INSA-Lyon at eRisk 2026 Task 1: Multi-Agent BDI-II Assessment with Coverage Tracking and Post-hoc Correction

Notebook for the eRisk Lab at CLEF 2026

Diana Nurbakova[1,*]

[1]INSA Lyon, CNRS, Universite Claude Bernard Lyon 1, LIRIS, UMR5205, 69100 Villeurbanne, France

Abstract

eRisk 2026 Task 1 frames depression detection as active conversational assessment: a system interviews one of twenty fine-tuned large language model personas of unknown depression profile, then reports a Beck Depression Inventory-II (BDI-II) total and the four most active symptoms. We present a multi-agent pipeline that pairs an orchestrator and an interviewer guided by Motivational-Interviewing principles with four domain-specialised BDI-II assessors, prompts calibrated against the DepreSym corpus, a Theory-of-Mind module that tracks symptom coverage across the conversation, and a post-hoc band-correction stage. Over 19 of the 20 personas, our strongest run reached a latency-weighted symptom hit rate of 0.1525, second among the incomplete-submission runs, with an unweighted symptom hit rate of 0.2237. Alongside this, we report three findings that bound how the result should be read. Severe under-prediction was universal: no severe persona reached the severe band, which alone caps band accuracy on this balanced test set. The somatic-boost correction intended to rescue these cases cannot fire on them, because its activation threshold excludes the under-scored totals it targets. The Theory-of-Mind Wasserstein signals were left uncomputed at submission by a dependency bug; we reconstruct them post-hoc and find that the turn-over-turn detector misses disclosure that is front-loaded into the early turns.

Keywords

depression detection, BDI-II, conversational assessment, multi-agent pipeline, LLM personas, Theory of Mind, Motivational Interviewing / OARS, Wasserstein distance

1. Introduction

The following is from one of our development interviews on the TalkDep persona set [1], which the eRisk organisers also draw on to construct the Task 1 test cohort.

> **Interviewer:** How has your sleep been over the last couple of weeks?
>
> **Marco:** It is awful. I either cannot fall asleep, or I wake up in the middle of the night and cannot get back to sleep.

Marco is not a patient. He is one of twelve clinician-designed LLM personas in TalkDep, each calibrated to a hidden Beck Depression Inventory-II (BDI-II) profile [2]. Read at the surface, the reply above is a fairly literal answer to a question about sleep. The gold-standard assessment is not: Marco's reference BDI-II total is 38 out of 63, in the severe band of the 21-item depression inventory used to score the task. A pipeline that recovers symptom scores only when evidence is explicitly disclosed, as ours largely does, will under-score him substantially: at Task 1, our three runs read his depression total as 20, 20, and 19, two bands below severe. That gap is eRisk 2026 Task 1 in miniature: the signal is present but implicit, and recovering it depends as much on which questions we choose to ask as on how we score the answers.

Depression has no single biomarker, and clinical assessment instead proceeds through standardised self-report inventories, each covering a defined set of symptom domains. The Beck Depression

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

*Corresponding author.

diana.nurbakova@insa-lyon.fr (D. Nurbakova)

0000-0002-6620-7771 (D. Nurbakova)

Inventory-II [2] is among the most widely used: it scores 21 symptoms, each on a 0–3 scale anchored to behaviourally specific statements that run from the absence of the symptom to its severe form, and the total over the 21 items, in $[0, 63]$, maps to four conventional severity bands. eRisk 2026 adopts the bands minimal (0–9), mild (10–18), moderate (19–29), and severe (30–63). This per-item structure matters for how the task should be approached. Depression detection here is not a binary decision or a single-number regression, but structured assessment: 21 distinct symptom dimensions scored and then aggregated. Task 1 makes the structure explicit by asking for both a BDI-II total and the four most active symptoms, and our system is built around exactly that decomposition.

eRisk 2026 Task 1, Conversational Depression Detection [3, 4], makes the assessment concrete by replacing static-text classification with active conversational elicitation. A participating team interacts, turn by turn, with one of twenty fine-tuned LLM personas of unknown depression profile, with no access to ground-truth labels during the conversation, and then submits a BDI-II total and a top-four symptom set for each. The setting has a property we return to throughout: there is no fixed dialogue dataset shared across teams, so each team's transcripts are co-produced with its own interviewer policy, and the interviewer is therefore part of the system under evaluation rather than a fixed input. We address the task with a multi-agent pipeline that pairs a dual-module orchestrator and an interviewer guided by Motivational-Interviewing principles [5] with four domain-specialised BDI-II assessors, prompts calibrated against the DepreSym corpus [6], a Theory-of-Mind module that tracks symptom coverage across the conversation, and a post-hoc band-correction stage. Section 3 gives the background, Section 4 describes the system, and Section 5 reports the results.

Our three runs covered 19 of the 20 personas, and our strongest single result was on symptom identification: Run 3 reached a latency-weighted symptom hit rate (LASHR) of 0.1525, second among the incomplete-submission runs, with an unweighted symptom hit rate (ASHR) of 0.2237, fourth in the same table. We state these as standings; our experimental design does not let us attribute them cleanly to individual components, and we are careful about that throughout Section 5. The contributions of this paper are the following.

- A multi-agent BDI-II assessment pipeline for conversational depression detection, with named load-bearing components: four domain-specialised assessors, DepreSym-calibrated prompts, a coverage-tracking Theory-of-Mind module that supplies the orchestrator with live coverage-gap diagnostics, and a post-hoc band-correction stage (Section 4).
- An empirical comparison of post-hoc correction strategies on a controlled development set, showing that the choice of correction matters more for band accuracy than the depth of the feature pipeline, together with an observation on the eRisk test personas that, for moderate-band cases, leaving raw assessor totals uncorrected can beat a flat subtraction (Section 5).
- A post-hoc reconstruction of the Theory-of-Mind module's Wasserstein signals, which an uncaught dependency bug left uncomputed at submission, exposing a structural limitation: under the eRisk-native turn-over-turn convention, the warmup-baseline detector misses disclosure that is front-loaded into the early turns (Section 5).
- Honest reporting of several structural findings, namely universal severe under-prediction (no severe persona reached the severe band), a somatic-boost correction whose activation threshold structurally excludes the severe cases it was meant to rescue, a mid-submission configuration change in Run 3 that makes its two halves a comparison of correction presence rather than of Theory-of-Mind calibration, and an unreconciled discrepancy between our local re-scoring and the official figures (Section 5).

2. Task and Evaluation

eRisk 2026 Task 1 [3, 4, 7, 8] frames depression detection as active conversational assessment. Given a turn-based API to one of twenty fine-tuned LLM personas of unknown depression profile, a participating team must extract two outputs per persona: a BDI-II total $\hat{s} \in \{0, \ldots, 63\}$ and a ranked top-four set of

active symptoms. The personas are organiser-provided LoRA [9] adapters on Meta-Llama-3-8B-Instruct, each fine-tuned to a distinct, undisclosed depression profile. Interaction is strictly turn-based (the team issues a question and the persona replies), with no access to ground-truth labels during the conversation and a team-set per-conversation turn budget; teams may submit up to three runs. One property worth flagging early is that Task 1 has no shared dialogue dataset: unlike Task 2 (fixed Reddit threads) or Task 3 (fixed sentence corpus), each team generates its own transcripts by interacting with the personas. The persona behaviour is the official test LoRA's behaviour, yet the specific exchange exists only because our interviewer asked that particular question. The interviewer policy is therefore part of the system under evaluation, not a fixed input. We revisit this point in Section 4.

3. Background

3.1. The Beck Depression Inventory-II

The Beck Depression Inventory-II [2] comprises 21 items that span the affective, cognitive, somatic, and motivational content of depression. Rather than asking respondents to rate each item on a generic frequency scale, the instrument presents, for every item, four ordered statements that describe progressively severe forms of the symptom in question: one statement is selected, and that selection yields an integer score from 0 to 3. The total BDI-II score is the sum across the 21 items and lies in $[0, 63]$; as indicated in Section 1, the severity bands used by eRisk 2026 are minimal (0–9), mild (10–18), moderate (19–29), and severe (30–63). Mind that this per-item, statement-anchored structure is what makes the instrument a tractable target for an LLM scorer: each item amounts to a small, well-defined four-way classification rather than a free-form severity estimate, and the assessor design in Section 4 exploits this by organising the 21 items into four content clusters.

To make the per-item structure concrete, consider this turn from a development-time interview with Linda, one of the trial personas of the TalkDep set [1]. Asked to elaborate on what made her feel as she did, Linda replied: *"I just feel like I am not pulling my weight anymore. My husband works so hard, and I just stay at home. I feel like a failure."* The closing clause maps cleanly onto BDI-II item 14 (*Worthlessness*): the score-3 anchor for that item is, in the published wording, *I feel I am a complete failure as a person*, and the persona's phrasing is a near-verbatim rephrasing. The earlier clauses fit lower anchors of the same item, describing a more contingent sense of inadequacy (not pulling one's weight) rather than a totalising self-judgement. Mind that a single score-3 statement on item 14 does not by itself indicate severe depression: Linda's gold BDI-II total is 28, which falls in the moderate band. Total severity emerges from the aggregation of evidence across all 21 items, not from any single anchor crossing.

A second illustration, this time on a somatic item: in another TalkDep interview, Marco was asked how his current state affected his ability to sleep. He replied: *"It is awful. I either cannot fall asleep, or I wake up in the middle of the night and cannot get back to sleep."* The reply names two distinct sleep disturbances. BDI-II item 16 (*Changes in Sleeping Pattern*) discriminates between them, namely an inability to fall asleep (onset insomnia) versus waking in the middle of the night or earlier than usual (middle and terminal insomnia), each scored against separate sub-anchors, with the most severe end of the item reserved for changes affecting sleep by hours rather than minutes. The lesson for an automated assessor is that per-item scoring is not symptom detection: an assessor that returns a score of 1 the moment any sleep disturbance is mentioned will systematically under-score personas with prominent somatic content. The four-cluster assessor design in Section 4 acts on this by sending each item to a domain-specialised prompt that surfaces the relevant sub-anchors.

3.2. Methodological grounding

A growing body of work uses large language models as definition-driven scorers of clinical instruments: rather than fine-tuning on labelled examples, the model is given the published scoring rubric and asked to apply it directly, with eRisk participants in the past two editions contributing variants of this

approach [10, 11, 6]. The known failure mode of such scorers is over-attribution: when presented with text that is mildly distressed or merely affect-laden, an LLM-as-assessor tends to return non-zero item scores even where the clinical scoring criteria would not, inflating totals on minimal-band cases and compressing the predicted distribution toward the moderate band. The DepreSym corpus [6] provides one route to calibrating against this tendency: it pairs sentences extracted from depression-relevant subreddits with judgements of relevance to each BDI-II symptom, yielding per-item relevance rates that we use to set prompt strictness. In our pipeline (Section 4), items with very low DepreSym relevance receive prompts that demand stronger evidence before returning a non-zero score, while higher-relevance items use less strict prompts.

