VANGUARD at eRisk 2026 Task 1: Conversational Depression Detection with LLM Personas

Notebook for the eRisk Lab at CLEF 2026

Anamaria Vișan[1], Ciprian-Octavian Truică[1,2], Adrian Paschke[3] and Elena-Simona Apostol[1,*]

[1]National University of Science and Technology Politehnica Bucharest, Splaiul Independenței 313, București 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania
[3]Fraunhofer Institute for Open Communication Systems, Berlin, 10589, Germany

Abstract

In this paper, we describe the VANGUARD team in the eRisk 2026 Task 1 (Conversational Depression Detection). Participants conduct multi-turn dialogs with twenty weekly LoRA adapters on `meta-llama/Meta-Llama-3-8B-Instruct` and submit conversation logs together with a BDI-II score and up to four key symptoms per persona. We implement three independent local runs: an Llama-as-judge assessor, a rule-based linguistic analyzer, and a hybrid fusion of both. Our archived submissions cover Run 3 for personas 1 to 18 and Runs 1 and 2 for personas 19 to 20. On the hybrid run, the mean estimated BDI is 6.2 (median 4.5). Organizer-reported scores in the lab overview [1] for our 18-persona submission (ADODL 0.732, rank 27 among incomplete submissions) and released golden labels (BDI MAE 16.9, symptom recall 0.18) are discussed in Section 4.6.

Keywords

depression, conversational agents, LLM personas, BDI-II, eRisk, mental health NLP

1. Introduction and task description

1.1. Motivation

Major depressive disorder is often undiagnosed; daily language about mood, sleep, energy, and self-worth can reveal risk before formal assessment. eRisk evaluates early mental-health inference from text [1]. Task 1 (2026) uses simulated patients, i.e., LoRA fine-tuned on Llama-3-8B-Instruct, so teams investigate evaluation strategies under controlled BDI-II-aligned severity [2, 3, 4].

1.2. Task formulation

Organizers release 20 personas as LoRA adapters in weekly pairs (February to April 2026) on Hugging Face. Teams may have unlimited-turn dialogs with each persona, but evaluation metrics penalize late decisions. Direct questions such as "Are you depressed?" are forbidden, and a mandatory system prompt must remain verbatim. For each run, teams upload `interactions_run{N}.json` (full dialogue) and `results_run{N}.json` with `bdi-score` and up to four `key-symptoms`. Up to three automated runs, plus an optional manual run, must be mutually independent.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
anamaria.visan@stud.acs.upb.ro (A. Vișan); ciprian.truica@upb.ro (C. Truică); adrian.paschke@fokus.fraunhofer.de (A. Paschke); 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)
0009-0002-7229-9649 (A. Vișan); 0000-0001-7292-4462 (C. Truică); 0000-0003-3156-9040 (A. Paschke); 0000-0001-6397-4951 (E. Apostol)

1.3. Core difficulties

Teams must elicit symptoms indirectly because personas answer in open-ended language while evaluators optimize symptom identification, severity estimation, and timeliness jointly. Running an 8B base model with LoRA adapters demands substantial GPU memory. Finally, the task is a research simulation and must not be treated as a clinical diagnosis.

2. State of the art

Our pipeline design is grounded in a structured synthesis of 51 papers on text-based depression detection, early risk, conversational assessment, and eRisk-style evaluation. The corpus spans systematic reviews, social-media classifiers, interview and chatbot systems, transformer models, and lab overviews. Below we distill themes most relevant to Task 1: multi-turn dialogue with simulated patients, BDI-II severity and symptom labels, and timeliness of decisions.

2.1. Social-media text and classical ML/DL

Cross-sectional studies show that language on social platforms carries a depressive signal when reliable labels exist. Classical pipelines combine handcrafted linguistic features with SVMs, random forests, or ensembles [5, 6]; systematic reviews report a shift toward RNNs, CNNs, and BERT-family encoders, with BiLSTM+attention and hybrid feature stacks in several surveys [7, 8]. Class imbalance, noisy self-labels, platform bias, and weak cross-domain generalization recur as limitations [9]. Multi-level frameworks map posts to severity bands (non-depressed through severe) using BERT, clustering, and SVMs [10]. Task 1 uses controlled, organizer-provided dialogues rather than observational timelines.

2.2. Linguistic markers and BDI-aligned severity

Lexical and stylistic cues can track depressive vs. non-depressive states in essays and posts, though markers depend on language and population. Clinical inventories (BDI-II, PHQ, HAMD) remain the reference for severity; automated systems often approximate them from text. eRisk's conversational pilot adopted BDI-II-grounded personas and metrics such as DCHR, ADODL, and latency-aware variants (LDCHR, LASHR) to reward accurate *and* early judgments [3, 2]. Task 1 keeps the same dual target: a scalar BDI estimate and up to four key symptoms drawn from the 21 BDI-II items.

