INSA-Lyon at eRisk 2026 Task 2: DUET, Dual-view Evidence Tracking for Contextualised Early Detection of Depression

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

We present DUET (Dual-view Evidence Tracking), our submission to eRisk 2026 Task 2, contextualised early detection of depression from a stream of Reddit threads. DUET is an incremental per-user accumulator that assembles a ≈2341-dimensional feature vector at every round. The vector combines target-text embeddings, community-channel signals (reply sentiment, concern phrases, thread topic similarity), multi-scale Wasserstein temporal-shift detection over symptom and emotion trajectories, a Mahalanobis-based drift score against control and depressed populations, a Theory-of-Mind module contrasting a self-view of the target user with an observer-view of how the community perceives them, and a Thompson-sampled per-symptom personalisation. The five submitted runs span the precision–recall–latency frontier across XGBoost, a multi-layer perceptron, and a stacked ensemble, with an ERDE-aware adaptive decision threshold. Runs 0 and 4 tied for rank 5 of 70 runs in the field on $F_1 = 0.80$, four of the five reached perfect top-10 ranking quality from round 100 onwards with NDCG@100 = 0.89 (joint rank 3 in the field), and our ERDE5-optimised run reached $\text{ERDE}_5 = 0.15$ at a median true-positive alert latency of 9 rounds.

Keywords

Early risk detection, Depression detection from social media, Theory of Mind, Wasserstein distance, BDI-II

1. Introduction

Depression detection from social media is clinically motivated: as platforms become a venue where mental-health concerns surface before they reach professional support, automated tools that flag at-risk users early can supplement clinician and moderator judgement [1, 2]. eRisk 2026 Task 2 [3, 4], contextualized early detection of depression, frames this as an online problem: a stream of Reddit threads is delivered round by round, and at every round a system must decide whether each target user is at risk, weighing latency against accuracy under an ERDE-style (Early Risk Detection Error) penalty for late true-positive alerts [5]. What distinguishes the 2026 edition from earlier eRisk task lines is the contextualised framing [6]: each round delivers not just the target's own writing but the entire thread it sits in, including the submission, all comments, the reply hierarchy, and timestamps. Often a target user's writing in a given round is brief or ambiguous. The diagnostic signal is then carried by the community's reaction, namely concern phrases and shifts in reply sentiment, rather than by what the user wrote. Subject Subject A in the test cohort exemplifies the pattern: the target contributes only a handful of words per writing on average, but receives substantial community concern across the stream. This implies that an automatic detection system needs to alert on this subject as soon as possible by giving more weight to the community's text rather than target's own text.

Earlier eRisk submissions converged on transformer-based representations of target-user writing, supplemented by hand-engineered features and post-buffer mechanisms for stable decision-making [7, 8, 9]. The contextualised 2026 task suggests three modelling choices that we adopt: (i) two-channel evidence, separating what the target user writes from how the community responds, so that diagnostic

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

*Corresponding author.

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

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

signal can be picked up from either side of the asymmetry described above; (ii) multi-scale Wasserstein distributional shift [10, 11], treating each per-round feature as a trajectory and asking how its empirical distribution changes over time, at three temporal scales; and (iii) a Theory-of-Mind (ToM; [12, 13]) contrast between the model's reconstruction of the target's self-view and its reconstruction of how the community perceives the target, with their disagreement as a separate diagnostic feature. The first follows from what the motivational pattern above is about. The second derives from the observation that depressive episodes often manifest as changes in the rate and pattern of certain symptoms rather than as their absolute level. And the third operationalises the clinical observation that reduced insight, namely a mismatch between self-presentation and how others see one, is a familiar feature of depressive disorders. The resulting pipeline is structurally analogous to knowledge tracing [14]: a per-user state is maintained across observations and updated incrementally with each new round, adapted here to a streaming mental-health-risk setting.

We present **DUET** (Dual-view Evidence Tracking), an incremental pipeline that fuses the three commitments into a ≈2341-dimensional per-user feature vector consumed by classifier variants spanning the precision–recall–latency frontier. Runs 0 and 4 reach $F_1 = 0.76$, the best of our portfolio. Run 1, the ERDE5-optimised variant, reaches $\text{ERDE}_5 = 0.15$ and $\text{ERDE}_{50} = 0.05$, the latter shared as the best ERDE_{50} in the field.

2. Task and Data

Task 2 of eRisk 2026 frames depression detection as an online, latency-sensitive classification problem over a stream of Reddit threads. We adopt the formalisation directly: given for each target user u a sequence of threads $T_u^{(t)}$ delivered round by round (each thread carrying the submission, all comments, the reply hierarchy, timestamps, and a flag marking which contributions are by u), we produce at every round a binary decision $d_u^{(t)} \in \{0, 1\}$ together with a continuous risk score $s_u^{(t)} \in [0, 1]$. Three constraints make the setting distinctive: a positive decision is irreversible (once $d_u^{(t)} = 1$, the alert persists for all $t' > t$); decisions are made online, i.e., round t is settled before round $t + 1$ is revealed; and the evaluation penalises late true-positive alerts exponentially. Mind that the per-round delivery is the full conversational thread, not an isolated target post: target writings appear alongside the entire community context that responds to them, and exploiting this asymmetry is the principal modelling opportunity of the task.

For training and cross-validation we use the eRisk 2025 Task 2 release [15]: 909 users, of which 102 depressed and 807 control, i.e., 11.2% positive. Evaluation is on the eRisk 2026 stream of 523 subjects with 91 positives (17.4%) delivered over 500 rounds [3, 4]. The defining property of both releases is an asymmetry between the two channels available at every round: target-authored text is sparse, with a median of 25 words per writing for positive subjects and 13 for controls, while the conversational context is voluminous, with a median of about 13,700 community comments per positive subject across the stream. Two further textual cues on the training release motivate the Layer-1 lexical features of Section 3, namely positive subjects use first-person singular pronouns at 2.2 times the control rate and negative-emotion words at 2.9 times the control rate. Three properties of the released data matter for interpreting our results. First, distinctive terms in the two classes are partly subreddit-driven (positives concentrate on drug- and condition-related vocabulary, controls on finance, cryptocurrency, and chess), so surface lexical cues alone risk encoding topical confounds rather than depression signal. This is one reason we model the community channel separately. Second, thread coverage decays sharply over the 500 rounds, from 522 subjects at round 0 down to 12 at round 500; the live system exploits this at runtime (Section 4). Third, one of the 523 golden subjects never receives a thread in the released stream, so the effective subject count for round-by-round prediction is 522.

Task 2 is scored on two families of metrics. The decision-based family penalises late true-positive alerts: ERDE_o multiplies the per-subject false-negative cost by a sigmoid latency penalty $\text{lc}_o(k) = 1 - (1 + e^{k-o})^{-1}$ evaluated at the alert round k and operating point o, and is reported at $o \in \{5, 50\}$, with ERDE_5 penalising lateness an order of magnitude more steeply than ERDE_{50}. Standard F_1 is also

