lexxy at eRisk 2026: Conversational Depression Detection via LLM Personas Using a Multi-Strategy Two-Agent Pipeline

Notebook for the eRisk Lab at CLEF 2026

Lakshika Singh Chouhan[1,*], Ritesh Kumar[1]

[1]Indian Institute of Information Technology Surat, Gujarat 394190, India

Abstract

Automated depression screening has traditionally operated over static text collections; eRisk 2026 Task 1 introduces a fundamentally different paradigm by requiring participating systems to *conduct* diagnostic interviews with LLM-simulated patient personas—LoRA-adapted variants of Meta-Llama-3-8B-Instruct—and infer Beck Depression Inventory-II (BDI-II) severity solely from the resulting conversation, without access to ground-truth labels during interaction. This paper describes the lexxy submission: a modular two-agent pipeline where a strategy-parameterised Conversational Agent elicits symptom-relevant responses across all 21 BDI-II dimensions and a turn-level Evaluation Agent scores each dimension through two sequential passes, combining direct lexical matching with inference over contextual cues.

Three submission runs implement distinct conversational framings—self-disclosure, empathy, and direct questioning—while sharing an identical underlying pipeline. Across all 20 official personas every run attains DCHR = 0.30 and ADODL in the range 0.841–0.856. The direct-questioning run achieves ASHR = 0.2875, placing 5th globally among all eRisk 2026 Task 1 participants. A key empirical finding is the ASHR–ADODL depth–breadth tension: emotionally framed strategies yield higher score calibration while direct questioning maximises symptom coverage. The system operates fully zero-shot, requiring no labelled training data at any stage.

Keywords

depression detection, conversational AI, BDI-II, LLM personas, LoRA, DCHR, ASHR, ADODL, eRisk 2026, CLEF

1. Introduction

More than 280 million people worldwide live with depression [1, 2], and it remains the single largest contributor to years lived with disability globally. Early intervention matters a great deal here: the longer an episode goes untreated, the lower the chances of full recovery, yet it is still common for people to wait 6–8 years before seeking professional help [3], even in health systems with good resources. This gap is part of why automated screening tools — systems that do not depend on a clinician being available — have attracted so much research interest.

The eRisk workshop series at CLEF [4] has been studying this kind of early risk detection since 2017, largely by analysing streams of social-media posts for signs of depression, anorexia, self-harm, and gambling disorder [5, 6, 7]. eRisk 2026 Task 1 changes that setup considerably: rather than working with text that already exists, participating systems now have to hold an open-ended interview with a LoRA-fine-tuned [8] Llama-3-8B patient persona [9] and infer BDI-II [10] severity purely from that conversation. No ground-truth score is available during the interaction itself, so there is no way to fall back on post-hoc lookup.

Our system, developed by Team lexxy, makes four contributions:

1. A modular two-agent architecture (Figure 1) that keeps dialogue management separate from clinical scoring, so each part can be improved on its own.

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

*Corresponding author.

lakshikasingh7062002@gmail.com (L. S. Chouhan); riteshkumar@iiitsurat.ac.in (R. Kumar)

2. A two-pass Evaluation Agent (Figure 2) that picks up symptom signals from explicit lexical markers as well as implicit pragmatic cues, which reduces systematic underscoring on mild-range personas by an estimated 5–8 BDI points.
3. Three strategy-parameterised runs used as a controlled ablation of conversational framing, which surfaced a previously undocumented ASHR–ADODL depth–breadth trade-off.
4. An ASHR of 0.2875 for the direct-questioning run, placing 5th globally in the Task 1 official evaluation.

2. Background and Related Work

2.1. BDI-II as a Clinical Instrument

Developed by Beck et al. [10], the BDI-II operationalises DSM-based major depressive episode criteria across 21 items rated 0–3, yielding totals from 0 to 63. eRisk 2026 maps this continuum to four severity bands: *minimal* (0–9), *mild* (10–18), *moderate* (19–29), and *severe* (30–63). Participating systems must predict both an integer BDI total and up to four key contributing symptoms. Among the 21 items, nine exhibit the strongest signal-to-noise ratio in conversational elicitation: Sadness, Loss of Pleasure, Loss of Energy, Sleep Changes, Pessimism, Loss of Interest, Appetite Changes, Concentration Difficulty, and Fatigue [11]; these receive elevated priority in the Symptom Planner.

