Awakened at eRisk 2026: A Multi-Agent LLM Architecture for Automated BDI-II Depression Screening via Simulated Clinical Interviews

Notebook for the eRisk Lab at CLEF 2026

Cosmin-Gabriel Becheru[1], Ciprian-Octavian Truică[1,2,*] and Elena-Simona Apostol[1]

[1]National University of Science and Technology POLITEHNICA Bucharest, Splaiul Independenţei 313, Bucureşti 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania

Abstract

This paper presents the system developed by team Awakened for Task 1 of the eRisk 2026 Lab at CLEF 2026. The task requires automatically conducting clinical interviews with simulated patients, estimating Beck Depression Inventory-II (BDI-II) scores, and identifying depressive symptom patterns. Our approach uses a multi-agent conversational architecture, where three LLM-based components: a Doctor agent, an Analyst agent, and a Patient simulator, each take on a separate role in the screening process. We submitted two runs, each illustrating a different architectural decision. In Run 1, all three agents run locally on the same model (Meta-Llama-3-8B-Instruct), with role specialization handled through LoRA adapter switching rather than separate model instances. Run 2 takes a composite approach: the Doctor and Analyst roles are offloaded to a larger model (Llama-3.3-70B-Versatile) accessed via the Groq API, while patient simulation remains local, using the fine-tuned Llama-3-8B with LoRA adapters, as required by task constraints. We go over how the prompts were set up to shape each agent's behavior, and how we built the heuristics that turn what the agent says in conversation into actual BDI-II scores. There are also early-stopping rules in place so the interview does not continue indefinitely without a clear endpoint. The two runs we did gave us a better picture of what you gain and what you lose when everything stays local versus when you hand things off to a stronger model on the outside, especially in how well the system reasons clinically and how close the scores end up to the real ones.

Keywords

depression detection, BDI-II, LLMs, multi-agent systems, simulated clinical interviews, eRisk, CLEF 2026

1. Introduction

Depression is a common mental health condition, affecting more than 280 million people worldwide according to the World Health Organization [1]. Early and accurate screening helps ensure early intervention, but many regions have limited access to mental health professionals [2]. The Beck Depression Inventory-II (BDI-II) [3] is a widely used self-report instrument with 21 items that measure how severe depressive symptoms are, with total scores from 0 (minimal) to 63 (severe depression).

The eRisk Lab, part of the CLEF evaluation forum, is now a benchmark for early risk prediction [4, 5]. In the 2026 edition, Task 1 asks participants to interact with 20 simulated patients. Each patient is represented by a Large Language Model (LLM) fine-tuned with patient-specific LoRA adapters [6]. The goal is to conduct a clinical interview with a set number of conversation turns, estimate a BDI-II score for each patient, and list key depressive symptoms.

The task comes with three challenges. First, systems operate under a limited interaction budget, meaning that clinically relevant information must be extracted within a limited number of dialogue turns. Second, the patient simulator is fixed (Llama-3-8B with LoRA adapters), so participants must design their interviewing strategy independently. Third, estimating a BDI-II score from free-text conversation

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
cosmin.becheru@upb.ro (C. Becheru); ciprian.truica@upb.ro (C. Truică); elena.apostol@upb.ro (E. Apostol)
https://sites.google.com/view/ciprian-octavian-truica (C. Truică); https://sites.google.com/view/elena-simona-apostol (E. Apostol)
0009-0002-7720-8048 (C. Becheru); 0000-0001-7292-4462 (C. Truică); 0000-0001-6397-4951 (E. Apostol)

requires mapping qualitative patient responses to quantitative symptom severity ratings, which involves a non-trivial degree of clinical reasoning.

This paper presents a multi-agent architecture developed to address Task 1. The proposed approach divides the screening process into three collaborating agents: (i) a Doctor agent that generates empathetic, clinically directed questions; (ii) an Analyst agent that evaluates each patient response for BDI-II symptom indicators; and (iii) the Patient simulator provided by the task organizers. Two runs were submitted to explore different configurations of this architecture, as detailed in Section 4.

The remainder of this paper is organized as follows: Section 2 reviews related work on LLM-based mental health applications. Section 3 describes the task definition. Section 4 presents the methodology. Section 5 reports and discusses the results. Section 6 concludes with future directions.