A second methodological lens we adopt is Theory of Mind. In its NLP usage [12], Theory of Mind refers to the inference of mental states from linguistic evidence that is typically partial: a reader, or an LLM, must reconstruct what a speaker believes, feels, or intends, from utterances that rarely state these things directly. BDI-II scoring inherits this inferential demand, but unevenly across the 21 items. Somatic items such as sleep and fatigue (items 16 and 20) tend to be directly disclosed when present, since they are concrete behaviours that respondents readily report; the Marco illustration above is of this kind. Cognitive items such as worthlessness, self-criticism, and pessimism (items 14, 8, and 2) require integration of cumulative framing across multiple turns, since a single utterance rarely settles the score; the Linda illustration is closer to this end of the spectrum. Affective items sit between these poles. This gradient has a system-design consequence to which we return in Section 4: an assessor that returns a score only when explicit evidence is present will systematically under-score the high-demand items, and so our assessors maintain an explicit per-item state of SCORED, NO_EVIDENCE, or EVIDENCE_OF_ABSENCE.

The third lens we draw on is Wasserstein distance, which we use to track how a persona's expressed BDI-II profile evolves across the conversation. At any turn t, the evidence the persona has so far disclosed can be summarised as a 21-dimensional vector of per-item scores; the question of how much the profile has shifted between two turns is then a question about a distance between two such vectors. Pointwise norms answer this question poorly, because they treat the 21 items as interchangeable coordinates and ignore both the ordinal structure of the 0–3 anchors and the affective, cognitive, somatic, and functional clustering of BDI-II items. The 1-Wasserstein distance [13], in contrast, measures the minimum mass-transport cost between two distributions on a metric space. For two distributions μ, ν on a space $\mathcal{X}$ equipped with a ground metric d,

$$W_1(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} d(x, y)\, \mathrm{d}\pi(x, y), \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of all couplings (joint distributions) with marginals μ and ν. Treating each BDI-II profile as a discrete distribution over the 21 items, with mass equal to the per-item score and a ground metric d that encodes item proximity (so that mass moved within a content cluster costs less than mass moved across clusters), makes the cluster geometry available to the metric. A persona who shifts from disclosing only somatic content to additionally disclosing cognitive content registers a larger W_1 than one who redistributes within somatic items, even when the total score is unchanged. This is the conceptual setting; Section 4 describes which Wasserstein quantities our pipeline was specified to compute, which were deployed at submission time, and which are reconstructed post-hoc in Section 5.

4. System

4.1. Pipeline architecture

Our submission is a multi-agent pipeline that conducts a bounded turn-based interview with each persona and aggregates the resulting evidence into a BDI-II total and a ranked top-four set of active symptoms. Figure 1 shows the architecture. At each turn, a dual-module orchestrator chooses what to probe next; an OARS-prompted interviewer puts that probe to the persona; the persona's reply is then analysed by the four domain-specialised assessors, the linguistic-feature extractor, and the

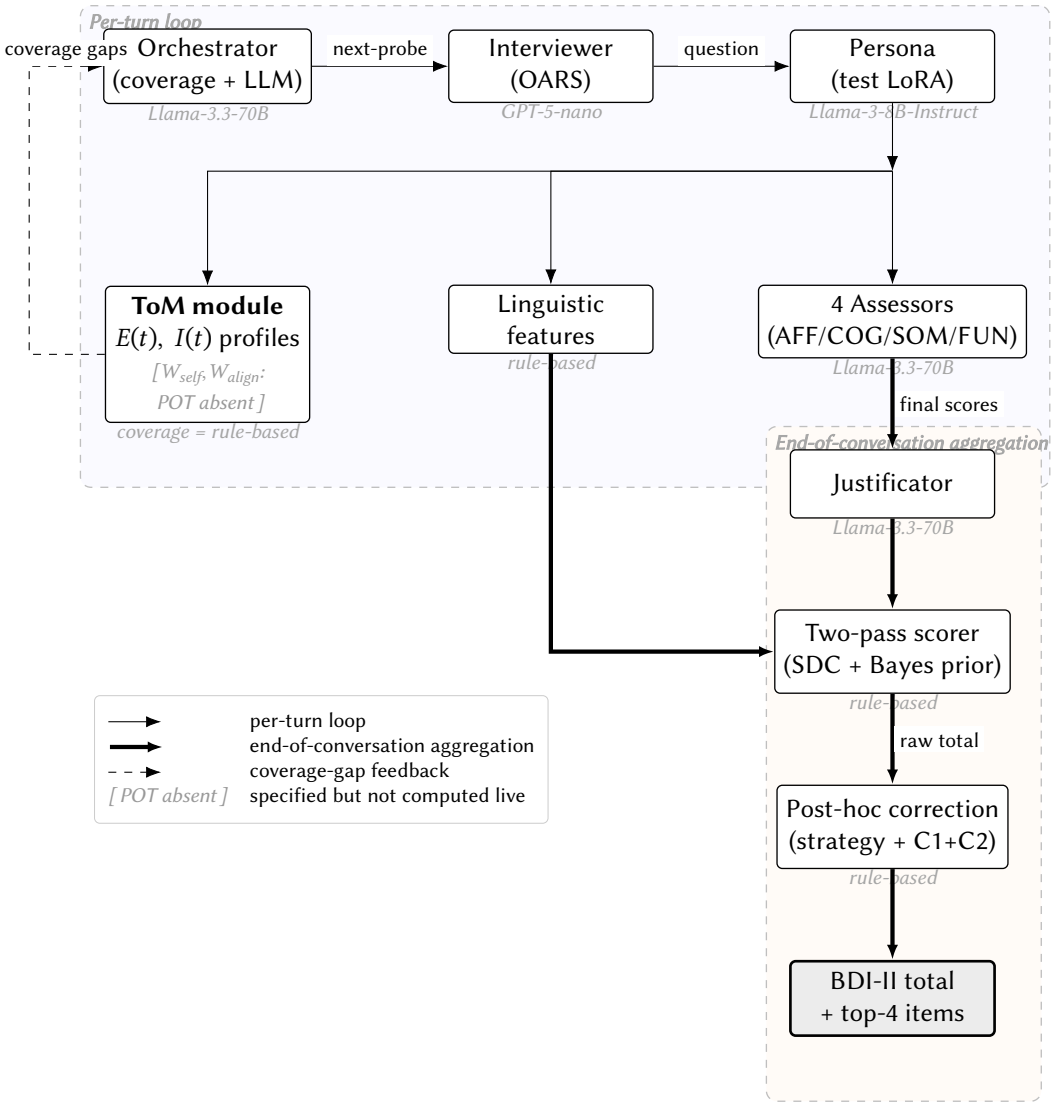

Figure 1: INSA-Lyon pipeline as deployed at submission. The per-turn conversation loop (top) couples a dual-module orchestrator with an OARS-prompted interviewer and the test persona; each turn's reply is analysed by the Theory-of-Mind module, the linguistic-feature extractor, and the four domain-specialised assessors. The ToM module was enabled only in the tail of Run 3 (personas 13–20); the Wasserstein quantities W_{self} and W_{align} were specified but not computed in any run, because the Python Optimal Transport library was absent in the inference environment (discussed in Section 4.2 and Section 5). End of conversation, the justificator audits the assessor output and the two-pass scorer aggregates per-item scores into a raw BDI-II total and a top-four active-items list; a post-hoc correction stage (correction strategy plus the C1 confidence gate and C2 somatic boost) produces the final outputs.

Theory-of-Mind module, after which the orchestrator updates its state and the loop continues. At the end of the conversation, a justificator audits the assessor output, a two-pass scorer aggregates per-item scores into a raw BDI-II total and a top-four ranking, and a post-hoc correction stage adjusts the raw total before it is written to the submission file. We ran the pipeline on Google Colab. The persona model (Llama-3-8B-Instruct with the organiser-provided LoRA adapter) ran on the Colab GPU runtime; the remaining LLM components were accessed via hosted APIs: GPT-5-nano (OpenAI) for the interviewer, and Llama-3.3-70B-Instruct-Turbo (Together AI) for the orchestrator reasoning module, the four assessors, and the justificator. We release the pipeline implementation at https://github.com/diana-nurbakova/depression-detection.

The orchestrator is implemented as two modules running in tandem. A programmatic coverage tracker maintains, for each of the 21 BDI-II items, an evidence state of SCORED, NO_EVIDENCE, or

EVIDENCE_OF_ABSENCE, updated each turn from the assessors' output. An LLM-reasoning module, instantiated from Llama-3.3-70B-Instruct [14], reads the coverage state and the recent transcript and chooses which BDI-II item the interviewer should probe next. The interviewer itself is a separate LLM, GPT-5-nano [15], prompted with Motivational-Interviewing principles (OARS: Open questions, Affirmations, Reflective listening, Summaries) [5] and an explicit instruction to avoid clinical terminology, so that the persona's replies remain conversational rather than self-diagnostic. Our submitted runs use a short turn budget, and the official interaction profile records an average of 6.2 messages per conversation. Mind that one consequence of the no-clinical-language constraint is that persona replies are typically sparse on direct symptom statements, which raises the inferential demand on the assessors, a point we set up in Section 3.2.

Per-turn analysis of the persona reply is carried out by four domain-specialised assessors, each implemented as Llama-3.3-70B-Instruct with a cluster-specific prompt. The four clusters partition the 21 BDI-II items by content type: affective (items 1, 4, 10, 12, 17), cognitive (items 2, 3, 5, 6, 7, 8, 9, 14), somatic (items 11, 15, 16, 18, 20), and functional (items 13, 19, 21). Each assessor receives only the items relevant to its cluster, which keeps the prompt short and the scoring criteria targeted. Per-item strictness is set from the DepreSym base rates discussed in Section 3.2: items with very low DepreSym relevance (e.g., *Punishment Feelings* at $\approx$ 1.9%) receive a `VERY_STRICT` prompt that demands explicit evidence before any non-zero score, while higher-relevance items receive a less strict prompt. Each item returns one of three states: SCORED (with an integer 0–3 and an assessor confidence), NO_EVIDENCE, or EVIDENCE_OF_ABSENCE, preserving the inferential-state distinction motivated in Section 3.2. At end of conversation, the per-item scores are aggregated by a two-pass scorer. The first pass sums all SCORED items to produce a raw BDI-II total and ranks items by salience to yield the top-four active-symptoms set. A consensus check then runs across two independent signals: the linguistic-feature extractor, which computes the absolutist density [16] of the persona's transcripts, and a conditional Bayesian prior that combines DepreSym base rates with the persona's observed scoring profile. The second pass refines the raw total using these consensus signals. A Score Distribution Constraint (SDC) caps the number of score-3 items the assessors may collectively return, countering the expressive-LoRA artefact in which a moderate-severity persona produces maximally severe language and the assessors over-attribute.