2.2. The eRisk Shared Task Series

Losada and Crestani [5] established the ERDE (Early Risk Detection Error) metric and the test collection methodology that has anchored subsequent editions. From 2017 onward [4, 6, 7], eRisk expanded to multiple clinical conditions while retaining the core paradigm of scoring decisions over passively collected post streams. Task 1 in 2026 breaks this convention: the evaluation environment is now *interactive*, which invalidates prior retrieval- and classification-based pipelines and necessitates entirely new metrics, described in Section 3. The full motivation, data collection protocol, and complete results for this edition are given in the official eRisk 2026 overview papers [12, 13], to which we refer readers for details beyond the scope of this system report.

2.3. Mental Health NLP and Dialogue Agents

CLPsych [14] pioneered shared evaluation of computational mental-health detection. Coppersmith et al. [15] established that elevated first-person pronoun rates and negative-affect vocabulary are reliable depression markers in social media—a finding directly reflected in the lexical component of our Evaluation Agent. Later surveys by Yates et al. [16] and Chancellor and De Choudhury [17] extended these insights to other conditions. A critical distinction separates all these prior systems from the present work: they *classify* text that already exists, whereas this pipeline must *generate* questions calibrated to surface clinical signals. Woebot [18] demonstrated that users engage meaningfully with automated conversational agents; Sharma et al. [19] showed that empathic responses deepen emotional disclosure, motivating run 2; and Moon's reciprocity norm [20] grounds the self-disclosure framing in run 1.

2.4. Parameter-Efficient Fine-Tuning

LoRA [8] enables efficient model adaptation by injecting low-rank matrices into frozen transformer layers. QLoRA [21] extends this to 4-bit NF4 quantisation, reducing memory by approximately $4\times$ with negligible accuracy loss. Naively loading 20 separate fine-tuned models would require roughly 320 GB of VRAM; the adapter-swapping architecture introduced in Section 4 reduces this to 15 GB by maintaining a single quantised base and hot-swapping lightweight adapters between personas.

3. Task Setup

Twenty LLM-simulated personas (IDs 1–20), released incrementally between 16 February and 20 April 2026 at `Anxo/erisk26-task1-patient-{ID}-adapter` on HuggingFace, constitute the evaluation set for Task 1. The organisers specify a mandatory verbatim system prompt:

> *"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. Do not mention you are an AI."*

Modification of this prompt is prohibited and constitutes a protocol violation. Submissions consist of `results_{run}.json` (predicted BDI score plus up to four key symptoms) and `interactions_{run}.json` (full conversation transcript). Up to three runs may be submitted per team.

3.1. Official Evaluation Metrics

Four metrics measure complementary aspects of system quality. **DCHR** (Diagnostic Category Hit Rate) measures exact severity-band agreement:

$$\text{DCHR} = \frac{1}{|U|} \sum_u \mathbf{1}[\hat{C}_u = C_u] \tag{1}$$

ADODL (Absolute Deviation of Ordinal Diagnostic Level) captures numerical score proximity, normalised by the maximum possible deviation of 63:

$$\text{ADODL} = \frac{1}{|U|} \sum_u \frac{63 - |\hat{s}_u - s_u|}{63} \tag{2}$$

ASHR (Active Symptom Hit Rate) measures the fraction of correctly identified key symptoms per persona:

$$\text{ASHR} = \frac{1}{|U|} \sum_u \frac{|\hat{K}_u \cap K_u|}{|K_u|} \tag{3}$$

LDCHR and **LASHR** apply a speed-based discount penalising decisions reached after many turns:

$$\text{penalty}(k) = -1 + \frac{2}{1 + e^{-p(k-1)}}, \quad p = \frac{\ln 3}{29}, \quad \text{speed}(k) = 1 - \text{penalty}(k) \tag{4}$$

Speed decays from 1.00 at turn 1 to 0.83 at turn 10 and 0.50 at turn 30 (Table 1), rewarding early confident decisions.

Table 1
Speed values at representative turn counts (Eq. 4).

k	1	5	10	15	20	25	30
speed(k)	1.00	0.93	0.83	0.74	0.66	0.57	0.50

4. System Description

4.1. Architecture Overview

The pipeline is organised as a turn-based loop across five components: a Persona Client, a Conversational Agent, an Evaluation Agent, a Symptom Planner, and a Risk Fusion stage (Figure 1, presented after the Persona Client is introduced below).