2.3. Ground truth and label construction

Reviews note disagreement on how "depression on social media" studies define ground truth (clinician surveys, structured interviews, self-declarations in posts, or community membership) [10]. Each choice adds bias (stigma, under-reporting, label noise). Simulated patients with known BDI profiles, as in Task 1, trade ecological validity for reproducible evaluation with shared test collections and transparent metrics, as in prior eRisk editions [11].

2.4. Conversational and interview-based assessment

Several studies move from isolated posts to dialogue. Conversation transcripts can be classified with sequence models (e.g., LSTM over turns) [7]. On psychiatric-interview corpora such as DAIC-WOZ, turn- or question-aware models that group exchanges by PHQ-8 domain improve detection from consultation text; large-scale counseling text analyses link topic, sentiment, and lexical patterns to symptom severity and progress [12]. Online therapy work reports similar patterns for severity, while progress is harder to track without dialogue-level features. Task 1 uses multi-turn probing and full transcript scoring rather than single-shot classification.

2.5. Chatbots and two-tier architectures

A common pattern couples an empathetic dialogue module with a parallel classifier: the bot gathers natural language while a separate model assigns disorder or risk labels [13]. DEPRA-style agents run structured clinical interviews (Hamilton/SIGH-D and IDS-C-style flows) and report feasibility for early screening in non-clinical trials [14]. Surveys mention mental-health chatbots as a future area but cite dataset scarcity, validation gaps, and uncertain uptake. Task 1 forbids direct "Are you depressed?" questions; our indirect, BDI-oriented script follows structured conversational elicitation used elsewhere.

2.6. Transformers, LLMs, and domain models

Transformer encoders (BERT, RoBERTa, MentalBERT) are widely used for social-media detection; domain-adapted models and hybrid lexical and neural fusion stay competitive on imbalanced data [8, 15]. Fine-tuned GPT-3.5 and Llama-2 classifiers report high accuracy on depressed vs. control posts. LLM-as-judge pipelines and structured JSON assessment (e.g., TalkDep-style pairwise severity ranking) can score full dialogues without hand-built features [16]. Task 1's Run 1/3 judges follow that path; Run 2 is lexical and rule-based; organizer LoRA personas on Llama-3-8B-Instruct fix the patient side [17].

2.7. Early detection and timeliness

Static labels are not enough for eRisk-style tasks: systems are also judged on *when* they alert [11]. Losada and Crestani introduced shared collections and ERDE-style costs that balance accuracy and delay on sequential writings; later work revises ERDE and proposes latency-aware F-scores, and streaming or neuro-symbolic setups target early alerts. Task 1 penalizes long dialogues before correct BDI/symptom estimates, in the same spirit as ERDE but over turn count rather than documents.

2.8. Design implications for Task 1

Our three runs combine neural judgment over full transcripts (LLM-as-judge and TalkDep-style assessment), a rule-based analyzer, and hybrid fusion when generation or parsing fails. We keep indirect questioning and BDI-II vocabulary; our 15-turn logs are shorter than many clinical interviews and therefore bias recall against timeliness. Partial persona coverage reflects compute limits and phased adapter releases, not a full 20×3 matrix.

2.9. Ethics and limitations

The reviewed literature repeats the same cautions: automated tools are not diagnostic devices, false positives can harm users, and privacy and stigma limit data collection [7, 8, 10]. Reproducibility and external validation are often weak. Task 1 stays challenge-only: simulated patients, no real user data, and outputs for organizer evaluation only.

3. Proposed solution

Team VANGUARD implemented Task 1 in a single modular notebook pipeline (DepressionDetectionPipeline) that loads weekly LoRA personas, conducts scripted indirect interviews, scores transcripts with one of three independent judges, and exports organizer-ready JSON under `task1-llms-results/persona{N}/`.

3.1. Architecture

The pipeline comprises four components that mirror the official task design. The PersonaLoader loads `meta-llama/Meta-Llama-3-8B-Instruct` once, applies each

Table 1
Archived Task 1 submissions (VANGUARD).

Run	Judge	Personas with both JSON files
1	Llama local	19, 20
2	Rule-based	19, 20
3	Hybrid local	1 to 18

`Anxo/erisk26-task1-patient-XX-adapter` via PeftModel, and uses NF4 4-bit quantization with optional 8-bit CPU-offload when GPU memory is insufficient. The ConversationEngine conducts multi-turn chat with the mandatory system prompt left verbatim and logs each role/message turn. The LLMJudgeEvaluator and DepressionAnalyzer produce BDI-II scores from LLM JSON, rule-based keywords and patterns, or a hybrid fallback. The OutputFormatter maps internal symptom names to official BDI-II labels and writes `interactions_run{N}.json` and `results_run{N}.json`.