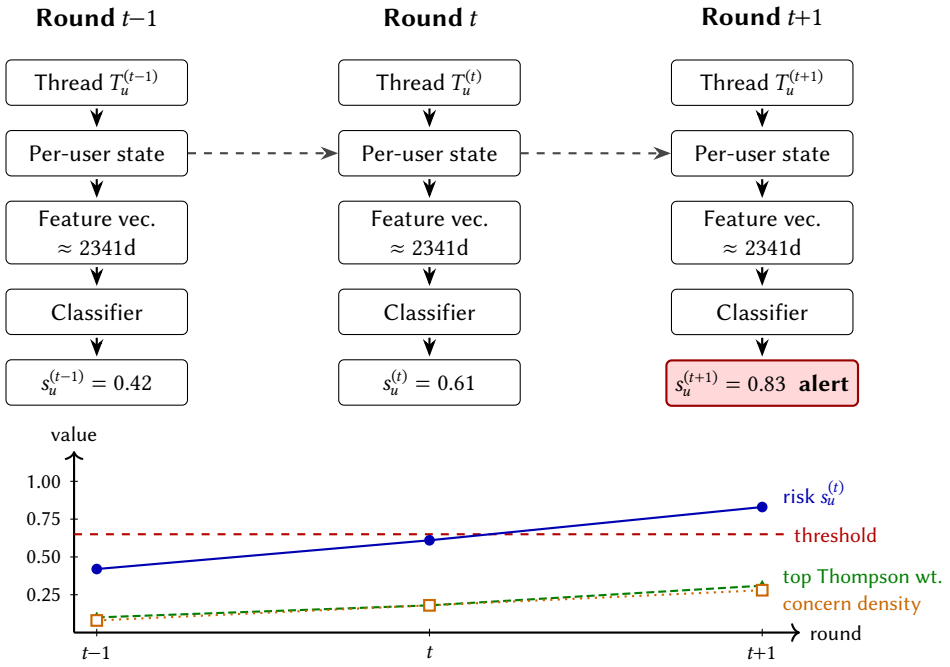

Figure 1: DUET architecture, unrolled across three consecutive rounds. Each column shows the same incremental loop: thread input, per-user state update, feature assembly, classifier, decision. Dashed horizontal arrows show the state carrying across rounds. The trace beneath the columns shows three illustrative quantities (the Layer-2 concern density, the top symptom weight from the Thompson bandit, and the risk score $s_u^{(t)}$) evolving across the three rounds and crossing the alert threshold at round $t + 1$. All values are constructed for illustration.

reported, together with F_{latency}, which discounts F_1 by a speed factor derived from the median true-positive alert latency. The ranking-based family scores how well the model identifies the highest-risk subjects without committing to a binary decision: P@10, NDCG@10, and NDCG@100 at the final round and at intermediate checkpoints. We refer to the eRisk overview for full formulae [3, 4]. These two families motivate different operating points of the same model: the steep early penalty of ERDE_5 drives our ERDE5-optimised run (R1) to alert aggressively, while ERDE_{50} and F_1 tolerate more evidence accumulation and motivate the conservative R0 and R4 configurations (Sections 3 and 5).

3. System

3.1. Overview

DUET (Dual-view Evidence Tracking) is an incremental per-user accumulator. At each round t, the new thread $T_u^{(t)}$ is folded into a per-user state. From that state we assemble a feature vector of approximately 2341 dimensions, and emit a risk score $s_u^{(t)}$ and a decision $d_u^{(t)}$ via a per-run classifier with an ERDE-aware adaptive threshold. The state carries forward unchanged into round $t + 1$, so DUET is structurally analogous to a knowledge-tracing model: per-user evidence accumulates incrementally rather than being recomputed from scratch [14]. Figure 1 unrolls one such loop across three consecutive rounds with an illustrative trace. Mind that the live submission ran two of the feature blocks described below under a different configuration than the designed pipeline; we report the engineering reality and its consequences in Section 4.

3.2. Feature layers

The feature vector is assembled from three layers, summarised in Table 1. Layer 1 encodes what the target user writes. Every text the target authors in a round is embedded by an ensemble of three

sentence transformers (`all-mpnet-base-v2`, `all-MiniLM-L12-v2`, and `all-distilroberta-v1`) [16], concatenated into a 1920-dimensional vector, and folded into an exponentially-decayed running mean across rounds with $\lambda = 0.95$. We add 189 BDI-II symptom-activation features (per-symptom cosine similarity to 21 Variant-C reference embeddings, summarised across rounds by maximum, mean, and seven distributional statistics, namely mean, variance, skewness, kurtosis, and the three quartiles) [17], and a 4-dimensional lexical block ratioing first-person-singular pronouns, negative-emotion words, absolutist words [18], and cognitive-process words against total word count.

Layer 2 encodes what the community does. This is the payoff of the asymmetry argued in Section 2. We compute three reply-sentiment features (mean, standard deviation, and linear-trend slope of VADER compound [19] on direct replies to the target across rounds), one concern-phrase ratio (the proportion of rounds in which at least one of 42 hand-curated phrases of community worry matches a non-target comment), three conversational-position features (the fraction of rounds in which the target is the submission author, the fraction in which the target contributes no text at all, and the mean reply depth of target turns), and 21 thread-topic-similarity features (cosine between thread title embedding and each BDI-II reference embedding, averaged across rounds), for 28 features in total. Mind that two further volume features (the mean and linear-trend slope of target text length) are computed by the extractor but excluded from the assembled vector.

Layer 3 captures semantic signal independent of the symptom dimensions. An emotion classifier [20] produces a per-round distribution over basic emotions; the cross-round mean and Shannon entropy of this distribution together form a 9-dimensional block (eight emotion means plus one entropy term; texts shorter than ten words fall back to a uniform distribution). Mind that the deployed block contains an internal vocabulary mismatch: the code reserves eight slots named for Plutchik's primary emotions, while the model is trained on Ekman's six emotions plus a neutral class. Six emotions are shared and populate their named slots correctly; the two Plutchik-only slots (anticipation, trust) are structurally zero in every vector because the model cannot emit them; the model's neutral score has no destination slot and is discarded before renormalisation. The pairing is internally consistent across training and inference, namely every classifier was trained on this 9-dimensional block as-is, but two slots are wasted and the achievable entropy range shifts to $[0, \log 6]$ rather than $[0, \log 8]$. We discuss this as a known design-time bug in Section 6. A BERTopic model [21] fitted on the training stream contributes 40 topic-distribution features plus one depression-proportion feature, defined as the cross-round mean of an indicator firing when the top-1 topic for a round belongs to a depression-associated set identified at training time. Mind that the BERTopic block ran for the first nineteen rounds of the live submission and was skipped from round 20 onwards under a memory-saving fix (Section 4).

3.3. Distributional shift: Wasserstein and Mahalanobis

DUET also computes two non-parametric summaries of the per-user trajectories assembled in the previous layers, capturing *distributional shift*: how the distribution of a feature changes over time within a single user. The first is the 1-dimensional Wasserstein distance W_1 between adjacent half-windows of each feature trajectory at three temporal scales. Letting $\hat{F}_A$ and $\hat{F}_B$ denote the empirical CDFs of two windows A and B, we use the closed-form 1-D distance

$$W_1(\hat{F}_A, \hat{F}_B) \;=\; \int_0^1 \left| \hat{F}_A^{-1}(q) - \hat{F}_B^{-1}(q) \right| dq$$

computed via the POT library [22]. The three scales compare two adjacent half-windows: a *short* scale ($[k{-}10, k{-}5]$ versus $[k{-}5, k]$, requiring at least 10 rounds of history), a *medium* scale ($[k{-}50, k{-}25]$ versus $[k{-}25, k]$, requiring at least 50 rounds), and a *long* scale ($[0, n/2]$ versus $[n/2, n]$, requiring at least 20 rounds), with k the current round and n the total rounds observed. These three scales are intended to capture acute deterioration, sub-chronic evolution, and chronic trajectory change respectively. We apply this across 21 BDI-II symptom activation series ($3 \times 21 = 63$ features), the emotion-entropy series (3 features), the 1920-dimensional sentence-transformer embedding (sliced W_1 over 50 random projections, 3 features), and the topic-entropy series (3 features), for 72 features in total. Mind that, on

Table 1
Feature blocks of DUET and their dimensions; total approximately 2341.

Block	Dim.
Layer 1 – Target text	**2113**
Sentence-transformer running mean	1920
BDI-II symptom activations	189
Lexical ratios	4
Layer 2 – Community	**28**
Reply sentiment (VADER)	3
Concern-phrase ratio	1
Conversational position	3
Thread topic similarity	21
Layer 3 – Semantic	**50**
Emotion (mean + entropy)	9
BERTopic + depression-proportion	41
Distributional (§3.3)	**77**
Wasserstein + Mahalanobis + combined	72 + 3 + 2
Theory of Mind (§3.4)	**47**
Thompson (§3.5)	**25**
Round counter	1

the training release, the three scales are not equally informative: 153 subjects fire on the short scale alone against only 5 that fire on the long scale alone, and the long-only case turns out to be a control subject that clears its firing threshold by a single hundredth. We return to this in Section 5.

