Overview of the CLPsych 2026 Shared Task: Capturing and Characterizing Mental Health Changes through Social Media Timeline Dynamics

Iqra Ali[1*], **Talia Tseriotou**[1*], **Guy Dvir**[3*], **Callum Chan**[6], **Yuxiang Zhou**[1], **Juan Antonio Lossio-Ventura**[4], **Ayal Klein**[3], **Aya Shamir**[3], **Dan Sayda**[3], **Anthony Hills**[1], **Aya Zirikly**[5,7], **Diana Inkpen**[6], **Dana Atzil-Slonim**[3], **Maria Liakata**[1,2]
[1]Queen Mary University of London, [2]The Alan Turing Institute ,
[3]Bar Ilan University, [4]National Institutes of Health,
[5]The George Washington University, [6]University of Ottawa , [7]Johns Hopkins University
{i.ali; t.tseriotou; m.liakata}@qmul.ac.uk

Abstract

We provide an overview of the CLPsych 2026 Shared Task, which focuses on capturing and characterizing mental health dynamics from social media timelines through structured modeling of self-states. This year advances the longitudinal paradigm set by prior CLPsych shared tasks (2022, 2025), by integrating fine-grained psychological representation using the MIND framework. The task is organized into three main components: **(1)** post-level identification of adaptive and maladaptive self-states through ABCD elements and sub-elements, along with estimation of their presence; **(2)** timeline-level detection of Moments of Change, including both abrupt switches and gradual escalations based on ABCd element and sub-element combinations; and **(3)** sequence-level modeling, involving summarization of change processes over time and identification of recurrent dynamic signatures.

1 Introduction

Computational linguistics and clinical psychology have increasingly converged to study how language reflects mental health, psychological functioning, and change over time (Montejo-Ráez et al., 2024; Villarreal-Zegarra et al., 2024; Hua et al., 2025). Within this space, CLPsych has played a central role by bringing together NLP researchers and mental health experts to develop methods for modeling meaning, behavior, and psychological signals in language (Calvo et al., 2017; Le Glaz et al., 2021; Zirikly and Dredze, 2022). This collaboration has shifted the field from static risk detection toward modeling mental health as a dynamic and longitudinal process (Tsakalidis et al., 2022a; Tseriotou et al., 2025). Such an evolution in perspective is crucial, as mental health unravels through fluctuating affective, behavioral, cognitive, and motivational processes shaped by context over time (Owen et al., 2024; Bucur et al., 2026). Social media provides a natural setting for observing these dynamics through temporally ordered self-disclosures of distress, coping, and interpersonal experiences (Garg, 2023). CLPsych shared tasks have reflected this progression. Early tasks focused on user-level classification, while CLPsych 2022 introduced the longitudinal notion of *Moments of Change* (MoC), distinguishing sudden *switches* from gradual *escalations* (Tsakalidis et al., 2022a). CLPsych 2025 extended this paradigm with an emphasis on explainability through evidence extraction and summarization of self-state processes (Tseriotou et al., 2025). Building on this, CLPsych 2026 advances longitudinal modeling by focusing on fine-grained psychological dynamics underlying mental health changes (Figure 1). The task is grounded in the MIND framework (Atzil-Slonim, 2025, 2026), which represents self-states as combinations of Affect, Behavior, Cognition, and Desire (ABCD), further decomposed into adaptive and maladaptive sub-components pertaining to different dimensions such as need and relation with self and others. Mental health change is modeled as shifts in the dominance and interaction of these self-states over time.

The shared task comprises three complementary components (Figure 2): (1) post-level modeling of adaptive and maladaptive self-states, (2) longitudinal detection of *Moments of Change* (MoC) , and (3) structured summarization of mental health dynamics. The dataset consists of expert-annotated Reddit timelines with self-state annotations, well-being scores, and change labels. Together, the tasks frame mental health modeling as structured, explainable, and longitudinal inference rather than isolated classification (Tseriotou et al., 2025; Atzil-Slonim, 2025, 2026). From a computational perspective, the shared task introduces challenges in multi-label classification, temporal modeling, and explainable summarization, while contributing to-

[*]Denotes equal contribution.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 389–421

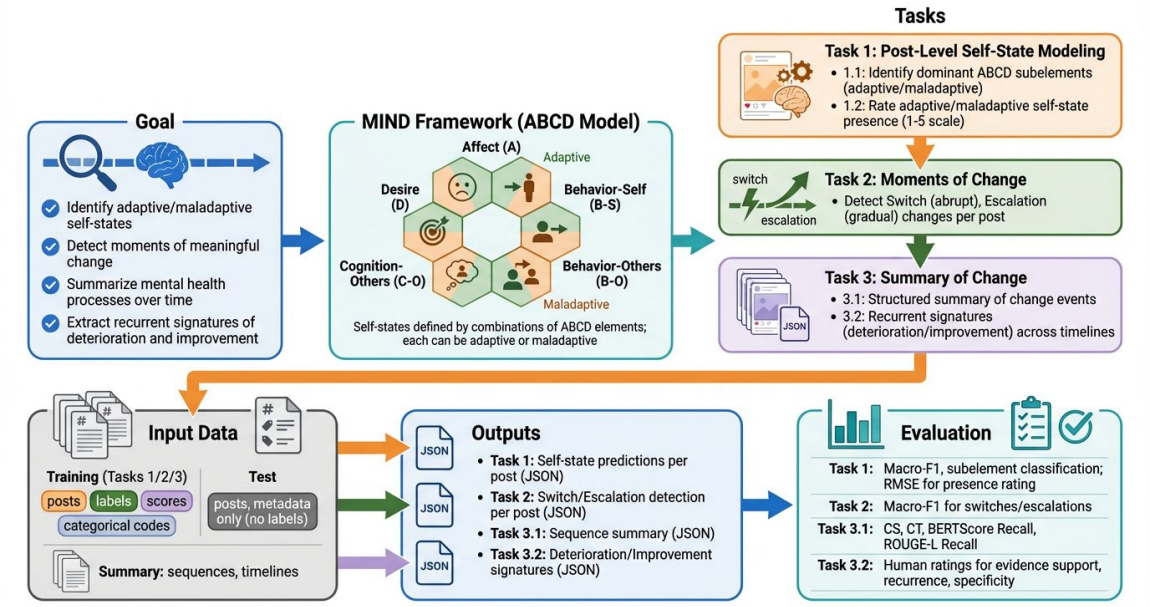

Figure 1: Overview of the CLPsych 2026 Shared Task for modeling mental health changes from longitudinal social media timelines. The framework is grounded in the MIND (ABCD) model and comprises three tasks: (1) post-level self-state identification and presence rating, (2) detection of moments of change (switches and escalations), and (3) longitudinal summarization of deterioration/improvement patterns. Systems are evaluated using classification, regression, correlation, summarization, and human-centered metrics including Macro-F1, RMSE, Spearman correlation, ROUGE recall, BERTScore recall, and evidence-based human judgments.

ward NLP systems that model mental health as dynamic and contextually grounded (Tsakalidis et al., 2022a; Tseriotou et al., 2025).

2 Related Work

From classification to structured representation of mental health dynamics: Early NLP work framed mental health as document- or user-level classification (Coppersmith et al., 2015; Shing et al., 2018; Zirikly et al., 2019; Sawhney et al., 2022b,a). Subsequent work shifted to longitudinal monitoring of mood changes (Hills et al., 2023; Tseriotou et al., 2024a) with a parallel line of work focused on interpretability through symptom-informed modeling (Nguyen et al., 2022), evidence-based prediction (Parapar et al., 2023; Garg, 2024; Chim et al., 2024), and explanation methods such as SHAP (Lundberg and Lee, 2017). CLPsych 2026 combines these two directions by modeling interacting self-state components in an interpretable manner.

LLMs for structured reasoning in mental health: LLMs enable flexible reasoning over mental health narratives, supporting classification (Amin et al., 2023), data augmentation (Liyanage et al., 2023), contextual inference (Xu et al., 2024a; Yang et al., 2023), and explanation (Xu et al., 2024b; Chim et al., 2024). However, hallucinations and reliability issues persist despite instruction tuning and Chain-of-Thought prompting (Li et al., 2023). This motivates structured, schema-constrained outputs, as adopted in CLPsych 2026 via the ABCD self-state framework and hybrid temporal modelling.

Grounding predictions in self-state evidence: Evidence extraction improves interpretability (Xu et al., 2024a; Yang et al., 2024; Xu et al., 2024c; Chim et al., 2024), but typically identifies supporting spans without linking them to structured psychological constructs. CLPsych 2026 aims to ground predictions in predefined ABCD components and their composition into adaptive and maladaptive self-states.

Structured summarization of change processes: Summarization has been applied to social media timelines (Sotudeh et al., 2022; Song et al., 2024, 2025), counseling (Srivastava et al., 2022), and clinical dialogues (Michalopoulos et al., 2022; Adhikary et al., 2024). However, capturing clinically meaningful temporal and psychological dynamics remains challenging (Klein et al., 2024; Asgari et al., 2024). CLPsych 2026 aims to advance this by requiring structured summaries of self-state config-

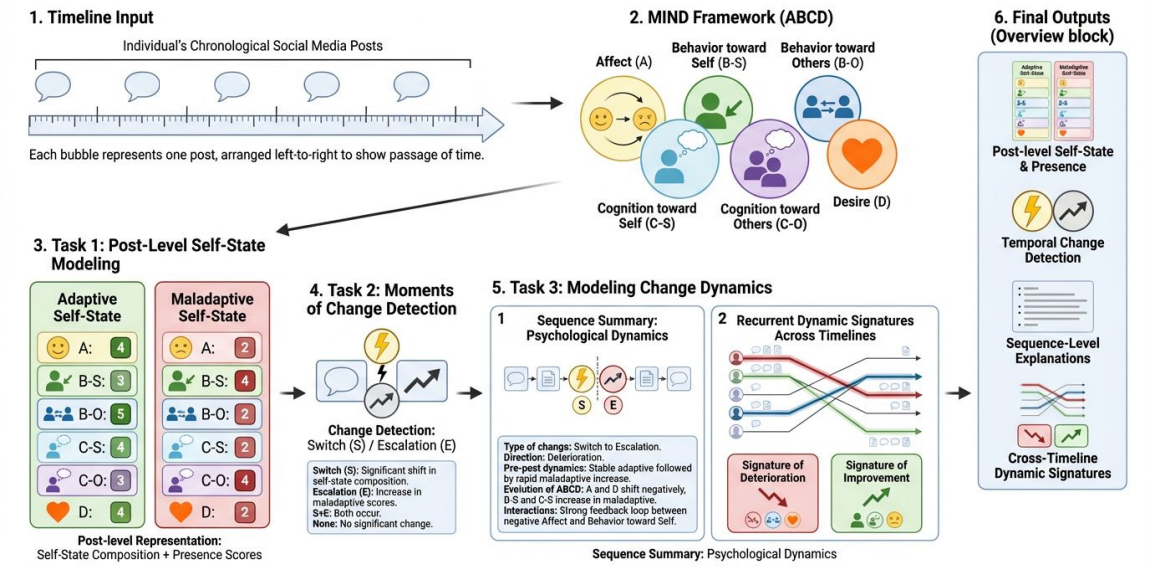

Figure 2: Illustration of the CLPsych 2026 shared task pipeline for modeling mental health dynamics from longitudinal social media timelines. The framework maps chronological posts to adaptive and maladaptive self-states using the MIND (ABCD) model, followed by detection of moments of change (switches and escalations) and sequence-level modeling of deterioration/improvement dynamics. Final system outputs include post-level self-state predictions, temporal change detection, sequence summaries, and cross-timeline mental health signatures.

urations, their evolution, and their role in change.

Modeling temporal dynamics and moments of change: Temporal modeling has progressed from RNNs and more standard sequential methods (Tseriotou et al., 2023; Bayram and Benhiba, 2022; Homan et al., 2022; Chim et al., 2025) to time-aware approaches such as Hawkes process models (Hills et al., 2023), hierarchical transformers (Hills et al., 2024), and temporally aware transformers (Tseriotou et al., 2024b). These are shown to improve detection of gradual and abrupt changes, aligning with Moments of Change (Tsakalidis et al., 2022a). CLPsych 2026 extends this line of work by jointly modeling temporal dynamics (i.e. switches and escalations) and structured self-state processes (i.e. ABCD elements).

Self-State Modeling and the MIND Framework for Dynamic Mental Health: Mental health is increasingly viewed as a dynamic system of interacting psychological components (Hofmann and Hayes, 2019; Bickman, 2020). The MIND framework (Atzil-Slonim, 2025, 2026) models this through self-states composed of Affect, Behavior, Cognition, and Desire (ABCD), where mental health change reflects shifts in the dominance and interaction of competing states over time (Stiles, 2001; Revelle, 2007; Bromberg, 2014; Beck et al., 2021; Lazarus and Rafaeli, 2023). Most NLP work models isolated components such as emotion (Mayer et al., 2024) or cognition (Singh et al., 2024), rather than their temporal interactions. Recent CLPsych shared tasks have been bridging this gap through longitudinal modeling of mental states (Tsakalidis et al., 2022a; Tseriotou et al., 2025).

3 Task Definition and Instructions

Task 1.1: ABCD subelement identification and self-state composition Given a post, participants identify dominant adaptive and/or maladaptive ABCD subelements and combine them into self-state compositions. For each element: Affect (A), Behavior toward others (B-O), Behavior toward self (B-S), Cognition toward others (C-O), Cognition toward self (C-S), and Desire (D), a single dominant subelement is selected per valence. Posts may contain adaptive, maladaptive, both, or no self-states. Categorization in Appendix Table 10.

Task 1.2: Self-state presence rating Participants estimate adaptive and maladaptive self-state presence on a 1–5 scale, reflecting psychological centrality rather than lexical frequency. Both states may score highly. Evaluated as a regression task.

Task 2: Identify Moments of Change (MoC) Task 2 focuses on timeline-level detection of state

changes. Given chronological post sequences, participants identify **Switches**, defined as sudden well-being changes of ≥ 2 between consecutive posts (or the nearest prior post within one week), and **Escalations**, defined as gradual intensification across consecutive posts. Well-being scores follow rescaled GAF criteria (1–10). Posts can denote Switch, Escalation, both, or neither.

Task 3.1: Sequence-level summarization Given a sequence surrounding a change event (e.g. Switch), participants generate a structured summary describing the evolution of self-state dynamics. The exact change point must be inferred. Summaries aim to capture the main psychological theme, intra- and inter-state dynamics, and, when applicable, the change type (Switch/Escalation), direction, and timing. **Task 3.2: Cross-timeline dynamic signatures** Participants identify recurrent *Signatures of Deterioration* and *Signatures of Improvement* across timelines based on the MIND framework. Each signature describes a recurring dynamic pattern (within- or between self-states) and is supported by at least two sequences (with identifiers). Individual task details can be found in Fig.2, while Fig.1 provides an overview of the entire CLPsych 2026 shared task.