2. Related Work

LLM-Based Mental Health Assessment. Recent studies have investigated the application of large language models (LLMs) in mental health contexts, including depression detection from social media text [7, 8], studies on the prevalence and impact of diagnosed and undiagnosed depression [9], and reviews of prognosis and treatment outcomes in major depression [10]. For example, models such as MentalBERT [11], pre-trained on mental health discourse from platforms such as Reddit, demonstrate improved performance on depression classification tasks.

Multi-Agent LLM Architectures. The Transformer architecture introduced by Vaswani et al. [12], and the few-shot learning abilities described by Brown et al. [13] made it possible to create LLM agents with specific roles. Building on this, Petrescu et al. [14] developed a framework that combines event detection and sentiment analysis. They showed that breaking down a complex NLP task into smaller, specialized parts and then combining the results can outperform a single model. Although class imbalance can highly influence performance [15], ensemble models manage to mitigate against this shortcoming. Our work applies this modular approach to BDI-II depression screening, with different agents for question generation, patient interaction, and symptom analysis.

Prior eRisk Editions. Since 2017, the eRisk Lab has organized tasks on early risk detection, starting with a depression test collection by Losada and Crestani [16]. Later overviews included topics like depression [4], pathological gambling [17], and anorexia [18]. Earlier editions mainly analyzed social media posts for signs of risk. In 2025, the focus shifted to early detection of depression using full conversation threads [5]. The 2026 Task 1 [19, 20] represents a further evolution by requiring active participation in the clinical interview process through simulated patient interaction.

3. Task Overview

Task 1 of eRisk 2026 requires participants to interact with 20 simulated patients, each implemented using a Meta-Llama-3-8B-Instruct model augmented with a patient-specific LoRA adapter hosted on Hugging Face (`Anxo/erisk26-task1-patient-00..19-adapter`). Each patient model runs under a standardized system prompt instructing it to behave realistically, maintain a natural speaking style with short answers and hesitations, and avoid disclosing that it is an AI.

For each patient, participants must conduct a multi-turn conversation, limited to a specified maximum number of turns, by generating doctor questions that address the BDI-II symptom domains. At the end of each conversation, the system generates a final output containing an estimated BDI-II score (an integer between 0 and 63) and a list of up to four key symptoms identified during the interaction, formatted according to the following JSON schema:

```
{
    "LLM": "<patient_id>",
```

```
    "bdi-score": <integer>,
    "key-symptoms": ["symptom1", "symptom2", ...]
}
```

Evaluation takes into account both the accuracy of the BDI-II score estimate and the relevance of the reported symptoms.

4. Methodology

4.1. System Architecture

We designed a multi-agent system with three functional roles that interact in a turn-based loop for each patient (see Figure 1):

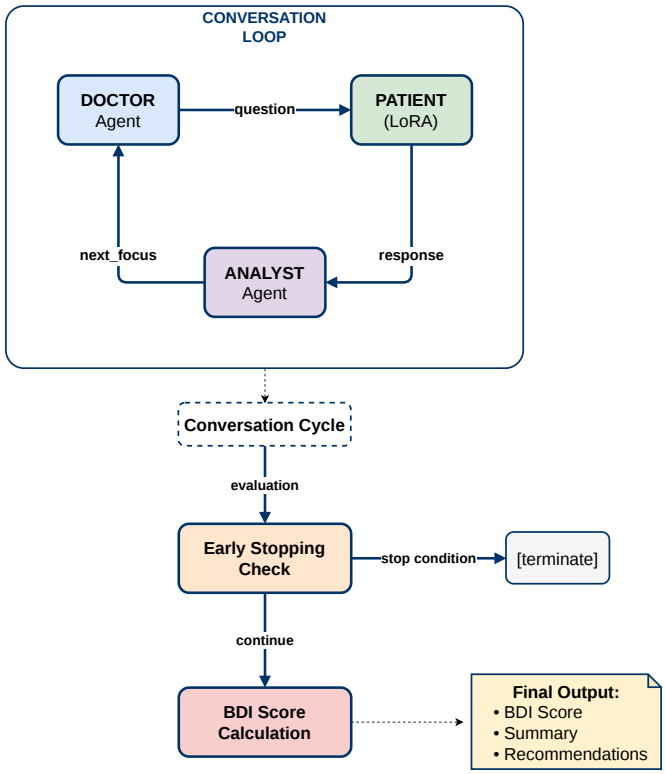

Figure 1: Multi-agent conversation loop. The Doctor generates clinically directed questions, the Patient responds, and the Analyst evaluates the response for indicators of depressive symptoms.