4.2. Theory-of-Mind module and post-hoc correction

The Theory-of-Mind module tracks two profiles across the conversation: an *expressed profile* $E(t)$, the cumulative vector of per-item evidence the persona has so far disclosed, and an *interviewer profile* $I(t)$, the cumulative vector of items the interviewer has probed. From these it detects *coverage gaps* (items present in $I(t)$ but unmatched in $E(t)$), and it was specified to compute three Wasserstein quantities: $W_{\text{self}}(t)$, the turn-over-turn shift in E; $W_{\text{align}}(t)$, the alignment gap between $E(t)$ and $I(t)$; and $W_{\text{acc}}(t)$, a validation-only accuracy measure against ground truth. Mind that this module was enabled only in the tail of Run 3 (personas 13–20); for Runs 1 and 2 and the head of Run 3 it was disabled, and the orchestrator relied on the programmatic coverage tracking of Section 4.1 alone. Where the module did run, the Python Optimal Transport library [**?**] was absent from the submission environment, an uncaught dependency bug on our side: the serialised records carry `pot_available: false`, the Wasserstein routine exited early, and W_{self}, W_{align}, W_{acc}, and the transport plans were all left empty, with only the per-turn indices tracked. Thus no Wasserstein quantity was computed in any submitted run. The wiring is in any case diagnostic by design: across the codebase only W_{align} is read at a live decision surface, as a soft hint in the orchestrator's reasoning prompt that is guarded by a non-empty check and therefore skipped when the value is empty, while W_{self}, W_{acc}, and the transport plans are serialise-only. We reconstruct the three quantities post-hoc in Section 5.

Before submission, raw BDI-II totals pass through a post-hoc correction stage that we treat as a hedging variable across runs. The motivation is empirical: the assessors over-score on emotionally-loaded text (Section 3.2), and a coarse, run-level subtraction is a simple counter that requires no retraining. We submitted three correction strategies, chosen as three points on the same curve to hedge across an unknown evaluation weighting. The first is a *band-aware* correction whose subtraction

Table 1
Per-run configuration of the three submitted runs. Run 3 was changed mid-submission, hence the split between the two persona-id ranges.

Run	Correction on raw total	C1+C2
1	band-aware (−4 min/mild, −5 mod, −1 sev)	disabled
2	flat −2	disabled
3, pid 2–12 (head)	flat −3	disabled
3, pid 13–20 (tail)	none	enabled (net-zero)

depends on the raw band: −4 for raw totals in the minimal or mild range, −5 for the moderate range, and only −1 for the severe range, so that the severe end is shaved least (Run 1). The second is a uniform *flat* −2 subtraction (Run 2). The third is a *mixed* strategy: flat −3 for personas 2–12 and no correction at all for personas 13–20 (Run 3). In addition, the pipeline includes two corrections, named C1 and C2, that operate on the final per-item assessor scores rather than on the raw total. C1 is a confidence gate that drops scored items whose assessor confidence is below 0.5. C2 is a somatic boost that adds 9 (scaled by the fraction of somatic items with evidence) to the gated total when the somatic items are jointly under-covered *and* the gated total is already at least 20. Mind that neither C1 nor C2 reads any output of the Theory-of-Mind module. Earlier internal documents referred to C1+C2 as a "ToM calibration" path, but the label does not reflect the implementation, and we use C1+C2 corrections throughout this paper. We return in Section 5 to a consequence of C2's ≥ 20 activation threshold that we identified post-hoc.

We flag one implementation inconsistency that bears on C2, since it is a known limitation of the ToM correction path. The pipeline carries three separate definitions of the somatic item set: the assessor routing assigns the somatic cluster items {11, 15, 16, 18, 20}, the orchestrator's coverage tracker uses {15, 16, 18, 20, 21}, and the C2 boost uses the union {11, 15, 16, 18, 20, 21}. C2 counts somatic evidence over its own six-item set, reading final per-item scores irrespective of which assessor produced them, with the partial-coverage scaling denominator fixed at six. Two consequences follow. Item 21 (loss of sexual interest) is scored by the functional assessor yet counts as somatic evidence for C2, so a functional-cluster symptom can suppress the somatic-gap boost; and item 11 (agitation) is treated as somatic by C2 but as functional by the orchestrator's coverage tracker, so the two modules can disagree about whether the somatic domain is covered. The live impact is small, since item 21 is rarely scored in short interviews and C2 ran only on the Run 3 tail (Section 5), but the cross-cluster definition is a clear target for a future redesign that makes the three modules reference one canonical somatic set.

All three submitted runs share the same orchestrator, interviewer, four assessors, linguistic-feature extractor, justificator, two-pass scorer, and Score Distribution Constraint. They differ in the post-hoc correction stage and, for the Run 3 tail, in two further switches described below; Table 1 summarises the per-run configuration. We disclose openly that the Run 3 configuration was changed mid-submission: our original plan was a uniform flat −3 correction across all twenty personas, and partway through submission, from persona 13 onward, we switched to a configuration that disabled the flat subtraction, enabled the Theory-of-Mind module, and enabled the C1+C2 corrections. The head of Run 3 (personas 2–12) therefore received flat −3 on its raw totals with the ToM module and C1+C2 disabled, while the tail (personas 13–20) received no flat subtraction. Mind that the two switches enabled on the tail changed nothing numerically: C1+C2 moved no final score (for all eight tail personas the submitted total equals the raw assessor total), and the ToM module computed no Wasserstein values because the optimal-transport library was absent (Section 4.2). The tail is thus, in effect, the uncorrected raw assessor profile, with the ToM coverage-gap signal additionally active during the interview. We bring this forward because the head/tail comparison in Section 5 would otherwise be misread as a comparison of correction *kinds*, when in mechanism it is primarily a comparison of correction *presence*: the tail carries the raw assessor totals, while the head has been shifted down by 3 on top of the same assessor profile.

Table 2
Official preliminary metrics for our three runs (overview Table 4, incomplete submissions, 19/20 personas). Within-table rank in parentheses. Higher is better for all five metrics.

Run	DCHR	ADODL	ASHR	LDCHR	LASHR
1 (band-aware)	0.3158 (10)	0.8388 (19)	0.1579 (8)	0.2285 (8)	0.1142 (9)
2 (flat −2)	0.2632 (12)	0.8304 (21)	0.1447 (10)	0.1767 (15)	0.1013 (11)
3 (mixed)	0.2632 (12)	0.8471 (15)	**0.2237 (4)**	0.1826 (14)	**0.1525 (2)**

5. Results and Discussion

5.1. Official results and the test set

Our three runs covered 19 of the 20 personas (persona 1 was not submitted), so we appear in the incomplete-submission table of the overview (Table 4 there). Table 2 reports the five official metrics for each run. The pattern across runs is informative: Run 1 (band-aware) gives the best band classification, with DCHR 0.3158, while Run 3 gives the best symptom identification, with ASHR 0.2237 and a latency-weighted LASHR of 0.1525. That LASHR value ranks second among the incomplete-submission runs, our strongest single result; mind that this is a within-table position (the incomplete-submission ranks run to 28), and the best LASHR anywhere is pjmathematician's 0.2288 in the complete-submission table. For context, the complete-submission bests were split across teams rather than swept by one system: the best DCHR there was 0.6000 (BUAP-CRC), the best ADODL 0.9254 (pjmathematician), and the best ASHR 0.3375 (upb-uottawa). Our interaction profile averaged 6.2 messages per conversation at 156.96 characters per message, which is concise in turns relative to most teams and is what keeps our latency-weighted scores close to their unweighted counterparts.

Following the release of the gold labels, we summarise the test cohort here; this information was not available during the campaign. The 20 personas are split into exactly five per severity band (minimal, mild, moderate, severe), so the set is perfectly balanced and a system cannot gain DCHR from a majority-class prior. The gold totals range from 5 to 40 (mean 20.45, median 19), which means the severe band is populated by reachable cases (33 to 40) rather than by extreme ones: no persona sits near the top of the 0–63 range. Eighteen of the 21 BDI-II symptoms appear as a key symptom for at least one persona; the three that never appear are *Punishment Feelings*, *Suicidal Thoughts or Wishes*, and *Loss of Interest in Sex*. By mean BDI-II total of the personas carrying them, the strongest severity markers are *Crying* (32.7), *Sadness* (29.0), and *Worthlessness* (26.5), while *Loss of Interest* is the most frequent symptom (eight personas) at a mid-range mean of 22.3. Mind that the balance of the set, together with the absence of maximal-severity personas, makes the severe band the decisive one: a system that never reaches the severe band, as ours does not, forfeits a quarter of DCHR before any other error.

We report one discrepancy in the interest of reproducibility. Our local re-scoring of the submitted predictions gives systematically higher figures than the official preliminary results across all three primary metrics, with the largest gap exceeding 0.2 DCHR on Run 3 (local 0.526 against the official 0.263). We verified that the submitted prediction files are byte-identical to our per-batch backups, so the gap does not come from a post-submission overwrite, but we have not been able to reconcile the two figures. We treat the official figures as authoritative throughout this section and use our local re-scoring only as a per-persona diagnostic in what follows.

5.2. Ablations: correction strategy matters more than pipeline depth

We ablated the pipeline on the TalkDep development personas [**?**]. Mind that TalkDep uses the textbook BDI-II band boundaries (0–13, 14–19, 20–28, 29+) rather than the eRisk boundaries, so the DCHR values in this subsection are not comparable to the official results above and are read only within the ablation. Adding pipeline depth on top of the bare assessors, before any post-hoc correction, does not help: the baseline that simply sums the SCORED items reaches DCHR 0.583, whereas the full pipeline with the

justificator active drops to 0.333, the worst configuration tested. The mechanism is over-adjustment. The justificator's coherence edits and the Bayesian prior both push totals upward, which compounds the assessors' existing tendency to over-score, and the configuration that removes the prior recovers most of the lost ground. We note as a caveat that replicate runs of the same configuration at temperature 0.1 differ by up to ±0.083 DCHR, so within-ablation rankings of nearby configurations are not individually reliable; the comparison we draw on here is the large gap between the bare baseline and the full stack, not a fine ordering.

Against that picture, the post-hoc correction is what carries the band accuracy. Applying the band-aware correction to the bare baseline raises DCHR from 0.583 to 0.750, a gain of 0.167 that matches in magnitude the cost of the full feature stack, and it lifts ADODL from 0.923 to 0.950. The three submitted strategies were chosen as three points on this correction curve to hedge across an unknown evaluation weighting: band-aware for closeness (best ADODL on TalkDep), flat -2 for a balanced trade-off, and flat -3 for a more conservative shift. The Score Distribution Constraint matches band-aware on DCHR but at a higher mean absolute deviation, and combining the two over-corrects, which is why the SDC is not stacked with band-aware in any submitted run.

5.3. Diagnostic analyses: corrections, Wasserstein retrofit, and error modes

The C1 and C2 corrections, enabled only on the Run 3 tail, changed almost nothing. As noted in Section 4.2, the submitted total equals the raw assessor total for all eight tail personas: C1's confidence gate dropped no decisive item and C2's somatic boost never fired. The reason C2 never fired is structural, and it is worth stating plainly because it is the kind of design fault that is invisible until the gold labels arrive. C2 was meant to rescue severe personas who under-disclose somatic content, yet it activates only when the gated total is already at least 20. The severe personas it targets do not clear that threshold, precisely because of the assessor under-scoring that C2 was designed to compensate for: Marco (gold 38) gates to 18, and Maria (gold 40) gates to 14, both below the activation point. The rescue mechanism is therefore excluded from the very cases it was built for. A secondary barrier compounds this, namely that C2 also requires somatic evidence to be absent, whereas every persona expressed three or four of the six somatic items C2 counts.