4.2. Persona Client

`SharedBaseModel` retains a single 4-bit NF4 quantised Llama-3-8B in GPU memory; `AdapterSwapClient` hot-swaps LoRA weights between personas, covering all 20 within 15 GB VRAM. Generation parameters: temperature 0.6, top-p = 0.9, maximum 256 new tokens per turn.

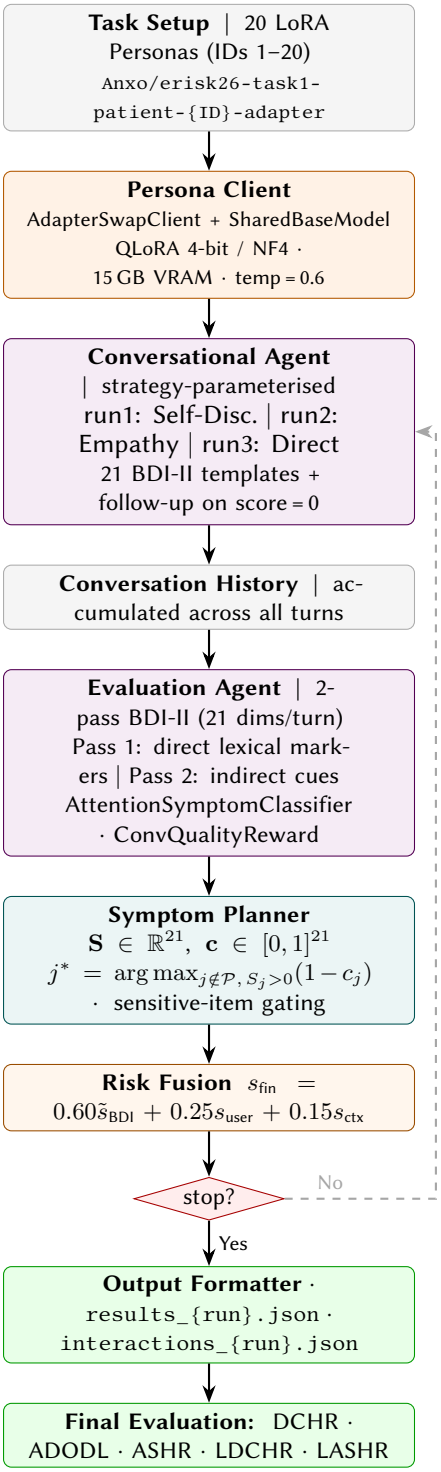

Figure 1: System pipeline (turn-based loop). Colour coding: orange = persona layer, purple = agent layer, teal = planning, green = output. The dashed arrow is the "No" branch returning to the Conversational Agent for the next turn. Early stopping fires when any of three conditions holds: (1) estimated BDI ≥ 19 and turn $>$ minimum; (2) mean confidence $\bar{c} \geq 0.85$; (3) all non-sensitive dimensions probed.

Figure 1 illustrates the complete turn-based pipeline. At each turn, the Persona Client obtains the

persona's reply, the Evaluation Agent re-scores all 21 BDI-II dimensions against the accumulated conversation, the Symptom Planner selects the next dimension to probe and checks stopping criteria, and the Conversational Agent formulates the next question. Iterations proceed for at most 21 turns, with early stopping permitted from turn 3 onward.

4.3. Conversational Agent

Each strategy provides one question template per BDI-II symptom dimension plus a follow-up probe triggered when the initial response scores 0 (Fix 2, Table 2).