Each conversation loop begins with the Doctor agent generating a concise question aimed at a specific clinical domain (e.g., sleep patterns, loss of interest, or fatigue). The Patient simulator then responds using the LoRA-adapted Llama-3-8B model, conditioned on the official system prompt and the full conversational history up to that point. Once the patient responds, the Analyst agent reviews the entire conversation. It estimates the severity and duration of any BDI-II symptoms observed until that point and proposes a direction for the next turn. Before the loop continues, an early-stopping check determines whether the conversation continues or is terminated early based on the clinical evidence accumulated so far.

4.2. Run 1: Fully Local Architecture

In Run 1, all three agents run on the same Meta-Llama-3-8B-Instruct base model, loaded locally on an NVIDIA RTX 3060 GPU with 12 GB VRAM in float16 precision. Role differentiation is handled through LoRA adapter toggling: when the patient is speaking, the adapter is active and produces patient-specific responses, while for the Doctor and Analyst turns, it is disabled via (`model.disable_adapter()`), bringing the model back to its base behavior.

The Doctor prompt instructs the model to ask one direct question per turn, with responses limited to 2 sentences or 20 words. The temperature is set to 0.6, `top_p=0.9`, and `max_new_tokens=40`. The Analyst prompt sets up the model as a neutral observer tasked with identifying BDI-II symptoms along with their severity and duration. The output the Analyst generates follows the following format:

```
STATUS: [Healthy / Potentially Depressed]
OBSERVATIONS: [Category | Severity 1-3 | Duration], ...
NEXT_FOCUS: [Directive for next question]
```

The Analyst temperature is set very low (0.01) to keep its output consistent with the expected format.

For scoring in Run 1, a rule-based heuristic is applied to everything the Analyst has accumulated throughout the conversation. Suicidal indicators add 30 points to the total. Core symptoms (i.e., sadness, loss of pleasure, pessimism, and worthlessness) contribute 6 points if marked as chronic and 3 otherwise. All remaining symptoms add 4 points if chronic and 2 if not. The final score is clamped to the valid BDI-II range of 0 to 63.

Early stopping can be triggered after at least 3 turns under two conditions: either the patient has been classified as healthy with no flags by the Analyst, or at least two chronic symptoms have already been confirmed. In both cases, continuing the conversation is unlikely to change the outcome.

4.3. Run 2: Hybrid Architecture (Local + Groq API)

Run 2 was motivated by a concrete limitation observed in Run 1: an 8B-parameter model simply lacks the capacity for the kind of nuanced clinical reasoning the task demands. To address this, we offloaded the Doctor and Analyst roles to Llama-3.3-70B-Versatile (a substantially more capable model) accessed through the Groq API. The patient simulation remained local on Llama-3-8B with LoRA adapters, as the task requires.

The main differences between the two runs are summarized in Table 1.

Table 1
Comparison of architectural choices between Run 1 and Run 2.

Aspect	Run 1 (Local)	Run 2 (Hybrid)
Doctor Model	Llama-3-8B (local)	Llama-3.3-70B (Groq API)
Analyst Model	Llama-3-8B (local)	Llama-3.3-70B (Groq API)
Patient Model	Llama-3-8B + LoRA (local)	Llama-3-8B + LoRA (local)
Analyst Output	Structured text (3 lines)	JSON with Chain-of-Thought
BDI Scoring	Rule-based heuristic	LLM-based final evaluation
Early Stopping	Rule-based (Chronic count)	LLM-recommended (`stop_recommended`)

In contrast to the first configuration, the Analyst agent in Run 2 generates a structured JSON output that includes explicit clinical reasoning through a Chain-of-Thought process:

```
{
  "reasoning": "string",
  "symptoms_found": [{"name": "string", "severity": 1, "duration": "
     string"}],
  "status": "Healthy / Potentially Depressed",
  "stop_recommended": true/false,
```

```
  "next_focus": "string"
}
```

In Run 2, the final BDI-II score is generated during the Final BDI Board step. The 70B model reads the full conversation transcript using an extra Groq API call and provides a holistic assessment in the following format:

```
{
  "bdi_score": <integer>,
  "key_symptoms": ["symptom1", "symptom2"]
}
```

We have added several reliability measures in Run 2. API failures are handled with exponential backoff retry logic for HTTP 429 and 503 errors. To save time, if the process is interrupted midway, already-completed patients are skipped on restart rather than reprocessed from scratch. For better memory management, LoRA adapters are loaded and unloaded as needed during the run to prevent VRAM usage from building up.

