Theory-Explicit Prompting for MIND Self-States: Hierarchical LLMs and Dynamic Signature Extraction in Mental Health Timelines

Pawan Kumar, Ankit Meshram, Shubham Jha, Loitongbam Gyanendro Singh
Department of Computer Science and Engineering,
Indian Institute of Technology Ropar, Punjab, India
{2024aim1007, 2024aim1002, 2024csm1019, gyanendro}@iitrpr.ac.in

Abstract

This paper presents a system for the CLPsych 2026 Shared Task on longitudinal mental health modeling from social media timelines, grounded in the MIND framework (Atzil-Slonim, 2025). MIND conceptualizes mental health as evolving self-states defined by **A**ffect, **B**ehavior, **C**ognition, and **D**esire (ABCD), providing a structured lens on mental health trajectories. The system centers on a *theory-explicit prompting framework* for structured sequence summarization (Task 3.1) and recurrent dynamic signature extraction (Task 3.2), encoding the full ABCD taxonomy directly into the LLM prompt to ensure clinically grounded, interpretable outputs. A three-stage pipeline infers a direction-of-change label per sequence, produces structured ABCD summaries with few-shot exemplar augmentation, and aggregates these summaries to derive cross-individual recurrent patterns. The system ranks *first* on deterioration-related recurrent signatures and *second* overall, achieving the top Fit and Specificity scores in Task 3.2, demonstrating the benefits of explicit clinical grounding for conceptual accuracy.

1 Introduction

Social media platforms, particularly Reddit, offer rich longitudinal records of self-disclosure that unfold over weeks, months, and years, enabling forms of continuous monitoring and early risk detection that single-point clinical assessments cannot easily provide (Coppersmith et al., 2014; Tsakalidis and Liakata, 2022). As a result, these platforms have become a key resource for computational mental health, supporting work on early warning signals, symptom trajectories, and crisis detection.

Prior work has examined diverse aspects of mental health–relevant language in user-generated content, including how linguistic and contextual properties vary across societal and non-societal topics (Singh and Singh, 2024), the extraction and summarization of suicidal ideation using large language models (Singh et al., 2024a), and the detection of critical moments of change through longitudinal and multi-task learning approaches (Singh et al., 2024b; Azim et al., 2022; Tsakalidis et al., 2022). While these methods demonstrate strong predictive performance, they typically target isolated subtasks and often remain weakly connected to established psychological theory, limiting interpretability, complicating clinical integration, and constraining our ability to model mental-state evolution in ways that are meaningful to clinicians.

The CLPsych 2026-25 Shared Task (Ali et al., 2026)(Tseriotou et al., 2025) addresses this gap by placing longitudinal mental health modeling within the MIND framework (Atzil-Slonim, 2025), which conceptualizes mental health as a dynamic configuration of *self-states* shaped by **A**ffect (A), **B**ehavior toward the self (B-S) and toward others (B-O), **C**ognition about the self (C-S) and about others (C-O), and **D**esire (D)–collectively referred to as ABCD. At any point in time, one self-state configuration may be dominant while others remain latent, and shifts between more adaptive and more maladaptive self-states are intended to reflect clinically meaningful patterns of deterioration or recovery. The task spans `40 Reddit-based user timelines` from mental health subreddits, split into 30 training and 10 test timelines (Ali et al., 2026), each consisting of chronologically ordered posts annotated according to the MIND scheme.

This paper concentrates on the sequence-level tasks of the CLPsych 2026 Shared Task (`Task 3`). Details on the task and the dataset are provided in Appendix A. `Task 3.1` requires generating structured natural-language summaries of the ABCD dynamics surrounding each detected change event (Switch or Escalation), while `Task 3.2` builds on these summaries to identify *recurrent* ABCD-grounded dynamic signatures that distinguish de-