4 Data

The CLPsych 2026 dataset extends the 40 gold-standard Reddit timelines from CLPsych 2025 (Tseriotou et al., 2025), originally derived from the Reddit-New dataset introduced in CLPsych 2022 (Tsakalidis et al., 2022b). It contains 30 training and 10 test timelines with post-level annotations of adaptive/maladaptive ABCD subelements, well-being scores, self-state presence scores, and MIND-grounded summaries. For CLPsych 2026, 36 additional posts were added to complete clinically meaningful windows around Moments of Change (MoC), including Escalation spans and pre-change context. The dataset supports three levels of modeling aligned with the shared tasks (Section 3; Table 11): post-level self-state modeling (Task 1), timeline-level MoC detection of *Switches* and *Escalations* (Task 2), and change-centered sequence modeling and summarization (Task 3) (Details in Appendix A.

5 Evaluation Metrics

Task 1.1: ABCD Classification: Element presence uses binary F1; subelements use multiclass F1, macro-averaged into Adaptive/Maladaptive scores.

$$\textbf{Score} = \frac{F1_{\text{adapt}} + F1_{\text{maladapt}}}{2}$$

Task 1.2: Presence Rating: We report MAE, RMSE, QWK, and Spearman; ranking is based on averaged RMSE (lower is better).

$$\textbf{Score} = \frac{\text{RMSE}_{\text{adapt}} + \text{RMSE}_{\text{maladapt}}}{2}$$

Task 2: Moments of Change: We compute F1 at post-level (pooled) and timeline-level (macro over timelines).

$$\textbf{F1}_{\text{post}} = \frac{F1_S^{\text{post}} + F1_E^{\text{post}}}{2}$$

$$\textbf{F1}_{\text{tl}} = \frac{1}{2}\left(\frac{1}{T}\sum_{t=1}^{T} F1_{S,t}^{\text{tl}} + \frac{1}{T}\sum_{t=1}^{T} F1_{E,t}^{\text{tl}}\right)$$

$$\textbf{Score} = \frac{F1_{\text{post}} + F1_{\text{tl}}}{2}$$

Task 3.1: Sequence Summarization: Evaluate summaries ($\leq$350 words) using: CS, CT, BERTScore Recall, and ROUGE-L Recall.

$$\textbf{CS} = \frac{1}{|S|}\sum_{s\in S}\frac{1}{|G|}\sum_{g\in G}(1 - \text{NLI}(\text{Contr} \mid g, s))$$

$$\textbf{CT} = \frac{1}{|S|}\sum_{s\in S}\max_{g\in G}\text{NLI}(\text{Contr} \mid g, s)$$

$$\textbf{BS} = \frac{1}{|G|}\sum_{g\in G}\max_{s\in S}\text{BERTScore}(g, s)$$

Task 3.2: Dynamic Signatures: Evaluate deterioration/improvement signatures via human ratings of Fit, Recurrence, and Specificity.

$$\textbf{Score} = 0.5 \cdot \text{Fit} + 0.5 \cdot \text{HMean}(\text{Rec}, \text{Spec})$$

6 Codabench Submissions

Participants submitted JSON predictions via Codabench for Tasks 1, 2, and 3.1 (max three submissions), covering all test instances without raw text. Codabench handled automatic evaluation and leaderboard display. Task 3.2 was submitted by email with deterioration and improvement signatures and evaluated manually (Appendix C).

7 Teams & Results

Participating Teams: 39 teams (93 participants) registered, with 26 having prior participation in CLPsych shared tasks. Of these, 20 teams submitted outputs for at least one task and 13 submitted system papers (Table 12). Participation varied by task: 17 teams for Tasks 1.1/1.2, 18 for Task 2, 13 for Task 3.1, and 9 for Task 3.2.

7.1 Baselines

We use `Llama-3.1-8B-Instruct` and `TempoFormer` baselines; details are provided in Appendix D. Most experiments with the Llama models use zero-shot prompting and follow the

prompt structures proposed by Chan et al. (2025) and Lossio-Ventura et al. (2025).

Task 1: A one-shot structured prompting baseline with `Llama-3.1-8B-Instruct`, predicts adaptive/-maladaptive subelements and presence jointly. The prompt includes a single demonstration example and enforces strict JSON output. Both subtasks involve a single post using explicit label definitions.

Task 2: We provide three baselines: (1) zero-shot prompting to derive well-being scores and detect switches/escalations, (2) a pipeline using Task 1 outputs to compute normalized well-being and detect changes via thresholds and monotonic trends, and (3) fine-tuning `TempoFormer` on temporal post sequences. The latter models longitudinal dynamics directly. Switch and escalation detection are handled separately.

Task 3.1: Two baselines are provided: (1) zero-shot prompting with `Llama-3.1-8B-Instruct` to generate summaries capturing themes and self-state dynamics, and (2) a pipeline that augments prompts with Task 1 and Task 2 predictions. This includes subelements, presence scores, and change labels per post. The enriched context improves temporal coherence and interpretability.

Task 3.2: A zero-shot prompting baseline with `Llama-3.1-8B-Instruct` generates one deterioration and one improvement signature. The model uses structured prompts over sequence summaries to extract themes and self-state dynamics. Outputs are concise (<90 words).

Team	R	AvgSub	EP-A	EP-M	EP-Avg	SE-A	SE-M
CUNY	1	**0.442**	**0.613**	**0.749**	**0.681**	**0.388**	0.496
StateOfMIND	2	0.441	0.554	0.706	0.630	0.346	**0.537**
Meronym Labs	3	0.420	0.542	0.735	0.638	0.323	0.517
USAI	4	0.410	0.549	0.736	0.642	0.333	0.487
Aurevia	5	0.381	0.546	0.680	0.613	0.328	0.434
MKC	6	0.361	0.496	0.681	0.588	0.295	0.427
McMasterNLP	7	0.351	0.541	0.666	0.603	0.317	0.385
ull	8	0.335	0.478	0.546	0.512	0.330	0.340
BLUE	9	0.318	0.358	0.643	0.501	0.188	0.448
NoviceTrio	10	0.313	0.514	0.606	0.560	0.281	0.344
Afrilan	11	0.297	0.517	0.600	0.559	0.286	0.308
psytechlab	12	0.274	0.439	0.429	0.434	0.263	0.285
DreamerNLplus	13	0.254	0.440	0.437	0.439	0.266	0.241
CtbuY	14	0.236	0.392	0.479	0.436	0.204	0.268
debju	15	0.195	0.286	0.355	0.320	0.148	0.242
DrosophilAI	16	0.179	0.000	0.642	0.321	0.000	0.358
CSE_IIT_Ropar	17	0.175	0.209	0.323	0.266	0.149	0.202
Baseline	–	0.247	0.392	0.605	0.498	0.156	0.338

Table 1: Task 1.1 results. R = rank; AvgSub = average subelement macro F1; EP = element presence; SE = subelement macro F1; A/M = adaptive/maladaptive. Higher scores indicate better performance. Systems are ranked by AvgSub.

Team	R	Avg	M	A	QWK	C-RMSE	Spr.	MAE
Meronym Labs	1	**0.917**	**0.833**	1.000	0.677	**0.920**	0.697	**0.667**
USAI	2	0.942	0.905	0.979	**0.730**	0.943	**0.752**	0.681
Aurevia	3	0.981	1.027	**0.935**	0.645	0.982	0.674	0.688
CUNY	4	0.989	0.866	1.112	0.682	0.997	0.690	0.688
StateofMIND	5	0.994	0.858	1.130	0.695	1.003	0.702	0.701
MKC	6	1.010	1.014	1.007	0.570	1.010	0.604	0.757
McMasterNLP	7	1.035	0.957	1.112	0.526	1.037	0.570	0.799
CSE_IIT_Ropar	8	1.146	1.230	1.061	0.367	1.149	0.418	0.833
DrosophilAI	9	1.154	1.027	1.280	0.571	1.161	0.562	0.819
debju	10	1.156	1.264	1.047	0.232	1.161	0.361	0.875
ull	11	1.171	1.236	1.106	0.355	1.173	0.395	0.875
NoviceTrio	12	1.255	1.344	1.167	0.553	1.258	0.578	0.944
DreamerNLplus	13	1.285	1.312	1.258	0.411	1.286	0.416	0.931
Afrilan	14	1.329	1.219	1.439	0.340	1.333	0.404	1.014
psytechlab	15	1.407	1.389	1.424	0.367	1.407	0.376	1.063
CtbuY	16	1.501	1.572	1.429	0.480	1.502	0.591	1.160
BLUE	17	1.606	1.667	1.546	0.230	1.607	0.423	1.194
Baseline	–	1.424	1.439	1.409	0.555	1.424	0.603	1.083

Table 2: Task 1.2: Avg = average RMSE; M/A = maladaptive/adaptive RMSE; QWK = quadratic weighted kappa; Spr. = Spearman correlation. Lower Avg, RMSE, C-RMSE, and MAE are better; higher QWK and Spr. are better. Bold and underline indicate best and second-best results.

7.2 Results

We summarize system performance based on best runs (see Appendix C for full details).

System Characteristics: Most systems used pipeline designs where earlier task outputs informed later stages, often combined with RAG or hybrid frameworks.

Model Characteristics: All teams used LLMs, with fine-tuned transformers for structured prediction and LLMs for reasoning and summarization. Some incorporated lightweight or hybrid models, while temporal modeling ranged from explicit sequence models to prompt-based methods.

Task 1.1: Table 1 shows ABCD composition results ranked by Avg Subelement F1. **CUNY** (0.442) and **StateOfMIND** (0.441) lead, followed by **Meronym Labs** and **USAI**. Element presence consistently outperforms subelement prediction, highlighting difficulty in fine-grained classification. Maladaptive elements are easier to detect than adaptive (e.g., 0.749 vs. 0.613 for **CUNY**). Adaptive subelements show the largest performance drop, reflecting subtle and context-dependent patterns.

Task 1.2: Table 2 reports presence rating results ranked by Avg RMSE. **Meronym Labs** (0.917) leads, followed by **USAI** and **Aurevia**. Performance varies by dimension: best maladaptive RMSE (0.833) from **Meronym Labs**, best adaptive (0.935) from **Aurevia**. **USAI** shows strongest agreement (QWK 0.730, Spearman 0.752). Overall errors remain high, indicating task difficulty.

Team	R	Comb.	P-Esc	P-Mac	P-Sw	T-Esc	T-Mac	T-Sw
USAI	1	**0.600**	0.694	**0.639**	**0.583**	0.611	0.561	0.510
CtbuY	2	0.588	0.706	0.611	0.516	0.690	0.564	0.439
JNLP	3	0.580	**0.738**	0.615	0.492	0.649	0.545	0.441
CUNY	4	0.572	0.714	0.581	0.448	0.709	0.563	0.416
MKC	5	0.554	0.615	0.587	0.558	0.527	0.521	**0.514**
Codezone	6	0.553	0.636	0.504	0.372	**0.719**	**0.602**	0.484
Aurevia	7	0.484	0.667	0.556	0.444	0.416	0.413	0.411
Meronym Labs	8	0.466	0.596	0.510	0.423	0.452	0.421	0.391
debju	9	0.447	0.511	0.510	0.510	0.302	0.385	0.467
DreamerNLplus	10	0.442	0.550	0.491	0.432	0.387	0.392	0.397
McMasterNLP	11	0.412	0.375	0.357	0.339	0.660	0.467	0.274
BLUE	12	0.403	0.447	0.439	0.431	0.348	0.367	0.386
NoviceTrio	13	0.385	0.414	0.405	0.396	0.338	0.365	0.392
DrosophilAI	14	0.383	0.524	0.433	0.341	0.324	0.334	0.344
psytechlab	15	0.372	0.409	0.371	0.333	0.507	0.372	0.237
CSE_IIT_Ropar	16	0.357	0.276	0.284	0.293	0.562	0.430	0.299
Afrilan	17	0.268	0.217	0.286	0.356	0.221	0.250	0.279
Cloud	18	0.260	0.465	0.264	0.063	0.380	0.257	0.133
TempoFormer	–	0.572	0.667	0.561	0.456	**0.736**	0.583	0.430
Weighted Sum + T1.1	–	0.365	0.407	0.383	0.359	0.343	0.346	0.350
Llama ZS	–	0.272	0.000	0.169	0.337	0.400	0.376	0.352

Table 3: Task 2 results. R = rank; Comb. = combined post/timeline macro F1; P/T = post/timeline-level evaluation; Esc = escalation; Sw = switch; Mac = macro F1. Higher scores indicate better performance. Systems are ranked by Comb.

Task 2: Table 3 shows MoC detection results ranked by combined F1. **USAI** (0.600) leads, followed by **CtbuY** and **JNLP**. Systems specialize: **JNLP** excels in escalation (0.738), **USAI** in switches (0.583). At timeline level, **Codezone** (Esc) and **MKC** (Switch) perform best. Joint modeling of post- and timeline-level dynamics remains challenging. The high performance of the Tempoformer baseline shows the importance of architectures incorporating temporal awareness by design.

Task 3.1: Table 4 reports summarization results ranked by AvgRank. **Meronym Labs** leads, followed by **DreamerNLplus** and **CUNY**. Metrics vary: **Aurevia** has highest consistency, **psytechlab** lowest contradiction, and **USAI** best ROUGE/BERTScore. High overlap does not guarantee better ranking, showing trade-offs between faithfulness and similarity. Generating both consistent and aligned summaries remains difficult.

Task 3.2: Table 5 reports explanation generation results. For improvement, **DreamerNLplus** leads (0.7608), with complementary strengths across teams (fit, recurrence, specificity). For deterioration, **CSE_IIT_Ropar** leads (0.7891), though variability is higher. Some systems achieve perfect scores in one dimension but underperform overall. Task 3.2 evaluated signatures using *Fit*, *Recurrence*, and *Specificity*, with final rankings combining Fit and the harmonic mean of Recurrence and Specificity. For improvement, **DreamerNLplus ranked first** through high Specificity and strong Recurrence, followed by **CSE-IIT-Ropar** and **MKC**. For

Team	R	CS	CT	R-L	BERT-R	CS_r	CT_r	$R\text{-}L_r$	$BERT\text{-}R_r$	AvgRank
MERONYM_LABS	1	0.801	0.659	0.266	0.345	3	3	6	4	**4.0**
DreamerNLplus	2	0.735	0.767	0.285	0.345	7	7	4	3	5.3
CUNY	3	0.789	0.714	0.292	0.295	5	5	3	9	5.5
NoviceTrio	4	0.705	0.771	0.318	0.341	8	8	2	5	5.8
USAI	5	0.681	0.849	**0.333**	**0.365**	10	13	1	1	6.3
Aurevia	5	**0.866**	0.625	0.185	0.226	1	2	11	11	6.3
MKC	7	0.654	0.834	0.284	0.359	11	10	5	2	7.0
psytechlab	8	0.857	**0.571**	0.078	0.147	2	1	13	13	7.3
JNLP	9	0.791	0.666	0.117	0.164	4	4	12	12	8.0
McMasterNLP	9	0.770	0.761	0.208	0.255	6	6	10	10	8.0
CSE_IIT_Ropar	11	0.688	0.812	0.242	0.306	9	9	8	8	8.5
ULL	12	0.585	0.846	0.262	0.320	13	11	7	6	9.3
CtbuY	13	0.615	0.848	0.232	0.317	12	12	9	7	10.0
Baseline-1	–	0.763	0.753	0.255	0.226	8	7	9	13	9.3
Baseline-2	–	0.767	0.745	0.269	0.235	7	6	6	11	7.5

Table 4: Task 3.1 results. R = overall rank; CS = consistency; CT = contradiction; R-L = ROUGE-L recall; BERT-R = BERTScore recall; CS_r/CT_r = consistency/contradiction ranks; $R\text{-}L_r$/$BERT\text{-}R_r$ = ROUGE-L/BERTScore recall ranks; AvgRank = average rank across evaluation metrics. Higher scores indicate better performance, while lower AvgRank indicates better overall ranking. Bold and underline indicate best and second-best results, respectively.

