average, yet still packed almost 490 characters into every message, suggesting dense, information-rich questioning.
- DS-GT followed an intermediate approach, with ≈ 21 messages per run and 783 characters per message, balancing breadth and depth of interaction.
- PJs-team produced long messages ($\approx 1\ 045$ characters) within a limited number of turns (≈ 8), delivering extended prompts.
- LT4SG employed a fixed sequence of ten short messages averaging only 41 characters, representing the most lightweight strategy.

4.3. Pilot Task: Evaluation Metrics

Based on evaluation metrics that have been developed from eRisk 2019 [48], 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.

4.4. Pilot task: Results

Table 13 presents the official runs, ranked by best ADODL. The strongest submission, *SINAI-UJA (run 1)*, achieves an ADODL of 0.93, meaning the predicted scores differ by less than five points on average from the clinician reference. Its DCHR of 0.58 shows that most of these small errors still fall within the

incorrect severity band. *DS-GT* attains comparable category accuracy (0.50) with only a modest drop in ADODL, rearching similar level reliability despite larger absolute score errors.
Across all teams, however, symptom recognition stays behind score estimation: even the best ASHR values hover below 0.30, indicating that systems often capture the global severity signal without isolating which symptoms drive it.

Table 13
Evaluation for Task 3 with teams ordered in terms of best ADODL. '*' indicates the manual runs (human-in-the-loop).

Team	Run	DCHR	ADODL	ASHR
SINAI-UJA	0	**0.66**	0.92	0.21
	1	0.58	**0.93**	**0.29**
	2	0.41	0.88	0.21
DS-GT	0	0.42	0.83	0.12
	1	0.50	0.89	0.27
	2	0.33	0.86	**0.29**
	3	0.50	0.84	0.25
ixa_ave	0*	0.33	0.80	0.25
	1	0.33	0.76	**0.29**
	2	0.33	0.83	0.21
	3	0.17	0.81	0.19
LT4SG	0	0.33	0.78	0.06
PJs-team	0	0.33	0.73	0.25

5. Participating Teams

Table 14 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 2025 working notes proceedings for the participants.
Lotu-ixa [42]. The Lotu-ixa team, affiliated with University of the Basque Country, in Spain, participated in task 2 proposed as part of eRisk CLEF this year. The team proposes a method to (i) apply a semantic relabelling process to the training data, (ii) then design and fine-tune a classification model, and (iii) finally combine risk signals derived from both the target user and the conversational context. For (i), a similarity score was computed for representative positive and negative examples, and then a percentile-based strategy determined the messages suitable for relabelling. The classifier (ii) was derived from XLM-RoBERTa and fine-tuned on the relabelled dataset from (i) using a binary cross-entropy loss function, with optimised hyperparameters explored via grid search. Finally, (iii) the team computed user risk, context risk and thread risk scores, calculating a binary decision based on these. The team performed five runs using different thread risk score settings. In their run #4, they obtain the best recall score (1.0), and runs #0-#3 yielded the best ERDE5 (0.05) among the participants of this task. For the ranking-based metrics, their approach demonstrates highly competitive performance in precision and NDCG (most of their runs achieve 1.0). Their approach has been competitive across all ranking metrics.

Table 14
Participation across the three eRisk 2025 tasks: number of submitted runs per team.

Team	Task 1 # runs	Task 2 # runs	Task 3 # runs
SonUIT [26]	5	0	0
ThinkIR [27]	5	0	0
Ixa_ave [28]	5	0	4
Synapse	3	0	0
PJs-team [29]	5	5	1
ELiRF-UPV [30]	5	5	0
COTECMAR-UTB [31]	3	2	0
COMFOR	1	0	0
LHS712-Team-1 [32]	5	0	0
Team-Gryffindor	1	0	0
INESC-ID [33]	5	0	0
UET-Psyche-Warriors [34]	5	5	0
NYCUNLP	5	5	0
BGU-Data-Science [35]	5	0	0
HULAT_UC3M [36]	5	0	0
RELAI	2	0	0
UniORNLP-dahlia	2	0	0
Lotu-ixa [42]	0	5	0
HIT-SCIR [43]	0	5	0
SINAI-UJA [44]	0	5	3
DS-GT [45]	0	2	4
Capy-team	0	5	0
HU [46]	0	5	0
FU-TU-DFKI [47]	0	1	0
LT4SG	0	0	1
Total runs	**67**	**50**	**13**