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 547–553

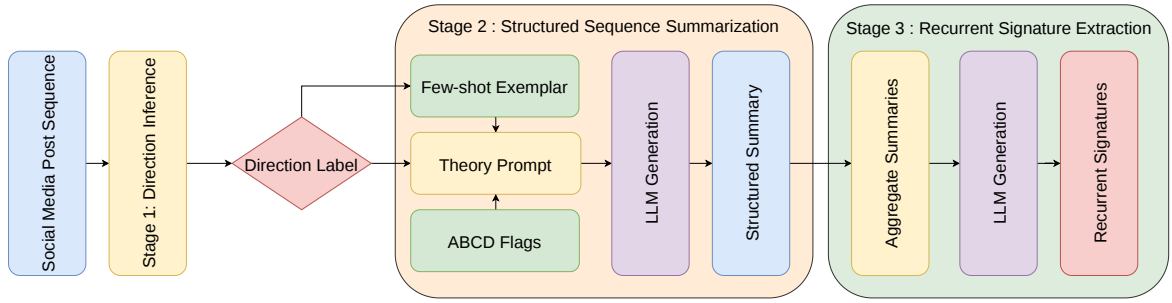

Figure 1: Three-stage pipeline for Task 3. **Purple** nodes denote LLM generation steps; **yellow** nodes denote prompt construction and aggregation; **grey** nodes denote external context inputs. Stage 1 infers a direction-of-change label from each social media post sequence. Stage 2 assembles a theory-explicit MIND/ABCD prompt-receiving both the post sequence and the direction label as input-with a matched few-shot exemplar and per-post annotations, and generates a structured summary (Task 3.1). Stage 3 aggregates training summaries by direction and extracts recurrent dynamic signatures across users (Task 3.2).

terioration from improvement across multiple individuals' trajectories. We address these tasks through a *theory-explicit prompting framework* that encodes the complete MIND/ABCD taxonomy directly into the system prompt to guide structured sequence summarization and signature extraction. Each summary is organized into three clinically motivated sections: *Central Theme*, *Within-State Dynamics*, and *Between-State Dynamics*, and uses formal ABCD abbreviations and relational vocabulary to support theory-aligned, clinically interpretable modeling of longitudinal mental health.

Building on this design, a three-stage pipeline drives the system: (1) a keyword-based heuristic infers a direction-of-change label (*deterioration*, *improvement*, or *mixed*) for each sequence; (2) this label conditions a two-part structured prompt, with a matched few-shot exemplar retrieved from training, that produces the Task 3.1 ABCD summary; and (3) these summaries are aggregated across users to perform cross-individual signature extraction in Task 3.2. Our pipeline achieves *first place* in identifying deterioration signatures and *second place* overall in Task 3.2, attaining the highest Fit Score (1.000 overall; 0.875 deterioration) and Specificity Score (0.938) of any participating team in the deterioration category. These results illustrate how theory-explicit prompting can bridge the gap between high-performing LLMs and psychologically interpretable characterizations of mental health trajectories.

2 Methodology: Summarization and Signature Extraction

Our sequence-level component uses a unified `theory-explicit prompt-based LLM pipeline` to address structured sequence summarization (Task 3.1) and recurrent pattern extraction (Task 3.2). All LLM generation steps use **Mistral-7B-Instruct-v0.2**, served via a locally hosted Ollama endpoint; no proprietary APIs were used at any stage, in compliance with the CLPsych 2026 Data Access Form. We encode the ABCD taxonomy, self-state definitions, and the required output format directly in the system prompt so that all generations are explicitly grounded in the MIND framework (Atzil-Slonim, 2025, 2026), following the role-prompting paradigm (Kong et al., 2024).

The pipeline has three stages: (1) infer a direction-of-change label for each sequence; (2) generate a structured summary for Task 3.1 conditioned on this label; and (3) aggregate Task 3.1 summaries over the training set to extract cross-individual signatures for Task 3.2. The full pipeline is illustrated in Figure 1.