4.4. Patient Simulation

Both runs use the same patient simulator configuration, as required. The base model is `meta-llama/Meta-Llama-3-8B-Instruct`, with patient-specific LoRA adapters loaded from `Anxo/erisk26-task1-patient-00..19-adapter`. The system prompt provided was *"You are a simulated patient. Act realistically based on your internal training. Ensure contextual realism. Avoid overly detailed or formal speech. Keep natural speaking style (e.g., short answers, hesitations, casual expressions). Do not mention you are an AI."*. Generation runs at `temperature=0.6` with `top_p=0.9` and `max_new_tokens=256`, and the full conversation history is maintained across turns through `tokenizer.apply_chat_template()`.

4.5. Maximum Conversation Turns

Both runs are configured with a maximum of 8 conversation turns per patient. With early stopping, most conversations concluded between turns 3 and 5 when sufficient clinical evidence was gathered.

5. Results and Discussion

5.1. Run 1 Results

Table 2 presents the BDI-II scores and key symptoms identified for all 20 patients in Run 1.

The average BDI-II score for all patients in Run 1 is 17.7 (sample SD = 5.99), with scores ranging from 10 to 31. Fatigue was detected as a symptom in 18 out of 20 patients. The two patients without a fatigue flag were patient 10 (Agitation) and patient 14 (Sleep Changes and Loss of Pleasure). This pattern suggests that the 8B model, when used as the Analyst, often identifies fatigue-related cues more than other symptom areas, which may lead to missing other symptoms.

5.2. Run 2 Results

For Run 2, the aggregated results obtained from running the hybrid configuration (Local Llama-3-8B + Groq Llama-3.3-70B), on all 20 simulated patients are shown in Table 3. Statistical analysis of Run 2 shows an average BDI-II score of 23.9, with a sample standard deviation of about 12.44. Scores ranged widely, from 0 for Patient 3 to a high of 46 for Patient 7. From a clinical perspective, the distribution of cases evaluated by the 70B model, using the standard BDI-II severity bands, is as follows: 3 patients with minimal depression (score 0–13), 3 with mild depression (score 14–19), 7 with moderate depression (score 20–28), and 7 with severe depression (score 29–63). Regarding symptom diversity, the hybrid

Table 2
Run 1 results — BDI-II scores and key symptoms for all 20 patients.

Patient	BDI Score	Key Symptoms
1	12	Fatigue
2	16	Fatigue, Sadness
3	21	Fatigue, Loss of Pleasure
4	22	Fatigue, Sadness, Loss of Pleasure
5	18	Fatigue, Agitation, Loss of Energy
6	16	Fatigue, Sleep Changes
7	24	Fatigue, Sleep Changes
8	31	Agitation, Loss of Concentration, Fatigue, Loss of Pleasure
9	10	Fatigue
10	12	Agitation
11	16	Fatigue, Sleep Changes
12	10	Fatigue
13	16	Fatigue, Appetite Changes
14	19	Sleep Changes, Loss of Pleasure
15	12	Fatigue
16	24	Fatigue, Sleep Changes, Loss of Appetite
17	18	Fatigue, Agitation
18	28	Fatigue, Loss of Pleasure
19	19	Fatigue, Loss of Pleasure
20	10	Fatigue

system eliminated the repetitive behavior identified in Run 1. Although Fatigue remains a frequent indicator (identified in 13 patients), a strong emergence of other key clinical symptoms is observed: Pessimism appears in 11 cases, Loss of Interest is successfully mapped in 14 patients, and Sadness in 10 cases. The larger model proves the capability to isolate severe manifestations, such as Suicidal Ideation, which was correctly found in Patients 4, 15, and 16.