The adapter path `patient-00` maps to submission id `"1"`, and `patient-19` maps to `"20"`.

3.2. Persona interaction

Generation follows organizer defaults (temperature 0.6, top-p 0.9, up to 256 new tokens per turn, left padding, EOS and Llama-3 end-of-turn terminators). The interview script has 15 indirect questions: nine rapport and BDI-oriented probes on mood, activities, sleep, appetite, work, social contact, and outlook, plus six follow-ups on impact, duration, support, coping, and closure. None ask whether the persona is depressed. For submission, we disabled early stopping and inline analysis so all 15 turns finish before judging, at the cost of timeliness; the codebase can still stop after five turns when rule-based confidence reaches 0.8.

3.3. Three independent runs

Runs are indexed 1 to 3. Each instantiates a fresh pipeline with no shared judge state.

Run 1 (Llama local) reuses the loaded model after the dialogue. It receives a structured prompt over all 21 BDI-II items and must return JSON with BDI score, severity, up to four key symptoms, confidence, and reasoning (192 judge tokens max). Run 2 (rule-based) counts keyword and pattern hits per symptom on assistant turns (0 to 3 per item), sums a BDI proxy, and picks top symptoms without further generation. Run 3 (hybrid) tries Run 1 parsing first and falls back to Run 2 when output is empty or invalid (logged as hybrid stabilization). TalkDep-style pairwise prompts exist in the repo but were not scored for submission.

3.4. Batch execution

Execution split into two phases. Phase 1 ran all three judges on the latest weekly pair (personas 20 and 19); Phase 2 walked personas 18 down to 1. We created a new DepressionDetectionPipeline per run and cleared the GPU cache between runs and personas. Exported files follow the organizer FTP layout in our local archive.

4. Results

4.1. Submission coverage

Table 1 lists archived outputs. Coverage is below the nominal 20×3 design: Phase 2 archived Run 3 for personas 1 to 18, while Phase 1 kept only Runs 1 and 2 for personas 19 to 20. Run 3 for personas 19 to 20 was never written to disk.

We archived 22 bundles in total: 18 from Run 3 and two personas $\times$ two runs for Runs 1 and 2. The next subsection reports Run 3.

Table 2
BDI-II estimates on the last two personas (Runs 1 and 2).

Persona	Run 1 BDI	Run 2 BDI	Run 1 key symptoms (sample)
19	0	1	(none listed)
20	14	3	Sadness, Irritability / Sadness, Agitation, Loss of Energy

Table 3
Task 1 vs. golden labels (all archived personas, n=20).

Metric	Value
BDI MAE (all personas)	15.95
BDI MAE (Run 3 only, n=18)	16.89
Mean symptom hits (of ≤ 4 predicted)	0.70
Mean symptom recall vs. GT set	0.18
Mean symptom precision	0.28

4.2. Run 3 (personas 1 to 18)

Mean estimated BDI is 6.2 (median 4.5, range 1 to 22). Fourteen personas sit in the minimal band (0 to 13); three are mild (14 to 19: personas 1, 10, 18); persona 11 reaches 22 (moderate). None are severe. Loss of Energy is the most common predicted symptom (11 personas), then Sadness (8), Changes in Sleeping Pattern (6), and Agitation (4). Every `interactions_run3.json` has 30 messages (15 user and 15 assistant turns), as expected for the fixed script.

4.3. Runs 1 and 2 (personas 19 to 20)

Persona 20 diverges sharply between judges (BDI 14 vs. 3). Over both personas, Run 1 averages 7.0 and Run 2 averages 2.0.

4.4. Operational notes

All runs used GPU execution with 4-bit weights. Without CUDA, the notebook is impractically slow; when VRAM was insufficient, weights could be offloaded to the CPU. Adapter swaps added first-token delay on smaller GPUs, which shaped our two-phase schedule.

4.5. Comparison to released ground truth

Organizers later published persona-level BDI-II scores, key symptoms, and a synonym map. We aligned archived runs (Run 3 on personas 1 to 18; Runs 1 and 2 on 19 to 20) through that map.

Golden BDI values often exceed our estimates (persona 1: 26 vs. 14; persona 5: 37 vs. 5).the Both judges look conservative and miss key symptoms of the organizer. Synonym mapping (e.g., “Loss of energy” vs. “Loss of Energy”) helps naming, but overlap stays low.