2.1 Stage 1: Direction Inference

Each sequence is labeled as *deterioration*, *improvement*, or *mixed* by scanning its gold summary (training) for deterioration keywords (e.g., *worsen*, *suicidal*) and improvement keywords (e.g., *recover*, *hopeful*). The keyword lists were derived through close reading of the training gold summaries combined with established clinical terminology from the MIND framework manual (Atzil-Slonim, 2025); no automated selection procedure was used. Full keyword lists are provided in Ap-

pendix A.3. The higher count determines the label; ties yield *mixed*. For test sequences, the same heuristic is applied to the post texts. This direction labels both frames, the kind of change described in Task 3.1, and groups Task 3.1 summaries when building signatures in Task 3.2.

2.2 Stage 2: Task 3.1 – Structured Sequence Summarization

Given a direction label, we construct a two-part prompt following the in-context learning paradigm (Brown et al., 2020; Liu et al., 2023): a fixed **system part** and a sequence-specific **user part**. The **system part** asks the model to act as a clinician trained in MIND, following the role-prompting paradigm (Kong et al., 2024), and to produce a three-part summary:

- **Central Theme**: the dominant ABCD dimension driving the change and how it evolves across the sequence.
- **Within-State Dynamics**: interactions *within* self-states (e.g., co-activation, reinforcement, suppression) between ABCD subelements.
- **Between-State Dynamics**: how adaptive and maladaptive self-states change in relative dominance, culminating in a Switch or unfolding through an Escalation.

The model must use formal third-person clinical prose, explicitly reference ABCD elements with *(A)*, *(B-S)*, *(B-O)*, *(C-S)*, *(C-O)*, *(D)*, state the direction of change, and not reproduce numeric scores. Outputs are capped at 350 words to match the shared task truncation limit.

The **user part** provides sequence identifiers and change types (Switch / Escalation), the inferred direction label, and for each post: its index, inferred self-state valence (adaptive / maladaptive / neutral), ABCD presence flags with scores, and post text truncated to 500 characters. The 500-character limit was chosen empirically: pilot runs showed that content beyond this threshold introduced no additional ABCD signal not already captured by the structured flags, while substantially increasing token count and generation latency. For training sequences, ABCD flags are drawn from the gold annotations; for test sequences, they are produced by our Task 1 system, meaning Task 3.1 test performance reflects end-to-end pipeline quality including Task 1 prediction noise.

Few-Shot Augmentation. We add a `few-shot exemplar` to the user part, following the in-context learning paradigm (Brown et al., 2020; Liu et al., 2023). For a sequence with change type c and direction d, we retrieve a training summary whose (change type, direction) best matches (c, d) (exact match; otherwise, direction-only; otherwise, most recent). It is inserted under `EXAMPLE SUMMARY (for style reference only)`, followed by `Now generate a summary for the following new sequence`, preventing contamination of the target input.

The exemplar acts as a stylistic and structural template: it encourages the correct use of ABCD abbreviations, relational language (e.g., *mutually reinforces*, *progressively suppresses*), and the three-part structure. Decoding uses temperature $\tau = 0.3$, chosen after pilot runs comparing $\tau \in \{0.0, 0.1, 0.3, 0.5\}$: $\tau = 0.0$ produced overly templated outputs that reproduced exemplar phrasing too closely, while $\tau \geq 0.5$ introduced unnecessary variation in clinical terminology. The 700-token limit comfortably accommodates the 350-word requirement with formatting overhead. Outputs are split into three sections via regex over numbered headers; if parsing fails, the full output is kept as a single summary.

2.3 Stage 3: Task 3.2 – Recurrent Dynamic Signature Extraction

Stage 3 takes as input the structured summaries produced by Stage 2 over the full training set. We use Stage 2-generated summaries rather than the gold reference summaries, so that the signatures reflect the actual output distribution of our pipeline rather than the gold annotation; this ensures Task 3.2 signatures are grounded in the same self-state characterisations the pipeline produces at test time.

Summaries are grouped by direction label (*deterioration*, *improvement*, *mixed*). A single `aggregated prompt` then presents all summaries, organized by group and truncated to 800 characters each.