SINAI-UJA [44]. SINAI-UJA team, from the University of Jaen (Spain), participated in tasks 2 and 3 of the eRisk 2025 challenge. For task 2, the team relied on the provided train and test sets, but also developed a new dataset for this task, to be able to have training data with context. The team fine-tuned RoBERTa and Mental RoBERTa models in different settings, optimised with Optuna, and performed five runs using different settings of model and parameter combinations. Their system was one of the top three in terms of efficiency, completing the task in less than 10 hours. Their system achieved a perfect recall (1.0) in all runs, but with the cost of having a low precision (0.17-0.24) and low F1 (0.29-0.39). In the ranking-based evaluation, the team performed competitively at early stages, aligning the the top-ranked teams. For Task 3, the SINAI team proposed a modular system composed of two collaborating LLMs: (1) is responsible for interacting with the user, and (2) does not interact with the user, but receives the conversation and analyses it and updates the state of the depressive symptoms. Moreover, this LLM reasons whether it needs more information or not, ending the conversation when needed. The team uses Llama-3.1-8B-Instruct model for both LLMs. They submitted three runs, with different prompt configurations, achieved the fastest interaction with an average of 6.54 messages per run, as well as achieving the best overall ADODL (0.93), ASHR (0.29) and DCHR (0.66), highlighting that the estimations were highly aligned with the BDI-II levels of the simulated personas and that their approach effectively identified key symptoms.

COTECMAR-UTB [31]. COTECMAR-UTB, affiliated with Universidad Tecnologica de Bolivar, in Colombia, participated in Tasks 1 and 2 proposed as part of eRisk CLEF this year. For Task 1, the team focused on high-confidence training data and balanced the data using EDA and SMOTE. The authors propose a pipeline that includes data preprocessing and cleaning, training ML models, including LR, SVM and BERT, among others. After that, they apply VADER to identify texts with negative sentiment and score the sentences. They submitted one run, achieving a middle-tier performance. For Task 2, the team trained an LSTM model to predict the risk of depression. The team submitted 2 runs, with moderate performance, achieving a best F1 of 0.40 and a Recall of 0.65. For the ranking-based metrics, the metrics have room for improvement, suggesting that the model had difficulties when prioritizing relevant messages.

HULAT_UC3M [36]. The HULAT-UC3M team, affiliated with Universidad Carlos III, from Madrid (Spain), participated in Task 1 proposed as part of eRisk 2025 challenge. The team proposed training a multi-classifier (SVM) to classify all the sentences into their corresponding symptoms, keeping only the ones with higher probabilities according to different thresholds; filtering sentences according to different criteria for each run; and scoring the sentences using either VADER or roberta-base-sentiment. The authors use the training data with unanimity to minimise noise. Their best run uses RoBERTa, selecting the top 1000 sentences based on confidence scores, achieving an AP of 0.018. Their two runs using high-confidence-based filtering had a positive impact on the performance, but the scoring method can be improved.

BGU-Data-Science [35]. BGU-Data-Science team, affiliated with Ben-Gurion University of the Negev, in Israel, participated in Task 1 of the eRisk 2025. The authors approached the task as a sentence ranking problem by computing the semantic similarity between user sentences and BDI-II symptom descriptions embedded using Sentence-BERT. The team performed query expansion and filtered out sentences that were not in the first person. For the first person filtering, the team employed three different methods: a basic filtering approach using first-person pronouns, a method using spaCy, and Claude Sonnet 3.7 to assess whether a sentence conveys the user's personal experience. The team achieved their best results with the baseline approach, using only embeddings from Sentence-BERT, which resulted in an AP score of 0.240. Although incorporating query expansion and first-person filtering did not yield the highest AP, it did achieve the highest P@10 compared to other runs from the team.