4.6. Organizer evaluation

The eRisk 2026 overview reports Task 1 metrics for all teams [1]. We submitted three partial runs: Run 1 and Run 2 on two personas each (Run 1 partly manual), Run 3 on eighteen personas with full automation. Our dialogs averaged 15.0 participant turns and 80.6 characters per turn, shorter than most entries in the table.

Table 4 reproduces organizer scores for incomplete submissions. (Metrics are computed only over submitted personas; ranks are within that table).

Run 2 reaches ADODL 0.9286 and DCHR 1.0000 on two personas only. Run 3 ranks 27th on ADODL among partial submissions (ASHR 0.1806). High scores on 2/20 personas need cautious reading; full

Table 4
VANGUARD Task 1 scores in the lab overview (incomplete submissions).

Run	Personas	DCHR	ADODL	ASHR	LDCHR	LASHR
1	2/20	0.5000	0.8810	0.1250	0.1533	0.0383
2	2/20	1.0000	0.9286	0.0000	0.3066	0.0000
3	18/20	0.2222	0.7319	0.1806	0.0681	0.0554

comparison requires all twenty personas on Runs 1 to 3, as in full-coverage teams (e.g., pjmathematician at ADODL 0.9254 over 20 personas).

5. Discussion and conclusion

Our approach offers a fully local pipeline, literature-grounded indirect questioning, and three independent scoring methods. Limitations include an incomplete run-by-persona matrix, a large gap between predicted and golden BDI and symptoms, and fixed 15-turn logs because submitted runs disabled early stopping, even though the task rewards timeliness. The rule-based judge often underestimates severity compared with the Llama judge on the same transcript. We plan to complete all runs for all personas, calibrate judges to organizer severity profiles, enable confidence-based early stopping, and refine symptom-focused prompting aligned with the BDI-II inventory.

We presented VANGUARD's Task 1 system with documented partial submissions. On the hybrid Across 18 personas, the mean estimated BDI is approximately 6.2.

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 GPT-4 to perform grammar and spelling checks, paraphrase, and reword. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] 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.

[2] A. T. Beck, R. A. Steer, G. Brown, Beck depression inventory–II, Psychological assessment (1996).

[3] 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.

[4] 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.

[5] R. Chiong, G. S. Budhi, S. Dhakal, F. Chiong, A textual-based featuring approach for depression detection using machine learning classifiers and social media texts, Computers in Biology and Medicine 135 (2021) 104499.

[6] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the international AAAI conference on web and social media, volume 7, 2013, pp. 128–137.

[7] D. William, D. Suhartono, Text-based depression detection on social media posts: A systematic literature review, Procedia Computer Science 179 (2021) 582–589.

[8] W. A. Gadzama, D. Gabi, M. S. Argungu, H. U. Suru, The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review, Personalized Medicine in Psychiatry 45 (2024) 100125.

[9] 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.

[10] A. Aldkheel, L. Zhou, Depression detection on social media: a classification framework and research challenges and opportunities, Journal of Healthcare Informatics Research 8 (2024) 88–120.

[11] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International conference of the cross-language evaluation forum for European languages, Springer, 2016, pp. 28–39.

[12] T. Althoff, K. Clark, J. Leskovec, Large-scale analysis of counseling conversations: An application of natural language processing to mental health, Transactions of the Association for Computational Linguistics 4 (2016) 463–476.

[13] N. Ghoshal, V. Bhartia, B. Tripathy, A. Tripathy, Chatbot for mental health diagnosis using data augmentation techniques and deep learning, International Journal of Embedded Systems 17 (2024) 171–182.

[14] P. Kaywan, K. Ahmed, A. Ibaida, Y. Miao, B. Gu, Early detection of depression using a conversational AI bot: A non-clinical trial, Plos one 18 (2023) e0279743.

[15] A. Q. Zulfa, S. A. I. Alfarozi, S. Kusrohmaniah, S. Wibirama, MentalBERT with Contrastive Learning for Mental Disorder Classification from Social Media: A Comparative and Explainable Study, in: 2026 International Conference on Current Research in Artificial Intelligence and Data Science (ICCRAIDS), volume 1, IEEE, 2026, pp. 1–7.

[16] X. Wang, A. Perez, J. Parapar, F. Crestani, TalkDep: Clinically Grounded LLM Personas for Conversation-Centric Depression Screening, in: Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 6554–6558. URL: https://doi.org/10.1145/3746252.3761617. doi:10.1145/3746252.3761617.

[17] K. Wang, Q. Tan, P. Zhang, Emotional Cause Analysis Based on In-Context Learning of Large Language Models, INNO-PRESS: Journal of Emerging Applied AI 2 (2026).