The model is instructed to report only signatures that occur in *at least two distinct user timelines*. For each signature, it must specify: (i) the ABCD interaction pattern (which subelements co-activate or suppress adaptive functioning); (ii) how adaptive vs. maladaptive dominance evolves before the change event; (iii) how the dynamics crystallize at the Switch or Escalation; and (iv) two to three training timeline IDs as evidence.

Team	CS↑	CT↓	R-L↑	BS↑	Rank
MERONYM_LABS	0.801	0.659	0.266	0.345	1
DreamerNLplus	0.735	0.767	0.285	0.345	2
CUNY	0.789	0.714	0.292	0.295	3
Ours	0.688	0.812	0.242	0.306	**11**
Baseline	0.765	0.749	0.262	0.231	–

Table 1: Results of Task 3.1 (selected top 3 teams vs. ours). CS = Consistency; CT = Contradiction; R-L = ROUGE-L Recall; BS = BERTScore Recall. Bold = best per column.

Signature extraction uses temperature $\tau = 0.2$, chosen to favour stable, high-level cross-individual patterns over idiosyncratic phrasing, with up to 1,500 new tokens to accommodate the six required signatures with their evidence citations. We parse outputs by splitting on deterioration signature and improvement signatures, then apply field-specific regex to extract each component. When parsing fails, we retain the raw output for manual inspection.

Team	Fit	Recur	Spec	Overall
Task 3.2 Deterioration extraction results				
Ours	**0.875**	0.563	**0.938**	**0.789**
Aurevia	0.688	0.813	0.563	0.676
DreamerNLplus	0.438	0.688	0.875	0.604
MeronymLabs	0.438	0.563	0.813	0.551
psytechlab	0.625	0.625	0.250	0.491
Baseline	0.563	0.375	0.438	0.483
Overall Task 3.2 competition results				
Ours	**1.000**	0.688	0.375	0.743
DreamerNLplus	0.625	0.813	1.000	**0.761**
MKC	0.750	0.563	0.938	0.727
McMasterNLP	0.688	0.375	0.750	0.594
psytechlab	0.688	1.000	0.250	0.544
Aurevia	0.375	0.625	0.500	0.465
Baseline	0.375	0.438	0.375	0.389

Table 2: Results for *Improvement* and *Deterioration*. Systems are ranked by *Overall*, which aggregates `Fit` (alignment with the ABCD framework), `Recur` (cross-user recurrence), and `Spec` (specificity vs. genericness). Our system ranks *first* in Deterioration and *second* in Improvement, with the highest `Fit` in both.

3 Results and Analysis

We focus our analysis on Task 3, where our system is most closely aligned with the MIND framework and achieves its strongest performance. Quantitative results are reported on the 10 test timelines provided by the shared task organizers, and we complement these with a qualitative analysis of the extracted deterioration signatures.

3.1 Task 3.1: Summarisation Quality

Table 1 reports the official Task 3.1 evaluation across four metrics: `Consistency` (CS)–absence of contradiction against gold summaries via NLI; `Contradiction` (CT)–maximum NLI contradiction probability (lower is better); `ROUGE-L Recall`–longest common subsequence recall; and `BERTScore Recall`–semantic token-level coverage. The final rank is based on the average score rank across all four metrics. Our system (rank 11) shows moderate CS (0.688) and BERTScore (0.306) with a relatively high CT (0.812), indicating that our ABCD-grounded summaries are internally coherent but diverge in phrasing from the gold annotations. The high CT score reflects phrasing divergence rather than factual contradiction—our summaries characterise the same clinical dynamics using different ABCD-grounded terminology. Systems with high ROUGE-L (e.g., USAI: 0.333) produce summaries with greater lexical overlap with the gold text, as ROUGE-L directly measures longest common subsequence recall against the reference. Our approach prioritises *conceptual* accuracy—correct identification and description of ABCD relational dynamics—over *lexical* reproduction of the gold phrasing. Evidence for this distinction lies in our Task 3.2 Fit Score (1.000): our summaries consistently use the required ABCD abbreviations, describe relational dynamics between subelements (e.g., *mutually reinforces*, *progressively suppresses*), and maintain the three required sections, even when specific phrasing diverges from the gold. This conceptual–lexical trade-off is a strong trade-off between the two tasks.