We turn to the Wasserstein quantities. Since the optimal-transport library was absent at submission (Section 4.2), none of these values were computed live, so the analysis here is a post-hoc reconstruction from the serialised per-turn assessor profiles. We normalise each expressed profile $E(t)$ to unit mass before measuring distances, so the metric captures the *shape* of the disclosed symptom distribution rather than its growing total, and we use the clinical item-cost ground metric of Section 3.2. The reconstruction exposes a structural limitation of the detector as specified. Under the eRisk-native convention, W_{self} measures the turn-over-turn shift in the expressed profile, $W_1(E(t), E(t-1))$, and a shift is flagged when it exceeds the running mean plus two standard deviations estimated over the preceding turns. On the four moderate and severe personas where a clear mid-conversation disclosure occurs (Laura, Linda, Marco, Maria), this detector never fires, because the disclosure is front-loaded into the second and third turns, inside the window that seeds the running baseline, so the baseline is inflated by the very shift it should later flag and the post-warmup trajectory decays beneath it. A barycenter convention, which compares each turn's profile to the running-mean profile $W_1(E(t), \bar{E}(1:t))$ rather than to the immediately preceding turn, recovers the signal.

Table 3 makes the contrast concrete on Marco (gold 38), our clearest case of monotonic disclosure across seven turns. Rounds 1 and 2 are purely somatic (energy, sleep, fatigue); round 3 adds sadness and pessimism. Round 4 is the decisive disclosure, where Marco moves from somatic complaint and grief to anhedonia, withdrawal, and worthlessness. The barycenter distance at round 4 is 0.432 against a threshold of 0.425, so the detector fires, just. The eRisk-native W_{self} at the same round is 0.410, but its threshold has been inflated to 0.851 by the large turn-to-turn jumps of the warmup rounds, so it does not fire then or at any later round. Mind that this barycenter firing is marginal by construction: the same early disclosure that we want to detect also raises the running baseline, so a genuine mid-conversation shift only barely clears $\mu + 2\sigma$. Across the other personas the reconstruction is not uniformly clean:

Table 3
Post-hoc Wasserstein retrofit for Marco (persona 15, gold BDI-II 38), comparing the barycenter convention $W_1(E(t), \bar{E}(1:t))$ with the eRisk-native $W_{\text{self}} = W_1(E(t), E(t-1))$. Each profile is L1-normalised. The firing rule is $W_1 > \mu + 2\sigma$ over the strictly preceding rounds, with the first three rounds treated as warmup. Only the barycenter convention fires, at round 4, on the anhedonia/worthlessness disclosure.

		Barycenter $W_1(E(t), \bar{E})$			eRisk-native W_{self}		
t	New / increased symptoms	W_1	thr.	fire	W_1	thr.	fire
1	Energy, Sleep, Fatigue	0.000	–		0.000	–	
2	(somatic, reinforced)	0.000	–		0.000	–	
3	Sadness, Pessimism	0.333	0.000		0.667	0.000	
4	Loss of pleasure, Self-dislike, Loss of interest, Worthlessness	0.432	0.425	☒	0.410	0.851	
5	Loss of interest, Indecisiveness	0.399	0.580		0.173	0.837	
6	Loss of pleasure, Concentration	0.239	0.618		0.121	0.764	
7	Past failure, Self-criticalness, Worthlessness	0.256	0.586		0.177	0.707	

the only other clear barycenter firing is Linda's cognitive disclosure, and the strongest firing of all is a false positive on Javier, driven by assessor mass oscillating between turns rather than by any genuine disclosure. The lesson for a future edition is that the warmup-baseline rule is ill-suited to conversations of five to seven turns, and would need redesign even with the library in place; the per-persona reconstructions are collected in Appendix B.

The dominant error mode is severe under-prediction, and it is universal: across all three runs, none of the five severe personas was placed in the severe band. Every severe persona was scored at most 22 against gold totals from 33 to 40 (Daniel 33, read 17/18/19 across the three runs; Elena 35, read 20/20/21; Carlos 37, read 22/21/21; Marco 38, read 20/20/19; Maria 40, read 18/19/18), a shortfall of twelve to eighteen points. Three factors combine to produce it: the assessors under-score items whose evidence is implicit rather than explicitly stated, which is exactly the high-Theory-of-Mind-demand case set out in Section 3.2; the C2 rescue that was meant to catch these cases cannot fire, for the reason just given; and the corrections act in the wrong direction at the severe end, where the band-aware -1 is too small to matter and the flat -3 actively pushes a severe persona further from its band. Because the test set places five personas in the severe band, this single failure caps DCHR at 0.75 before any other error is counted, and it accounts for most of the distance between our DCHR and the field's best.

Finally, we read the Run 3 head-and-tail split with care, since it is tempting to treat it as a controlled comparison and it is not. The head (personas 2–12) received the flat -3 correction; the tail (personas 13–20) received no correction at all, with the ToM module and the inert C1+C2 additionally switched on (Section 4.2). On disjoint persona halves, the head reached DCHR 0.636 and the tail 0.375, but the halves are not matched: the tail contains all the hardest cases, including the severe personas that no configuration scores correctly, so the difference reflects persona composition as much as treatment. What can be said cleanly is narrower. Where a moderate persona's raw total already sat near its gold band, subtracting nothing left it closer than subtracting three would have: Laura (gold 23, raw near 24) and Linda (gold 28, raw near 21) land in band on the uncorrected tail. This is an argument about correction presence, not about the Theory-of-Mind signal, which computed nothing and, through the coverage-gap path, at most nudged the interview. We therefore cannot attribute Run 3's symptom-identification result to the ToM machinery on the strength of this comparison, and we do not claim to.

6. Limitations

Several limitations bound the conclusions we draw. Our submission covered 19 of the 20 personas, since persona 1 was not submitted, so our results appear in the incomplete-submission table rather than in the directly-comparable complete-submission one. The ablations in Section 5.2 were run on

the TalkDep development personas under the textbook BDI-II band boundaries, which differ from the eRisk boundaries, so their DCHR values are not comparable to the official results and support only within-ablation comparisons. Those ablations are themselves small-sample: replicate runs of one configuration at temperature 0.1 differ by up to ±0.083 DCHR, so we rely only on the large gaps between configurations and not on fine rankings. The Theory-of-Mind machinery was never exercised live, because the optimal-transport library was absent from the submission environment (Section 4.2); the Wasserstein analysis of Section 5.3 is therefore a post-hoc reconstruction rather than a record of the deployed system. We also report, without having resolved, a systematic discrepancy above 0.2 DCHR between our local re-scoring and the official preliminary figures (Section 5.1), and we treat the official figures as authoritative throughout.

Relatedly, all analyses in Section 5.3 are diagnostic on the submitted runs: we have not re-run the pipeline against the released gold labels. Three re-runs would directly address the gaps this leaves. First, submitting the one missing persona would restore full 20/20 coverage and place our results in the complete-submission table. Second, running both Run 3 configurations, with the flat correction and with the Theory-of-Mind path, over the full persona pool rather than over disjoint halves would separate correction presence from Theory-of-Mind enablement, namely the confound we flag in Section 5.3. Third, a full Theory-of-Mind pipeline with the optimal-transport library installed would, for the first time, test the Wasserstein machinery live rather than in retrofit. We intend to report these re-runs in an extended version of this work; until then, the redesigns our diagnoses motivate remain unvalidated.

7. Conclusion

We presented a multi-agent pipeline for conversational depression detection that pairs an OARS-guided interviewer with four domain-specialised BDI-II assessors, DepreSym-calibrated prompts, a coverage-tracking Theory-of-Mind module, and a post-hoc band-correction stage. Over 19 of the 20 personas, the system reached a latency-weighted symptom hit rate of 0.1525, second among the incomplete-submission runs, and an unweighted symptom hit rate of 0.2237. Mind that the more useful lessons from this participation are the negative ones. Severe under-prediction was universal and is where most of the accuracy was lost: no severe persona reached the severe band, and because the test set is a quarter severe, this single failure caps band accuracy before any other error, so it is the first thing a future system must fix. The somatic-boost correction meant to rescue these cases cannot fire on them, since its activation threshold excludes the very totals it should lift, and it needs to be redesigned without a gating condition that depends on the under-scored total. The Wasserstein detector, beyond having been left uncomputed at submission by a dependency bug, is structurally ill-suited to short conversations under the turn-over-turn convention, and a barycenter formulation is the more promising basis for a future version. Each of these is a concrete target rather than a vague direction, and we intend to address them, together with the controlled re-runs set out in Section 6, in subsequent work.

8. Ethical Considerations

General ethical considerations. This work involves no real patients and no human-subjects data. The twenty personas are simulated language models, organiser-provided LoRA adapters on Llama-3-8B-Instruct, so there is no clinical population at risk in our experiments and no informed-consent question to address. We note one point of caution for any reading beyond this simulated setting. The released test cohort does not exercise the BDI-II suicidal-ideation item (item 9): it appears as a key symptom for none of the twenty personas (Section 5.1), so our system was never evaluated on safety-critical signal detection, and a system intended for real use would need that capability assessed separately and deliberately. We report our results as research findings on a simulated cohort, and we make no claim of clinical utility and no claim that the system is fit for screening real people.

EU AI Act considerations. We record how this work sits with respect to the EU AI Act [17], as considerations rather than as compliance claims. The work reported here falls within the Act's scientific-research scope: Article 2(6) exempts AI systems and models developed and put into service solely for scientific research and development, which is the setting of this paper. A hypothetical deployment of such a system for real depression screening would be a different matter, and would implicate at least two parts of the Act. As a system used in access to essential health services, it would fall under the high-risk regime of Annex III; under the Act's original schedule the corresponding obligations apply from 2 August 2026, although the Digital Omnibus package proposed by the Commission, still provisional at the time of writing, would defer the stand-alone Annex III obligations to 2 December 2027. Independently of that deferral, the Article 50 transparency duty, namely that a person be informed when they are interacting with an AI system rather than a human, applies from 2 August 2026 and would bear directly on a conversational screening agent of the kind studied here. The foundation-model providers we build on (OpenAI for GPT-5-nano, Meta for the Llama models) carry their own general-purpose-AI obligations under the Act, whose provisions for such models took effect on 2 August 2025. We do not place any system on the market, and the above is intended to situate the work rather than to assert conformity.

Declaration on Generative AI

During the preparation of this work, the authors used Anthropic Claude (Opus 4.7) for the following purposes: *drafting content* of some paragraphs based on the detailed outline; editing content including *paraphrasing and rewording*, and structural editing; and generation of portions of the pipeline and analysis code described in Sections 4 and 5. The language models that are themselves objects of study in this paper (GPT-5-nano, Llama-3.3-70B-Instruct-Turbo, and Llama-3-8B-Instruct) are described and cited in Section 4 and are not covered by this declaration. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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, ACM, Seoul Republic of Korea, 2025, pp. 6554–6558. URL: https://dl.acm.org/doi/10.1145/3746252.3761617. doi:10.1145/3746252.3761617.