INESC-ID [33]. The INESC-ID team, affiliated with University of Lisboa, in Portutal, participated in the first task of the eRisk Lab. Although this task is framed as an information retrieval challenge, the authors approach it as a regression or classification problem. The team explored several methods, including fine-tuned foundation models (DeBERTa-v3-large), unsupervised similarity based approaches, and LLM-based classification using GPT-4o-Mini. The authors make use of the training data provided for this task to train and validate their approaches. The DeBERTa model was finetuned for regression to predict a relevancy score ranging from 0 to 1, while the other two methods were framed as binary classification tasks. The best-performing run of the team was an ensemble approach that combined outputs from all the methods, achieving the highest scores AP, R-PREC, and P@10.

HU [46]. The HU team, affiliated with the Habib University, in Pakistan, participated in Task 2 from the eRisk 2025 challenge. The runs submitted by the team cover a wide range of approaches, including transformer-based models (ModernBERT) with time-aware loss or data augmentation strategies, Llama 3.1 summarization with BERT classification, a zero-shot model using Llama-4-Scout-17B, and a simple threshold approach. The best performing method, using Llama 3.1 for summarization and BERT classification with an incorporated alert policy (run #1), achieved an F1 score of 0.75, ranking 3rd out of 12 teams in decision-based evaluation. In ranking-based evaluation, the same run obtains a perfect score of 1.00 in P@10 and NDCG@10 after one writing.

FU-TU-DFKI [47]. The FU-TU-DFKI team is affiliated with three different organizations from Germany: the Freie Universität Berlin, University of Hannover, and the Technical University of Berlin. They participated in Task 2 of eRisk 2025. The authors conducted two pilot studies that focused on the linguistic analysis of the dataset provided for the task. The first study examined the use of first-person singular pronouns and the verbs commonly associated with them. The second study involved a concept analysis of the keywords found in the data. The insights gained from these studies helped inform their proposed method. The team's hybrid system combines a transformer-based model (MentalBERT) with linguistically informed features, such as the use of first person pronouns and associated verbs, as well as other relevant keywords. In addition, the system incorporates metadata, including late-night posting frequency and the sentiment of the posts. The team achieved modest results by processing only 449 out of a total of 1 280 user threads.

ThinkIR [27]. The ThinkIR team comes from two organizations in India, the Indian Institute of Science Education and Research Kolkate, and the Vellore Institute of Technology. ThinkIR submitted five runs for Task 1. Four rely on classical IR ranking with different query expansion strategies, namely kNN word embedding expansion, pseudo relevance feedback (PRF), and GPT generated prompt reformulations, while the remaining one uses a RoBERTa based multi label classifier. The best run, which involves RoBERTa fine tuning, achieved an AP of 0.068, R Precision of 0.157, P@10 of 0.409, and NDCG of 0.228, leading every metric among their runs. The experiments confirm that transformer fine-tuning outperformed all classical expansion methods, although PRF on the top ten documents still produced competitive rankings.

Ixa_ave [28]. The ixa_ave team is affiliated with the HiTZ Basque Center for Language Technology, from the University of the Basque Country (Spain). ixa_ave took part in task 1 and the inaugural pilot Task at eRisk 2025. For task 1 they fine-tuned multilingual BERT, appending a 21-dimensional vector of cosine-similarity scores to each sentence and predicting with a 21-head classifier. They tried two similarity-based data-reduction ideas: (i) skip training sentences whose similarity to any BDI-II item exceeds $\beta = 0.5$, and (ii) at inference keep only sentences whose similarity is at least $\theta \in 0.3, 0.5$. Among the five submitted runs, base_filter30 ($\theta = 0.3$, no training pruning) was best, reaching AP = 0.102 under majority voting. In the Pilot Task they compared a manual questionnaire interview (run 0) with three LLM agents: GPT-4-long (run 1), GPT-4-short (run 2) and Falcon-11B (run 3). Both GPT-4 variants matched the human baseline on DCHR = 0.33, whereas Falcon obtained worse results, with 0.17.

UET-Psyche-Warriors [34]. The UET-Psyche-Warriors team is affiliated with the VNU University of Engineering and Technology, in Vietnam. The authors participated in Task 1 and Task 2 of eRisk CLEF 2025 challenge. For task 1, they explored both semantic similarity-based ranking and a machine learning approach using a multi-task DepRoBERTa model fine-tuned for symptom detection and severity estimation. Their best run (Run 4) achieved an NDCG of 0.623 and an AP of 0.339, ranking second overall. For Task 2, the team implemented a multi-stage system combining sentence-level severity scoring with rule-based aggregation strategies. Run 2, which incorporated temporal accumulation with a bonus heuristic, achieved their best results with an F1 score of 0.73 and a latency-aware F1 of 0.68, placing them fourth overall.