The Mahalanobis block scores how far the user's current feature profile sits from the control and depressed population centres. We fit two multivariate Gaussians to the training set after PCA to 50 dimensions with Ledoit-Wolf shrinkage [23], one on the 807 control users and one on the 102 depressed users, and at test time report three scalars per user: the Mahalanobis distance D_M^{ctrl} to the control centre, the distance D_M^{dep} to the depressed centre, and their difference $D_M^{\text{ctrl}} - D_M^{\text{dep}}$. The discriminative quantity is the difference: a user closer to the depressed centre than to the control centre has a positive value, and a user closer to control has a negative one.[1]

3.4. Theory of Mind module

Theory of Mind, namely the capacity to attribute distinct mental states to oneself and to others [12], has been instrumented in computational pragmatics for several decades, and more recently used to probe whether large language models can solve false-belief tasks [13]. In the depression literature, reduced insight is associated with a dissociation between how a user presents their own state and how the community around them perceives that state: a target may describe themselves as fine, while replies from peers cluster around concern phrases and supportive interventions. DUET instruments this dissociation as a single dual-mentalising step. We prompt a large language model twice per thread, once to produce a self-view that scores the target's BDI-II symptoms from the target's own writings alone, and once to produce an observer-view that scores the same symptoms from how the community responds to the target. The mean per-item absolute difference between the two views is our *insight gap*; it is dimensionless, bounded in $[0, 2]$ when symptom scores are on a 1–3 integer scale, and constitutes a single feature whose magnitude reflects exactly the disagreement between target self-presentation and community perception.

In practice, this is implemented as follows. First, each thread is formatted as a structured transcript with target turns and target-author indicators marked `[TARGET]` and community turns marked with

[1] A 2-dimensional combined-distributional block is also appended to the assembled feature vector: its first scalar duplicates D_M^{ctrl}, and its second is the unweighted mean of the 72-dimensional Wasserstein vector. We report this for completeness; the duplication is an assembler-level redundancy, not a designed feature.

role and reply depth; a four-class priority truncation governs which content survives a 2000-token budget (P1 = target posts plus their direct replies, kept in full; P2 = other posts on a branch that includes the target, kept if budget allows; P3 = one-hop-further posts, truncated to 100 characters; P4 = unrelated branches, dropped). Second, we issue two separate calls to `meta-llama/Llama-3.3-70B-Instruct` [24] via the HuggingFace Inference API [25], each with a two-message (`system` + `user`) shape, no assistant primer, and no few-shot examples. Both system prompts inline the same 21 BDI-II Variant-C symptom definitions (Appendix A) and a JSON output schema; they are kept byte-identical across all calls of each type. The self-view prompt asks the model to assess the target from the target's own writings alone. In contrast, the observer-view prompt asks how the community perceives the target from the formatted transcript.

The 47-dimensional ToM block is then assembled from the two parsed responses. Self-view per-symptom scores in $\{1, 2, 3\}$ occupy slots $[0:21]$ (each divided by 3 to fall in $[0.333, 1]$), observer-view scores occupy $[21:42]$, the self- and observer-view depression probabilities and the insight gap (computed as the mean per-item absolute difference between self- and observer-view scores) occupy $[42, 43, 44]$, and the observer concern level normalised by 3 and the community-response category occupy $[45, 46]$. The community-response category is mapped to a scalar in $[0, 1]$ by the encoding {concern: 1.0, support: 0.8, advice: 0.6, mixed: 0.5, normalisation: 0.3, casual: 0.0}; mind that this collapses a six-way categorical into an ordinal community-concern axis, namely a modelling assumption inherited from the codebase that we have not tested against alternative encodings (Section 6). On silent threads where the target contributes no text, the self-view call is skipped and only the observer-view call fires; slots $[0:21]$ and slot $[42]$ remain at their default zero. JSON outputs are parsed by a layered fallback (direct parse, then markdown-fenced extraction, then outermost {...} substring); persistent parse failures leave the corresponding sub-vector at its default zero rather than triggering a retry.

Both calls use temperature 0.1, a maximum of 2048 output tokens, and a 120-second timeout. Server-side JSON enforcement (e.g., `response_format`) is not used; output discipline is enforced at the prompt level alone, and recovery is the parse fallback described above. Calls are sequential per (user, round, thread); no batching across users, rounds, or threads is performed at the LLM step. The Ollama back-end is also configured in the codebase as an alternative local inference path, but the training-feature build used the HuggingFace endpoint exclusively.

Table 2 shows the designed ToM output for two constructed contrastive users: User A, whose self- and observer-view scores match (an acknowledged presentation, insight gap 0.24), and User B, whose self-view minimises while the community notices (a reduced-insight presentation, insight gap 0.95). The pattern that survives in User B's $|\Delta|$ column is clinically intuitive: the affective and cognitive items (sadness, loss of pleasure, worthlessness, concentration) diverge by 2 between self and observer, while the somatic items (sleep, fatigue, appetite) diverge by less, consistent with users acknowledging physical symptoms more readily than mental ones.

Mind that the table above shows the ToM block as designed and as built at training; the live submission ran a degraded version of the module, with consequences for the no-ToM ablation we report in Section 5.

3.5. Thompson-sampled symptom personalisation

Not every BDI-II symptom is equally informative for every user. To personalise symptom importance per user, we maintain a 21-armed Beta bandit, with arms initialised at $\text{Beta}(1, 1)$. At each round, we mark a symptom as active if its cosine activation against the symptom reference embedding exceeds $\tau = 0.3$; the corresponding Beta posterior is updated by incrementing α_i on activation and β_i otherwise. The expected weight $\alpha_i/(\alpha_i + \beta_i)$, normalised to sum to one across symptoms, gives a personalised importance profile. From this we expose 25 features: a Thompson-weighted symptom score, the Shannon entropy of the weight distribution, the active-symptom count, the full 21-dimensional weight vector, and the mean posterior uncertainty across arms. Mind that the entropy and uncertainty terms are themselves informative at early rounds: a flat distribution and high uncertainty signal that the model has not yet identified which symptoms matter for this particular user.

Table 2
Illustrative output of the designed ToM module for two constructed contrastive users. User A: acknowledged presentation, self- and observer-view scores aligned. User B: reduced-insight presentation, self minimises while the community notices. All values are illustrative; the users do not exist. The module shown here ran at training and in Phase 1 of the live submission; Section 4 reports the inference-time behaviour.

		User A			User B		
#	BDI-II item	self	obs.	|Δ|	self	obs.	|Δ|
1	Sadness	3	3	0	1	3	2
2	Pessimism	2	3	1	1	2	1
3	Past failure	2	2	0	1	2	1
4	Loss of pleasure	2	2	0	1	3	2
5	Guilt	1	1	0	1	2	1
6	Punishment	1	1	0	1	1	0
7	Self-dislike	2	2	0	1	3	2
8	Self-criticism	2	2	0	1	2	1
9	Suicidal thoughts	1	1	0	1	1	0
10	Crying	1	2	1	1	1	0
11	Agitation	1	1	0	1	1	0
12	Loss of interest	2	3	1	1	3	2
13	Indecisiveness	1	2	1	1	2	1
14	Worthlessness	2	2	0	1	3	2
15	Loss of energy	3	3	0	2	3	1
16	Sleep changes	3	3	0	2	3	1
17	Irritability	2	2	0	2	2	0
18	Appetite changes	2	2	0	2	2	0
19	Concentration	2	3	1	1	3	2
20	Fatigue	3	3	0	2	3	1
21	Sexual interest	2	2	0	1	1	0
Self depression prob.		0.72	—		0.18	—	
Observer concern level		—	3		—	3	
Community response		—	concern		—	concern	
Insight gap				**0.24**			**0.95**

3.6. Classifiers and run portfolio

We train four classifier variants on the assembled feature vector. The primary model is XGBoost [26] (`max_depth`=6, 300 trees, learning rate 0.1, `scale_pos_weight` set automatically to compensate for the 11.2% positive rate). The secondary model is a two-layer MLP ($D \to 256 \to 64 \to 1$, batch normalisation and dropout 0.3 after each hidden layer, Adam optimiser [27] with learning rate 10^{-3} and BCEWithLogits loss weighted by class imbalance). A reference RBF-SVM with balanced class weights is included for comparison. A stacked ensemble combines XGBoost, the MLP, and the SVM via 5-fold stratified out-of-fold predictions fed into a logistic-regression meta-learner. All models consume features after `StandardScaler` normalisation; the entire classification stack is built with `scikit-learn` [28].