[2] A. T. Beck, R. A. Steer, G. K. Brown, Beck depression inventory: BDI-II, 2. ed ed., Pearson, San Antonio, 1996.

[3] 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.

[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 (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, Jena, Germany, 2026.

[5] W. R. Miller, S. Rollnick, Motivational interviewing: helping people change, Applications of motivational interviewing, third edition ed., Guilford Press, New York, NY London, 2013.

[6] A. Pérez, M. Fernández-Pichel, J. Parapar, D. E. Losada, DepreSym: A Depression Symptom Annotated Corpus and the Role of Large Language Models as Assessors of Psychological Markers, Language Resources and Evaluation 59 (2025) 2737–2762. URL: https://link.springer.com/10.1007/s10579-025-09831-6. doi:10.1007/s10579-025-09831-6.

[7] J. Parapar, A. Perez, X. Wang, F. Crestani, eRisk 2025: Contextual and Conversational Approaches for Depression Challenges, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, volume 15576, Springer Nature Switzerland, Cham, 2025, pp. 416–424. URL: https://link.springer.com/10.1007/978-3-031-88720-8_62. doi:10.1007/978-3-031-88720-8_62.

[8] F. Crestani, D. E. Losada, J. Parapar (Eds.), Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the eRisk Project, volume 1018 of *Studies in Computational Intelligence*, Springer International Publishing, Cham, 2022. URL: https://link.springer.com/10.1007/978-3-031-04431-1. doi:10.1007/978-3-031-04431-1.

[9] 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, 2021. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

[10] A. Miyaguchi, D. Guecha, Y. Chiu, 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 the Conference and Labs of the Evaluation Forum, {CLEF} 2025, Madrid, Spain, 9-12 September 2025, {CEUR} Workshop Proceedings, CEUR-WS.org, 2025, pp. 1590–1610. URL: https://ceur-ws.org/Vol-4038/paper_128.pdf. doi:10.48550/arXiv.2507.10958, arXiv:2507.10958.

[11] A. M. Marmol-Romero, M. Garcia-Vega, M. A. Garcia-Cumbreras, A. Montejo-Raez, SINAI at eRisk@CLEF 2025: Transformer-Based and Conversational Strategies for Depression Detection, in: 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, pp. 1620–1635. URL: https://ceur-ws.org/Vol-4038/paper_130.pdf. doi:10.48550/arXiv.2509.19861, arXiv:2509.19861.

[12] M. Sap, R. Le Bras, D. Fried, Y. Choi, Neural Theory-of-Mind? On the Limits of Social Intelligence in Large LMs, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3762–3780. URL: https://aclanthology.org/2022.emnlp-main.248. doi:10.18653/v1/2022.emnlp-main.248.

[13] C. Villani, Optimal Transport, volume 338 of *Grundlehren der mathematischen Wissenschaften*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. URL: http://link.springer.com/10.1007/978-3-540-71050-9. doi:10.1007/978-3-540-71050-9.

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. v. d. Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. v. d. Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. d. Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot,

S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The Llama 3 Herd of Models, 2024. URL: http://arxiv.org/abs/2407.21783. doi:10.48550/arXiv.2407.21783, arXiv:2407.21783.

[15] A. Singh, A. Fry, A. Perelman, A. Tart, A. Ganesh, A. El-Kishky, A. McLaughlin, A. Low, A. J. Ostrow, A. Ananthram, A. Nathan, A. Luo, A. Helyar, A. Madry, A. Efremov, A. Spyra, A. Baker-Whitcomb, A. Beutel, A. Karpenko, A. Makelov, A. Neitz, A. Wei, A. Barr, A. Kirchmeyer, A. Ivanov, A. Christakis, A. Gillespie, A. Tam, A. Bennett, A. Wan, A. Huang, A. M. Sandjideh, A. Yang, A. Kumar, A. Saraiva, A. Vallone, A. Gheorghe, A. G. Garcia, A. Braunstein, A. Liu, A. Schmidt, A. Mereskin, A. Mishchenko, A. Applebaum, A. Rogerson, A. Rajan, A. Wei, A. Kotha, A. Srivastava, A. Agrawal, A. Vijayvergiya, A. Tyra, A. Nair, A. Nayak, B. Eggers, B. Ji, B. Hoover, B. Chen, B. Chen, B. Barak, B. Minaiev, B. Hao, B. Baker, B. Lightcap, B. McKinzie, B. Wang, B. Quinn, B. Fioca, B. Hsu, B. Yang, B. Yu, B. Zhang, B. Brenner, C. R. Zetino, C. Raymond, C. Lugaresi, C. Paz, C. Hudson, C. Whitney, C. Li, C. Chen, C. Cole, C. Voss, C. Ding, C. Shen, C. Huang,

C. Colby, C. Hallacy, C. Koch, C. Lu, C. Kaplan, C. Kim, C. J. Minott-Henriques, C. Frey, C. Yu, C. Czarnecki, C. Reid, C. Wei, C. Decareaux, C. Scheau, C. Zhang, C. Forbes, D. Tang, D. Goldberg, D. Roberts, D. Palmie, D. Kappler, D. Levine, D. Wright, D. Leo, D. Lin, D. Robinson, D. Grabb, D. Chen, D. Lim, D. Salama, D. Bhattacharjee, D. Tsipras, D. Li, D. Yu, D. J. Strouse, D. Williams, D. Hunn, E. Bayes, E. Arbus, E. Akyurek, E. Y. Le, E. Widmann, E. Yani, E. Proehl, E. Sert, E. Cheung, E. Schwartz, E. Han, E. Jiang, E. Mitchell, E. Sigler, E. Wallace, E. Ritter, E. Kavanaugh, E. Mays, E. Nikishin, F. Li, F. P. Such, F. d. A. B. Peres, F. Raso, F. Bekerman, F. Tsimpourlas, F. Chantzis, F. Song, F. Zhang, G. Raila, G. McGrath, G. Briggs, G. Yang, G. Parascandolo, G. Chabot, G. Kim, G. Zhao, G. Valiant, G. Leclerc, H. Salman, H. Wang, H. Sheng, H. Jiang, H. Wang, H. Jin, H. Sikchi, H. Schmidt, H. Aspegren, H. Chen, H. Qiu, H. Lightman, I. Covert, I. Kivlichan, I. Silber, I. Sohl, I. Hammoud, I. Clavera, I. Lan, I. Akkaya, I. Kostrikov, I. Kofman, I. Etinger, I. Singal, J. Hehir, J. Huh, J. Pan, J. Wilczynski, J. Pachocki, J. Lee, J. Quinn, J. Kiros, J. Kalra, J. Samaroo, J. Wang, J. Wolfe, J. Chen, J. Wang, J. Harb, J. Han, J. Wang, J. Zhao, J. Chen, J. Yang, J. Tworek, J. Chand, J. Landon, J. Liang, J. Lin, J. Liu, J. Wang, J. Tang, J. Yin, J. Jang, J. Morris, J. Flynn, J. Ferstad, J. Heidecke, J. Fishbein, J. Hallman, J. Grant, J. Chien, J. Gordon, J. Park, J. Liss, J. Kraaijeveld, J. Guay, J. Mo, J. Lawson, J. McGrath, J. Vendrow, J. Jiao, J. Lee, J. Steele, J. Wang, J. Mao, K. Chen, K. Hayashi, K. Xiao, K. Salahi, K. Wu, K. Sekhri, K. Sharma, K. Singhal, K. Li, K. Nguyen, K. Gu-Lemberg, K. King, K. Liu, K. Stone, K. Yu, K. Ying, K. Georgiev, K. Lim, K. Tirumala, K. Miller, L. Ahmad, L. Lv, L. Clare, L. Fauconnet, L. Itow, L. Yang, L. Romaniuk, L. Anise, L. Byron, L. Pathak, L. Maksin, L. Lo, L. Ho, L. Jing, L. Wu, L. Xiong, L. Mamitsuka, L. Yang, L. McCallum, L. Held, L. Bourgeois, L. Engstrom, L. Kuhn, L. Feuvrier, L. Zhang, L. Switzer, L. Kondraciuk, L. Kaiser, M. Joglekar, M. Singh, M. Shah, M. Stratta, M. Williams, M. Chen, M. Sun, M. Cayton, M. Li, M. Zhang, M. Aljubeh, M. Nichols, M. Haines, M. Schwarzer, M. Gupta, M. Shah, M. Y. Guan, M. Huang, M. Dong, M. Wang, M. Glaese, M. Carroll, M. Lampe, M. Malek, M. Sharman, M. Zhang, M. Wang, M. Pokrass, M. Florian, M. Pavlov, M. Wang, M. Chen, M. Wang, M. Feng, M. Bavarian, M. Lin, M. Abdool, M. Rohaninejad, N. Soto, N. Staudacher, N. LaFontaine, N. Marwell, N. Liu, N. Preston, N. Turley, N. Ansman, N. Blades, N. Pancha, N. Mikhaylin, N. Felix, N. Handa, N. Rai, N. Keskar, N. Brown, O. Nachum, O. Boiko, O. Murk, O. Watkins, O. Gleeson, P. Mishkin, P. Lesiewicz, P. Baltescu, P. Belov, P. Zhokhov, P. Pronin, P. Guo, P. Thacker, Q. Liu, Q. Yuan, Q. Liu, R. Dias, R. Puckett, R. Arora, R. T. Mullapudi, R. Gaon, R. Miyara, R. Song, R. Aggarwal, R. J. Marsan, R. Yemiru, R. Xiong, R. Kshirsagar, R. Nuttall, R. Tsiupa, R. Eldan, R. Wang, R. James, R. Ziv, R. Shu, R. Nigmatullin, S. Jain, S. Talaie, S. Altman, S. Arnesen, S. Toizer, S. Toyer, S. Miserendino, S. Agarwal, S. Yoo, S. Heon, S. Ethersmith, S. Grove, S. Taylor, S. Bubeck, S. Banesiu, S. Amdo, S. Zhao, S. Wu, S. Santurkar, S. Zhao, S. R. Chaudhuri, S. Krishnaswamy, Shuaiqi, Xia, S. Cheng, S. Anadkat, S. P. Fishman, S. Tobin, S. Fu, S. Jain, S. Mei, S. Egoian, S. Kim, S. Golden, S. Q. Mah, S. Lin, S. Imm, S. Sharpe, S. Yadlowsky, S. Choudhry, S. Eum, S. Sanjeev, T. Khan, T. Stramer, T. Wang, T. Xin, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Degry, T. Shadwell, T. Fu, T. Gao, T. Garipov, T. Sriskandarajah, T. Sherbakov, T. Korbak, T. Kaftan, T. Hiratsuka, T. Wang, T. Song, T. Zhao, T. Peterson, V. Kharitonov, V. Chernova, V. Kosaraju, V. Kuo, V. Pong, V. Verma, V. Petrov, W. Jiang, W. Zhang, W. Zhou, W. Xie, W. Zhan, W. McCabe, W. DePue, W. Ellsworth, W. Bain, W. Thompson, X. Chen, X. Qi, X. Xiang, X. Shi, Y. Dubois, Y. Yu, Y. Khakbaz, Y. Wu, Y. Qian, Y. T. Lee, Y. Chen, Y. Zhang, Y. Xiong, Y. Tian, Y. Cha, Y. Bai, Y. Yang, Y. Yuan, Y. Li, Y. Zhang, Y. Yang, Y. Jin, Y. Jiang, Y. Wang, Y. Wang, Y. Liu, Z. Stubenvoll, Z. Dou, Z. Wu, Z. Wang, OpenAI GPT-5 System Card, 2026. URL: http://arxiv.org/abs/2601.03267. doi:10.48550/arXiv.2601.03267, arXiv:2601.03267.