Self-Disclosure (run 1). Opens with a first-person anecdote to exploit the reciprocity norm [20]: *"I've had stretches where everything felt grey. How has your mood been?"*

Empathy (run 2). Validates emotional state before eliciting content, following motivational interviewing principles [19]: *"Sometimes life can feel really heavy. How has your mood been?"*

Direct (run 3). Poses unframed BDI-domain questions for maximum dimensional coverage efficiency: *"How would you describe your mood over the past two weeks?"*

Rationale for the self-disclosure opening. Run 1 was included as a controlled ablation of Moon's reciprocity norm [20], which predicts that a brief, low-intensity disclosure from the interviewer lowers the persona's threshold for reciprocal self-disclosure and can surface affect-laden language useful for calibration (Section 6). We acknowledge the reviewer's concern that this framing departs from standard clinical-interview practice, where a counsellor or therapist is trained to withhold personal affect so as not to model or reinforce negative emotion in the patient. Two design choices mitigate this risk: the disclosed anecdote is a single, generic, past-tense sentence ("I've had stretches where everything felt grey") rather than an ongoing or escalating emotional narrative, and it appears only once, at the first turn, rather than being repeated or intensified over the conversation. We do not observe evidence in the transcripts of personas mirroring the interviewer's specific affective language beyond the immediate reply. Nevertheless, we agree this framing is better suited to a research ablation than to a deployed screening assistant, and we treat run 1 accordingly, as a lower bound illustrating the coverage cost of prioritising rapport over systematic probing rather than as a recommended interaction style.

4.4. Evaluation Agent

After every turn, all 21 BDI-II dimensions are re-scored from the full conversation history through two sequential passes (Figure 2).

Pass 1 (direct). Matches surface-level lexical markers: for example, "*lie awake for hours*" maps to Sleep = 2–3, while "*feel completely worthless*" maps to Worthlessness = 3.

Pass 2 (indirect). Handles contextual cues after resolving negations and hedges: "*haven't been going out*" yields Loss of Interest = 1–2. Here a range, rather than a single point score, is emitted whenever the contextual evidence is compatible with more than one adjacent BDI-II severity level for that item—e.g., withdrawal language alone cannot distinguish a mild reduction in activity (1) from a more pronounced loss of interest (2) without an explicit statement of degree. The Score Merge stage (Figure 2) resolves each range to a single value by taking its mid-point for the running BDI total, while the width of the range is propagated as an inverse contribution to that dimension's confidence c_j: wider ranges yield lower confidence, which in turn makes the dimension more likely to be re-probed by the Symptom Planner (Eq. 5). Without this pass, personas expressing mild-range symptoms through hedged or elliptical language are underscored by approximately 5–8 BDI points.

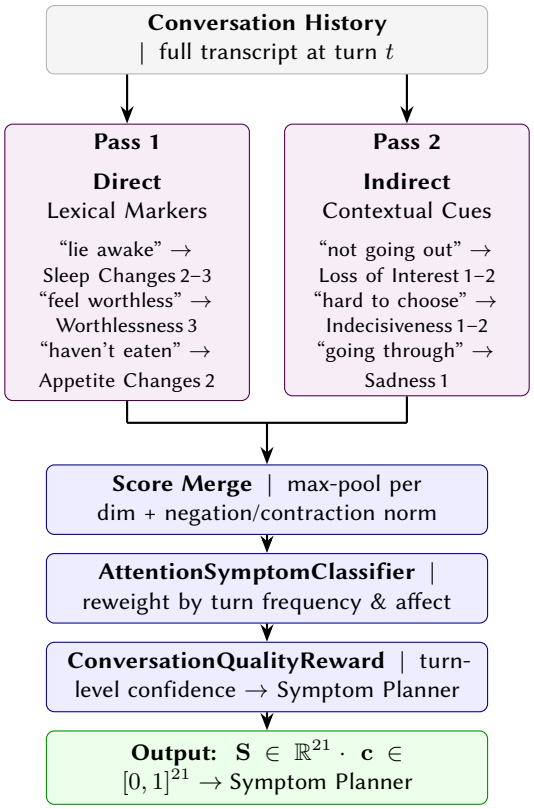

Figure 2: Evaluation Agent: two-pass scoring architecture. Pass 1 extracts explicit symptom markers from the conversation history; Pass 2 infers implicit contextual cues after negation normalisation. The T-bar join aggregates both passes before *AttentionSymptomClassifier* reweights scores by conversational salience, and *ConversationQualityReward* forwards a turn-level confidence signal to the Symptom Planner.