At inference, the per-round risk score $s_u^{(t)}$ is compared against an adaptive threshold that tracks the ERDE latency penalty:

$$\theta(k) \;=\; \theta_{\text{init}} - \big(\theta_{\text{init}} - \theta_{\text{floor}}\big) \cdot \text{lc}_o(k),$$

with $\text{lc}_o(k)$ the sigmoid latency cost of Section 2 at the same operating point o used for scoring. The threshold drops monotonically over rounds, namely it is strict early and lenient late. An alert is emitted, and persists, the first round in which $s_u^{(t)} \geq \theta(k)$ has held for t_{con} consecutive rounds. Mind that this calibration favours evidence accumulation over early sensitivity. On the training simulation the mean alert round for our conservative configuration (R0) was 158, an order of magnitude later than the test-time median true-positive latency we observed (Section 5).

Table 3 summarises the five submitted runs. R0 (FULL_XGBOOST_ERDE50) is the conservative XGBoost baseline; R1 (FULL_XGBOOST_ERDE5) is the aggressive early-alerting variant, with a lower starting threshold, a steeper schedule, and single-round confirmation; R2 (FULL_NN_ERDE50) substitutes the MLP under the conservative policy; R3 (FULL_ENSEMBLE_BALANCED) deploys the stacked ensemble at an

Table 3
Run-portfolio configuration. θ_{init} and θ_{floor} are the upper and lower threshold endpoints; o is the ERDE operating point that the threshold schedule tracks; t_{con} is the number of consecutive rounds the score must remain above $\theta(k)$ before the alert latches.

Run	Name	Classifier	θ_{init}	θ_{floor}	o	t_{con}
R0	FULL_XGBOOST_ERDE50	XGBoost	0.85	0.45	50	2
R1	FULL_XGBOOST_ERDE5	XGBoost	0.70	0.35	5	1
R2	FULL_NN_ERDE50	MLP	0.85	0.45	50	2
R3	FULL_ENSEMBLE_BALANCED	ensemble	0.80	0.40	30	2
R4	NO_TOM_XGBOOST_ERDE50	XGBoost	0.85	0.45	50	2

intermediate operating point; R4 (NO_TOM_XGBOOST_ERDE50) is identical to R0 with the 47-dimensional ToM block zeroed and is intended as the ToM ablation. Mind that R4 does not measure what its name suggests. We discuss the confound in Section 5.

4. Implementation and Experimental Setup

4.1. Offline configuration

For the offline portion of the pipeline we use the eRisk 2025 Task 2 release as the training and validation corpus [15], namely 909 users with 102 labelled depressed and 807 labelled control (an 11.2% positive rate). Cross-validation uses 5-fold stratified splits that preserve this rate per fold, and all classifiers use random seed 42 where applicable. The BERTopic model is fit on the training threads with target $n_{\text{topics}} = 40$, UMAP reduction to 5 dimensions, and HDBSCAN with `min_cluster_size` = 50; depression-associated topics are identified at fit time as those with at least 10 documents and a depression density of at least 1.5× the corpus base rate. The Mahalanobis covariances are fit on the same training set after PCA to 50 components with Ledoit-Wolf shrinkage, separately for the control and depressed sub-populations. The only feature block that requires its own dedicated offline build is the Theory of Mind module: Layers 1–3 and the distributional features can extract directly from raw threads in the live pipeline, but the ToM block depends on synchronous LLM calls with a 120-second per-call timeout. We therefore pre-computed the 47-dimensional ToM features once on the training corpus, by formatting every thread as described in Section 3.4 and issuing the two LLM calls per thread via the HuggingFace InferenceClient (`meta-llama/Llama-3.3-70B-Instruct`, temperature 0.1, `max_tokens` 2048, 120-second timeout). The resulting 909×47 array is what the shipped classifiers train on. Mind that this is the rich Option C ToM signal of Section 3.4: every dimension populated by the LLM, every slot semantically grounded in the prompts of Appendix A. The fact that the live pipeline does not reproduce this signal is the engineering story of the next subsection and the analytical story of Section 5.

4.2. Live engineering and two code regimes

The live submission interacted with the eRisk 2026 server in rounds and ran from April 17 to April 23, 2026, completing all 500 rounds across all five runs. Two code changes were deployed mid-run. The first, between rounds 19 and 20 on April 19, altered two of the feature blocks described in Section 3, namely the BERTopic block and the ToM observer-view. The second, between rounds 98 and 99 on April 23, replaced the per-user non-batched encoder calls with a single batched per-round call to the sentence-transformer ensemble, without further altering the feature configuration. The live pipeline therefore ran in three operational regimes, summarised in Table 4, only the first boundary of which is a feature-regime boundary.

Phase 1 (rounds 0–19) executed the pre-fix code: per-user non-batched embedding of every text, encoding of the `other_comments` needed to populate the ToM observer-view, and the BERTopic transform producing a real 41-dimensional topic block. Per-round wall-clock averaged approximately 100 minutes,

Table 4
The three live-submission operational regimes. The first boundary (rounds 19↔20) is a feature-regime change; the second (rounds 98↔99) is an encoder-batching change that did not alter the feature configuration.

	Phase 1	Phase 2a	Phase 2b
Rounds	0–19	20–98	99–500
Dates	Apr 17–19	Apr 19–23	Apr 23–23
Per-round wall-clock	≈ 100 min	≈ 42–88 min	≈ 3 min
Sentence-transformer encoding	per-user	per-user	batched per round
`other_comments` encoded	yes	no	no
Observer-view	populated	`None`	`None`
BERTopic transform	active (real 41-d)	skipped (41 zeros)	skipped (41 zeros)
R0 alerts issued in phase	50 of 101	51 of 101 (combined)	

dominated by the non-batched encoding of community comments. Phase 2a (rounds 20–98) executed the first post-fix code, skipping the encoding of `other_comments` (so the observer-view was `None` throughout) and the BERTopic transform (so the 41-dimensional topic block was zero throughout). Per-round wall-clock fell modestly, from approximately 88 minutes at round 20 to approximately 42 minutes at round 98, the gradual decline reflecting cohort thinning rather than the code change itself. Phase 2b (rounds 99–500) executed the second post-fix code, with the batched per-round encoder in place. Per-round wall-clock dropped by roughly an order of magnitude to approximately 3 minutes per round, and the remaining 402 rounds completed in under 24 hours. The feature configuration in Phase 2b was identical to Phase 2a; only the operational cost changed.

The per-round wall-clock did not penalise the official scores: eRisk latency is round-indexed, namely an alert at round k carries the same latency cost regardless of how long the server took to deliver round k, so the wall-clock move from approximately 100 minutes to approximately 3 minutes per round changed how long the submission took us, not where alerts fell relative to the round index. The metric-relevant consequence of the two deployments is the feature-regime blend: 50 of Run 0's final 101 alerts fired in Phase 1 with the richer feature set, and 51 fired across Phases 2a and 2b combined under the degraded configuration. The implications for Run 4 (the no-ToM ablation, additionally affected by a slot-`[42]` semantic drift) are discussed in Section 5.