ELiRF-UPV [30]. The ELiRF-UPV team, affiliated with Polytechnic University of Valencia, in Spain, participated in Tasks 1 and 2 of the eRisk 2025 challenge. For Task 1, the team developed an adapter architecture over pre-trained sentence similarity models, incorporating attention over reference embeddings derived from both cluster centroids and the BDI-II question-answer pairs. For Task 2, they explored three approaches: a classical SVM classifier, a Longformer fine-tuned on user-level data, and a task-adapted Longformer model trained using a data augmentation strategy designed to simulate early detection conditions. Their best-performing system in Task 2 was a Linear SVM using TF-IDF features, ranking 6th overall in the competition.

HIT-SCIR [43]. The HIT-SCIR is affiliated with the Harbin Institute of Technology, in the Univervisty of Harbin (China). They participated in task 2 of the CLEF 2025 eRisk Lab. Their proposal focuses on contextualized early detection of depression on social media, utilizing a multi-stage framework. Their approach addresses the challenge of limited interactive context in training data by employing LLMs for contextual data augmentation. Specifically, they use LLMs to simulate social interactions, generating comments for original user posts and then summarizing these comments to create a rich semantic context. A core component of their system is a psychiatric scale-guided risky post screening module, which identifies depression-related information from user post histories. This module calculates a risk score for each post based on its cosine similarity with symptom descriptions from established psychological scales, like the BDI-II. Posts with higher risk scores are then filtered for depression risk detection. The detection itself uses MentalBERT, a BERT variant optimized for mental health texts, to generate post embeddings, and a Transformer with attention mechanisms to model inter-post interactions and generate user features. The entire screening and detection process is trained end-to-end using a Straight-Through Estimator (STE). For early detection testing, a dynamic risky post queue and different alerting strategies are employed. The team submitted five runs with varying operational parameters for their dynamic user-level early risk assessment strategy, using a voting ensemble of their top three performing models. This integrated approach led to strong performance, achieving first rank in several evaluation metrics, including F1 (0.85 for HIT-SCIR-4), ERDE50 (0.03 for HIT-SCIR-4), and Flatency (0.82 for HIT-SCIR-2 and HIT-SCIR-4). They also achieved first place in the majority of ranking-based metrics, such as P@10 and NDCG@10 across almost all writings evaluations.

PJs-team [29]. The PJs-team, affiliated from Netaji Subhas University of Technology , from India, presented distinct approaches for three tasks. In the task 1, the team used finetuned bi-encoders (e.g., DistilRoBERTa, e5-small) with CoSENTLoss and their ensemble using Reciprocal Rank Fusion (RRF). They also employed finetuned cross-encoders ('ModernBERT-large', 'ModernBERT-base') with BinaryCrossEntropyLoss for reranking, and reranker ensembles using majority voting or scaled mean averaging. The cross-encoder ensemble run gave their top scores (AP 0.279, P@10 0.800). In task 2, they presented a two-stage pipeline first filters each new post with a custom DistilRoBERTa sentence-transformer against four early BDI cues (pessimism, punishment feelings, self-dislike, indecisiveness). High-scoring texts or users previously flagged are analysed by an ensemble of four hosted LLMs (Claude 3.7 Sonnet, Amazon Nova Pro, Llama 3-70B, Claude 3.5 Haiku). Majority vote delivers the final decision. The single-model Sonnet run achieved the team's best F1 = 0.71 with low ERDE@5 = 0.09. In task 3, they built a single LLM agent (Claude Sonnet) driven by a long system prompt embedding the full BDI-II questionnaire. The agent chats about movies to elicit emotions and updates the 21 BDI scores each turn, ending when all scores are set. On the pilot evaluation it reached ADODL 0.73 and DCHR 0.33.