5.3. Discussion

Run 1 limitations. In Run 1, the main impediment was that the 8B model has limited clinical reasoning when operating without its LoRA adapter. The Doctor agent often asks questions that are either too long or sound mechanical, even though there are prompt limits in place. The Analyst has an inclination to label cases as Fatigue, no matter what the patient actually says. The rule-based BDI scoring method is consistent and repeatable, but it misses the subtle differences in severity that a human clinician would notice.

Run 2 improvements. The hybrid system evaluated in Run 2 demonstrates a clear qualitative leap over Run 1, directly validated by the official preliminary results of the eRisk 2026 competition [19, 20]. Both runs were officially submitted and evaluated on the full set of 20 personas. For Run 1 the official scores were DCHR = 0.2500 (rank 9), ADODL = 0.8325 (rank 17), ASHR = 0.1000 (rank 16), LDCHR = 0.2167 (rank 9), and LASHR = 0.0807 (rank 13) [19, 20]. For Run 2 the official scores were DCHR = 0.5000 (rank 3), ADODL = 0.8802 (rank 12), ASHR = 0.2625 (rank 7), LDCHR = 0.3374 (rank 2), and LASHR = 0.1709 (rank 4) [19, 20].

The most significant performance aspect of our architecture was interaction optimization (early decision-making): Awakened recorded an average of only 5.3 messages per conversation, making it one of the fastest and most efficient strategies in the competition [19, 20]. This efficiency allowed Run 2 to achieve a latency-penalized LDCHR of 0.3374, placing the run directly 2nd globally in terms of evidence-gathering efficiency, while the LASHR of 0.1709 placed it 4th globally [19, 20]. Together with the 3rd-place DCHR, these scores confirm that the Chain-of-Thought logic implemented on the 70B model extracted much more accurate indicators within a reduced conversational budget than the fully local 8B pipeline.

Table 3
Run 2 results — BDI-II scores and key symptoms identified by the 70B holistic Final BDI Board across all 20 patients.

Patient (LLM)	BDI Score	Identified Key Symptoms
1	26	Fatigue, Insomnia, Loss of Interest, Pessimism
2	26	Sadness, Fatigue, Loss of Interest, Pessimism
3	0	[None]
4	31	Sadness, Insomnia, Loss of Appetite, Suicidal Ideation
5	31	Fatigue, Loss of Interest, Changes in Appetite, Sleep Disturbances
6	41	Sadness, Loss of Interest, Insomnia, Pessimism
7	46	Sadness, Fatigue, Loss of Interest, Pessimism
8	22	Sadness, Insomnia, Loss of Interest, Fatigue
9	26	Sadness, Fatigue, Loss of Interest, Insomnia
10	14	Sadness, Loss of Interest, Self-Doubt
11	14	Fatigue, Loss of Interest, Pessimism
12	22	Fatigue, Insomnia, Loss of Interest, Pessimism
13	38	Sadness, Fatigue, Pessimism, Loss of Interest
14	31	Sadness, Insomnia, Fatigue, Pessimism
15	26	Fatigue, Loss of Interest, Sleepiness, Suicidal Ideation
16	38	Sadness, Pessimism, Loss of Interest, Suicidal Ideation
17	6	Fatigue, Variable Energy
18	14	Pessimism, Lack of motivation, Feeling stuck, Dissatisfaction
19	4	[None]
20	22	Fatigue, Loss of Motivation, Pessimism, Loss of Interest

Shared limitations. Both runs leave room for optimization in future scenarios. Despite the excellent latency performance of the model in Run 2, a small number of conversations stopped prematurely (as seen with Patient 19, which stopped at a score of 4 with no key symptoms mapped), indicating that the LLM-based early stopping decision can sometimes be too conservative. Furthermore, reliance on an external API (Groq) introduces network risks that, in a real clinical environment, should be mitigated by using heavily quantized massive local models (e.g., 70B in 4-bit).

6. Conclusions and Future Work

We presented a multi-agent LLM architecture for automated BDI-II depression screening through simulated clinical interviews, submitted as part of eRisk 2026 Task 1. Our two runs demonstrate the feasibility and trade-offs of fully local versus hybrid approaches to this novel task.

A few things stood out across these findings. The Doctor/Analyst/Patient separation proved to be a useful way to structure the interviews, keeping each component's responsibility clear. The gap between the 8B and 70B models was noticeable mostly in the Analyst role, where the larger model produced more varied and clinically grounded symptom assessments. On the scoring side, the rule-based approach in Run 1 was consistent but somewhat rigid, while the LLM-based holistic assessment in Run 2 handled edge cases more gracefully. In terms of efficiency, the hybrid system averaged only 5.3 messages per conversation, resulting in competitive latency-aware metrics and a 2nd-place result in LDCHR [19, 20].