5. Results and Analysis

5.1. Decision-based metrics

Table 5 reports the decision-based metrics for the five submitted runs alongside their field rank. Runs 0 and 4 are tied for rank 5 of 70 systems in the field on $F_1 = 0.80$; the field leader (HUGETIME) reaches $F_1 = 0.83$, and a cluster of three systems (UNED-GELP, NoirByte, MindAILab) sit immediately above us at $F_1 = 0.82$ [3, 4]. Run 1, our ERDE5-optimised variant, reaches $F_1 = 0.78$ with the best early-detection profile of our portfolio: $\text{ERDE}_5 = 0.15$, $\text{ERDE}_{50} = 0.05$, and a median true-positive alert latency of 9 rounds against 13 for Run 0. Run 3 (the stacked ensemble at the intermediate operating point) reaches $F_1 = 0.75$. Run 2 (the MLP under the conservative policy) is the recall-extreme outlier of the portfolio at $F_1 = 0.62$, capturing 95% of positive subjects at precision 0.46, namely the operating point a recall-first triage workflow would want.

The latency advantage of Run 1 is concrete enough to illustrate with two representative subjects. Across the 28 true positives where Run 1 alerts at least 5 rounds earlier than Run 0, the median gap is 21 rounds. The median case is Subject B: Run 0 alerts at round 50, Run 1 at round 29, exactly 21 rounds earlier. A larger gap arises with Subject A, whose round-by-round writing is sparse but who receives substantial community concern across the stream: Run 0 alerts at round 41, Run 1 at round 4, a gap of 37 rounds. This is the same subject we used in Section 1 to motivate the community-channel argument, and the gap shows that an ERDE-aware threshold can capitalise on the community signal early enough to matter. Mind that the official decision-based metrics reflect a blended pipeline: 50 of

Table 5
Official decision-based metrics for the five INSALyon runs on eRisk 2026 Task 2 [3, 4]. Field rank is on F_1 within the full field of 70 runs.

Run	Classifier	P	R	F_1	ERDE_5	ERDE_{50}	lat. TP	$F_{\text{lat.}}$
R0 (full_xgboost_erde50)	XGBoost	0.76	0.85	0.80	0.16	0.07	13.0	0.76
R1 (full_xgboost_erde5)	XGBoost	0.69	0.89	0.78	0.15	0.05	9.0	0.75
R2 (full_nn_erde50)	MLP	0.46	0.95	0.62	0.14	0.08	7.5	0.60
R3 (full_ensemble_balanced)	ensemble	0.63	0.92	0.75	0.16	0.06	11.5	0.72
R4 (no_tom_xgboost_erde50)	XGBoost	0.75	0.85	0.80	0.16	0.07	14.0	0.76
Field leader F_1 (HUGETIME)				0.83				
Field tie at rank 2 (UNED-GELP, NoirByte, MindAILab)				0.82				

Table 6
Official ranking-based metrics for the five INSALyon runs at round 500 [3, 4].

Run	P@10	NDCG@10	NDCG@100
R0	1.00	1.00	0.89
R1	1.00	1.00	0.89
R2	1.00	1.00	0.85
R3	1.00	1.00	0.89
R4	1.00	1.00	0.89
Field leader (DeepCare)			0.91
Field rank 2 (MindAILab)			0.90

Run 0's 101 alerts fired in Phase 1 of the live submission (rounds 0–19, with BERTopic active and the observer-view embedding live), and 51 fired across Phases 2a and 2b combined under the degraded feature configuration (Section 4.2).

5.2. Ranking-based metrics

Table 6 reports the ranking-based metrics. All five runs reach a perfect top-10 ranking, namely P@10 = NDCG@10 = 1.00, by round 100 and maintain this through round 500. The four full-feature runs (R0, R1, R3, R4) reach NDCG@100 = 0.89 at round 500, tied for rank 3 of 70 systems in the field; only DeepCare (0.91) and MindAILab (0.90) rank higher, and we are tied with five NoirByte runs at 0.89 [3, 4]. The MLP-only configuration (R2) reaches 0.85, separating it from the rest of our portfolio on ranking quality just as it does on decision-based precision. The ranking quality stabilises early in the stream, namely the four full-feature runs reach NDCG@100 = 0.84 at 100 writings, climb to 0.87 at 250 writings, and settle at 0.89 by 500 writings; the top-10 metrics (P@10 and NDCG@10) saturate at 1.00 by the 100-writings checkpoint and stay there.

5.3. Finding 1: XGBoost carries, and the Wasserstein scales are unequal

Two observations explain the rank ordering across runs. First, gradient-boosted feature selection handles the sparse, high-dimensional vector better than the alternatives we tried. On local re-scoring across the 523 test subjects, Runs 0 and 4 produce 77 true positives against 24–25 false positives; Run 3 (the stacked ensemble) holds 49 false positives at slightly lower recall; Run 2 (the MLP under the same threshold schedule) over-alerts to 102 false positives, namely more than the 91 total positives in the test set. The MLP and the ensemble do not benefit from the threshold schedule that was tuned for XGBoost, and we revisit this in Section 5.4.

Second, the multi-scale Wasserstein design of Section 3.3 is asymmetric in practice. On the mean-symptom trajectory in the training simulation, the short scale fires for 153 subjects without the long scale firing, while the long scale fires for only 5 subjects without the short scale firing. Mind that the

five long-only cases are not strongly discriminative: the canonical example, subject Subject C, is a control subject that clears its firing threshold by 0.010, namely a single hundredth above the bar. The short scale, by contrast, fires informatively on positives such as subject Subject D (17 short-scale fires across the stream and no long-scale fires; Run 0 alerts at round 116). We carry the three-scale design through to the live submission because we do not have the data to remove the long scale safely, but the asymmetry suggests it is a candidate for removal or re-weighting (Section 7).

5.4. Finding 2: Train-to-test generalisation suggests a conservative calibration

The decision-based metrics of Section 5.1 are noticeably better than what we observed in 5-fold cross-validation on the training data. XGBoost lifted from CV $F_1 \approx 0.70$ to test $F_1 = 0.80$ (a +0.10 absolute jump), and the stacked ensemble lifted from CV $F_1 \approx 0.63$ to test $F_1 = 0.75$ (+0.12). The MLP was the only configuration that did not lift, in fact moving in the opposite direction (CV ≈ 0.65, test $F_1 = 0.62$, namely -0.03); on the test cohort it reaches recall 0.95 at precision 0.46, i.e., it generalised by alerting indiscriminately. We read this pattern as a sign that the 2025-derived feature distribution and the threshold schedule generalised better to the 2026 test cohort than CV alone would have predicted, with the MLP a clear exception.

A second observation concerns alert latency. On the eRisk 2026 test stream the median true-positive alert latency is 13 for Run 0 and 9 for Run 1, namely Run 1 alerts a median of four rounds earlier than Run 0 on true positives, at the cost of seven additional false positives (36 against 24). Mind that this is consistent with how the threshold schedule was calibrated: $\theta(k)$ favours evidence accumulation over early sensitivity (Section 3.6), and a more aggressive early-alert policy buys earlier detection at the price of precision. A natural follow-up is to optimise the threshold directly against ERDE_5 rather than F_1, since the conservative calibration likely cost us ERDE_5 without a corresponding F_1 benefit, and the trade-off is on the table for the next iteration (Section 7).

5.5. Finding 3: The no-ToM ablation is confounded by a train-test feature mismatch

Run 4 was intended to ablate the 47-dimensional ToM block, namely it is identical to Run 0 except that the `no_tom` mask zeroes the ToM slot. The official metrics report the two runs as effectively indistinguishable. Local re-scoring puts the difference at a single subject (one additional false positive in Run 4) and the metric difference is at most 0.01 across F_1, ERDE_5, ERDE_{50}, and F_{latency}. The intuitive reading, namely *ToM features do not help XGBoost*, is not supported by the design and we report this transparently.