LHS712-Team-1 [32]. The LHS712Team comes from School of Information & Department of Learning Health Sciences, in the University of Michigan, USA. The authors participated in task 1, and benchmarked a wide spectrum of ten runs, covering: (i) Classical baselines, Logistic Regression and SVM coupled with CountVectorizer or TF-IDF features. (ii) Domain specific embeddings, ClinicalBERT and Sentence-BERT sentence vectors fed into Linear-SVC or LR classifiers. (iii) They also fine-tuned BERT, with a "[SYMPTOM] [SEP] sentence" formulation finetuned for five epochs, where a symptom keyword filter first pruned the 17 million sentence test set to keep inference tractable. (iv) A method based on hybrid retrieval, where BM25 selects candidates that are reranked by SBERT cosine similarity. Finally, the fine-tuned BERT with unanimous-label training was their top performer, yielding AP 0.078, R-Prec 0.169, P@10 0.344 and NDCG 0.287 on the official unanimity evaluation, well above their traditional baselines.

DS-GT [45]. The DS-GT team from the Georgia Institute of Technology, in USA, participated in the task two and the pilot task. In task 2, the team contrasted two pipelines: Voting Classifier combining engineered features (TF-IDF, VADER sentiment, LIWC-style counts, posting-gap timings) in a soft vote ensemble of Random Forest, SGD-LogReg and Gradient Boosting. Here, lightGBM + temporal attention where MentalRoBERTa sentence embeddings feed a linearly-weighted recency mechanism and a sparse "depression-indicator" content matrix before classification. Both runs achieved recall = 1.0 but low precision (P = 0.11, F1 = 0.20) and identical $ERDE_5$ = 0.12, with the embedding-based model yielding far better ranking scores (P@10 = 0.90, NDCG@10 = 0.92 on the 1-writing cut). In the pilot Task, a unified prompt-engineering framework used several LLMs (Claude 3.7 Sonnet, GPT-4o, Gemini Flash/Pro) to conduct $\approx$ 20 turn interviews, outputting structured JSON with item-level BDI-II scores and key symptoms. The best run (Claude Sonnet) placed second overall (DCHR 0.50, ADODL 0.89, ASHR 0.27). Exploratory analysis showed strong cross-model consistency (R2 = 0.91 between label level and BDI score) but wide variance on appetite and agitation cues.

SonUIT [26]. The SonUIT team is affiliated with the University of Information Technology (UIT), in Vietnam, and participated in task 1. Their system uses a two-stage pipeline: (i) Filtering, where they build averaged all-MiniLM-L6-v2 embeddings for each BDI-II symptom and pull the top 1000 sentences per symptom via cosine similarity. (ii) Reranking, where the candidate set is optionally resorted with BM25, a cross-encoder, or larger embedding models (bge-large-en-v1.5 and text-embedding-3-large). Five runs explored raw vs. pre-processed text and the different rerankers. Their configuration #2 (pre-processed text + embedding filter) posted the team's best scores and placed within the top-three teams on every evaluation metric (MAP = 0.334, R-Prec = 0.392, P@10 = 0.790, NDCG@1000 = 0.613).

6. Conclusions

This paper provided an overview of eRisk 2025, the ninth edition of the eRisk lab, which moved toward two new tasks that require richer conversational understanding and interactive settings. The Task 1, which was the final edition of the sentence-ranking challenge for BDI-II symptoms, attracted 67 runs from 17 teams. Task 2 introduced full-thread context for the first time in early detection of depression. In this task, we received 50 runs from 12 teams, and showed that models able to exploit dialogue structure can issue accurate alerts after remarkably few turns, although a clear trade-off persists between earliness and recall. The pilot task went a step further, replacing static corpora with live interaction against LLM-driven personas. Despite the absence of training data, five teams submitted 13 runs; top systems achieved near-perfect BDI-II score estimation yet still struggled to pinpoint the specific symptoms that reflect those scores, highlighting the difficulty of symptom-level grounding in open conversation.
Taken together, the 130 runs submitted this year confirm both the community's engagement and the practicality of evaluation settings that approach real conversational use cases. Three broad lessons emerge: adding even modest context improves detection, timeliness must remain a core metric. Moreover, clinician-guided LLM personas, despite having a lot of room for improvement, are able to create realistic yet privacy-preserving frameworks. Future eRisk editions will continue to shift toward dialogue-centric tasks and deeper integration of LLM capabilities to keep pace with how people communicate online and how assistive technologies are deployed.