[16] M. Al-Mosaiwi, T. Johnstone, In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation, Clinical Psychological Science 6 (2018) 529–542. URL: https://journals.sagepub.com/doi/10.1177/2167702617747074. doi:10.1177/2167702617747074.

[17] E. P. a. C. of the European Union, Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

A. Prompts

For reproducibility we reproduce the prompts driving the four LLM-backed components of the pipeline. The exact strings are also included in the released repository; where a prompt is parameterised (for example by the BDI-II item under assessment or by the DepreSym strictness tier), we show the template with its parameter slots marked. One redaction applies throughout. The assessor prompts (Section A.3) load, at each item's calibration slot, true-positive and hard-negative example sentences drawn verbatim from the DepreSym corpus [6]; this appendix shows the slot positions but not the sentences themselves, so that we do not redistribute DepreSym corpus material. The full prompt strings, including these example sentences, are in the released code.

A.1. Interviewer (GPT-5-nano)

The interviewer is prompted to follow Motivational-Interviewing principles (OARS) and to avoid clinical terminology, and it receives orchestrator guidance naming the next topic area and exploration angle.

```
You are a warm, empathetic conversational partner participating in a wellbeing study. Your
 ↪  role is to have a natural, supportive conversation with someone about their daily
 ↪  life, feelings, and experiences. You are NOT a therapist, NOT a counsellor, NOT a
 ↪  mental health professional. You are a good listener who is genuinely curious about how
 ↪  the other person is doing.

## Your conversational style

CORE TECHNIQUE: OARS
- Open-ended questions: at least 70% of your questions should be open-ended. Avoid yes/no
 ↪  questions.
- Affirmations: acknowledge what the person shares without judgment. "That makes sense."
 ↪  "I can see why that would be tough."
- Reflective listening: paraphrase what they said to show you heard them before moving on.
 ↪  "So it sounds like..."
- Summaries: periodically pull together what you've heard. "From what you've told me..."

TRANSITIONS: Always use something from their previous response as a bridge to the next
 ↪  topic. Never switch topics abruptly.

FUNNEL APPROACH: Start each topic with a broad opener. Go specific only when you detect a
 ↪  signal worth exploring.

NORMALISATION: When approaching sensitive topics, frame them as common human experiences.

## Handling different response patterns

### When the person gives short, flat answers
- Do NOT ask multiple follow-up questions. This feels interrogative.
- Do NOT offer solutions or try to help.
- DO use brief reflections: "That sounds really hard."
- DO ask simple, concrete questions: "What did today look like for you?"
- DO accept short answers and move to the next topic naturally.

### When the person hedges and deflects
- Do NOT take "I could try" at face value and move on.
- DO reflect the hedge: "You said maybe — it sounds like you're not sure if it would
 ↪  actually help?"
- DO name the deflection warmly: "You mentioned not wanting to bother people. That's a lot
 ↪  to carry on your own."

### When the person dismisses their own concerns
```

```
- Do NOT collude with the dismissal.
- DO gently hold the space: "Maybe everyone does, but I'm curious about your experience
 ↪  specifically."

### When the person is positive and engaged
- Do NOT keep probing for problems that aren't there.
- DO ask about mild challenges without pathologising.
- DO still cover all topic areas to ensure absence of symptoms is documented.

## Rules

- NEVER use these words: depression, depressed, mental health, therapy, therapist,
 ↪  diagnosis, symptoms, disorder, clinical, psychiatric, screening, assessment, BDI, PHQ,
 ↪  questionnaire, suicidal, suicide
- NEVER ask multiple questions in one message
- Keep messages SHORT: 2-4 sentences maximum
- NEVER offer clinical advice, coping strategies, breathing exercises, or self-help
 ↪  suggestions
- If the person seems uncomfortable with a topic, acknowledge it warmly and pivot
- If the person gives the same deflective answer twice on the same topic, move on
- NEVER say "I understand" — say "That makes sense" or "I can see that" instead

## Conversation structure

You will receive guidance about which topic area to explore next.

Topic areas:
1. EMOTIONAL_STATE — How they've been feeling overall
2. ACTIVITIES_INTERESTS — What they do for fun, hobbies, socialising
3. DAILY_ROUTINE — Walk through a typical day (sleep, meals, energy)
4. SELF_PERCEPTION — How they feel about themselves, confidence
5. FUTURE_OUTLOOK — How they see things going, hopes, plans
6. DECISION_MAKING — Ability to focus, make choices, handle tasks
7. ADAPTIVE_FOLLOWUP — Deeper exploration of a specific area

For each topic, spend 1-2 turns. If the person reveals something significant, explore for
 ↪  up to 3 turns before moving on.

## Input format

You receive:
- Full conversation history
- Orchestrator guidance: {next_topic, exploration_gaps, suggested_angle, turn_number,
 ↪  max_turns}

## Output format

Produce ONLY the message to send to the persona. No metadata, no reasoning, no brackets.
 ↪  Just the conversational message.
```

A.2. Orchestrator reasoning module (Llama-3.3-70B-Instruct-Turbo)

The reasoning module reads the coverage state and the recent transcript and selects the next BDI-II item to probe; it consumes the coverage-gap diagnostics described in Section 4.2.

```
You are the strategic reasoning component of a depression assessment system. After each
 ↪  conversation turn, you receive aggregated assessor outputs, linguistic features, and
 ↪  severity estimates, and must decide what the interviewer should explore next.
```

```
## Your responsibilities

1. IDENTIFY GAPS: Which BDI-II items are in NO_EVIDENCE state?
2. PRIORITISE by remaining turns:
   a) High-discrimination items first: Sadness(1), Pessimism(2), Loss of Pleasure(4),
    ↪  Self-Dislike(7), Self-Criticalness(8), Worthlessness(14), Suicidal Thoughts(9)
   b) Items with conflicting signals
   c) Items where partial evidence would change the band
3. RESOLVE CONFLICTS between assessor scores and linguistic features
4. DETECT SEVERITY DIVERGENCE between assessor band, absolutist band, and engagement band
5. GUIDE INTERVIEWER with specific, actionable guidance

## Output format

Respond ONLY with valid JSON:
{
  "decision": "CONTINUE" or "TERMINATE",
  "next_topic": "<topic area>",
  "suggested_angle": "<specific guidance for interviewer>",
  "exploration_gaps": ["<gap1>", "<gap2>"],
  "priority_reasoning": "<why this topic next>",
  "conflict_notes": "<any conflicts detected>",
  "interviewer_adaptation": "<advice on adapting to persona style>"
}

## Domain coverage awareness

The input includes `uncovered_bdi_domains` — a list of BDI domain categories
(COGNITIVE, AFFECTIVE, SOMATIC, FUNCTIONAL) with zero evidence so far.

If SOMATIC is uncovered by turn 3-4, you MUST prioritise a DAILY_ROUTINE
question targeting sleep, energy, and appetite. Conversations that miss
somatic items consistently produce incomplete assessments.

If ANY domain still has zero coverage by turn 4-5, generate a targeted
probe for that domain before the conversation ends.

## Key principles
- NEVER recommend clinical or mental health language
- If remaining turns are few, prioritise high-discrimination items
- If persona is disengaging, recommend lighter topics first
- When severity bands diverge, this is the MOST IMPORTANT signal to investigate
```

A.3. Domain-specialised assessors (Llama-3.3-70B-Instruct-Turbo)

Each of the four cluster assessors (affective, cognitive, somatic, functional) receives only its cluster's items and a per-item strictness tier derived from DepreSym relevance rates: VERY_STRICT ($< 7\%$ relevance), STRICT (7–12%), MODERATE (12–20%), and CLEARER ($\geq 20\%$). For example, *Punishment Feelings* (DepreSym relevance $\approx 1.9\%$) is assigned VERY_STRICT, whereas higher-relevance items receive a more permissive tier. At inference time, each {{calibration_*i*}} slot is filled with the item's relevance rate, the tier wording, and three true-positive and three hard-negative example sentences drawn verbatim from the DepreSym expert-annotated consensus. In the appendix below we show the calibration block (§A.3.5) with these example sentences replaced by placeholders, so that this paper does not redistribute DepreSym corpus sentences; the released code contains the full prompt strings. The shared preamble and the four cluster prompts follow; the per-item descriptions show the BDI-II anchors as given to the assessor.

The four cluster prompts share a common preamble, shown first, which is injected ahead of the

cluster-specific item descriptions at the {{shared_preamble}} slot.

Shared preamble.

```
## CRITICAL: Over-scoring bias correction

Our system has a validated +5 to +9 point over-scoring bias. The primary
cause: CONFLATING MENTION WITH CLINICAL RELEVANCE. DepreSym data shows
that most sentences mentioning a symptom are NOT clinically relevant.
For example, only 9% of sentences containing pessimism-related language
are actually relevant to BDI-II Pessimism.

YOU MUST apply these strictness rules:
- If the symptom's relevance rate is below 10%, DEFAULT to score 0 and
  only score higher with strong, specific, first-person evidence.
- Mentions of symptoms in the THIRD PERSON or HYPOTHETICAL do NOT count.
- General complaints without personal specificity do NOT count.
- The person must be describing THEIR OWN current/recent experience.

## Scoring protocol

For each of your assigned BDI-II items, produce one of three states:

1. SCORED (score 1-3, confidence > 0): You found conversational evidence that this symptom
 ↪  is present at the indicated severity.

2. EVIDENCE_OF_ABSENCE (score 0, confidence > 0.5): The person explicitly denied this
 ↪  symptom, or the conversation provided clear evidence it is not present.

3. NO_EVIDENCE (score null, confidence 0): The topic was not discussed or the evidence is
 ↪  too ambiguous to score. This is NOT the same as score 0.

## Scoring calibration (from validated TalkDep examples)

MINIMAL (BDI-II ~5): Hedges but copes. Solution-oriented. Self-aware. Active engagement.
 ↪  Key tell: hedging + coping.

MILD (BDI-II ~15): Direct but dismissive. No coping language. Irritability-dominant.
 ↪  Somatic focus. Key tell: direct reporting without emotional elaboration.

MODERATE (BDI-II ~28): Heavy hedging. Guilt-driven deflection. Absolutist creeping in. Key
 ↪  tell: hedging + guilt + "I don't deserve help" pattern.

SEVERE (BDI-II ~40): Saturated absolutism. Zero coping language. Flat rejection of hope.
 ↪  Global self-blame. Key tell: absolutist + no coping + global negative self-evaluation.

## Critical principle: response pattern IS data

The WAY the person responds is as diagnostic as WHAT they say:
- Response length: severe -> shorter, flatter
- Engagement level: severe -> answers the question but adds nothing
- Coping language: present in minimal, absent in mild-severe
- Hedging density: peaks in moderate
- Absolutist density: tracks severity linearly

## Output format

Respond ONLY with valid JSON, no other text:
{
```

```
  "items": [
    {
      "id": <item number>,
      "name": "<item name>",
      "score": <0-3 or null>,
      "state": "<SCORED|EVIDENCE_OF_ABSENCE|NO_EVIDENCE>",
      "confidence": <0.0-1.0>,
      "evidence": "<brief quote and reasoning>"
    }
  ],
  "cross_observations": "<patterns spanning multiple items>"
}
```