Two compounding implementation drifts confound the comparison. First, the live ToM block runs the Option A embedding-only path (Section 3.4), so the 21 self-view symptom scores, 21 observer-view symptom scores, observer depression probability, insight gap, observer concern level, and community-response category are all zero at inference, whereas the classifier was trained on Option C features in which each of these slots was populated by the LLM (Section 4.1). Second, and more subtly, slot `[42]` carries a different quantity in training and at inference: at training, it is the self depression probability that Prompt 1 returns, namely a bounded value in $[0, 1]$; at inference, it is the L2 norm of the cross-round mean of the round-level mean target-text embeddings, which is unbounded and has no $[0, 1]$ calibration. The classifier therefore sees its trained "self depression probability" feature replaced by an out-of-distribution semantic-mass proxy. Together, the structural zeroing of slots $[0:42]$ and $[43:47]$ and the semantic drift at slot `[42]` mean that Runs 0 and 4 feed the same near-zero, partially semantically-misaligned 47-dimensional block to the same classifier. The `no_tom` mask zeroes a block that is already mostly zero, and the contrast collapses.

We can quantify what is lost in offline cross-validation. With Option C ToM features present, the stacked ensemble gains +3.3 percentage points in F_1 relative to a no-ToM baseline on the training data. That gain is on training-time Option C features and cannot be transferred to the live submission under the present configuration: a clean test of ToM's contribution would require wiring Option C into the live `run_pipeline`, which we did not do (Section 7; Section 6).

6. Limitations

The most consequential limitation of DUET as evaluated is the gap between the ToM module as designed and as built at training (Section 4.1) and the ToM module that ran at live inference (Section 4.2). The classifiers were trained on the rich Option C signal in which all 47 dimensions were populated by the LLM, but live inference produces a near-zero block with one slot (`[42]`) that carries an unbounded semantic-mass proxy where the training-time pipeline put a bounded probability. As a direct consequence, Run 4 cannot be read as a clean ablation of the ToM contribution: Run 0 and Run 4 feed effectively the same block to the same classifier, the mask zeroes a block that is already mostly zero, and the two runs differ by a single subject (Section 5.5). The +3.3 percentage points in F_1 that ToM appeared to deliver in cross-validation is, in this configuration, undelivered; we cannot from these results conclude that ToM contributes nothing, only the more modest claim that we have not yet tested its contribution in the live setting.

Two engineering realities shape what the official metrics can tell us. The live submission ran in two code regimes whose blend is reflected in the reported scores: Phase 1 with BERTopic active and the observer-view embedding live (rounds 0–19), and Phase 2 with both disabled (rounds 20–500); 50 of Run 0's 101 alerts fired under the richer Phase 1 configuration (Section 4.2). The adaptive threshold schedule was calibrated for evidence accumulation rather than early sensitivity (Section 3.6): training mean alert rounds (158 for Run 0, 114 for Run 1) were an order of magnitude later than the test medians (13 and 9), namely the schedule was conservative for the cohort we faced, and a more aggressive early-alert policy would likely improve ERDE_5 without material F_1 cost. A second calibration finding concerns the multi-scale Wasserstein design: the long scale fires rarely and does so on a control example that barely clears its threshold (Section 5.3); it is a candidate for removal or re-weighting that we have not yet evaluated. The wall-clock pressure of the live run drove two mid-stream engineering changes (Section 4.2). Neither was a limitation of the official metric, which is round-indexed, but together they produced the regime split.

Two further limitations concern the scope of what we have evaluated. DUET is evaluated on a single test cohort, namely 523 English-language Reddit subjects with a 17.4% positive rate, drawn from a specific time period and platform (Section 2). The training data is the eRisk 2025 Task 2 release, used here as a 2025-to-2026 proxy despite a meaningfully different positive rate (11.2% versus 17.4%); we have not assessed how the system generalises to other platforms, other languages, other demographics, or any clinical population. Two design-choice sensitivities are similarly untested. The BDI-II "Variant C" symptom phrasings (Appendix A) drive every cosine-based feature in Layer 1 and provide the reference embeddings against which the Wasserstein and thread-topic features are computed. We have not evaluated sensitivity to alternative formulations. The ordinal encoding of `community_response_type` into a $[0, 1]$ scalar (Section 3.4) is similarly a modelling assumption inherited from the codebase: we have not tested it against alternatives such as a one-hot representation, a learned embedding, or removal. Both are appendix-inspectable but empirically un-probed. A separate kind of design-time issue concerns the Layer-3 emotion block (Section 3.2): the code reserves eight slots named for Plutchik's primary emotions, while the model is trained on Ekman's six emotions plus a neutral class, so two emotion slots are structurally zero and the model's neutral score is dropped before renormalisation. The mismatch is internally consistent across training and inference, namely every classifier was trained on the 9-dimensional block in this configuration, but was not corrected for the submission because doing so would have required retraining every classifier on a different feature matrix.

7. Conclusion

We have presented DUET (Dual-view Evidence Tracking), an incremental per-user accumulator for contextualised early detection of depression on a stream of Reddit threads. DUET maintains a per-user state across rounds and folds new threads into a ≈2341-dimensional feature vector that combines target-text embeddings, community-channel signals, multi-scale Wasserstein temporal-shift detection, a

designed dual-mentalising Theory-of-Mind module, and a Thompson-sampled symptom personalisation. Two of our five submitted runs tied for rank 5 of 70 systems on $F_1 = 0.80$, four of the five runs reached perfect top-10 ranking quality from round 100 onwards with NDCG@100 = 0.89 at joint rank 3 in the field, and our ERDE5-optimised run reached $\text{ERDE}_5 = 0.15$ at a median true-positive alert latency of 9 rounds.

Three directions follow most directly from the present findings. The first is wiring the designed Option C ToM module into the live pipeline: this would eliminate the train-test feature mismatch documented in Section 5.5 and turn Run 4 into an interpretable test of the ToM contribution, namely the experiment we intended to run but could not given the operational constraints of the live submission. The second is optimising the decision threshold directly against ERDE_5 rather than against F_1, since the conservative calibration cost ERDE_5 without a corresponding F_1 benefit on the test cohort (Section 5.4). The third is engineering the live pipeline for per-round wall-clock constraints from day one, so the full designed feature set is available under operational pressure rather than being progressively disabled during the run (Section 4.2). A smaller follow-up is the multi-scale Wasserstein design: the long scale fires rarely and not class-discriminatively (Section 5.3), a candidate for removal or re-weighting in the next iteration.

Ethics Acknowledgement

Mental-health prediction from social-media text raises ethical concerns about consent, privacy, and downstream use; we address each explicitly here. The eRisk competition data is provided under the lab's participation agreement and is anonymised, namely it contains no original Reddit usernames, post identifiers, or other re-identification handles; DUET treats users as anonymous subject IDs throughout the pipeline; in this paper we further refer to specific subjects by anonymous tags (Subject A, B, C, D) rather than dataset IDs, and present only summary statistics or behavioural characterisations of their content as permitted by the data-use agreement. Training-time ToM feature extraction sent the formatted threads to the HuggingFace Inference API as part of the Option C pipeline (Section 3.4); the formatted threads contain only the anonymised competition content, and the API provider's data-handling policy applies to those calls. The model's JSON responses to these prompts may contain quoted or paraphrased text from the target user, but these responses are consumed only to produce the 47-dimensional Theory-of-Mind feature block and are not reproduced, persisted in any published artefact, or shared. Live inference uses no external LLM call. DUET is a research artefact and not a deployment-ready clinical instrument. Any operational use would require informed consent from subjects, independent clinical validation against trained raters, and human review of every alert. Consistent with this framing, no user text is reproduced verbatim in this paper; every illustrative subject is referenced by abstracted behavioural profile, and Figure 1's example trace is a constructed scenario, not a real case.