Team	Improvement					Deterioration				
	R	Fit	Rec	Spec	Ov.	R	Fit	Rec	Spec	Ov.
DreamerNLplus	1	0.6250	0.8125	**1.0000**	**0.7608**	3	0.4375	0.6875	0.8750	0.6038
CSE_IIT_Ropar	2	**1.0000**	0.6875	0.3750	0.7426	1	0.8750	0.5625	**0.9375**	**0.7891**
MKC	3	0.7500	0.5625	0.9375	0.7266	7	0.1875	0.1875	0.5000	0.2301
McMasterNLP	4	0.6875	0.3750	0.7500	0.5938	8	0.1875	0.1875	0.3125	0.2109
psytechlab	5	0.6875	**1.0000**	0.2500	0.5437	6	0.6250	0.6250	0.2500	0.4911
Aurevia	6	0.3750	0.6250	0.5000	0.4653	2	0.6875	0.8125	0.5625	0.6761
MeronymLabs	7	0.2500	0.2500	0.5625	0.2981	4	0.4375	0.5625	0.8125	0.5511
CtbuY	8	0.2500	0.2500	0.0000	0.1250	5	**1.0000**	**1.0000**	0.0000	0.5000
CUNY	9	0.0000	0.0000	0.2500	0.0000	9	0.0000	0.0000	0.3125	0.0000
Baseline	–	0.3750	0.4375	0.3750	0.3894	–	0.5625	0.3750	0.4375	0.4832

Table 5: Task 3.2 results. R = rank; Fit = fit score; Rec = recurrence score; Spec = specificity score; Ov. = overall score. Left block reports improvement prediction performance, while the right block reports deterioration prediction performance. Higher scores indicate better performance. Systems are ranked independently within each setting by Ov.

deterioration, **CSE-IIT-Ropar ranked first** with the highest Specificity and strong Fit, followed by **Aurevia** and **DreamerNLplus**.

8 Performance Analysis

Figure 4 shows state distributions, frequent ABCD sub-elements, and co-occurrence patterns.

Task 1 Performance Analysis: Distribution of Adaptive and Maladaptive Self-State Predictions: As shown in Table 6, maladaptive self-states received higher presence scores than adaptive states (2.83 vs. 2.18), with maladaptive predictions skewed toward higher scores (4–5) and adaptive predictions concentrated around lower scores, particularly score 2 (39.4%).

Element-Level and Sub-Element Prediction Trends: Adaptive predictions were mainly associated with **B-O**, **D**, and **B-S**, while maladaptive

State	Mean	Std	Score (%)				
			1	2	3	4	5
$^+$	2.18	0.95	26.3	39.4	25.4	7.8	1.2
$^-$	2.83	1.43	27.8	14.5	18.5	25.1	14.1

Table 6: Adaptive ($^+$) and maladaptive ($^-$) self-state presence score distributions.

predictions were dominated by **A**, **D**, and **C-S** (Table 7) (Stiles, 2001; Bromberg, 2014; Hofmann and Hayes, 2019). **B-O** (Behaviour towards Others) was the most frequent adaptive element (44.99%), whereas maladaptive **A** (Affect) was the dominant maladaptive element (50.93%).

Elements		Frequent Sub-elements	
Elem.	%	Sub-element	%
A^+	26.4	B-O(1)$^+$	40.8
B-O$^+$	45.0	B-S(1)$^+$	32.9
B-S$^+$	32.9	D(1)$^+$	18.6
C-O$^+$	20.5	C-O(1)$^+$	17.5
C-S$^+$	16.8	C-S(1)$^+$	16.3
D$^+$	38.8	D(5)$^+$	13.6
A^-	50.9	A(5)$^+$	12.3
B-O$^-$	20.2	C-S(2)$^-$	42.6
B-S$^-$	28.7	C-O(2)$^-$	32.0
C-O$^-$	37.3	B-S(2)$^-$	28.7
C-S$^-$	42.8	A(4)$^-$	22.4
D$^-$	43.6	D(6)$^-$	18.2
		D(2)$^-$	17.9
		B-O(2)$^-$	14.9

Table 7: Element-level and frequent sub-element prediction distributions. $^+$ Adaptive, $^-$ Maladaptive.

Co-occurrence and Combination Analysis: Adaptive and maladaptive self-states frequently co-occurred, with 58.1% of posts containing both and 14.4% containing neither.

8.1 Task 2 System Performance Analysis: Switch and Escalation Detection

Table 16 shows the type of model used by teams. Escalation prediction consistently outperforms Switch prediction across evaluation metrics (Table 8). Mean Escalation F1 reached 0.53 versus 0.41 for Switches, and the best Escalation F1 (0.74) exceeds the best Switch F1 (0.58). Most systems favor recall over precision, particularly for Switch prediction, resulting in a larger precision–recall gap. From a clinical perspective, the stronger predictability of Escalations relative to Switches may reflect important differences in the temporal organization of psychological change processes. Escalations appear to involve gradual amplification and accumulation of maladaptive self-state dynamics over time, producing more stable linguistic and behavioral traces that can be detected computationally. By contrast, Switches may correspond to more sudden reorganizations in self-state dominance, potentially triggered by external events, interpersonal ruptures, acute stressors, or sudden shifts in appraisal and affect regulation. This distinction is clinically meaningful because many deterioration trajectories in depression and related conditions are thought to emerge through progressively self-reinforcing cycles involving negative self-referential thinking, hopeless affect, withdrawal, and reduced engagement with adaptive goals or relationships (Hofmann and Hayes, 2019). Such gradual accumulation processes may naturally generate more coherent temporal signatures than rapid transitions. The findings suggest that timeline-based modeling may be particularly well-suited for detecting slowly consolidating maladaptive trajectories before more acute shifts occur.

Metric	Switch	Escalation
Mean Precision	0.35	0.57
Mean Recall	0.54	0.61
Mean F1	0.41	0.53
Median F1	0.43	0.54
Std. F1	0.12	0.15
Best F1	0.58	0.74
Mean Accuracy	0.67	0.73
Precision Range	0.09–0.55	0.23–0.80
Recall Range	0.05–1.00	0.17–1.00
F1 Range	0.06–0.58	0.22–0.74
P–R Gap	0.18	0.05

Table 8: Task 2 performance summary across submissions.

Presence Scores and Behavioural Predictions: Behavioural prediction rates vary across adaptive and maladaptive presence levels (Table 9; Figure 3). Adaptive states peak at moderate presence levels (2–3) before declining, whereas maladaptive states show a monotonic increase in behavioural instability, with Escalation rates rising from 4.34% at score 1 to 50.29% at score 5. These findings suggest that stronger maladaptive states are associated with escalating behavioural trajectories, while stronger adaptive states correspond to lower instability. Clinically, the adaptive and maladaptive patterns appear to reflect distinct temporal dynamics. Maladaptive states show a strong monotonic increase in both Switch and Escalation rates as presence scores increase, consistent with clinical models describing progressively self-reinforcing cycles of negative affect, hopelessness, and maladaptive self-referential processing in depression (Hayes et al., 2015). By contrast, adaptive states show higher Switch and Escalation rates at moderate presence levels, followed by declines at stronger levels. One possible interpretation is that moderate adaptive states may characterize more transitional or mixed phases in which adaptive and maladap-

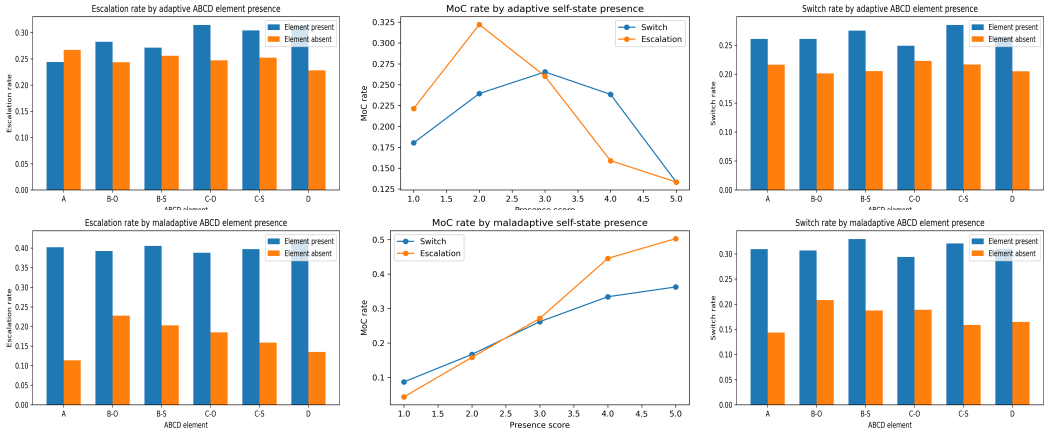

Figure 3: Relationship between adaptive/maladaptive self-state presence and Task 2 behavioral outcomes (switches and escalations). The plots show how the presence and intensity of individual ABCD self-state elements correlate with moments of change, including abrupt switches and gradual escalations, across adaptive and maladaptive self-state configurations. Higher maladaptive self-state presence is associated with increased switch and escalation rates, while adaptive self-state presence exhibits more stable behavioral dynamics.

tive processes coexist, whereas strongly adaptive states may reflect reduced deterioration dynamics (Bonanno and Burton, 2013).

State	Score	Switch (%)	Escalation (%)
$^+$	1	18.04	22.15
$^+$	2	23.94	32.22
$^+$	3	26.53	26.02
$^+$	4	23.84	15.89
$^+$	5	13.33	13.33
$^-$	1	8.68	4.34
$^-$	2	16.67	15.81
$^-$	3	26.18	27.13
$^-$	4	33.42	44.55
$^-$	5	36.26	50.29

Table 9: Switch and Escalation rates across adaptive ($^+$) and maladaptive ($^-$) presence scores.

8.2 Task 3.2 Team Meta-Analysis

Characteristics of High-Performing Signatures
Strong Self-State Dynamics: The strongest systems consistently described: maladaptive dominance, adaptive emergence, suppression between self-states, reflective dialogue, coexistence and transition dynamics. Self-state terminology exhibited one of the strongest positive correlations with final score ($r = 0.5425$). For example, the top deterioration system (*CSEIITRopar*) models explicit feedback loops, tracks suppression of adaptive processes, describes temporal progression into deterioration.
Mechanistic Temporal Progression: Strongest systems described change as a *dynamic process*, not a static emotional state.

A common deterioration trajectory across high-performing systems (CSEIITRopar, DreamerNLplus, Aurevia, and MeronymLabs) described maladaptive progression as a reinforcing cascade from self-critical or detached cognitions (C-S/C-O), to depressive affect (A), to hopeless desire states (D), followed by collapse of self-care and relational engagement (B-S/B-O), culminating in maladaptive dominance (MD):

$$\text{C-S/C-O} \rightarrow \text{A} \rightarrow \text{D} \rightarrow \text{B-S/B-O collapse} \rightarrow \text{MD}$$

Improvement: Weakening maladaptive dominance, emergence of adaptive dialogue, self-care and connection behaviors, restoration of competence or belonging. This signature performed strongly because it: explicitly models transition, describes adaptive takeover rather than simple symptom reduction.
Generic Systems Performed Poorly: Such as CtbuY and CUNY scored poorly despite mentioning ABCD concepts because they lacked: temporal progression, coherent dynamic structure. Signature resembled keyword extraction rather than dynamic modeling : *Adaptive self-state, Depressive affect, Self-criticism, Maladaptive states*.
Temporal Dynamics and Change Events: Several high-performing systems explicitly modeled **Change Events (CE)**. The psytechlab submissions were particularly notable for capturing pre-/post-CE dynamics, silence and reflection patterns, and transitions between coexistence and dominance.

8.3 Key Findings

Recovery trajectories are characterized by weakening maladaptive states, particularly self-critical cognition (*C-S*), depressive affect (*A*), hopeless-

ness (*D*), and interpersonal detachment (*C*-*O*), followed by adaptive loops involving self-care (*B*-*S*), relating behaviors (*B*-*O*), positive affect, adaptive expectations, and more accepting self-perception. **Improvement generally reflects coexistence and reflective integration of adaptive and maladaptive states rather than abrupt replacement.** By contrast, deterioration forms **self-reinforcing maladaptive cycles** in which self-critical cognition amplifies depressive affect and hopelessness while suppressing interpersonal engagement and self-care. Multiple systems also identified interpersonal detachment as a precursor of worsening trajectories, consistent with models of depressive maintenance involving negative self-referential processing, hopelessness, and social withdrawal (Blatt, 2004; Shahar, 2015). Notably, **these clinically meaningful themes emerge consistently across methodologically diverse systems**, suggesting that temporally organized self-state interactions capture psychologically meaningful longitudinal dynamics.

9 Conclusion

We introduced the CLPsych 2026 Shared Task on modeling mental health changes in social media timelines through self-state prediction, change detection, and sequence-level summarization. Results showed strong performance from LLM-based and pipeline approaches, especially when combining structured prompting, retrieval, ensembling, and temporal modeling. Maladaptive self-states were consistently easier to detect than adaptive ones, likely due to stronger and more explicit negative emotional signals (Baumeister et al., 2001; Atzil-Slonim et al., 2024). Similarly, Escalations were more predictable than Switches, suggesting that gradual deterioration yields clearer temporal patterns than abrupt transitions (Hayes et al., 2015). High-performing systems captured clinically meaningful dynamics, including maladaptive feedback loops involving self-critical cognition, hopeless affect, and behavioral withdrawal, while improvement trajectories reflected gradual integration of adaptive and maladaptive states rather than abrupt replacement (Stiles, 2001; Bromberg, 2014). These findings highlight the value of temporally grounded self-state modeling for longitudinal mental health analysis.