4.5. Symptom Planner

The planner maintains score vector $\mathbf{S} \in \mathbb{R}^{21}$ and confidence vector $\mathbf{c} \in [0,1]^{21}$. At each turn the next dimension to probe is the unprobed dimension with non-zero estimated score and lowest confidence:

$$j^* = \underset{j:\, S_j > 0,\, j \notin \mathcal{P}}{\arg\max}\ (1 - c_j) \tag{5}$$

Items 9 (Suicidal Thoughts) and 21 (Loss of Interest in Sex) are gated to preserve conversational safety. Stopping fires when any of the three conditions listed in Figure 1 is satisfied.

4.6. Risk Fusion

A weighted linear combination aggregates three complementary signals:

$$s_{\text{final}} = 0.60\, \tilde{s}_{\text{BDI}} + 0.25\, s_{\text{user}} + 0.15\, s_{\text{context}} \tag{6}$$

where $\tilde{s}_{\text{BDI}} = S_{\text{total}}/63$ and s_{context} encodes hopelessness, withdrawal, and anhedonia cues across the transcript. s_{user} operationalises the first-person negative-affect density following Coppersmith et al. [15]: every persona turn is tokenised and scanned for first-person singular markers ("I", "me", "my", "myself"); for each such marker, a local window of ± 5 tokens is checked against a negative-affect lexicon (sadness, worthlessness, fatigue, hopelessness, and related terms drawn from the BDI-II item vocabulary). s_{user} is the proportion of first-person markers with at least one co-occurring negative-affect term, aggregated across the full conversation and rescaled to $[0, 1]$; a higher density indicates that negative affect is being expressed in direct self-reference rather than in general or third-person terms.

4.7. Iterative Pipeline Refinement

Table 2 documents five targeted fixes applied across pipeline versions v1 to v15.

Table 2
Five targeted refinements from pipeline v1 to v15.

Fix	Description
1	Replaced closed yes/no starters with open, topic-anchored questions
2	Per-symptom follow-up probe triggered when initial response scores 0
3	Early exit when estimated BDI ≥ 19, conserving turns for uncertain dimensions
4	Randomised symptom probing order, independently seeded per run
5	Adaptive stopping when mean confidence $\bar{c} \geq 0.85$, avoiding over-probing

5. Experimental Setup

Hardware. All experiments ran on Google Colab with an NVIDIA T4 (15 GB VRAM) under 4-bit NF4 quantisation. Dependencies: `transformers`, `peft`, `bitsandbytes`, `accelerate`, and `torch`.

Run design. Three independent runs process all 20 personas, sharing no information across runs. The upper bound on conversation length is 21 turns; early stopping is permitted from turn 3 onward.

Evaluation protocol. Official ground-truth labels were provided by the task organisers after submission. No labelled development data was available during system design; the entire pipeline operates zero-shot.

6. Results

6.1. Official Scores

Table 3 reports all five official metrics for each submission run.

Table 3
Official eRisk 2026 Task 1 results for team lexxy (20 personas). ↑ = higher is better. **Bold** = best within our three runs. Δ_{ASHR} = percentage gain relative to run 1.

Run	Strategy	DCHR↑	ADODL↑	ASHR↑	Δ_{ASHR}	LDCHR↑	LASHR↑
run1	Self-Disclosure	0.3000	**0.8556**	0.1875	—	0.0137	0.0122
run2	Empathy	0.3000	0.8492	0.2500	+33.3%	0.0144	0.0162
run3	Direct	0.3000	0.8413	**0.2875**	+53.3%	**0.0186**	**0.0196**

Run 3 ASHR = 0.2875 ranks 5th globally among all eRisk 2026 Task 1 submissions.

6.2. ASHR Comparison across Strategies

ASHR improves monotonically across runs (0.188 → 0.250 → 0.288). The 53.3% relative gain from run 1 to run 3 confirms that unframed BDI-domain coverage maximises symptom breadth. Empathy occupies an intermediate position (ASHR = 0.250, +33.3%), suggesting partial but not full compensation for the disclosure-depth advantage of emotional framing.

6.3. Category Accuracy and Score Proximity (DCHR, ADODL)

DCHR = 0.30 across all three runs—equivalent to six of 20 personas correctly categorised—signals that category accuracy is constrained by threshold sensitivity rather than conversational framing. A

predicted score of 18 when the ground truth is 19 incurs a full category penalty despite a numeric error of only one point; this ceiling is structural and strategy-agnostic.