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 3 and 5. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, M. Mitchell, CLPsych 2015 Shared Task: Depression and PTSD on Twitter, in: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for

Computational Linguistics, Denver, Colorado, 2015, pp. 31–39. URL: http://aclweb.org/anthology/W15-1204. doi:10.3115/v1/W15-1204.

[2] 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.

[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, 2026.

[5] D. E. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use, in: N. Fuhr, P. Quaresma, T. Gonçalves, B. Larsen, K. Balog, C. Macdonald, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, volume 9822, Springer International Publishing, Cham, 2016, pp. 28–39. URL: http://link.springer.com/10.1007/978-3-319-44564-9_3. doi:10.1007/978-3-319-44564-9_3.

[6] 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.

[7] S. G. Burdisso, M. Errecalde, M. Montes-y Gómez, A text classification framework for simple and effective early depression detection over social media streams, Expert Systems with Applications 133 (2019) 182–197. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419303525. doi:10.1016/j.eswa.2019.05.023.

[8] J. M. Loyola, S. Burdisso, H. Thompson, L. C. Cagnina, M. Errecalde, {UNSL} at eRisk 2021: {A} Comparison of Three Early Alert Policies for Early Risk Detection, in: Proceedings of the Working Notes of {CLEF} 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021, volume 2936 of *{CEUR} Workshop Proceedings*, CEUR-WS.org, 2021, pp. 992–1021. URL: https://ceur-ws.org/Vol-2936/paper-81.pdf.

[9] M. Trotzek, S. Koitka, C. M. Friedrich, Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences, IEEE Transactions on Knowledge and Data Engineering 32 (2020) 588–601. URL: https://ieeexplore.ieee.org/document/8580405/. doi:10.1109/TKDE.2018.2885515.

[10] G. Peyré, M. Cuturi, Computational Optimal Transport with Applications to Data Sciences, Foundations and Trends® in Machine Learning 11 (2019) 355–607. URL: https://www.emerald.com/ftmal/article/11/5-6/355/1332396/Computational-Optimal-Transport-with-Applications. doi:10.1561/2200000073.

[11] 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.

[12] D. Premack, G. Woodruff, Does the chimpanzee have a theory of mind?, Behavioral and Brain Sciences 1 (1978) 515–526. URL: https://www.cambridge.org/core/product/identifier/S0140525X00076512/type/journal_article. doi:10.1017/S0140525X00076512.

[13] M. Kosinski, Evaluating large language models in theory of mind tasks, Proceedings of the National Academy of Sciences 121 (2024) e2405460121. URL: https://pnas.org/doi/10.1073/pnas.2405460121. doi:10.1073/pnas.2405460121.

[14] A. T. Corbett, J. R. Anderson, Knowledge tracing: Modeling the acquisition of procedural knowledge, User Modelling and User-Adapted Interaction 4 (1995) 253–278. URL: http://link.springer.com/10.1007/BF01099821. doi:10.1007/BF01099821.

[15] J. Parapar, A. P{\'{e}}rez, X. Wang, F. Crestani, Overview of eRisk 2025: Early Risk Prediction on the Internet (Extended Overview), 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. 1449–1473. URL: https://ceur-ws.org/Vol-4038/paper_117.pdf.

[16] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3980–3990. URL: https://www.aclweb.org/anthology/D19-1410. doi:10.18653/v1/D19-1410.

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

[18] 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.

[19] C. Hutto, E. Gilbert, VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text, Proceedings of the International AAAI Conference on Web and Social Media 8 (2014) 216–225. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14550. doi:10.1609/icwsm.v8i1.14550.

[20] j-hartmann/emotion-english-distilroberta-base · Hugging Face, 2022. URL: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base.

[21] M. Grootendorst, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, 2022. URL: https://arxiv.org/abs/2203.05794. doi:10.48550/ARXIV.2203.05794.

[22] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, T. Vayer, POT: Python Optimal Transport, Journal of Machine Learning Research 22 (2021) 1–8. URL: http://jmlr.org/papers/v22/20-451.html.

[23] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, Journal of Multivariate Analysis 88 (2004) 365–411. URL: https://linkinghub.elsevier.com/retrieve/pii/S0047259X03000964. doi:10.1016/S0047-259X(03)00096-4.

[24] 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. van der 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. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de 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. Srini-

vasan, 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. De 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. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, 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: https://arxiv.org/abs/2407.21783. doi:10.48550/ARXIV.2407.21783.

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6. doi:10.18653/v1/2020.emnlp-demos.6.

[26] T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, 2016, pp. 785–794. URL: https://dl.acm.org/doi/10.1145/2939672.2939785. doi:10.1145/2939672.2939785.
[27] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2017. URL: http://arxiv.org/abs/1412.6980. doi:10.48550/arXiv.1412.6980, arXiv:1412.6980.
[28] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: http://jmlr.org/papers/v12/pedregosa11a.html.

A. Appendix: ToM System Prompts

This appendix reproduces verbatim the two system prompts and the thread-transcript format used by the designed ToM module of Section 3.4, namely Prompt 1 (self-view) and Prompt 2a (observer-view, independent). These are the prompts that built `tom_features.npz`, the training feature matrix the shipped classifiers learned from (Section 4.1). The full source is available in the project repository at `src/erisk_task2/tom/prompts.py`.

A.1. BDI-II Variant C symptom definitions

The 21 symptom definitions below are inlined verbatim into both prompts. They originate in `layer1.py` (`SYMPTOM_DESCRIPTIONS`) and are formatted into the prompts at `ToMModule.__init__` time, so the resulting prompt strings are byte-identical across every call to the same prompt type.

```
1-Sadness: Persistent feelings of sorrow, unhappiness, or
   emotional pain.
2-Pessimism: Discouragement and hopelessness about the future.
3-Past failure: Feeling like a failure, seeing many
   disappointments.
4-Loss of pleasure: Reduced enjoyment of activities,
   hobbies, social life.
5-Guilty feelings: Excessive guilt, self-blame for
   bad things.
6-Punishment feelings: Expectation of punishment, sense that
   bad things are deserved.
7-Self-dislike: Self-criticism, disappointment in
   oneself as a person.
8-Self-criticalness: Harsh self-judgment for all
   faults and mistakes.
9-Suicidal thoughts: Thoughts of ending one's life,
   death wishes.
10-Crying: Increased tearfulness, uncontrollable
   emotional outbursts.
11-Agitation: Restlessness, irritability, inability
   to stay still or relax.
12-Loss of interest: Social withdrawal, apathy,
   not caring about things.
13-Indecisiveness: Difficulty making decisions,
   putting off choices.
14-Worthlessness: Profound sense of having no value,
   being useless.
```

```
15-Loss of energy: Fatigue, everything takes extra effort.
16-Sleep changes: Insomnia, oversleeping, or disrupted
   sleep patterns.
17-Irritability: Short temper, easily frustrated or angered.
18-Appetite changes: Eating much more or less,
   weight gain or loss.
19-Concentration difficulty: Brain fog, difficulty
   focusing, forgetfulness.
20-Tiredness/fatigue: Constant exhaustion, lack of
   motivation due to tiredness.
21-Loss of interest in sex: Reduced libido, no sexual desire.
```

The line wrapping above is editorial: the codebase emits each definition as a single unbroken line.

A.2. Prompt 1 — Self-view system prompt

Issued when the target user has at least one own contribution in the thread (`thread.has_target_text == True`). The system message below is sent as a single `role:"system"` message with the BDI-II definitions of Appendix A.1 inlined where indicated; the corresponding user message is the concatenation of every text the target authored in the thread, in chronological order, joined by blank lines.