3.2 Task 3.2: Quantitative Performance

Tables 2 report the official Task 3.2 results. Scores assess: `Fit` (alignment with the ABCD clinical framework), `Recurrence` (cross-individual pattern grounding), and `Specificity` (non-genericity of the description).

MIND Framework Alignment Our perfect Fit Score (1.000 overall; 0.875 deterioration) reflects the effectiveness of theory-explicit prompting: encoding the ABCD taxonomy directly into the system part ensures that signatures closely match the clinical framework used for evaluation, outperforming all other teams in this dimension.

Fit–Specificity Trade-off Our lower Specificity Score in the overall ranking (0.375) reflects a trade-off inherent to our aggregation strategy. Presenting all training summaries simultaneously encourages cross-individual abstraction at the cost of per-case granularity. This contrasts with our high deterioration Specificity (0.938), suggesting that deterioration patterns are both more framework-aligned *and* more distinctive, while improvement patterns are harder to characterize specifically.

Task 3.1 vs. Task 3.2 Contrast A notable pattern emerges: our system performs relatively modestly on Task 3.1 (rank 11) but strongly on Task 3.2 (rank 2). This is consistent with the hypothesis that our approach optimizes for ABCD framework fidelity rather than lexical overlap with gold summaries. Task 3.1 is partly evaluated on ROUGE-L (lexical recall), which disadvantages summaries that correctly characterize the clinical dynamics but use different phrasing. Task 3.2 is evaluated on Fit, which directly rewards framework alignment, where our approach excels. This suggests that theory-explicit prompting is most beneficial for tasks requiring *conceptual* rather than *lexical* accuracy.

Deterioration vs. Improvement Asymmetry. Our strong deterioration performance relative to improvement suggests that maladaptive change follows more stereotyped ABCD patterns (self-criticism, hopelessness, and avoidance forming closed loops) that are robustly recurrent across individuals, while improvement signatures are more heterogeneous and harder to characterize across individuals. This asymmetry has implications for clinical monitoring: automated deterioration detection may generalize more reliably, while improvement characterization may require more individualized modeling.

4 Conclusion

We presented a theory-explicit prompting pipeline for longitudinal mental health modeling grounded in the MIND framework (Atzil-Slonim, 2025), encoding the ABCD taxonomy directly into the system prompt to produce clinically interpretable summaries and recurrent signatures. Our system ranked first in deterioration signature extraction and second overall in Task 3.2, achieving the highest Fit Score (1.000) across all teams–demonstrating that grounding NLP systems in established psychological theory is a performance-relevant design decision, particularly for tasks where conceptual accuracy matters more than lexical surface overlap. The contrasting Task 3.1 performance highlights a broader tension between framework fidelity and lexical proximity to reference annotations, a challenge that remains open for clinically grounded generation.

Ethics Statement

This work uses a dataset of social media posts from mental health subreddits, handled in compliance with the CLPsych 2026 data access agreement. No individual users are identified, and all post examples in this paper have been paraphrased as required by the shared task guidelines. The systems described are research prototypes not intended for clinical deployment without human oversight; automated characterization of mental health states carries risks of misclassification with serious consequences for vulnerable individuals. All LLM generation steps use **Mistral-7B-Instruct-v0.2**, served via a locally hosted Ollama endpoint. No proprietary APIs were used at any stage, in compliance with the CLPsych 2026 Data Access Form requirement prohibiting closed-source model use.

Limitations

Our dataset of 40 Reddit timelines limits statistical reliability; the direction inference heuristic may mislabel ambiguous sequences, introducing noise into Task 3.2 aggregation, and the 90-word signature limit constrains the expressiveness of ABCD interaction descriptions. For test sequences, per-post ABCD flags are produced by our Task 1 system rather than gold annotations, meaning Task 3.1 test performance reflects end-to-end pipeline noise. LLM outputs are non-deterministic; we report single inference passes without ensemble averaging. All results are specific to English-language Reddit posts from mental health subreddits and may not generalize to other languages, platforms, or clinical populations.