Limitations

As in the vast majority of prior work leveraging social media for individual-level mental health assessments, this year's shared task involves individuals who generated content in self-selected online communities. The present tasks were conducted using social media posts made on various mental health-related subreddits in the English language, by users who willingly self-disclosed their thoughts and feelings. Generalization of the approaches presented in this work to other contexts and in other languages remains an open area of research.

Annotation was performed over 40 relatively short timelines due to the annotation load for clinical experts. This potentially hinders the performance of smaller supervised models, still leaving open questions around their true potential. Additionally, although the well-being score was annotated on the post-level with full timeline content visibility, the longitudinal manifestation of individuals' well-being remains underexplored. Since the annotation process involved selection of the most salient available adaptive and maladaptive spans for each ABCD element, this task does not yet explore the more nuanced selection of additional evidence spans and their connection to one another.

Although the dynamic evolution of self-states was, to some extent, addressed in this work with respect to summarization, there is still need to explore such dynamic progression through the lenses of other tasks such as monitoring and dialogue tracking. Finally, multimodality, which provides important cues especially in the clinical setting in terms of the manifestation of self-states, remains for now a future direction.

Ethics

This year's tasks explored the prediction of well-being scores from online posts of users over time, as well as the extraction of adaptive and maladaptive evidence spans and further summarization of self-state information at the post and timeline levels. This multi-task framework is grounded in the MIND scheme (Slonim, 2024) that views human experience as consisting of self-states fluctuating over time. Each self-state constitutes of identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire (ABCD).

While the evidence extraction and summaries provide some guidance with respect to ABCD elements and maladaptive and adaptive states, this

cannot be used for diagnostic purposes, especially without the involvement of human experts. Adaptive and maladaptive evidence extracted by such models should be reviewed by clinical experts or used to augment their capacity by efficiently presenting information to them.

Additionally, the task cannot make any claims about the potential evidence providing explanations for well-being scores. Rather, it forms a research direction towards making causal links between the two, paving the way towards language models that can better reason along their decision making process.

In terms of data, even though we are using publicly available content from Reddit, we prohibited its redistribution and the use of any third-party LLMs that would require sending (part of) the information to the provider's servers, to ensure protection of the sensitive content.

Acknowledgements

This work was supported by a UKRI/EPSRC Turing AI Fellowship to Maria Liakata (grant ref EP/V030302/1) and the Alan Turing Institute (grant ref EP/N510129/1). This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK (KP0016) as a Keystone project lead by Maria Liakata. Diana Inkpen was supported by the Natural Science and Engineering Research Council of Canada. Juan Antonio Lossio-Ventura was supported by the National Institute of Mental Health Intramural Research Program (ZICMH002968). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the author(s) and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services. The shared task organizers would like to express their gratitude to the anonymous users of Reddit whose data feature in this year's shared task dataset; to the clinical experts from Bar-Ilan University who annotated the data for all tasks; to all team members for their participation; and to ACL for its support for CLPsych.

References

Prottay Kumar Adhikary, Aseem Srivastava, Shivani Kumar, Salam Michael Singh, Puneet Manuja, Jini K Gopinath, Vijay Krishnan, Swati Kedia Gupta, Koushik Sinha Deb, and Tanmoy Chakraborty. 2024. Exploring the efficacy of large language models in summarizing mental health counseling sessions: benchmark study. *JMIR Mental Health*, 11:e57306.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *IEEE Intelligent Systems*, 38(2):15–23.

Anson Antony, Gautam Vijay Kumar, and Annika Marie Schoene. 2026. Agentic pipelines meet retrieval-augmented icl: A zero-training approach to mental health modeling. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Elham Asgari, Nina Montana-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, and Dominic Pimenta. 2024. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *medRxiv*, pages 2024–09.

Dana Atzil-Slonim. 2025. Multimodal intrapersonal and interpersonal dynamics (MIND): A transtheoretical coding manual. Preprint, Bar-Ilan University.

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.

Dana Atzil-Slonim, Juan Martin Gomez Penedo, and Wolfgang Lutz. 2024. Leveraging novel technologies and artificial intelligence to advance practice-oriented research. *Administration and Policy in Mental Health and Mental Health Services Research*, 51(3):306–317.

Roy F Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs. 2001. Bad is stronger than good. *Review of general psychology*, 5(4):323–370.

Ulya Bayram and Lamia Benhiba. 2022. Emotionally-informed models for detecting moments of change and suicide risk levels in longitudinal social media data. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 219–225, Seattle, USA. Association for Computational Linguistics.

Aaron T Beck, Molly R Finkel, and Judith S Beck. 2021. The theory of modes: Applications to schizophrenia and other psychological conditions. *Cognitive Therapy and Research*, 45:391–400.

Leonard Bickman. 2020. Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision

mental health. *Administration and Policy in Mental Health and Mental Health Services Research*, 47(5):795–843.

Amirmohammad Ziaei Bideh, Shameed Charlomar Job, Ava Yahyapour, and Alla Rozovskaya. 2026. Cuny at clpsych 2026: A pipeline approach to classification and summarization of mental health change. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Sidney J Blatt. 2004. *Experiences of depression: Theoretical, clinical, and research perspectives.* American Psychological Association.

George A Bonanno and Charles L Burton. 2013. Regulatory flexibility: An individual differences perspective on coping and emotion regulation. *Perspectives on psychological science*, 8(6):591–612.

Philip M Bromberg. 2014. *Standing in the spaces: Essays on clinical process trauma and dissociation*. Routledge.

Ana-Maria Bucur, Marcos Zampieri, Tharindu Ranasinghe, and Fabio Crestani. 2026. A survey on multilingual mental disorders detection from social media data. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–918, Rabat, Morocco. Association for Computational Linguistics.

Igor Buyanov, Nafisa Valieva, and Ekaterina Mazurina. 2026. psytechlab at clpsych 2026: Utilising natural language processing methods and large language models for social media text analysis. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Callum Chan, Sunveer Khunkhun, Diana Inkpen, and Juan Antonio Lossio-Ventura. 2025. Prompt engineering for capturing dynamic mental health self states from social media posts. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 256–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Jenny Chim, Julia Ive, and Maria Liakata. 2025. Evaluating synthetic data generation from user generated text. *Computational Linguistics*, 51(1):191–233.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190, St. Julians, Malta. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. 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*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Duc Minh Do, Tin Trung Pham, Vu Tran, and Minh Le Nguyen. 2026. Prompt-based modeling of moments of change and change summaries in mental health timelines. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Muskan Garg. 2023. Mental health analysis in social media posts: A survey: M. garg. *Archives of Computational Methods in Engineering*, 30(3):1819–1842.

Muskan Garg. 2024. Wellxplain: Wellness concept extraction and classification in reddit posts for mental health analysis. *Know.-Based Syst.*, 284(C).

Adele M Hayes, Carly Yasinski, J Ben Barnes, and Claudi LH Bockting. 2015. Network destabilization and transition in depression: New methods for studying the dynamics of therapeutic change. *Clinical psychology review*, 41:27–39.

Anthony Hills, Adam Tsakalidis, and Maria Liakata. 2023. Time-aware predictions of moments of change in longitudinal user posts on social media. In *Advanced Analytics and Learning on Temporal Data*, pages 293–305, Cham. Springer Nature Switzerland.

Anthony Hills, Talia Tseriotou, Xenia Miscouridou, Adam Tsakalidis, and Maria Liakata. 2024. Exciting mood changes: A time-aware hierarchical transformer for change detection modelling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12526–12537, Bangkok, Thailand. Association for Computational Linguistics.

Stefan G Hofmann and Steven C Hayes. 2019. The future of intervention science: Process-based therapy. *Clinical Psychological Science*, 7(1):37–50.

Stephanie Homan, Marion Gabi, Nina Klee, Sandro Bachmann, Ann-Marie Moser, Martina Duri', Sofia Michel, Anna-Marie Bertram, Anke Maatz, Guido Seiler, Elisabeth Stark, and Birgit Kleim. 2022. Linguistic features of suicidal thoughts and behaviors: A systematic review. *Clinical Psychology Review*, 95:102161.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, David A Clifton, et al. 2025. Large language models in mental health care: a scoping review. *Current Treatment Options in Psychiatry*, 12(1):27.

Kyomin Hwang, Hyeonjin Kim, Hyunho Lee, and Nojun Kwak. 2026. Team mkc at clpsych 2026: Capturing and characterizing mental health changes through social media timeline dynamics. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Ayal Klein, Jiayu Song, Jenny Chim, Liran Keren, Andreas Triantafyllopoulos, Björn W Schuller, Maria Liakata, and Dana Atzil-Slonim. 2024. Clinical insights from social media: Assessing summaries of large language models and humans.

Pawan Kumar, Ankit Meshram, Shubham Jha, and Loitongbam Gyanendro Singh. 2026. Theory-explicit prompting for mind self-states: Hierarchical llms and dynamic signature extraction in mental health timelines. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Gal Lazarus and Eshkol Rafaeli. 2023. Modes: Cohesive personality states and their interrelationships as organizing concepts in psychopathology. *Journal of Psychopathology and Clinical Science*, 132(3):238.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. 2023. Compressing context to enhance inference efficiency of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6342–6353, Singapore. Association for Computational Linguistics.

Chandreen Liyanage, Muskan Garg, Vijay Mago, and Sunghwan Sohn. 2023. Augmenting Reddit posts to determine wellness dimensions impacting mental health. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 306–312, Toronto, Canada. Association for Computational Linguistics.

Juan Antonio Lossio-Ventura, Callum Chan, Arshitha Basavaraj, Hugo Alatrista-Salas, Francisco Pereira, and Diana Inkpen. 2025. 5cNLP at BioLaySumm2025: Prompts, retrieval, and multimodal fusion. In *Proceedings of the 24th Workshop on Biomedical Language Processing (Shared Tasks)*, pages 215–231, Vienna, Austria. Association for Computational Linguistics.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 47684777, Red Hook, NY, USA. Curran Associates Inc.

Tobias Mayer, Neha Warikoo, Amir Eliassaf, Dana Atzil-Slonim, and Iryna Gurevych. 2024. Predicting client emotions and therapist interventions in psychotherapy dialogues. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1463–1477, St. Julian's, Malta. Association for Computational Linguistics.

George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arturo Montejo-Ráez, M. Dolores Molina-González, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Luis Joaquín García-López. 2024. A survey on detecting mental disorders with natural language processing: Literature review, trends and challenges. *Comput. Sci. Rev.*, 53(C).

Abir Naskar and Mike Conway. 2026. Hierarchical multi-stage modeling of adaptive and maladaptive self-states in social media timelines. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

David Owen, Amy J Lynham, Sophie E Smart, Antonio F Pardiñas, and Jose Camacho Collados. 2024. Ai for analyzing mental health disorders among social media users: Quarter-century narrative review of progress and challenges. *Journal of Medical Internet Research*, 26:e59225.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Alina Ponomareva, Nina Stekacheva Sancho, and Karina Litvinova. 2026. Stateofmind at clpsych 2026 shared task. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Federico Ravenda, Volodymyr Karpenko, Antonietta Mira, and Andrea Raballo. 2026. P2p - from posts to patterns: An llm ensemble approach to mental health dynamics detection. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical*

Psychology. Association for Computational Linguistics.

William Revelle. 2007. Experimental approaches to the study of personality. *Handbook of research methods in personality psychology*, pages 37–61.

Nathan Roll, Irene Yi, Sufian Aldogom, Grace Brown, Eric Basile, Isaac Gutterman, Lakshika Tennakoon, and Ammar Ahmed. 2026. Team aurevia at clpsych 2026: Local healthcare nlp for schema-constrained self-state modeling. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Ramit Sawhney, Shivam Agarwal, Atula Tejaswi Neerkaje, Nikolaos Aletras, Preslav Nakov, and Lucie Flek. 2022a. Towards suicide ideation detection through online conversational context. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 17161727, New York, NY, USA. Association for Computing Machinery.

Ramit Sawhney, Atula Neerkaje, and Manas Gaur. 2022b. A risk-averse mechanism for suicidality assessment on social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 628–635, Dublin, Ireland. Association for Computational Linguistics.

Golan Shahar. 2015. *Erosion: The psychopathology of self-criticism*. Oxford University Press.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36.

Gopendra Vikram Singh, Sai Vardhan Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal LLM-based detection and reasoning framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22546–22570, Miami, Florida, USA. Association for Computational Linguistics.

Dana Atzil Slonim. 2024. Self-other dynamics (sod): A transtheoretical coding manual.

Jiayu Song, Mahmud Akhter, Dana Atzil Slonim, and Maria Liakata. 2025. Temporal reasoning for timeline summarisation in social media.

Jiayu Song, Jenny Chim, Adam Tsakalidis, Julia Ive, Dana Atzil-Slonim, and Maria Liakata. 2024. Combining hierachical VAEs with LLMs for clinically meaningful timeline summarisation in social media. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14651–14672, Bangkok, Thailand. Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. Mentsum: A resource for exploring summarization of mental health online posts.

Aseem Srivastava, Tharun Suresh, Sarah Peregrine, Lord, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering.

William B Stiles. 2001. Assimilation of problematic experiences. *Psychotherapy: Theory, Research, Practice, Training*, 38(4):462.

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. 2022a. 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, 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. 2022b. 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.

Talia Tseriotou, Ryan Chan, Adam Tsakalidis, Iman Munire Bilal, Elena Kochkina, Terry Lyons, and Maria Liakata. 2024a. Sig-networks toolkit: Signature networks for longitudinal language modelling. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 223–237, St. Julians, Malta. Association for Computational Linguistics.

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.

Talia Tseriotou, Adam Tsakalidis, Peter Foster, Terence Lyons, and Maria Liakata. 2023. Sequential path signature networks for personalised longitudinal language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5016–5031, Toronto, Canada. Association for Computational Linguistics.

Talia Tseriotou, Adam Tsakalidis, and Maria Liakata. 2024b. TempoFormer: A transformer for temporally-aware representations in change detection. In *Proceedings of the 2024 Conference on Empirical*

Methods in Natural Language Processing, pages 19635–19653, Miami, Florida, USA. Association for Computational Linguistics.

David Villarreal-Zegarra, C Mahony Reategui-Rivera, Jackeline García-Serna, Gleni Quispe-Callo, Gabriel Lázaro-Cruz, Gianfranco Centeno-Terrazas, Ricardo Galvez-Arevalo, Stefan Escobar-Agreda, Alejandro Dominguez-Rodriguez, and Joseph Finkelstein. 2024. Self-administered interventions based on natural language processing models for reducing depressive and anxiety symptoms: Systematic review and meta-analysis. *JMIR Mental Health*, 11.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024a. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024b. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024c. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 44894500, New York, NY, USA. Association for Computing Machinery.