System message (verbatim):

```
You are a clinical psychologist analyzing social media
text for indicators of depression. Assess the user's
mental state based SOLELY on their own writings. Report
only symptoms where you find textual evidence. Respond
with valid JSON only.

BDI-II Symptom Definitions:
  [21-symptom block from Appendix A.1, inlined verbatim]

Output format:
{
  "active_symptoms": {
    "<symptom_name>": {"score": <1-3>,
                       "evidence": "<brief quote or paraphrase>"},
    ...
  },
  "depression_probability": <0.0-1.0>,
  "overall_impression": "<1-2 sentences>"
}

Severity scale: 1=mild/possible, 2=moderate/clear,
3=severe/strong.
If no indicators found, return:
{"active_symptoms": {}, "depression_probability": 0.0,
 "overall_impression": "No indicators observed."}
```

User message template (verbatim):

```
The following texts were written by a single user in
an online discussion thread:

{target_user_texts}
```

The placeholder {target_user_texts} is replaced at call time by "\n\n".join(thread.target_texts), namely every text the target authored in the current thread, joined by blank lines, in chronological order.

A.3. Prompt 2a — Observer-view system prompt, independent

Issued for every thread, including silent threads where Prompt 1 is skipped. The system message inlines the BDI-II definitions of Appendix A.1 and asks the model to attend to the [TARGET] markers in the formatted transcript described in Appendix A.4. The corresponding user message is the priority-truncated transcript itself.

System message (verbatim):

```
You are a clinical psychologist observing an online
discussion. Assess how OTHER PEOPLE in the conversation
perceive and respond to a specific target user marked
with [TARGET]. Focus on what the community's reactions
reveal about the target user's mental state. Respond
with valid JSON only.

BDI-II Symptom Definitions:
  [21-symptom block from Appendix A.1, inlined verbatim]

Output format:
{
  "perceived_symptoms": {
    "<symptom_name>": {
       "score": <1-3>,
       "observer_signal": "<what in others' responses
                           suggests this>"
    },
    ...
  },
  "observer_concern_level": <0-3>,
  "community_response_type": "<concern|support|advice|
                              normalization|casual|mixed>",
  "depression_probability": <0.0-1.0>,
  "key_observation": "<1-2 sentences>"
}
```

User message: the priority-truncated transcript built by format_thread(), described in Appendix A.4.

The downstream encoding of the community_response_type categorical into a scalar in $[0, 1]$ is {concern: 1.0, support: 0.8, advice: 0.6, mixed: 0.5, normalization: 0.3, casual: 0.0}; this scalar occupies slot [46] of the ToM block (Section 3.4).

A.4. Thread transcript schema

The formatted transcript serves as the entire user message for Prompt 2a. It is built by format_thread() (thread_formatter.py) and applies a four-class priority truncation governed by a 2000-token budget (interpreted as approximately 8000 characters via a 4-chars-per-token heuristic, minus header and footer).

Priority classes.

P1 Target's own posts, and direct replies to the target. Kept in full. If even P1 overflows the budget, the offending line is truncated to the remaining budget (only if more than 50 characters remain) with a trailing `...`, and iteration stops.

P2 Other nodes inside reply branches that include the target. Kept in full if budget allows, omitted otherwise.

P3 Non-target nodes one hop further out from a target branch. Truncated to the first 100 characters with a `... [+{N} more chars]` suffix, or omitted if budget is exhausted.

P4 Nodes in branches with no target participation. Always omitted.

Per-node format. Each kept node is one of the following (the `[TARGET]` suffix appears only when the corresponding author is the target):

```
[POST] <author>[TARGET]: <body>
[REPLY to POST by <parent_author>[TARGET]]
    <author>[TARGET]: <body>
[REPLY to <parent_author>[TARGET]]
    <author>[TARGET]: <body>
[COMMENT] <author>[TARGET]: <body>
```

Envelope:

```
=== THREAD (Round <N>) ===
Title: <thread.title or "[no title]">

<P1 lines>
<P2 lines>
<P3 lines>

=== END THREAD ===
```

Silent-thread short-circuit. When the target contributed no text to the thread, the formatter emits the envelope above with body content replaced by:

```
[TARGET did not contribute text in this thread]
```

In this case, Prompt 1 is not sent (`self_view = None`) and only Prompt 2a is issued with the placeholder transcript above.

A.5. Dormant prompts (defined but not used at the default config)

The codebase defines three further prompts that are not invoked under the default configuration. None of them was used to build `tom_features.npz` or in the live submission. We reproduce them here for a complete record of what the codebase contains.

A.5.1. Prompt 2b — Observer-view, chained

Available when `chained=True`; the default config sets this to `False`, so Prompt 2b never fired in the training-feature build. Differs from Prompt 2a in that it ingests Prompt 1's parsed self-view response in the user message before producing observer scores, and its output schema includes an explicit `perspective_gap` object and an `insight_assessment` field.

System message (verbatim):

```
You are a clinical psychologist comparing two
perspectives on a social media user: how they present
themselves versus how others perceive them. The user is
marked with [TARGET]. Respond with valid JSON only.

BDI-II Symptom Definitions:
  [21-symptom block from Appendix A.1, inlined verbatim]

Output format:
{
  "perceived_symptoms": {
    "<symptom_name>": {"score": <1-3>,
                       "observer_signal": "<evidence>"},
    ...
  },
  "observer_concern_level": <0-3>,
  "community_response_type": "<concern|support|advice|
                              normalization|casual|mixed>",
  "depression_probability": <0.0-1.0>,
  "perspective_gap": {
    "self_higher": ["<symptoms user rates higher than
                     observers>"],
    "observer_higher": ["<symptoms observers detect but
                         user doesn't express>"],
    "alignment": "<aligned|user_minimizes|
                  user_exaggerates|mixed>"
  },
  "insight_assessment": "<1-2 sentences>"
}
```

User message template (verbatim):

```
A previous assessment of this user's OWN writings found:
{self_view_json}

Full conversation thread:
{formatted_thread}
```

The placeholder `{self_view_json}` is replaced at call time by the raw JSON of Prompt 1's parsed response, re-serialised via `json.dumps(self_view)`.

A.5.2. Prompt 3 — Severity assessor (BDI-II total score)

Defined in `prompts.py` but no call site in the Task 2 pipeline references it. Estimates depression severity on the 0–63 BDI-II total scale, explicitly conditioned on a base-rate prior.

System message (verbatim, with the BDI-II definitions inlined where `{symptom_definitions}` appears):

```
You are a depression screening tool. Given a user's
accumulated social media posts, estimate depression
severity on the BDI-II scale (0-63). Be calibrated: most
social media users are NOT depressed (base rate ~10%).
Respond with valid JSON only.
```

```
BDI-II severity categories:
  Minimal: 0-13
  Mild: 14-19
  Moderate: 20-28
  Severe: 29-63

{symptom_definitions}

Output format:
{
  "bdi_total": <0-63>,
  "severity": "<minimal|mild|moderate|severe>",
  "trajectory": "<stable|worsening|improving|
                 fluctuating>",
  "active_symptoms": ["<symptom1>", "<symptom2>", ...],
  "confidence": <0.0-1.0>
}
```

User message template (verbatim):

```
Posts by a single user, chronological order. Total:
{n_posts} posts over {current_round} rounds.

{accumulated_texts}
```

A.5.3. Prompt 4 — Response-category classifier (Option B)

Invoked only when `method="option_b"`, which the default config does not select. Classifies a single reply to a target text into one of seven categories. Output is plain text (no JSON), matched case-insensitively against the seven labels.

System message (verbatim):

```
Classify the reply into exactly one category reflecting
what it reveals about the replier's perception of the
original poster. Respond with the category label only,
nothing else.

Categories:
CONCERN - Expresses worry about the user's wellbeing
ADVICE - Suggests help, therapy, or coping strategies
EMOTIONAL_SUPPORT - Empathy, validation, comfort
NORMALIZATION - Suggests the situation is normal
SHARED_EXPERIENCE - Relates personal similar experience
PRACTICAL_SUPPORT - Offers specific resources
CASUAL - No mental health signal
```

User message template (verbatim):

```
Target user wrote: "{target_text}"
Someone replied: "{reply_text}"
```

The placeholders `{target_text}` and `{reply_text}` are each truncated to 200 characters before substitution.