ADODL values between 0.84 and 0.86 translate to a mean absolute error of approximately $63 \times 0.15 \approx 9.5$ BDI points, confirming that relative severity is correctly ordered across personas even where exact category assignment fails.

6.4. The ASHR–ADODL Depth–Breadth Trade-off

The most notable finding from the three-run ablation is that ASHR and ADODL move in strictly opposite directions. Self-disclosure produces the highest ADODL (0.856) but the lowest ASHR (0.188); direct questioning delivers the highest ASHR (0.288) but the lowest ADODL (0.841); empathy sits between both. Affect-laden responses supply rich calibration cues for the numerical BDI total while tending to dwell on one or two dominant themes, leaving many dimensions unexplored. Direct questions distribute coverage more evenly but elicit terser responses with weaker calibration signal. This trade-off has not been previously documented in the eRisk literature and may generalise to other multi-dimensional conversational assessment tasks.

7. Discussion

Strength: symptom identification. Run 3's ASHR of 0.2875 (rank 5) validates that systematic, unframed BDI-domain coverage maximises symptom breadth. The 53.3% relative gain from run 1 to run 3 confirms that emotional framing, while psychologically grounded, trades symptom breadth for rapport.

Strength: score accuracy. ADODL of 0.84–0.86 reflects close score proximity to ground truth across all strategies. High ADODL combined with moderate DCHR reveals a specific gap: the system correctly ranks relative severity but lands near BDI boundaries without crossing them. A lightweight post-hoc recalibration—mapping s_{final} to severity bands via held-out boundary offsets—could substantially improve DCHR without modifying the conversational or scoring components.

Technical advantages. Decoupling the Conversational Agent from the Evaluation Agent allows each to be optimised independently. `AttentionSymptomClassifier` amplifies repeatedly-mentioned or high-affect symptoms without conflating them with overall severity, preserving dimensional resolution. Memory-efficient serving through QLoRA 4-bit NF4 and adapter hot-swapping keeps the entire 20-persona evaluation within a single 15 GB consumer GPU.

Practical advantages. The pipeline operates fully zero-shot: no labelled development examples, no fine-tuning, and no annotation of conversations. The three-run ablation delivers actionable depth–breadth insight at no additional infrastructure cost, using the same base model with only question templates and probing order differing across runs. Sensitive-item gating (items 9 and 21) ensures the system respects conversational norms and avoids distressing exchanges.

Performance advantages. Across five official metrics, results are consistent and competitive: DCHR = 0.30 is stable across all strategies; ADODL = 0.84–0.86 reflects close score accuracy; ASHR = 0.2875. The ASHR–ADODL trade-off is a novel empirical finding that can guide strategy selection in future conversational clinical assessment systems.

Why indirect scoring matters. Patient personas rarely assert symptoms directly. Depression surfaces through hedged affect ("things are sort of okay, I guess"), temporal stasis ("nothing has changed, really"), and implicit social withdrawal cues. Omitting Pass 2 leaves these signals undetected, compounding to a systematic 5–8 BDI-point underestimate on mild-range personas.

Ethical considerations. All interactions involved LLM-simulated personas; no real patient data was collected or analysed. The pipeline avoids leading questions, symptom suggestion, and explicit references to suicidality or other sensitive topics through sensitive-item gating (items 9 and 21).

8. Conclusion and Future Work

For eRisk 2026 Task 1, we built a two-agent pipeline that interviews LLM-simulated patient personas across all 21 BDI-II dimensions using three separately parameterised conversational strategies. The direct-questioning run reaches ASHR = 0.2875 (rank 5), and score accuracy stays consistent across strategies at ADODL = 0.841–0.856. The main contributions are the modular two-agent architecture, the memory-efficient QLoRA adapter serving setup, and the two-pass indirect scoring mechanism, which corrects for systematic underestimation on mild-range personas. The three-run ablation also surfaces a depth–breadth trade-off between ASHR and ADODL that, as far as we know, has not been reported before: direct questioning gains 53.3% in symptom coverage over self-disclosure, but at the cost of some score calibration accuracy.