Hongyi Zhang, Derron Li, Scarlett Cleary, Aadi Sanghani, Akshay Krishna Sirigana, Brian Miguel Pimentel, Kelsey Isman, Kian Omoomi, Vasudha Varadarajan, Charles Welch, and Allison Claire Lahnala. 2026. Mcmasters of change: Predicting well-being states and transitions from longitudinal language. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Mengjia Zhang and Yi Yang. 2026. A multi-strategy framework for mental health change detection in social media timelines. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Maryia Zhyrko, Daisy Monika Lal, Erik van Mulligen, and Lifeng Han. 2026. Dreamernlplus: Interpretable modeling of mental health dynamics from social media timelines using hybrid rule-based and rag methods. In *Proceedings of the Eleventh Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39, Seattle, USA. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

Appendix A Data Annotation

The dataset combines MIND annotations from CLPsych 2025 with additional annotations introduced for CLPsych 2026. In CLPsych 2025, clinical psychology Masters students annotated posts using the MIND framework, labeling adaptive/-maladaptive ABCD subelements, self-state presence scores, well-being scores, and post-level summaries under supervised reconciliation procedures (Tseriotou et al., 2025). Here, 36 additional posts were annotated to complete change-centered sequences for Tasks 2 and 3. For Task 3.1, gold sequence summaries were generated using Qwen-2.5-32B conditioned on post text, MIND summaries, ABCD annotations, self-state scores, and MoC labels, then manually reviewed and revised by a clinical expert. The summaries capture intra- and inter-state dynamics, relative self-state dominance, change direction, and pre-change context surrounding Switches and Escalations.

A.1 MIND Framework

Table 10 shows categories and sub-categories within the MIND framework.

Appendix B Data Statistics

Table 11 shows data statistics.

Appendix C Participant Submissions

This section presents an overview of the registration process (§C.1), individual systems (§C.2) from each participating team and provides an overview of methods (§C.3) and (§C.4) contains performance analysis across each task.

C.1 Registration Process

The registration and participation procedure comprised three stages: (a) completion of individual and team registration via an online form, (b) review and signing of a data sharing agreement, and (c) receipt of access instructions for the training data, which was provided as a password-protected compressed file. During the registration phase, the organizing team also assisted participants seeking collaborators for team formation. The data sharing agreement required teams to maintain the dataset in secure, password-protected private storage and prohibited both explicit and implicit data sharing through third-party platforms, including proprietary large language model services. For *Codabench-based* participation, each team was required to nominate one designated representative and share that users Codabench ID with the organizers prior to submission. Unless stated otherwise, teams were allowed up to three submissions per task on Codabench. Tasks 1, 2, and 3.1 were submitted through Codabench, whereas Task 3.2 was submitted separately via email in the required JSON format for human evaluation. Although participants could choose to make their scores visible on the Codabench leaderboard, these leaderboard positions were not treated as official rankings, since the final shared task rankings were determined using the shared tasks task-specific evaluation procedures. Table 12 summarizes the participating teams and their submission activity.

C.2 Individual Team Submissions

- CUNY (Bideh et al., 2026) employed a modular pipeline using locally served open-weight LLMs including Gemma 3, Qwen 3.5, and GPT-OSS. Tasks 1.1 and 1.2 used ensemble-based ICL prompting with majority voting, achieving the top-ranked Task 1.1 result. Task 2 used SVM and Random Forest classifiers trained on upstream predictions such as presence scores and inter-post deltas. Task 3.1 applied label-augmented ICL for improved summarization, with the team ranking 1st on Task 1.1 and 3rd on Task 3.1.

- StateOfMIND (Ponomareva et al., 2026) used a retrieval-augmented in-context learning ensemble with two open-weight LLMs and three sequential prompts for Tasks 1.1 and 1.2. Their framework combined unified, adaptive-focused, and affect-focused prompting strategies with deterministic aggregation of ABCD predictions. The system leveraged post-level retrieval and ensemble inference to improve subelement and presence prediction. Their best submission ranked 2nd for Task 1.1 and 5th for Task 1.2.

- Meronym Labs (Antony et al., 2026) proposed a zero-training retrieval-augmented in-context learning framework using frozen LLMs such as Qwen 3.5-27B and Gemma 4-31B with ChromaDB retrieval. Static and dynamically retrieved examples were integrated into prompts for ABCD classification and presence prediction. Task 2 used timeline-level prompting with automated prompt optimization, while Task 3 introduced a multi-agent contradiction-aware summarization pipeline with NLI reranking using DeBERTa-v3-large. The

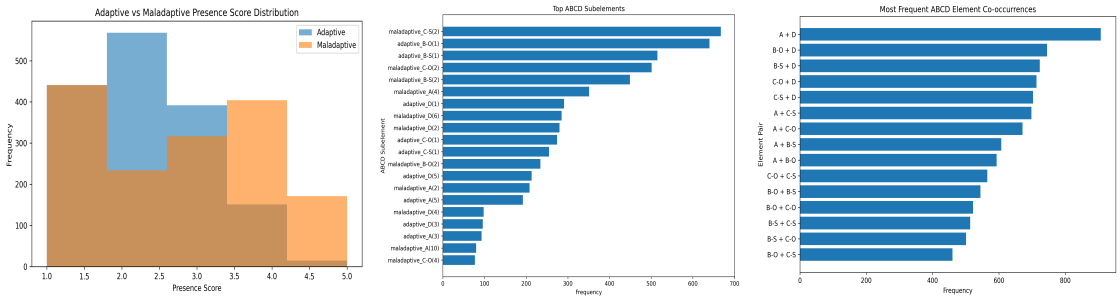

Figure 4: Task 1 analysis: state distributions, frequent ABCD sub-elements, and co-occurrence patterns.

Category		Sub-Categories	
		Adaptive Example	Maladaptive Example
Affect	Type of emotion expressed by a person.	Calm/Laid back, Emotional Pain/Grieving, Content/Happy, Vigor/Energetic, Justifiable Anger/Assertive Anger, Proud	Anxious/Tense/Fearful, Depressed/Desperate/ Hopeless, Mania, Apathetic/Don't care/Blunted, Angry (Aggressive, Disgust, Contempt), Ashamed/Guilty
Behavior	**Behavior of the self with the Other (BO)** The person's main behavior(s) toward the other.	Relating behavior, Autonomous behavior	Fight or flight behavior, Overcontrolled/controlling behavior
	Behavior toward the Self (BS) The person's main behavior(s) toward the self.	Self-care behavior	Self-harm/Neglect/ Avoidance behavior
Cognition	**Cognition of the Other (CO)** The person's main perceptions of the other.	Perception of the other as related, Perception of the other as facilitating autonomy/ competence needs	Perception of the other as detached or over attached, Perception of the other as blocking autonomy needs
	Cognition of the Self (CS) The person's main self-perceptions.	Self-acceptance and self-compassion	Self-criticism
Desire	The person's main desire, need, intention, fear or expectation.	Relatedness, Autonomy and adaptive control, Competence, Self-esteem, Self-care	Expectation that relatedness need will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met

Table 10: ABCD elements (Categories) with explanations, and their sub-categories.

Statistic	Train	Test	Overall
Timelines	30	10	40
Posts	373	92	465
Avg. Tok./Post	118.4	120.2	118.8
Median Tok./Post	47.0	81.0	60.0
WB Available	236	73	309
WB Missing	137	19	156
WB Mean	5.24	5.29	5.25
WB Std.	1.87	2.15	1.94
Switches	78	21	99
Escalations	75	24	99
Both	27	11	38
Neither	247	58	305
ADT A	43	21	64
ADT B-O	125	38	163
ADT B-S	84	23	107
ADT C-O	50	14	64
ADT C-S	52	15	67
ADT D	118	36	154
ADT Total	472	147	619
MAL A	161	44	205
MAL B-O	36	11	47
MAL B-S	59	22	81
MAL C-O	114	29	143
MAL C-S	135	30	165
MAL D	110	38	148
MAL Total	615	174	789
Deterioration	49	15	64
Improvement	25	04	29

Table 11: CLPsych 2026 dataset statistics.

framework avoided any task-specific fine-tuning while supporting structured psychological summarization.

- USAI (Ravenda et al., 2026) proposed an ensemble-based framework combining multiple

Team	1.1	1.2	2	3.1	3.2	Paper	System Summary
Aurevia	1	1	1	1	1	✓	✓
BLUE	1	1	1	–	–	×	✓
Codezone	–	–	1	–	–	×	✓
CSE_IIT_Ropar	1	1	1	1	1	✓	×
CUNY	1	1	1	1	1	✓	✓
CtbuY	1	1	1	1	1	✓	✓
DreamerNLplus	1	1	1	1	1	✓	✓
JNLP	–	–	1	1	–	✓	✓
McMasterNLP	1	1	1	1	1	✓	✓
Meronym Labs	1	1	1	1	1	✓	×
MKC	1	1	1	1	1	✓	✓
NoviceTrio	1	1	1	1	–	✓	✓
psytechlab	1	1	1	1	1	✓	✓
StateOfMIND	1	1	–	–	–	✓	✓
ULL	1	1	–	1	–	×	✓
USAI	1	1	1	1	–	✓	×
Total	**14**	**14**	**14**	**13**	**9**	**13**	**13**

Table 12: Team participation across tasks, paper submission status, and availability of system summaries.

open-source LLMs for Tasks 1, 2, and 3.1. Weighted voting and weighted averaging were used for ABCD subelement classification and presence estimation, respectively. Task 2 incorporated an additional refinement stage for improving temporal consistency in Switch and Escalation detection. For Task 3.1, few-shot prompting was applied using predicted annotations and change labels to generate sequence summaries.

- Aurevia (Roll et al., 2026) developed a privacy-conscious local healthcare NLP pipeline using TF-IDF retrieval, schema-constrained prompting with Qwen2.5-72B-Instruct, ordinal calibration, and de-

terministic post-processing. The system combined non-neural retrieval methods with local LLM-based JSON prediction for ABCD classification and presence estimation. Task 2 used handcrafted temporal signals and rule-based calibration, while Task 3 emphasized contradiction-aware retrieval-augmented summarization. Aurevia ranked 3rd for Task 1.2, 5th for Task 3.1, and achieved the best Task 3.1 consistency score.

- MKC (Hwang et al., 2026) developed a unified LLM-based framework using LoRA fine-tuned Qwen3 embedding models with 5-fold cross-validation for Tasks 1 and 2. Weighted Cross-Entropy and weighted Huber losses addressed severe class imbalance, while sliding-window temporal modeling supported Switch and Escalation detection. For summarization, the team fine-tuned Qwen3/Qwen3.5 models using supervised instruction tuning and generated psychological signatures in a zero-shot manner using Qwen3.5-9B. Their framework integrated both discriminative and generative timeline modeling.

- McMasterNLP (Zhang et al., 2026) proposed a multi-stage framework combining representation learning, sequence modeling, and prompt-based summarization. Task 1 used a dual-encoder architecture integrating MentalRoBERTa embeddings with LLM-generated ABCD archetype similarity features for subelement and presence prediction. Task 2 employed BiLSTM-based sequence models for detecting switches and escalations in well-being timelines. Task 3 used prompt-engineered Qwen3.5-9B summarization, with the system ranking 7th in both Tasks 1.1 and 1.2.

- NoviceTrio (Naskar and Conway, 2026) introduced a hierarchical multi-stage framework combining a multi-task transformer encoder with a four-stage instruction-tuned LLM pipeline. Stage 1 jointly predicted state presence and intensity, while Stage 2 performed element-conditioned subelement classification and BIO-based evidence extraction using structural masking. A staged Qwen3-4B LoRA pipeline sequentially handled state detection, evidence extraction, and refined presence estimation. RoBERTa achieved an 8.3% Macro-F1 improvement over baseline, while Qwen3 delivered the strongest overall performance.

- PSYTECHLAB (Buyanov et al., 2026) proposed a privacy-preserving framework combining locally deployed language models with task-specific neural architectures. Their system used ModernBERT for ABCD subelement classification and BiLSTM networks for temporal modeling of moments of change. Sequence summarization was handled using zero-shot LLM prompting. The framework emphasized secure local processing of sensitive mental health timelines.

- DreamerNLplus (Zhyrko et al., 2026) developed a hybrid interpretable framework spanning all subtasks. Task 1.1 combined Ollama-based LLM data augmentation with DeBERTa fine-tuning for ABCD classification, while Task 1.2 used Random Forest regression over structured feature vectors. Task 2 applied few-shot prompting with locally deployed Llama 3.1 for temporal change detection. For Tasks 3.1 and 3.2, the team explored both deterministic rule-based summarization and RAG-based signature extraction pipelines.

- CtbuY (Zhang and Yang, 2026) proposed a multi-strategy fusion framework integrating zero-shot prompting, QLoRA fine-tuning, and hierarchical TextCNN classification for Task 1. Task 2 employed Mental-BERT embeddings with log-scale temporal encoding and a three-layer BiLSTM trained using Focal Loss. Task 3.1 used dynamic RAG retrieval with Qwen-2.5-7B, while Task 3.2 combined ABCD transition statistics, K-Means clustering, and thematic distillation for signature generation. The framework emphasized robust temporal and retrieval-based reasoning.

- CSE_IIT_Ropar (Kumar et al., 2026) proposed a theory-explicit prompting framework grounded directly in the MIND ABCD taxonomy. Their system used a three-stage sequence-level pipeline involving heuristic deterioration/improvement labeling, structured clinical-style summarization, and aggregated signature extraction. LLM prompts explicitly encoded ABCD dynamics and enforced a structured summary format including central themes and within-/between-state dynamics. The framework emphasized psychologically interpretable prompting strategies.

- JNLP (Do et al., 2026) proposed a prompt-based in-context learning framework using locally deployed Qwen2.5-72B-Instruct-GPTQ-Int8 with vLLM inference. The system avoided task-specific fine-tuning and instead relied on few-shot and balanced few-shot prompting for Task 2 change detection and Task 3.1 summarization. Balanced demonstrations improved Task 2 Macro-F1 to 0.580,

while Task 3 summaries achieved strong consistency scores. Their framework demonstrated the effectiveness of prompt-based longitudinal reasoning in low-resource mental health settings.

- ULL ULL proposed a hybrid contrastive learning and regression framework for psychological state modeling. Task 1.1 used a SetFit classifier with an MPNet backbone trained using highlighted evidence spans and contrastive learning for ABCD classification. Task 1.2 used Ridge regression over sentence-transformer and TF-IDF features for presence estimation. For Task 3.1, the team combined a LoRA fine-tuned BART summarizer with retrieval-augmented prompting and rule-based validation.

- BLUE BLUE developed a hybrid zero-/few-shot LLM framework using DeepSeek-R1-32B and Mixtral-8x7B through Ollama. Dense retrieval with BGE-M3 embeddings retrieved annotated examples for ABCD classification under the MIND framework. Mixtral handled presence estimation and wellbeing trajectory prediction, while Task 2 Switch and Escalation labels were derived using rule-based analysis over predicted trajectories. The system combined retrieval-augmented prompting with temporal reasoning.