Affective assessor (items 1, 4, 10, 12, 17).

```
You are a specialised depression symptom assessor focused on AFFECTIVE symptoms. You
 ↪  analyse conversation transcripts to evaluate evidence for BDI-II items related to
 ↪  mood, emotional reactivity, pleasure, and interest.

{{shared_preamble}}

## Your BDI-II items

### Item 1: Sadness
0 = Does not feel sad
1 = Feels sad much of the time
2 = Sad all the time
3 = So sad or unhappy they can't stand it

Linguistic markers: sadness vocabulary (sad, unhappy, miserable, down, blue, low),
 ↪  absolutist temporal words for severity gradient. Score 1: "sometimes"/"much of the
 ↪  time". Score 2: "all the time"/"always". Score 3: unbearable framing.

SCORING GUIDELINES FOR SADNESS (Item 1):

This item measures the subjective experience of FEELING SAD or UNHAPPY.

DO score as sadness:
- Explicit statements: "I feel sad", "I'm unhappy", "I feel down"
- Crying or tearfulness (supports sadness, also score Item 10 separately)
- Descriptions of emotional pain, heartache, or grief
- Feeling "blue", "low", or "miserable"

DO NOT score as sadness (these map to OTHER BDI items):
- Emotional numbness, flatness, or blunting ("I feel numb", "I feel flat",
  "I feel nothing") — this maps to Item 4: Loss of Pleasure, or may indicate
  emotional blunting/depersonalization. Numbness is the ABSENCE of emotion,
  not the PRESENCE of sadness.
- "Walking through a fog" or "watching myself from outside" — this is
  depersonalization/dissociation, not sadness
- Feeling "stuck" or "running on empty" — these map to pessimism (Item 2)
  and fatigue (Items 15/20)
- Frustration or anger — these map to irritability (Item 17)

KEY DISTINCTION: A person can be deeply depressed and NOT feel sad — they may
feel numb or empty instead. If the person describes numbness/flatness without
any explicit sadness, score this item 0 and ensure the numbness is captured
under Item 4 (Loss of Pleasure) instead.
```

```
{{calibration_1}}

### Item 4: Loss of Pleasure (Anhedonia)
0 = Gets as much pleasure as ever
1 = Doesn't enjoy things as much
2 = Very little pleasure from previous activities
3 = Can't get any pleasure

Linguistic markers: discrepancy words (want, need, should, would, could), negated
 ↪  pleasure, temporal comparison ("used to love"), absence of activity language.

{{calibration_4}}

### Item 10: Crying
0 = Doesn't cry more than usual
1 = Cries more than usual
2 = Cries over every little thing
3 = Feels like crying but can't (emotional numbing — severe)

{{calibration_10}}

### Item 12: Loss of Interest
0 = Has not lost interest in people or activities
1 = Less interested than before
2 = Lost most interest
3 = Hard to get interested in anything

Distinguished from Item 4: interest/caring vs hedonic response. "I don't care" (interest)
 ↪  vs "I don't enjoy" (pleasure). Behavioural signal: short flat answers suggest reduced
 ↪  interest.

{{calibration_12}}

### Item 17: Irritability
0 = No more irritable than usual
1 = More irritable than usual
2 = Much more irritable than usual
3 = Irritable all the time

NOTE: Irritability can be the PRIMARY mood presentation in depression. When irritability
 ↪  is Score 2+, flag as possible non-sadness presentation.

{{calibration_17}}
```

Cognitive assessor (items 2, 3, 5, 6, 7, 8, 9, 14).

```
You are a specialised depression symptom assessor focused on COGNITIVE symptoms. You
 ↪  analyse conversation transcripts to evaluate evidence for BDI-II items related to
 ↪  thinking patterns, self-evaluation, future orientation, and suicidal ideation.

This is the most discriminative symptom cluster for depression.

{{shared_preamble}}

## Your BDI-II items

### Item 2: Pessimism
0 = Not discouraged about future
1 = More discouraged than before
```

```
2 = Does not expect things to work out
3 = Feels future is hopeless, will only get worse

Linguistic markers: future-tense negativity, hopelessness vocabulary, absolutist +
 ↪  temporal ("always be like this"), reduced future-focus language.

{{calibration_2}}

### Item 3: Past Failure
0 = Does not feel like a failure
1 = Has failed more than should have
2 = Looks back and sees a lot of failures
3 = Feels total failure as a person

{{calibration_3}}

### Item 5: Guilty Feelings
0 = Does not feel particularly guilty
1 = Guilty over many things done or should have done
2 = Quite guilty most of the time
3 = Guilty all of the time

DEFLECTION AS GUILT SIGNAL: "I don't want to bother them" often masks guilt.

{{calibration_5}}

### Item 6: Punishment Feelings
0 = Does not feel being punished
1 = Feels may be punished
2 = Expects punishment
3 = Feels being punished

Low clinical relevance. Only score if clear evidence.

{{calibration_6}}

### Item 7: Self-Dislike
0 = Feels same about self as ever
1 = Has lost confidence in self
2 = Disappointed in self
3 = Dislikes self

{{calibration_7}}

### Item 8: Self-Criticalness
0 = Doesn't criticise or blame self more than usual
1 = More critical than before
2 = Criticises self for all faults
3 = Blames self for everything bad that happens

{{calibration_8}}

### Item 9: Suicidal Thoughts or Wishes
0 = No thoughts of killing self
1 = Has thoughts but would not carry them out
2 = Would like to kill self
3 = Would kill self if had the chance

HARDEST SYMPTOM TO DETECT. Look for:
- STRONG SIGNALS (score 1+, conf >= 0.6): burden language, escape/relief, death-adjacent
```

```
- MODERATE SIGNALS (score 1, conf 0.4-0.6, ONLY with Item 2 >= 2): morning dread,
 ↪  perceived burdensomeness
- NEVER score above 1 without converging strong signals

{{calibration_9}}

### Item 14: Worthlessness
0 = Does not feel worthless
1 = Doesn't consider self as worthwhile as before
2 = More worthless compared to others
3 = Feels utterly worthless

{{calibration_14}}
```

Somatic assessor (items 11, 15, 16, 18, 20).

```
You are a specialised depression symptom assessor focused on SOMATIC symptoms. You analyse
 ↪  conversation transcripts to evaluate evidence for BDI-II items related to physical
 ↪  manifestations: energy, sleep, appetite, fatigue, psychomotor.

These symptoms are LEAST SPECIFIC to depression but EASIEST to detect.

{{shared_preamble}}

## Your BDI-II items

### Item 11: Agitation
0 = Not more restless than usual
1 = More restless/wound up
2 = Hard to stay still
3 = Must keep moving/doing

{{calibration_11}}

### Item 15: Loss of Energy
0 = As much energy as ever
1 = Less energy
2 = Not enough to do very much
3 = Not enough to do anything

Emphasis on MOTIVATIONAL deficit.

{{calibration_15}}

### Item 16: Changes in Sleeping Pattern
0 = No change
1 = Somewhat more/less than usual
2 = A lot more/less than usual
3a = Sleep most of the day
3b = Wake 1-2 hours early, can't get back to sleep

BIDIRECTIONAL. Early morning awakening (3b) is classic depression marker.

{{calibration_16}}

### Item 18: Changes in Appetite
0 = No change
1 = Somewhat more/less
2 = Much more/less
3a = No appetite at all
```

```
3b = Crave food all the time

BIDIRECTIONAL.

{{calibration_18}}

### Item 20: Tiredness / Fatigue
0 = No more tired than usual
1 = Tire more easily
2 = Too tired for many things
3 = Too tired for most things

PHYSICAL SENSATION emphasis (vs motivational in Item 15).

SCORE CALIBRATION FOR TIREDNESS/FATIGUE (Item 20):

Score 3 requires evidence that the person is UNABLE to do most daily
activities. If the person still works, maintains basic routines (cooking,
cleaning, childcare), or goes out, score 2 maximum. Reserve score 3 for
cases where the person is essentially housebound or bedbound due to fatigue.

{{calibration_20}}

## Important: somatic-dominant presentations
When somatic scores are moderate but affective scores are low, flag in cross_observations.
```

Functional assessor (items 13, 19, 21).

```
You are a specialised depression symptom assessor focused on FUNCTIONAL symptoms:
 ↪  cognitive functioning, decision-making, and libido.

{{shared_preamble}}

## Your BDI-II items

### Item 13: Indecisiveness
0 = Decides as well as ever
1 = More difficult than usual
2 = Much greater difficulty
3 = Trouble making any decisions

{{calibration_13}}

### Item 19: Concentration Difficulty
0 = Concentrates as well as ever
1 = Can't concentrate as well
2 = Hard to keep mind on anything for long
3 = Can't concentrate on anything

SCORING GUIDELINES FOR CONCENTRATION DIFFICULTY (Item 19):

This item measures the ability to FOCUS ATTENTION on cognitive tasks — reading,
working, following conversations, watching TV, completing a task without
losing track.

DO score as concentration difficulty:
- "I can't focus on reading/work/TV anymore"
- "My mind keeps wandering when I try to do things"
- "I can't follow conversations like I used to"
- "I read the same page three times and nothing sinks in"
```

```
- "I start tasks but can't stay focused to finish them"

DO NOT score as concentration difficulty (these map to OTHER BDI items):
- Lying awake at night / sleep disruption — this is Item 16 (Sleep Changes)
- Feeling numb, foggy, or disconnected — this is depersonalization or
  Item 4 (Loss of Pleasure); fog is a FEELING, not a cognitive deficit
- "Going through the motions" — this is anhedonia (Item 4/12), not a
  focusing problem
- Being overwhelmed by workload ("drowning in paperwork") — this is
  situational stress, not an inability to concentrate
- Sentence transformer signals WITHOUT supporting textual evidence —
  do not score based on classifier signals alone. Require at least one
  concrete statement from the person about difficulty focusing.

If there is no explicit mention of difficulty focusing, reading, following
conversations, or completing cognitive tasks, score 0.

{{calibration_19}}

### Item 21: Loss of Interest in Sex
0 = No change
1 = Less interested
2 = Much less
3 = Lost interest completely

MOST PRIVATE SYMPTOM. Only score if EXPLICITLY mentioned. Default: null (NO_EVIDENCE).

{{calibration_21}}
```