Acknowledgements

We thank the organizers of the CLPsych 2026 Shared Task for the annotated dataset and evaluation infrastructure, the MIND framework team for publicly releasing the ABCD annotation scheme, and the CSE Department of IIT Ropar for providing computational resources used in this research.

References

Iqra Ali, Talia Tseriotou, Guy Dvir, Callum Chan, Yuxiang Zhou, Juan Antonio Lossio-Ventura, Ayal Klein, Aya Shamir, Dan Sayda, Anthony Hills, Aya Zirikly, Diana Inkpen, Dana Atzil-Slonim, and Maria Liakata. 2026. Overview of the clpsych 2026 shared task: Capturing and characterizing mental health changes through social media timeline dynamics. In *Proceedings of the 11th Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Dana Atzil-Slonim. 2025. Multimodal intrapersonal and interpersonal dynamics (MIND): A transtheoretical coding manual.

Dana Atzil-Slonim. 2026. Leveraging theoretical and technological innovations to study the mechanisms that underlie therapeutic change in psychotherapy. In Louis G. Castonguay, Dana Atzil-Slonim, Michael Barkham, and Wolfgang Lutz, editors, *Practice-Based Evidence in the Psychological Therapies: Toward Policy Implications for Research, Training, and Clinical Guidelines*. Oxford University Press, New York.

Tayyaba Azim, Loitongbam Gyanendro Singh, and Stuart E Middleton. 2022. Detecting moments of change and suicidal risks in longitudinal user texts using multi-task learning. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 213–218.

Tom Brown and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, pages 51–60.

Aobo Kong and 1 others. 2024. Better zero-shot reasoning with role-play prompting. In *Proceedings of NAACL 2024*.

Pengfei Liu and 1 others. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Loitongbam Gyanendro Singh, Junyu Mao, Rudra Mutalik, and Stuart E Middleton. 2024a. Extracting and summarizing evidence of suicidal ideation in social media contents using large language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 218–226.

Loitongbam Gyanendro Singh, Stuart E. Middleton, Tayyaba Azim, Elena Nichele, Pinyi Lyu, and Santiago De Ossorno Garcia. 2024b. ConversationMoC: Encoding conversational dynamics using multiplex network for identifying moment of change in mood and mental health classification. In *Proceedings of the Machine Learning for Cognitive and Mental Health Workshop (ML4CMH 2024), co-located with AAAI 2024*, CEUR Workshop Proceedings, pages 42–56.

Loitongbam Gyanendro Singh and Sanasam Ranbir Singh. 2024. Characteristics of opinions in the societal and non-societal domains. *Social Network Analysis and Mining*, 14(1).

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, Jeff Leintz, and Maria Liakata. 2022. Overview of the CLPsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198, Seattle, USA. Association for Computational Linguistics.

Adam Tsakalidis and Maria Liakata. 2022. Mental health longitudinal modelling on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1–15.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Iqra Ali, Cian Kennedy, Guneet Singh Kohli, Anthony Hills, Ayah Zirikly, Dana Atzil-Slonim, and Maria Liakata. 2025. Overview of the CLPsych 2025 shared task: Capturing mental health dynamics from social media timelines. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 193–217, Albuquerque, New Mexico. Association for Computational Linguistics.

A Appendix

A.1 Task Overview

The CLPsych 2026 Shared Task (Ali et al., 2026) focuses on longitudinal mental health modeling within the MIND framework (Atzil-Slonim, 2025). This paper addresses the sequence-level tasks: structured summarization (Task 3.1) and recurrent signature extraction (Task 3.2), described below.