- Codezone Codezone proposed the Temporal-Clinical Graph Network (TCGN) for Task 2 change detection. Their framework combined transformer-based post encoders, a Temporal Difference Encoder over GAF-rescaled signals, and a Graph Attention Network operating on dynamic self-state transition graphs. A BiGRU with Clinical Rule-Augmented Contrastive Learning modeled longitudinal changes for Switch and Escalation detection. The graph-enhanced temporal architecture substantially outperformed sequential baselines.

C.3 Overview

We outline methods used in the *best run per team* in Table 13. For a complete picture of each team's approaches, including ablations, we refer the reader to the respective system description papers. We categorize each system based on the following properties:

- **LLM**: Uses a large language model.
- **PLM**: Uses a pretrained language model.
- **ML**: Uses traditional machine learning algorithms (e.g., random forest), excluding feature engineering techniques.
- **RAG**: Uses retrieval-augmented generation with automatic retrieval (excluding manually selected examples).
- **Pipeline**: Uses outputs from one task as inputs to another task (not joint prediction within a single prompt).
- **Domain**: Incorporates domain knowledge beyond the shared task guidelines.
- **Temporal**: Explicitly models cross-post dependencies (e.g., sequence models, previous posts, or timeline reasoning), excluding intra-post context.

Table 13 summarizes the modeling strategies adopted by each participating team. LLM-based approaches dominate the competition, with 14 teams using at least one large language model in their best-performing system. Pipeline-based systems are also highly prevalent (11 teams), indicating that many participants decomposed the shared task into sequential subtasks rather than relying on joint modeling. Temporal reasoning is explicitly modeled by 12 teams, particularly for Task 2, while six teams use retrieval-augmented generation (RAG). Traditional machine learning approaches remain relatively uncommon, appearing in only three systems.

Team	LLM	PLM	ML	RAG	Pipeline	Domain	Temporal
Aurevia	✓				✓	✓	✓
BLUE	✓			✓	✓		✓
Codezone		✓				✓	✓
CSE_IIT_Ropar	✓		✓				✓
CtbuY	✓	✓		✓	✓	✓	✓
CUNY	✓		✓		✓		✓
DreamerNLplus	✓	✓	✓	✓	✓		✓
JNLP	✓						✓
McMasterNLP	✓						
Meronym Labs	✓			✓	✓		✓
MKC	✓				✓		✓
NoviceTrio		✓			✓		
psytechlab	✓	✓			✓		✓
StateOfMIND	✓			✓	✓		
ull	✓	✓	✓	✓	✓		
USAI	✓				✓		✓
Total	14	6	3	6	11	3	12

Table 13: Methods used in each team's best submission. Pipeline refers to cases where outputs from one task are used in another task.

LLMs. Large language models (LLMs) form the backbone of nearly all high-performing systems in this shared task, either as primary predictors or as components within hybrid pipelines. Across the *best runs per team*, we observe clear trends in model selection, context utilization, and scaling behavior (Figures 5–7). First, model family diver-

sity has increased substantially compared to prior shared tasks (Chim et al., 2024; Tseriotou et al., 2025). While Llama-based models are used in several submissions, the strongest-performing systems are predominantly built on alternative open-weight model families such as Qwen, DeepSeek, and Gemma. In particular, top-ranked teams including CUNY, StateOfMIND, Meronym Labs, and USAI rely heavily on these newer model families, often in combination with other models in ensemble settings. This indicates a shift away from Llama-centric dominance toward a more diverse ecosystem of competitive LLMs compared to previous shared task trends.

Second, context length has expanded significantly (Figure 6). Whereas prior work was typically constrained to 32K tokens, several systems in this shared task report context windows exceeding 100K tokens, with some models supporting up to 1M tokens. However, most high-performing systems do not rely on extremely long contexts; instead, they employ targeted context construction strategies such as retrieval-augmented generation (RAG), structured prompting, or selective use of recent posts. This suggests that effective context selection is more important than raw context length for modeling psychological signals.

Third, model scale contributes to performance but is not the sole determining factor (Figure 7). Larger models (50B+ parameters), such as those used by CUNY, Aurevia, and USAI, achieve strong results across tasks, particularly for presence estimation (Task 1.2). However, mid-sized models (10–50B), including those used by StateOfMIND and Meronym Labs, remain highly competitive when combined with retrieval, ensembling, or calibration strategies. This highlights that inference-time design choices, including prompting, aggregation, and post-processing, play a critical role alongside model size.

Finally, we observe widespread adoption of retrieval-augmented and ensemble-based LLM pipelines among top-performing systems. Rather than relying on a single model pass, leading approaches combine multiple reasoning steps, retrieved examples, or model outputs to improve robustness and calibration. This trend reinforces the importance of structured inference strategies in complex, multi-task psychological modeling settings. Overall, these findings indicate that while LLMs are central to performance, success depends less on any single model family and more on how models are orchestrated through retrieval, prompting, and aggregation strategies.

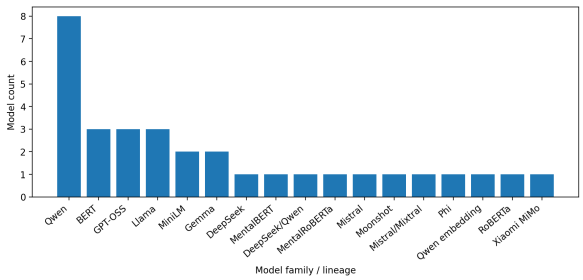

Figure 5: Model families used in the best runs of official submissions (including LLMs and PLMs). While Llama-based models are present, top-performing systems are predominantly based on alternative open-weight models such as Qwen, DeepSeek, and Gemma, indicating a shift toward a more diverse LLM ecosystem.

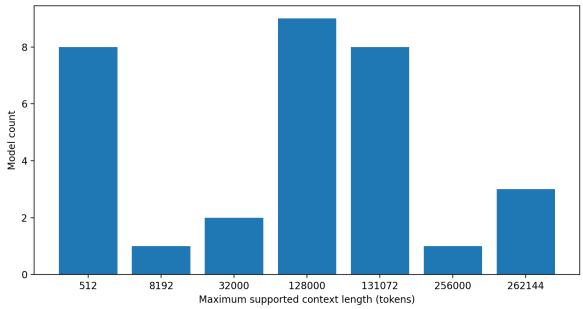

Figure 6: Maximum context length of LLMs used in best runs. Although several systems support very long contexts (100K+ tokens), most high-performing approaches rely on selective context construction (e.g., retrieval or recent posts) rather than fully utilizing the maximum available context.

C.4 Performance Analysis

We analyze system performance across Task 1.1 (subelement classification), Task 1.2 (presence estimation), Task 2 (moment-of-change detection), and Task 3 (summarization and explanation generation), focusing on how modeling choices influence leaderboard outcomes based on the best submission per team.

Overall Trends. Across tasks, we observe substantial variation in performance ceilings and system rankings. For **Task 1.1**, the best-performing system (CUNY) achieves an Average Subelement Macro F1 of 0.442. For **Task 1.2**, Meronym Labs achieves the lowest AvgRMSE of 0.917, followed closely by USAI (0.942) and Aurevia (0.981). In **Task 2**, USAI achieves the highest combined macro

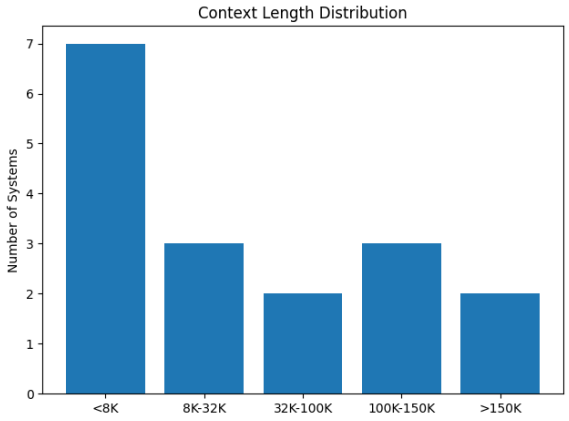

Figure 7: Model size distribution (in parameters) across best-performing systems. While larger models (50B+) achieve strong performance, mid-sized models (10–50B) remain competitive when combined with retrieval, ensembling, and calibration strategies, indicating that system design plays a key role alongside scale.

F1 of 0.600, followed by CtbuY (0.588) and JNLP (0.580). For **Task 3.1**, Meronym Labs ranks first overall based on average rank across evaluation metrics, while DreamerNLplus and CUNY rank second and third, respectively. For **Task 3.2**, DreamerNLplus achieves the best overall score for improvement explanations (0.7608), whereas CSE_IIT_Ropar achieves the best deterioration explanation score (0.7891). These results illustrate increasing task complexity from structured classification and regression toward temporally grounded generation and explanation.

Task 1.1: Subelement Classification. Task 1.1 exhibits a relatively wide spread in system performance, with AvgSub scores ranging from 0.175 to 0.442. Top-performing systems include CUNY (0.442), StateofMIND (0.441), Meronym Labs (0.420), USAI (0.410), and Aurevia (0.381). These systems primarily rely on LLM-based inference combined with ensembling, prompting, or retrieval strategies. Mid-performing systems such as MKC (0.361), McMasterNLP (0.351), and ull (0.335) remain competitive despite using smaller or more specialized architectures. Lower-performing systems such as CSE_IIT_Ropar (0.175), DrosophilAI (0.179), and debju (0.195) highlight the difficulty of fine-grained adaptive and maladaptive subelement prediction. Performance differences are particularly pronounced for adaptive subelements, which consistently achieve lower scores than maladaptive subelements across systems. This suggests that adaptive signals are more heterogeneous and difficult to identify than maladaptive indicators.

Task 1.2: Presence Estimation. Task 1.2 reveals a different ranking structure from Task 1.1. Meronym Labs achieves the best AvgRMSE (0.917), followed by USAI (0.942), Aurevia (0.981), CUNY (0.989), and StateofMIND (0.994). These results indicate that strong classification performance does not necessarily translate into optimal regression calibration. Systems also specialize across evaluation metrics. USAI achieves the strongest agreement with gold annotations, obtaining the highest Combined Quadratic Weighted Kappa (0.730) and Spearman correlation (0.752), whereas Meronym Labs achieves the lowest prediction error overall. In contrast, systems such as BLUE (1.606 AvgRMSE), CtbuY (1.501), and psytechlab (1.407) perform substantially worse, suggesting that regression-based presence estimation is highly sensitive to calibration and score normalization.

Team-Level Prediction Characteristics: Team-level statistics showed substantial variability in prediction coverage and intensity (Table 14). Adaptive coverage ranged from 34.78% to 98.91%, while maladaptive coverage ranged from 48.91% to 100.00%. Mean adaptive presence scores ranged from 1.42 to 3.51, compared to 1.89 to 3.70 for maladaptive states.

Metric	Min	Max
Adaptive Coverage (%)	34.8	98.9
Maladaptive Coverage (%)	48.9	100.0
Adaptive Mean Presence	1.42	3.51
Maladaptive Mean Presence	1.89	3.70
Adaptive Avg. Elements	0.00	3.62
Maladaptive Avg. Elements	0.90	4.55

Table 14: Team-level prediction range summary.

Statistical Significance Analysis: Welchs t-test confirmed a significant difference between adaptive and maladaptive presence scores (Table 15), with maladaptive states receiving higher scores overall.

Metric	Value
t-statistic	-13.38
p-value	1.3×10^{-39}
ADT Mean	2.18
MAL Mean	2.83

Table 15: Welchs t-test comparing adaptive and maladaptive presence scores.

Relationship Between Tasks 1.1 and 1.2. Performance across Task 1.1 and Task 1.2 demonstrates a moderate inverse relationship. Systems

Team	Model Type
USAI	LLM
CtbuY	Small LM
JNLP	LLM
CUNY	ML
MKC	Small LM
Codezone	Small LM
Aurevia	LLM
Meronym Labs	LLM
debju	LLM
DreamerNLplus	ML
McMasterNLP	Small LM
BLUE	LLM
NoviceTrio	Small LM
DrosophilAI	LLM
psytechlab	Small LM
CSE_IIT_Ropar	Small LM
Afrilan	ML
Cloud	LLM
TempoFormer	Small LM
Llama ZS	LLM

Table 16: Broad model categories used by Task 2 systems.

achieving stronger AvgSub scores generally obtain lower AvgRMSE values, although the relationship is not perfectly linear. For example, CUNY ranks first in Task 1.1 but fourth in Task 1.2, whereas Meronym Labs improves from third place in Task 1.1 to first place in Task 1.2. Similarly, USAI performs strongly across both tasks despite relying on different aggregation strategies. These findings indicate that classification and regression subtasks require partially distinct optimization objectives and inference strategies.

Task 2: Moment-of-Change Detection. Task 2 demonstrates the importance of temporal reasoning and cross-post contextual modeling. USAI achieves the highest combined macro F1 (0.600), followed closely by CtbuY (0.588), JNLP (0.580), MKC (0.554), and Codezone (0.553). In contrast, weaker systems such as Cloud (0.260) and Afrilan (0.268) perform substantially worse, highlighting the difficulty of detecting Switches and Escalations from longitudinal timelines. Performance varies considerably across subtasks. JNLP achieves the best post-level escalation F1 (0.738), Codezone achieves the strongest timeline escalation F1 (0.719), and MKC achieves the best timeline switch F1 (0.514). These differences indicate that systems specialize in distinct temporal reasoning capabilities. Architectures explicitly modeling timeline structure and temporal dependencies generally outperform static or post-independent approaches. Table 16 shows broad model categories used by Task 2 systems.

Task 3.1: Summarization Quality. Task 3.1 introduces a multi-dimensional evaluation framework combining consistency, contradiction, ROUGE-L recall, and BERTScore recall. Meronym Labs achieves the best overall ranking, followed by DreamerNLplus and CUNY. However, performance differs substantially across individual metrics. Aurevia achieves the highest consistency score (0.866), psytechlab achieves the lowest contradiction score (0.571), USAI achieves the best ROUGE-L recall (0.333), and USAI also achieves the highest BERTScore recall (0.365). Systems optimized for lexical overlap metrics do not necessarily achieve the best faithfulness metrics, illustrating an important trade-off between semantic similarity and factual consistency. These findings emphasize that summarization quality in mental health timelines cannot be captured adequately using a single evaluation metric.

Task 3.2: Explanation Generation. Task 3.2 evaluates explanation generation for improvement and deterioration scenarios. For improvement explanations, DreamerNLplus achieves the highest overall score (0.7608), followed by CSE_IIT_Ropar (0.7426) and MKC (0.7266). For deterioration explanations, CSE_IIT_Ropar performs best overall (0.7891), followed by Aurevia (0.6761) and DreamerNLplus (0.6038). Systems demonstrate complementary strengths across evaluation dimensions. CSE_IIT_Ropar performs particularly well on fit and specificity, psytechlab and CtbuY achieve perfect recurrence scores in some settings, and DreamerNLplus achieves the strongest specificity for improvement explanations. However, variance across systems remains high, indicating the difficulty of jointly optimizing fit, recurrence, and specificity.