For future work, interview questions could be designed to cover all 21 BDI-II items more methodically, rather than relying on the Analyst to deduce coverage from open-ended responses. The conversation flow could also be made more adaptive, adjusting dynamically as symptom patterns emerge rather than following a fixed sequence of focus areas. The early stopping logic needs further calibration because, in some cases, it terminated conversations before key symptoms had surfaced. Combining rule-based and LLM-based scoring into an ensemble could offer a middle ground between consistency and nuance. Finally, running quantized versions of larger models locally (such as a 4-bit 70B) could improve reasoning quality without introducing API dependencies.

Acknowledgments

The research presented in this paper is supported in part by The Academy of Romanian Scientists, through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV).

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checks, paraphrasing, and writing-style improvements, and ChatGPT for light language-editing support and formatting assistance. These tools were used solely to improve clarity, readability, and presentation. After using these tools, the authors carefully reviewed, edited, and verified the content as needed and take full responsibility for the publication's content.

References

[1] World Health Organization, Depressive disorder (depression), 2025. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] A. M. Boerema, M. Ten Have, A. Kleiboer, R. de Graaf, J. Nuyen, P. Cuijpers, A. T. F. Beekman, Demographic and need factors of early, delayed and no mental health care use in major depression: A prospective study, BMC Psychiatry 17 (2017) 367. doi:10.1186/s12888-017-1531-8.

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

[4] 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, Proceedings, Lecture Notes in Computer Science, Springer, 2022, pp. 233–256. doi:10.1007/978-3-031-13643-6_18.

[5] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of eRisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9–12, 2025, Proceedings, Lecture Notes in Computer Science, Springer, 2025, pp. 242–265. doi:10.1007/978-3-032-04354-2_15.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[7] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13, ACM, 2013, pp. 47–56. doi:10.1145/2464464.2464480.

[8] T. Zhang, A. M. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: A narrative review, npj Digital Medicine 5 (2022) 46. doi:10.1038/s41746-022-00589-7.

[9] A. Handy, R. Mangal, T. S. Stead, R. L. Coffee, L. Ganti, Prevalence and impact of diagnosed and undiagnosed depression in the united states, Cureus 14 (2022) e28011. doi:10.7759/cureus.28011.

[10] C. Kraus, B. Kadriu, R. Lanzenberger, C. A. Zarate, S. Kasper, Prognosis and improved outcomes in major depression: A review, Translational Psychiatry 9 (2019) 127. doi:10.1038/s41398-019-0460-3.

[11] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: Proceedings of the Thirteenth Language Resources

and Evaluation Conference, LREC 2022, European Language Resources Association, 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, volume 30, 2017.

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[14] A. Petrescu, C.-O. Truică, E.-S. Apostol, A. Paschke, EDSA-Ensemble: An event detection sentiment analysis ensemble architecture, IEEE Transactions on Affective Computing 16 (2025) 555–572. doi:10.1109/TAFFC.2024.3434355.

[15] C.-O. Truică, C. A. Leordeanu, Classification of an imbalanced data set using decision tree algorithms, University Politehnica of Bucharest Scientific Bulletin - Series C Electrical Engineering and Computer Science 79 (2017) 69–84.

[16] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5–8, 2016, Proceedings, Lecture Notes in Computer Science, Springer, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

[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, Proceedings, Lecture Notes in Computer Science, Springer, 2023, pp. 294–315. doi:10.1007/978-3-031-42448-9_22.

[18] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of eRisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Lecture Notes in Computer Science, Springer, 2024, pp. 73–92. doi:10.1007/978-3-031-71908-0_4.

[19] 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.

[20] A. Perez, J. Parapar, X. Wang, F. Crestani, Overview of eRisk 2026 Early Risk Prediction on the Internet: Symptom Ranking and Conversational Approaches for Depression and ADHD, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 17th International Conference of the CLEF Association, CLEF 2026, Jena, Germany, September 21–24, 2026, Proceedings, volume To be published of *Lecture Notes in Computer Science*, Springer, 2026.