Task 3.1 – Sequence Summarisation. Given a chronologically ordered window of posts surrounding a detected change event (Switch or Escalation), systems generate a structured natural language summary describing how self-state dynamics evolve across the sequence and culminate in (for a Switch) or unfold through (for an Escalation) the change. Summaries describe: (1) the central recurring psychological theme in ABCD terms; (2) the dynamics within each self-state, including how ABCD subelements interact; and (3) the relationship between adaptive and maladaptive self-states, including patterns of dominance, suppression, or reflective dialogue. The direction of change (improvement or

deterioration) must be stated explicitly (Ali et al., 2026).

Task 3.2 – Recurrent Signature Extraction. Building on Task 3.1, systems identify and summarise *recurrent* patterns of change across multiple timelines. Teams extract three signatures of deterioration and three signatures of improvement, each describing a recurring dynamic pattern in ABCD and self-state structure that leads up to and culminates in the change event. These signatures must be grounded in observable self-state dynamics across multiple individuals, rather than a single case (Ali et al., 2026).

A.2 Dataset

The shared task dataset comprises `40 Reddit-based user timelines` drawn from mental health-related subreddits, split into 30 training timelines and 10 test timelines (Ali et al., 2026). Each timeline is a chronologically ordered sequence of posts annotated according to the MIND scheme (Atzil-Slonim, 2025). Training annotations include sequence-level summaries around Switch and Escalation events covering Central Theme, Within-State Dynamics, and Between-State Dynamics. Test data contain only post text and metadata; all annotations are withheld.

A.3 Direction Inference Keyword Lists

The following keywords were used for direction inference (Stage 1). Lists were derived through close reading of training gold summaries and MIND framework terminology (Atzil-Slonim, 2025); no automated selection procedure was used.

Deterioration keywords: *worsen*, *worsening*, *escalat*, *collapse*, *deteriorat*, *maladaptive dominant*, *suicid*, *hopeless*, *despair*, *suppressing adaptive*.

Improvement keywords: *recover*, *stabilise*, *stabiliz*, *adaptive dominant*, *positive shift*, *hopeful*, *improvement*, *adaptive gain*, *reflective dialogue*, *re-emerg*.

A.4 Prompt Sketches

Prompt sketches for Stages 2 and 3 are shown below. All examples have been paraphrased to protect user privacy. `[ABCD TAXONOMY]` denotes the full MIND subelement taxonomy (omitted for space); the complete taxonomy follows Atzil-Slonim (2025).

Stage 2 — System Prompt (abridged).

You are a clinical psychologist trained in the MIND framework. Given a sequence of social media posts surrounding a `{change_type}` event reflecting a `{direction}` in well-being, produce a structured clinical summary in three parts:

1. **Central Theme**
2. **Within-State Dynamics**
3. **Between-State Dynamics**

Use formal third-person prose. Reference all ABCD elements as: **(A)**, **(B-S)**, **(B-O)**, **(C-S)**, **(C-O)**, **(D)**. State the direction of change explicitly. Do not reproduce numeric presence scores. Maximum 350 words.
`[ABCDTAXONOMY]`

Stage 2 — User Prompt (abridged).

EXAMPLE SUMMARY *(for style reference only):*
`{gold_training_summary}`

Now generate a summary for the following new sequence.

Timeline: `{timeline_id}`
Sequence: `{sequence_id}`
Type: `{change_type}`
Direction: `{direction}`

Post 1 [*adaptive*; **A**:(5) hopeful, **D**:(1) relatedness]:
`{post_text[:500]}`

Post 2 [*maladaptive*; **A**:(4) depressed, **C-S**:(2) self-critical]:
`{post_text[:500]}`

. . .

Stage 3 — Aggregated Prompt (abridged).

You are a clinical psychologist. Below are summaries of change sequences, grouped by direction. Identify recurrent ABCD-grounded dynamic signatures appearing in **at least two distinct timelines**. For each signature, describe:

(i) The ABCD interaction pattern

(ii) How adaptive/maladaptive dominance shifts

(iii) How dynamics crystallise at the change event

(iv) 2–3 supporting timeline IDs

DETERIORATION SUMMARIES:
[Timeline X]: `{summary[:800]}`
[Timeline Y]: `{summary[:800]}`
. . .
IMPROVEMENT SUMMARIES:
. . .