Method-Level Trends. Method-level analysis reveals several important trends (Tables 17, 18). RAG-based systems achieve higher average Task 1.1 performance (0.334 AvgSub) than non-RAG systems (0.307), but worse Task 1.2 performance (1.246 vs. 1.128 AvgRMSE), indicating that retrieval improves classification more consistently than regression calibration (Table 17). Pipeline-based systems also outperform non-pipeline systems in Task 1.1 (0.330 vs. 0.292 AvgSub), but achieve slightly worse regression performance in Task 1.2 (1.185 vs. 1.141 AvgRMSE), as shown in Table 17. Temporal systems perform substantially better in Task 2 (0.498 combined macro F1) than

Method	Teams	Task 1.1 AvgSub ↑	Task 1.2 AvgRMSE ↓
RAG	BLUE, CibuY, DreamerNLplus, Meronym Labs, StateOfMIND, ull	**0.334**	1.246
No RAG	Afrilan, Aurevia, CSE_IIT_Ropar, CUNY, DrosophilAI, MKC, McMasterNLP, NoviceTrio, psytechlab, USAI, debju	0.307	**1.128**
Pipeline	Aurevia, BLUE, CibuY, CUNY, DreamerNLplus, MKC, Meronym Labs, NoviceTrio, psytechlab, StateOfMIND, ull, USAI	**0.354**	1.120
No Pipeline	Afrilan, CSE_IIT_Ropar, DrosophilAI, McMasterNLP, debju	0.210	**1.099**
Temporal	Aurevia, BLUE, Codezone, CSE_IIT_Ropar, CibuY, CUNY, DreamerNLplus, JNLP, MKC, Meronym Labs, psytechlab, USAI	0.354	1.121
Non-temporal	Afrilan, DrosophilAI, McMasterNLP, NoviceTrio, StateOfMIND, ull, debju	0.282	**1.201**

Table 17: Method-level performance on Task 1.1 (Average Subelement Macro F1) and Task 1.2 (Average RMSE), computed using the method taxonomy from Table 13.

Approach	Teams	Task 2 Combined F1 ↑
Temporal	Aurevia, BLUE, Codezone, CSE_IIT_Ropar, CibuY, CUNY, DreamerNLplus, JNLP, MKC, Meronym Labs, psytechlab, USAI	**0.498**
Non-temporal	Afrilan, Cloud, DrosophilAI, McMasterNLP, NoviceTrio, debju	0.359

Table 18: Impact of temporal modeling on Task 2 using the temporal categorization from Table 13. Explicit temporal modeling improves moment-of-change detection performance.

Model	AvgSub	EP-A	EP-M	EP-Avg	SE-A	SE-M
`Llama-3.3-70B-Instruct`	**0.391**	**0.524**	**0.722**	**0.623**	**0.292**	**0.489**
`Llama-3.1-8B-Instruct`	0.247	0.392	0.605	0.498	0.156	0.338

Table 19: Baseline Task 1.1 results. AvgSub = average subelement macro F1; EP = element presence; SE = subelement; A = adaptive; M = maladaptive.

Model	AvgRMSE	RMSE-M	RMSE-A	QWK	C-RMSE	Spr.	MAE
`Llama-3.3-70B-Instruct`	**1.244**	**1.291**	**1.196**	**0.591**	**1.244**	**0.609**	**0.910**
`Llama-3.1-8B-Instruct`	1.424	1.439	1.409	0.555	1.424	0.603	1.083

Table 20: Baseline Task 1.2 results. AvgRMSE = average RMSE; C-RMSE = combined RMSE; Spr. = Spearman correlation. Lower is better for RMSE metrics; higher is better for QWK and Spr.

non-temporal systems (0.359), supporting the importance of explicit timeline modeling for moment-of-change detection (Table 18).

Effect of Model Size. Larger models generally achieve stronger performance across tasks. Systems using models larger than 50B parameters, including CUNY, Aurevia, and USAI, consistently rank among the best-performing systems in Tasks 1 and 2. However, mid-sized models (10–50B) such as those used by StateofMIND and Meronym Labs remain highly competitive when combined with retrieval, ensembling, or calibration strategies. In contrast, smaller models (<1B) consistently underperform relative to larger systems. These findings suggest that model scale contributes substantially to performance, but inference-time strategies such as retrieval, prompting, and aggregation remain equally important.

Summary. Overall, the results demonstrate that (1) Tasks 1.1 and 1.2 are related but require distinct optimization strategies, (2) Task 2 strongly benefits from temporal reasoning and timeline-aware modeling, (3) Task 3 introduces complex multi-objective generation challenges balancing faithfulness and semantic similarity, (4) retrieval and pipeline strategies provide task-dependent benefits, (5) larger models improve performance but are insufficient alone, and (6) ensemble approaches are particularly effective for structured prediction tasks but less consistently beneficial for generation tasks.

Appendix D Baselines

This section outlines implementation details of our baseline models (§7.1). In LLM-based methods, we employ `Llama-3.1-8B-Instruct`. We also perform ablation experiments using the larger `Llama-3.3-70B-Instruct` model to explore the effects of model size on Task performance. For the text generation Task 3.1, we also perform experiments varying temperature. Unless otherwise indicated, all experiments on generative LLMs use a temperature of 1.0 and a top_p of 0.7.

D.1 Task 1.1

The prompt for Task 1.1 instructs the model to identify predominant adaptive and maladaptive subelements across the six ABCD dimensions: Affect (A), Behavior toward Other (BO), Behavior toward Self (BS), Cognition toward Other (CO), Cognition toward Self (CS), and Desire (D). We frame this as a classification problem with label definitions embedded directly in the prompt. The labeling protocol enforces sparsity constraints: each element-valence pair may have at most one dominant sub-element, with null designations used when no element is present. Scores for both Llama models are shown in Table 19.

D.2 Task 1.2

The prompt for Task 1.2 is shared with Task 1.1 and instructs the model to assign integer presence scores ranging from 1 to 5 for both adaptive and maladaptive self-states overall. A constraint requires that any all-null valence configuration receive a score of 1. The prompt explicitly defines "presence" as psychological centrality rather than mere lexical frequency. Scores for both Llama models are shown in Table 20.

D.3 Task 2

Baseline 1 uses zero-shot prompting to elicit well-being scores, which are used to predict switches, and escalation labels. Using the baseline predic-

Method	Comb.	P-Esc	P-Mac	P-Sw	T-Esc	T-Mac	T-Sw
TempoFormer	**0.572**	**0.667**	**0.561**	**0.456**	**0.736**	**0.583**	0.430
Weighted Sum + T1.1 (70B)	0.394	0.378	0.407	0.436	0.335	0.380	0.425
Weighted Sum + T1.1 (8B)	0.365	0.407	0.383	0.359	0.343	0.346	0.350
Llama-70B ZS	0.317	0.000	0.219	0.438	0.400	0.416	**0.432**
Llama-8B ZS	0.272	0.000	0.169	0.337	0.400	0.376	0.352

Table 21: Baseline Task 2 results. Comb. = combined macro F1; P = post-level; T = timeline-level; Esc = escalation; Mac = macro F1; Sw = switch.

tions for Task 1.1, **Baseline 2** calculates well-being using the following formula, clamping the score between 1 and 10 inclusive:

Well-being $= \text{Round}((1 - mal\_ratio) \cdot 9 + 1 + adap\_bonus)$

Where:

mal_ratio = $\frac{\text{mal\_score}}{\text{weight\_sum}}$

mal_score = weighted sum of maladaptive elements identified in Task 1.1.

The weights used in the weighted sum are as follows: 'A'=2.0, 'BS'=1.8, 'D'=1.5, 'CS'=1.2, 'CO'=1.0, 'BO'=0.8.

weight_sum $= 2.0 + 1.8 + 1.5 + 1.2 + 1.0 + 0.8$

adap_bonus = $0.5\cdot$ count of adaptive elements identified in Task 1.1.

The TempoFormer (**Baseline 3**) was fine-tuned with a window size of 5 for Switches and 10 for Escalations (based on differences in the guideline definitions), over 4 epochs with a learning rate of 1e-5, using focal loss to address class imbalance. The first 6 layers of the model were frozen to account for the small number of training samples. Scores for all baseline experiments using both Llama models and TempoFormer are shown in Table 21.

D.4 Task 3.1

Baseline 1 employs zero-shot prompting on `Llama-3.1-8B-Instruct` using a temperature of 0.2. **Baseline 2** was formulated as a pipeline by using the model from Baseline 1 and feeding the predictions from the top performing baselines from Task 1 and Task 2 (TempoFormer) within the prompt. Additional experiments were performed for each approach, varying the temperature and using the larger `Llama-3.3-70B-Instruct` model. These results are presented in Table 22.

D.5 Task 3.2

The Task 3.2 Baseline is described in full detail by §7.1.

Model	Temp	CS	CT	R-L	BERT-R
Llama-8B	*0.2*	*0.763*	*0.753*	*0.255*	*0.226*
	0.4	0.774	0.694	0.246	**0.237**
	0.6	0.810	0.665	0.247	0.231
	0.8	0.791	0.704	0.255	0.230
	1.0	0.792	0.705	0.238	0.235
Llama-70B	0.2	0.819	0.659	0.222	0.222
	0.4	0.821	**0.649**	0.223	0.224
	0.6	0.819	0.660	0.226	0.229
	0.8	0.816	0.663	0.219	0.228
	1.0	**0.824**	0.679	0.220	0.225
Llama-8B Pipeline	*0.2*	*0.763*	*0.745*	***0.269***	*0.235*
	0.4	0.738	0.764	0.260	0.227
	0.6	0.722	0.801	0.250	0.231
	0.8	0.730	0.796	0.256	0.226
	1.0	0.756	0.745	0.246	0.227
Llama-70B Pipeline	0.2	0.759	0.700	0.235	0.229
	0.4	0.760	0.728	0.231	0.227
	0.6	0.757	0.742	0.233	0.227
	0.8	0.771	0.708	0.234	0.225
	1.0	0.762	0.768	0.234	0.221

Table 22: Baseline Task 3.1 results. CS = consistency; CT = contradiction; R-L = ROUGE-L; BERT-R = BERTScore. Scores used as official baselines have been *italicized*.

Appendix E Prompts used for Tasks 1, 2, and 3

This section presents the prompts used in the baseline methods. Listing 2 provides the prompt for Tasks 1.1 and 1.2 baselines. Listing 3 provides the prompt for Tasks 2 baseline for Well-being scores. Listing 4 provides the prompt for Tasks 2 baseline for escalation labeling. Listing 5 provides the prompt for Tasks 3.1 Baseline 1. Listing 6 provides the prompt for Tasks 3.1 Baseline 2. Listing 7 provides the prompt for Tasks 3.2 baseline.

Appendix F Examples of Expected Input and Output for Tasks 1, 2, and 3

This section provides illustrative examples for each task to clarify the expected input and output formats. An example of identifying adaptive and maladaptive self-state compositions is shown in Listing 8. An example of assigning adaptive and maladaptive presence scores is shown in Listing 9. An example of escalation and switch label prediction across posts is shown in Listing 10. An example of generating a structured psychological summary over a sequence is shown in Listing 11. Examples of recurrent dynamic patterns of deterioration and improvement are shown in Listing 12.

```
You are writing a GOLD Task 3.1 Sequence Summary. Follow instructions exactly.

Task 3.1: Summarising sequences surrounding change events

The goal of Task 3.1 is to generate a structured summary describing patterns of self-state dynamics and their progression over
       time within a sequence of posts surrounding a change (Switch and/or Escalation). Each sequence includes posts leading
      up to the change, as well as posts marking the change itself. The summary must describe how psychological change
      processes evolve across the sequence, and how they culminate in (Switch), or unfold through (Escalation), the identified
       change event. The direction of the change (improvement/deterioration) must also be indicated.

Switch: A switch reflects a substantial change in well-being between two consecutive posts. A switch occurs when |Wellbeing(t)
        Wellbeing(t1)|  2.

Escalation: An escalation refers to a gradual intensification of mood across a sequence of consecutive posts.

Use:
- post text
- change labels
- self-state presence scores
- ABCD subcategory compositions

The summary must describe:
- the central psychological theme across the sequence
- within-state ABCD dynamics
- between-state dynamics
- relative adaptive/maladaptive dominance
- how dynamics evolve before and during the change

When referencing ABCD elements, use:
(A), (B-S), (B-O), (C-S), (C-O), (D)

STYLE RULES:
- Write ONE coherent paragraph only.
- Use process-oriented relational language.
- Do not list ABCD labels independently.
- Integrate ABCD elements into relational dynamics.

HARD CONSTRAINTS:
- Output must be EXACTLY ONE SINGLE PARAGRAPH.
- Output MUST start EXACTLY with:
  Sequence summary:
- Explicitly state:
  - improvement OR deterioration
  - whether change culminates in a Switch or unfolds through an Escalation
- Anchor where change occurs:
  - early / mid / late within the sequence
- Mention self-state dominance verbally only.
- DO NOT output numeric presence scores.
- DO NOT invent content.

Narrative structure:
1. Introduce the central psychological theme.
2. Describe pre-change adaptive/maladaptive dynamics.
3. Describe relative state dominance and interactions.
4. Describe cross-state ABCD dynamics.
5. Describe how dynamics intensify or shift.
6. Describe the configuration at the change point.

--- SEQUENCE BLOCK ---
{sequence_block}

Return ONLY the final summary paragraph.
```

Listing 1: Prompt for Task 3.1 sequence summarization baseline.

```
You are labeling ONE social media post for the CLPsych 2026 Shared Task.

Perform both tasks below.

TASK 1.1  Dominant ABCD subelements and self-state composition
Identify which predefined ABCD subelements are meaningfully expressed in the post.

For each ABCD element:
- A = Affect
- BO = Behavior toward Other
- BS = Behavior toward Self
- CO = Cognition toward Other
- CS = Cognition toward Self
- D = Desire

Evaluate adaptive and maladaptive expressions separately.

Within a single element and valence:
- choose at most ONE dominant subelement

Within a single element:
- it is allowed to output both one adaptive and one maladaptive subelement
  if both are clearly expressed

If an element is not expressed for a given valence, use null.

TASK 1.2  Presence scores
Assign:
- adaptive_presence
- maladaptive_presence

using this scale:
1 = Not present
2 = Somewhat present
3 = Moderately present
4 = Much present
5 = Highly present

Important rules:
- If the adaptive self-state is not expressed at all (all adaptive element values are null),
  adaptive_presence MUST be 1.
- If the maladaptive self-state is not expressed at all (all maladaptive element values are null),
  maladaptive_presence MUST be 1.
- Presence reflects psychological centrality and influence, not mere word frequency.

Allowed labels (must match exactly):

Adaptive:
A: {ALLOWED["A_adaptive"]}
BO: {ALLOWED["BO_adaptive"]}
BS: {ALLOWED["BS_adaptive"]}
CO: {ALLOWED["CO_adaptive"]}
CS: {ALLOWED["CS_adaptive"]}
D: {ALLOWED["D_adaptive"]}

Maladaptive:
A: {ALLOWED["A_maladaptive"]}
BO: {ALLOWED["BO_maladaptive"]}
BS: {ALLOWED["BS_maladaptive"]}
CO: {ALLOWED["CO_maladaptive"]}
CS: {ALLOWED["CS_maladaptive"]}
D: {ALLOWED["D_maladaptive"]}

Return ONLY one JSON object with exactly this structure:
{{
  "task1_1": {{
    "adaptive": {{
      "A": null,
      "BO": null,
      "BS": null,
      "CO": null,
      "CS": null,
      "D": null
    }},
    "maladaptive": {{
      "A": null,
      "BO": null,
      "BS": null,
      "CO": null,
      "CS": null,
      "D": null
    }}
  }},
  "task1_2": {{
    "adaptive_presence": 1,
    "maladaptive_presence": 1
  }}
}}

Do not output any explanation, prose, markdown, comments, or extra keys.

{ONE_SHOT_EXAMPLE}

### Input Post:
{{post_text}}

### Output JSON:
```