A few directions look promising for future work. One is speed-aware stopping — reformulating the Symptom Planner's objective as a marginal utility function that weighs expected confidence gain against the LDCHR/LASHR speed penalty, which could allow earlier termination without losing coverage. Another is boundary recalibration: since ADODL stays consistently high, a lightweight post-hoc adjustment at the four BDI severity boundaries might improve DCHR without touching the conversational or scoring components at all. A third direction is combining strategies — starting with direct BDI-domain probes for coverage and switching to empathic framing once all dimensions have been initialised, which could in principle combine the ASHR advantage of run 3 with the ADODL advantage of run 1. Finally, LDCHR and LASHR could be used more deliberately in future submissions, for instance through a confidence-gated stopping rule that stops earlier for high-severity personas and spends the saved turns on under-explored mild-range dimensions.

Acknowledgments

We thank the eRisk 2026 task organisers for releasing the persona adapters, evaluation framework, and official ground-truth labels. All experiments were conducted on Google Colab (T4 GPU) under the Meta Llama 3 community licence.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) in order to: Drafting content, Grammar and spelling check, Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] H. A. Whiteford, L. Degenhardt, J. Rehm, et al., Global burden of disease attributable to mental and substance use disorders, The Lancet 382 (2013) 1575–1586. doi:10.1016/S0140-6736(13)61611-6.

[2] World Health Organization, Depressive disorder (depression), 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[3] G. Thornicroft, Most people with mental illness are not treated, The Lancet 370 (2007) 807–808. doi:10.1016/S0140-6736(07)61392-0.

[4] D. E. Losada, F. Crestani, J. Parapar, eRisk 2017: CLEF lab on early risk prediction on the internet: Experimental foundations, in: Proceedings of the 8th International Conference of the CLEF Association (CLEF 2017), volume 10456 of *Lecture Notes in Computer Science*, Springer, 2017, pp. 346–360. doi:10.1007/978-3-319-65813-1_30.

[5] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Proceedings of the 7th International Conference of the CLEF Association (CLEF 2016), volume 9822 of *Lecture Notes in Computer Science*, Springer, 2016, pp. 28–39. doi:10.1007/978-3-319-44564-9_3.

[6] D. E. Losada, F. Crestani, J. Parapar, Overview of eRisk: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction — CLEF 2018, volume 11018 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 343–361.

[7] D. E. Losada, F. Crestani, J. Parapar, eRisk 2019: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction — CLEF 2019, volume 11696 of *Lecture Notes in Computer Science*, Springer, 2019, pp. 340–357.

[8] 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, 2022. arXiv:2106.09685, published at ICLR 2022.

[9] Meta AI, Introducing Meta Llama 3: The most capable openly available LLM to date, 2024. URL: https://ai.meta.com/blog/meta-llama-3/.

[10] A. T. Beck, R. A. Steer, G. K. Brown, Manual for the Beck Depression Inventory-II, Psychological Corporation (1996). Second edition.

[11] D. J. Kupfer, E. Frank, Therapeutic targets in late-onset depression, Psychopharmacology Bulletin 32 (1996) 643–651.

[12] 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.

[13] 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.

[14] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, CLPsych 2015 shared task: Depression and PTSD on Twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2015), Association for Computational Linguistics, 2015, pp. 31–39. doi:10.3115/v1/W15-1204.

[15] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in Twitter, in: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2014), Association for Computational Linguistics, 2014, pp. 51–60.

[16] A. Yates, A. Cohan, N. Goharian, Depression and self-harm risk assessment in online forums, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017), Association for Computational Linguistics, 2017, pp. 2968–2978. doi:10.18653/v1/D17-1322.

[17] S. Chancellor, M. De Choudhury, Methods in predictive techniques for mental health status on social media: A critical review, npj Digital Medicine 3 (2020) 43. doi:10.1038/s41746-020-0233-7.

[18] K. K. Fitzpatrick, A. Darcy, M. Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial, volume 4, 2017, p. e19. doi:10.2196/mental.7785.

[19] A. Sharma, A. S. Funk, T. Althoff, A computational approach to understanding empathy expressed in text-based mental health support, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Association for Computational Linguistics, 2020, pp. 5263–5276. doi:10.18653/v1/2020.emnlp-main.425.

[20] Y. Moon, Intimate exchanges: Using computers to elicit self-disclosure from consumers, Journal of Consumer Research 26 (2000) 323–339. doi:10.1086/209566.

[21] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, 2023. arXiv:2305.14314, published at NeurIPS 2023.