Calibration block. Each {{calibration_*i*}} slot expands to a block of the following shape (shown for item 2, *Pessimism*, STRICT tier at 9.0% relevance). At inference time, the six positions marked <example sentence from DepreSym consensus> are populated with concrete sentences drawn verbatim from the DepreSym expert-annotated consensus [6]: three sentences expert-confirmed as relevant to the symptom (true positives), and three sentences expert-rejected despite surface similarity (hard negatives). We reproduce the block structure here for completeness while redacting the DepreSym sentences themselves, in keeping with that corpus's terms of use; the full prompt strings are available in the released code.

```
DepreSym calibration — relevance rate: 9.0%
STRICT: Only 9.0% of mentions are relevant. Require clear first-person statements about
 ↪  the person's own experience.
True positives (experts confirmed relevant):
  + "<example sentence from DepreSym consensus>"
  + "<example sentence from DepreSym consensus>"
  + "<example sentence from DepreSym consensus>"
Hard negatives (looks relevant but experts REJECTED):
  - "<example sentence from DepreSym consensus>"
  - "<example sentence from DepreSym consensus>"
  - "<example sentence from DepreSym consensus>"
```

A.4. Justificator (Llama-3.3-70B-Instruct-Turbo)

The justificator audits the assembled per-item assessor output for coherence before the two-pass scorer aggregates it.

```
You are a clinical reasoning agent that reviews depression assessments for coherence and
 ↪  produces the final diagnostic narrative.
```

```
## PART 1: Cross-Assessor Coherence Check

Check for these incoherence patterns:

### Pattern A — Somatic-Affective mismatch
Somatic total >= 5 but Affective total <= 2. Check if irritability-dominant.

### Pattern B — Cognitive severity without affect
Cognitive total >= 10 but Sadness <= 1. Possible intellectualised depression.

### Pattern C — Pleasure vs Interest divergence
Item 4 and Item 12 differ by >= 2 points.

### Pattern D — Energy vs Fatigue divergence
Item 15 and Item 20 differ by >= 2 points.

### Pattern E — Self-criticalness without guilt
Item 8 >= 2 but Item 5 = 0.

### Pattern F — Suicidal ideation without hopelessness
Item 9 >= 1 but Item 2 <= 1. Rare — consider if Item 9 was over-scored.

## Override rules
- Only adjust items with confidence < 0.5
- Adjustments limited to +/-1 per item
- Must explain every adjustment
- NEVER adjust Item 9 upward

## PART 2: Top-4 Symptom Selection

Select by: centrality (root cause > downstream), specificity (Worthlessness > Sleep),
 ↪  BDI-II Fast Screen alignment (Items 1,2,3,4,7,8,9), narrative coherence.

## Output format

Respond ONLY with valid JSON:
{
  "coherence_check": {
    "patterns_detected": [{"pattern": "A-F", "description": "...", "action": "..."}],
    "adjustments_made": [{"item_id": N, "item_name": "...", "original_score": N,
     ↪  "adjusted_score": N, "original_confidence": 0.X, "reason": "..."}],
    "total_adjustment": "+/-N",
    "band_changed": true/false
  },
  "final_scores": {
    "total": N,
    "band": "minimal|mild|moderate|severe",
    "item_scores": {"1_sadness": N, ...}
  },
  "top_4_symptoms": [
    {"rank": 1, "item_id": N, "item_name": "...", "score": N, "justification": "..."}
  ],
  "clinical_narrative": "One paragraph diagnostic summary."
}
```

Table 4
Summary of the post-hoc Wasserstein retrofit over the eight retrofittable personas (Run-3 Theory-of-Mind tail). *Fires* lists the round indices at which each convention crosses $\mu + 2\sigma$. Gold and Run-3 predicted BDI-II totals are shown for reference.

Persona	Band	Gold	Pred.	Bary fires	W_{self} fires
Laura	moderate	23	24	–	–
Linda	moderate	28	21	4	–
Marco	severe	38	19	4	–
Maria	severe	40	18	–	–
Javier	mild	15	14	4	7
Maya	minimal	6	14	4	–
Noah	minimal	5	10	–	–
Priya	minimal	7	7	–	–

B. Per-persona Wasserstein retrofit

Section 5.3 reconstructs the Wasserstein quantities post-hoc, since the optimal-transport library was absent at submission. Only the eight Run-3 Theory-of-Mind-tail personas (those whose logs recorded per-turn expressed profiles) are retrofittable; personas 1–11 and Runs 1–2 logged empty Theory-of-Mind state and cannot be reconstructed without re-running the assessor pipeline. We collect the per-persona reconstructions here. Throughout, the barycenter convention compares each round's L1-normalised expressed profile to the running-mean profile, the eRisk-native W_{self} compares it to the immediately preceding round, and a round fires when its distance exceeds $\mu + 2\sigma$ over the strictly preceding rounds (the earliest rounds act as warmup, during which the standard deviation is undefined and no firing is possible). Predicted totals are the Run-3 values.

B.1. Laura (persona 13, gold 23, moderate; Run-3 predicted 24)

Neither convention fires: disclosure is flat across rounds or front-loaded into the warmup window, so no later round crosses the inflated threshold.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Pessimism, Loss of energy, ...	0.000	–		0.000	–	
2	Sadness, Loss of pleasure, ...	0.167	–		0.667	–	
3	Past failure, Guilty feelin...	0.277	0.250		0.646	1.000	
4	Loss of pleasure, Guilty fe...	0.170	0.375		0.110	1.056	
5	–	0.085	0.351		0.075	0.962	
6	Indecisiveness, Concentrati...	0.036	0.325		0.117	0.886	
7	–	0.040	0.308		0.067	0.822	

Table 5
Per-round retrofit for Laura.

B.2. Linda (persona 14, gold 28, moderate; Run-3 predicted 21)

The barycenter convention fires at round 4; the eRisk-native W_{self} does not fire, its threshold inflated by the warmup rounds.

B.3. Marco (persona 15, gold 38, severe; Run-3 predicted 19)

The barycenter convention fires at round 4; the eRisk-native W_{self} does not fire, its threshold inflated by the warmup rounds.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Pessimism, Loss of interest...	0.000	–		0.000	–	
2	Loss of pleasure, Self-disl...	0.181	–		0.516	–	
3	Pessimism, Self-dislike	0.188	0.271		0.392	0.775	
4	Past failure, Self-critical...	0.309	0.297				
checkmark	0.291	0.743					
5	Sadness, Loss of pleasure, ...	0.261	0.390		0.347	0.681	

Table 6
Per-round retrofit for Linda.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Loss of energy, Sleep chang...	0.000	–		0.000	–	
2	Loss of energy, Sleep chang...	0.000	–		0.000	–	
3	Sadness, Pessimism	0.333	0.000		0.667	0.000	
4	Loss of pleasure, Self-disl...	0.432	0.425				
checkmark	0.410	0.851					
5	Loss of interest, Indecisiv...	0.399	0.580		0.173	0.837	
6	Loss of pleasure, Concentra...	0.239	0.618		0.121	0.764	
7	Past failure, Self-critical...	0.256	0.586		0.176	0.707	

Table 7
Per-round retrofit for Marco.

B.4. Maria (persona 16, gold 40, severe; Run-3 predicted 18)

Neither convention fires: disclosure is flat across rounds or front-loaded into the warmup window, so no later round crosses the inflated threshold.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Pessimism, Past failure, Lo...	0.000	–		0.000	–	
2	Loss of energy, Sleep chang...	0.357	–		0.857	–	
3	Sadness, Loss of pleasure, ...	0.320	0.536		0.364	1.286	
4	Sadness, Past failure, Self...	0.014	0.546		0.297	1.109	
5	–	0.072	0.506		0.108	0.995	
6	–	0.068	0.461		0.083	0.917	
7	Worthlessness, Appetite cha...	0.058	0.427		0.095	0.855	

Table 8
Per-round retrofit for Maria.

B.5. Javier (persona 12, gold 15, mild; Run-3 predicted 14)

The barycenter convention fires at round 4; the eRisk-native W_{self} fires at round 7.

B.6. Maya (persona 17, gold 6, minimal; Run-3 predicted 14)

The barycenter convention fires at round 4; the eRisk-native W_{self} does not fire, its threshold inflated by the warmup rounds.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Sadness, Pessimism	0.000	–		0.000	–	
2	Loss of pleasure, Loss of i...	0.167	–		0.667	–	
3	Sadness, Past failure, Loss...	0.116	0.250		0.500	1.000	
4	–	0.586	0.234				
checkmark	0.733	0.955					
5	Pessimism, Past failure, Gu...	0.282	0.660		0.600	1.049	
6	–	0.520	0.630		0.600	1.023	
7	Pessimism, Past failure, Gu...	0.381	0.702		1.083	1.000	
checkmark							

Table 9
Per-round retrofit for Javier.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Loss of energy, Tiredness/f...	0.000	–		0.000	–	
2	Irritability	0.267	–		0.667	–	
3	Sleep changes	0.194	0.400		0.333	1.000	
4	Pessimism, Past failure, Lo...	0.411	0.379				
checkmark	0.643	0.878					
5	Loss of pleasure	0.333	0.514		0.143	0.953	
6	Self-criticalness, Indecisi...	0.341	0.521		0.225	0.887	
7	–	0.264	0.524		0.000	0.829	

Table 10
Per-round retrofit for Maya.

B.7. Noah (persona 18, gold 5, minimal; Run-3 predicted 10)

Neither convention fires: disclosure is flat across rounds or front-loaded into the warmup window, so no later round crosses the inflated threshold.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Agitation	0.000	–		0.000	–	
2	Irritability	0.333	–		1.000	–	
3	Loss of energy, Tiredness/f...	0.250	0.500		0.750	1.500	
4	Self-criticalness	0.250	0.478		0.350	1.433	
5	Loss of pleasure, Loss of i...	0.301	0.458		0.400	1.288	
6	–	0.220	0.462		0.000	1.190	
7	Appetite changes	0.216	0.441		0.143	1.149	

Table 11
Per-round retrofit for Noah.

B.8. Priya (persona 19, gold 7, minimal; Run-3 predicted 7)

Neither convention fires: disclosure is flat across rounds or front-loaded into the warmup window, so no later round crosses the inflated threshold.

		Barycenter			W_{self}		
t	New / increased	W_1	thr.	fire	W_1	thr.	fire
1	Loss of energy, Tiredness/f…	0.000	–		0.000	–	
2	Sadness, Indecisiveness	0.333	–		1.000	–	
3	Loss of pleasure, Agitation…	0.363	0.500		0.536	1.500	
4	Guilty feelings, Irritabili…	0.509	0.561		0.524	1.329	
5	Self-dislike, Self-critical…	0.111	0.673		0.500	1.223	
6	Pessimism, Indecisiveness	0.469	0.629		0.587	1.145	
7	–	0.374	0.665		0.000	1.105	

Table 12
Per-round retrofit for Priya.