7. 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.

8. Declaration on Generative AI

During the preparation of this manuscript, generative AI tools were employed solely for light editing purposes, including proofreading, grammar correction, vocabulary improvement, and overall language polishing. All substantive ideas, analyses, experiments, and written content were created by the co-authors without direct text generation from any AI model.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] N. M. Son, D. V. Thin, Sonuit eRisk2025: Enhanced depression detection on social media via filtering and re-ranking, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[27] S. Adhikary, J. Das, D. Roy, Thinkir at eRisk 2025: Early detection and risk assessment of depression using transformer models, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[28] A. Varela, M. Oronoz, A. Casillas, A. Pérez, Detection of depression with symptom similarity: Data reduction and llm personas, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[29] P. Vachharajani, Transformer ensembles and llm-powered approaches for depression symptom analysis and contextualized early risk detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[30] A. C. Segarra, V. A. Esteve, A. M. Marco, L.-F. H. Oliver, Elirf-upv at eRisk 2025: New approaches to the detection and early detection of symptoms and signs of depression, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[31] L. F. M. Cardona, J. M. S. Loaiza, E. A. P. D. Castillo, J. C. M. Santos, J. E. S. Castañeda, Cotecmar-utb at eRisk 2025: Semantic-centroid symptom ranking and early depression detection using adaptive decision rule, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum,

Madrid, Spain, September 9-12, 2025.

[32] A. Benloucif, Y. Nannapuraju, S. Bellam, Y. Hu, Z. Zhao, V. Vydiswaran, Lhs712team-1 at eRisk@clef 2025: Searching for depression symptoms using various natural language processing algorithms, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[33] D. A. Nunes, E. Ribeiro, Inesc-id @ eRisk 2025: Exploring fine-tuned, similarity-based, and prompt-based approaches to depression symptom identification, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[34] T.-P. Mai, M.-H. L. H., D.-L. Tran, D.-C. Can, H.-Q. Le, Uet@eRisk2025: Severity estimation for depression symptoms searching and early risk detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[35] N. Munz, E. Aharon, A. Segal, K. Gal, Semantic retrieval of bdi symptoms in user writings, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[36] J. C. Molina, P. M. Fernandez, Hulat-uc3m at task 1@eRisk 2025: Detecting depression using machine learning approaches, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[37] 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.

[38] D. Otero, J. Parapar, Á. Barreiro, Beaver: Efficiently building test collections for novel tasks, in: Proceedings of the First Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020), Samatan, Gers, France, July 6-9, 2020, 2020.

[39] D. Otero, J. Parapar, Á. Barreiro, The wisdom of the rankers: a cost-effective method for building pooled test collections without participant systems, in: SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, Virtual Event, Republic of Korea, March 22-26, 2021, 2021, pp. 672–680.

[40] 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).

[41] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: WSDM, ACM, 2018, pp. 495–503.

[42] X. Larrayoz, A. Casillas, A. Pérez, Leveraging conversational context and semantic relabeling for early depression detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[43] Y. Zi, B. Wang, Y. Zhao, B. Qin, Hit-scir@eRisk2025: Exploring the potential of a learnable screening model and risk post buffer-based framework for contextualized early prediction of depression on social media, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[44] A. M. Mármol-Romero, M. García-Vega, M. Ángel García-Cumbreras, A. Montejo-Ráez, Sinai at eRisk@clef 2025: Transformer-based and conversational strategies for depression detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[45] D. Guecha, Y. Chiu, A. Miyaguchi, S. Gaur, Ds@gt at eRisk 2025: From prompts to predictions, benchmarking early depression detection with conversational agent based assessments and temporal attention models, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[46] M. Saad, M. Abbas, A. U. Chaudhry, F. Alvi, A. Samad, Contextualized early detection of depression – hybrid and time-aware approaches: Hu at eRisk task 2 2025, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[47] E. Kara, R. E. M. Peña, L. Raithel, Fu-tu-dfki@eRisk 2025: A linguistically informed but overdiagnosing approach to early depression detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, Madrid, Spain, September 9-12, 2025.

[48] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk 2019 early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10, Springer, 2019, pp. 340–357.