Listing 2: Prompt for Task 1.1 and Task 1.2 baselines.

```
You will be given ONE social media post from a single user.

Estimate the user's wellbeing on a 1-10 scale, where:
- 10 = excellent psychological and social functioning, minimal symptoms
- 1  = severe dysfunction or persistent risk

Wellbeing reflects overall functioning across:
- emotional distress
- hopelessness or suicidality
- motivation and coping
- social/interpersonal functioning
- daily functioning or impairment
- coherence/stability of sense of self

Important guideline:
- The scale is clinically anchored, not relative within the same individual.
- If the post does not provide strong evidence for major change, estimate
    conservatively.
- Clear evidence is required to justify strong improvement or deterioration.
- If evidence is partial or ambiguous, make only a minimal adjustment.

Return ONLY a single JSON object exactly in this format:
{"wellbeing": <integer from 1 to 10>}

Post:
"{post_text}"
```

Listing 3: Well-being Prompt for Task 2 Baseline 1.

```
You are given a chronologically ordered timeline of posts from one individual.

Your task is to identify Escalation events.

Definition:
An Escalation is a gradual intensification of mood over a SEQUENCE of
consecutive posts. It occurs when mood progressively shifts from neutral
or mildly valenced toward a more extreme state across multiple posts.

An escalation may reflect:
- gradual deterioration ("d")
- gradual improvement ("i")

Important distinctions:
- Escalation is gradual, not abrupt.
- Escalation is independent from Switch events.
- Only mark spans that clearly show progressive intensification across
  at least 2 consecutive posts.

Output requirements:
Return ONLY one JSON object with exactly one key:
{
  "escalations": [
    {"start": <int>, "end": <int>, "direction": "i" or "d"},
    ...
  ]
}

Rules:
- start and end are inclusive indices in the timeline order shown below.
- Only include spans that are clearly gradual.
- Keep at most {max_escalations} escalation spans.
- Do not include explanation or extra keys.

Timeline:
{timeline_block}
```

Listing 4: Escalation Prompt for Task 2 Baseline 1.

```
### Task
Your task is to generate a structured summary describing patterns of self-state dynamics and their
     progression over time within a sequence of posts surrounding a change (Switch and/or Escalation).

### Definitions
- Switch: A switch reflects a substantial and sudden change in well-being between two consecutive posts. A
     switch occurs when |Wellbeing(t)  Wellbeing(t1)|  2. The change may reflect either improvement or
     deterioration.
- Escalation: An escalation refers to a gradual intensification of mood over a sequence of consecutive posts
     . It occurs when an individuals mood progressively shifts from neutral or mildly valenced, toward a
     more extreme state (very negative or very positive) across a span of posts.

### Summary Content
The summary should include references, only when they are evident in the data, to the following aspects:
1. Central recurring theme across the posts: Describe the central dynamic psychological theme and change
     trajectory characterizing the change process across the sequence, articulated in terms of ABCD
     subelements: Affect (A); Behavior toward self (B-S) and Behavior toward others (B-O); Cognition toward
     self (C-S) and Cognition toward others (C-O); and Desire (D), grounded in concrete content from the
     posts. Explain how this theme evolves across the sequence. The description of the theme should reflect
     the relational dynamics between these ABCD subelements within and between self-states, as defined in
     the sections below. The theme should be described across the stages of the change process within the
     sequence, making clear how the theme appears before the change and how it develops as the change
     unfolds (when the change is that of escalation) or when it culminates (when the change is that of
     switch).
2. Dynamics within the Adaptive and Maladaptive self-states and their presence: Describe how present each
     self-state is and how its relative presence changes throughout the sequence as part of the change
     process. Presence refers to how strongly each self-state is expressed or dominant at different points
     in the sequence, whereas dynamics refer to the interactions between the ABCD subelements within that
     self-state. Where present, describe the adaptive and/or the maladaptive self-states in terms of Affect,
      Behavior, Cognition and Desire, expressing these ABCD subelements through explicit relational dynamics
      between them within the same self-state. If a self-state is described, relational dynamics between its
      ABCD subelements MUST also be described. Dynamics within a self-state are relational patterns between
     two or more subelements within the same self-state. These dynamics may be directional or reciprocal,
     such as co-activation, mutual reinforcement, exacerbation of one element by another, amplification of
     one element by another, or other structured interactions. This list is not exhaustive; other forms of
     dynamics may also be identified where conceptually appropriate. These relational dynamics between
     subelements within a self-state should be described as they unfold across the sequence, noting how they
      change over time.
3. Relationship between the adaptive and maladaptive self states and their relative presence: Describe how
     the adaptive and maladaptive self-states relate to one another and how that changes throughout the
     sequence. Describe how the relative presence and dominance of the adaptive and maladaptive self-states
     shifts across the sequence. This may include: one self-state dominating the other, suppressing or
     silencing the other, or both self-states coexisting through reflective dialogue. Use presence scores to
      support comparison over the sequence, but do not print numeric scores in the output. Additionally,
     examine whether dynamics occur between ABCD subelements across opposite self-states (suppression/
     attenuation, reflective dialogue, dominance competition, resilience or other structured interactions).
     If such crossself-state dynamics are present in the sequence, they MUST be described.

### Summary Guidelines
- The identity of the change event (Switch or Escalation) should be explicitly stated in the summary
- The direction of the change (improvement/deterioration) must be indicated - several changes in direction
     of change may be present in a sequence.
- It should be explicitly stated what dynamics characterized the pre-change phase, and how these dynamics
     culminated in the change event (in the case of a Switch) or unfolded through the change process (in the
      case of an Escalation).
- It should be explicitly stated in the summary when the change event occurs. Please note that a Switch
     event occurs within a single post, whereas an escalation unfolds across multiple posts.

### Guidelines for the Output
- The generated summary must not exceed 350 words.
- The summary should be a paragraph of text, not in a list. Do not add sections or titles.
- Generate only the summary as described, do not add any extra text or explanations.

### Input
{{INPUT_SEQUENCE_POST_TEXT_ONLY}}

### Output
```

Listing 5: Prompt for Task 3.1 Baseline 1.

```
### Task
Your task is to generate a structured summary describing patterns of self-state dynamics and their
     progression over time within a sequence of posts surrounding a change (Switch and/or Escalation).

### Input Description
Along with the post text, the adapative and maladaptive presence scores, and composition are give. It is
     also given whether if a switch or escalation is present for each post.

### Definitions
- Switch: A switch reflects a substantial and sudden change in well-being between two consecutive posts. A
     switch occurs when |Wellbeing(t)  Wellbeing(t1)|  2. The change may reflect either improvement or
     deterioration.
- Escalation: An escalation refers to a gradual intensification of mood over a sequence of consecutive posts
     . It occurs when an individuals mood progressively shifts from neutral or mildly valenced, toward a
     more extreme state (very negative or very positive) across a span of posts.

#### Presence Scale
1 = Not present
2 = Somewhat present
3 = Moderately present
4 = Much present
5 = Highly present

### Summary Content
The summary should include references, only when they are evident in the data, to the following aspects:
1. Central recurring theme across the posts: Describe the central dynamic psychological theme and change
     trajectory characterizing the change process across the sequence, articulated in terms of ABCD
     subelements: Affect (A); Behavior toward self (B-S) and Behavior toward others (B-O); Cognition toward
     self (C-S) and Cognition toward others (C-O); and Desire (D), grounded in concrete content from the
     posts. Explain how this theme evolves across the sequence. The description of the theme should reflect
     the relational dynamics between these ABCD subelements within and between self-states, as defined in
     the sections below. The theme should be described across the stages of the change process within the
     sequence, making clear how the theme appears before the change and how it develops as the change
     unfolds (when the change is that of escalation) or when it culminates (when the change is that of
     switch).
2. Dynamics within the Adaptive and Maladaptive self-states and their presence: Describe how present each
     self-state is and how its relative presence changes throughout the sequence as part of the change
     process. Presence refers to how strongly each self-state is expressed or dominant at different points
     in the sequence, whereas dynamics refer to the interactions between the ABCD subelements within that
     self-state. Where present, describe the adaptive and/or the maladaptive self-states in terms of Affect,
      Behavior, Cognition and Desire, expressing these ABCD subelements through explicit relational dynamics
      between them within the same self-state. If a self-state is described, relational dynamics between its
      ABCD subelements MUST also be described. Dynamics within a self-state are relational patterns between
     two or more subelements within the same self-state. These dynamics may be directional or reciprocal,
     such as co-activation, mutual reinforcement, exacerbation of one element by another, amplification of
     one element by another, or other structured interactions. This list is not exhaustive; other forms of
     dynamics may also be identified where conceptually appropriate. These relational dynamics between
     subelements within a self-state should be described as they unfold across the sequence, noting how they
      change over time.
3. Relationship between the adaptive and maladaptive self states and their relative presence: Describe how
     the adaptive and maladaptive self-states relate to one another and how that changes throughout the
     sequence. Describe how the relative presence and dominance of the adaptive and maladaptive self-states
     shifts across the sequence. This may include: one self-state dominating the other, suppressing or
     silencing the other, or both self-states coexisting through reflective dialogue. Use presence scores to
      support comparison over the sequence, but do not print numeric scores in the output. Additionally,
     examine whether dynamics occur between ABCD subelements across opposite self-states (suppression/
     attenuation, reflective dialogue, dominance competition, resilience or other structured interactions).
     If such crossself-state dynamics are present in the sequence, they MUST be described.

### Summary Guidelines
- The identity of the change event (Switch or Escalation) should be explicitly stated in the summary
- The direction of the change (improvement/deterioration) must be indicated - several changes in direction
     of change may be present in a sequence.
- It should be explicitly stated what dynamics characterized the pre-change phase, and how these dynamics
     culminated in the change event (in the case of a Switch) or unfolded through the change process (in the
      case of an Escalation).
- It should be explicitly stated in the summary when the change event occurs. Please note that a Switch
     event occurs within a single post, whereas an escalation unfolds across multiple posts.

### Guidelines for the Output
- The generated summary must not exceed 350 words.
- The summary should be a paragraph of text, not in a list. Do not add sections or titles.
- Generate only the summary as described, do not add any extra text or explanations.

### Input
{{INPUT_SEQUENCE_POST_TEXT_AND_PIPELINE}}

### Output
```

Listing 6: Prompt for Task 3.1 Baseline 2.

```
### Task
Your goal is to identify and summarize recurrent dynamic patterns of deterioration (
    a switch towards lower well-being, or an escalation of deterioration) or
    improvement (a switch towards higher well-being, or an escalation of improvement
    ) that recur across multiple sequences.

### Input Description
You will be given a series of post sequence summaries. These summaries describe
    patterns of self-state dynamics and their progression over time within a
    sequence of posts surrounding a change (Switch and/or Escalation). You will also
     be given their sequence_id and timeline_id.

### Definitions
- Switch: A switch reflects a substantial and sudden change in well-being between
    two consecutive posts. A switch occurs when |Wellbeing(t)  Wellbeing(t1)|  2.
    The change may reflect either improvement or deterioration.
- Escalation: An escalation refers to a gradual intensification of mood over a
    sequence of consecutive posts. It occurs when an individuals mood progressively
    shifts from neutral or mildly valenced, toward a more extreme state (very
    negative or very positive) across a span of posts.
- Signature: Describes a recurrent dynamic pattern (ABCD + self-state relations)
    observed across individuals that leads to and culminates as the change occurs.

### Pattern Identification
Across timelines, identify the most frequent patterns in:
- Central Theme: Recurrent themes of deterioration or improvement articulated in
    ABCD elements
- Dynamics within self-states: Recurrent ABCD configurations (separately for the
    maladaptive and adaptive self states), dynamics among subelements within the
    same self-state, changes in presence of a self state across the sequence leading
     to and through the change event.
- Dynamics between self-states: Patterns in the balance between self-states in terms
     of presence over time (patterns of dominance versus coexistence), is there a
    dialogue/reflection between the self states or silencing dominance and how does
    this change over time, dynamics between subelements from different self-states
    over time.

### Output Content

#### Signature Summary
- Based on the identified patterns, you must extract both extract 1 signature of
    deterioration and 1 signature of improvement.
- A signature does not have to include both within-state and between-state dynamics;
     it may be based on recurrent dynamics within a self-state, between self-states,
     or a combination of both. The more details a signature contains, the higher it
    will be scored; therefore, a combination of details will yield a higher score.

### List of Evidence
- Along with each signature summary, list the sequence_id and timeline_id of
    supporting evidence.
- In the case of a signature of improvement, identify which sequence_id and
    timeline_id exhibits this signature.
- In the case of a signature of deterioration, identify which sequence_id and
    timeline_id exhibits this signature.

### Guidelines for the Output
- Separate your response into 2 sections: '### Signature of Deterioration' and '###
    Signature of Improvement'.
- Output the signature summary, then the list of evidence.
- Each signature summary should not exceed 90 words.

### Input
{{INPUT_SEQUENCE_SUMMARIES}}

### Output
```

Listing 7: Prompt for Task 3.2 baseline.

```
Example Input post:
Im searching for someone who could help keep me accountable with daily exercise and
    prevent late-night binge eating. I have a heart issue that involves low blood
    pressure (it doesnt increase with activity) along with a high heart rate. I need
     to do light activities, such as walking, every day to maintain my strength and
    energy. If I skip even one day, it becomes very difficult to get out of bed the
    following day, so staying consistent is essential for my health.

Im not overweight, but I sometimes overeat at night and would really value support
    in avoiding that pattern. When I consume a large amount of carbs or calories at
    once, it seems to worsen my symptoms. I feel more lightheaded and notice my
    heart rate climbing. Because of that, I really need to prevent those episodes,
    but at times my depression and lack of motivation take over and I end up
    thinking "forget it" and doing it anyway.

Having someone to talk to during those moments would make a big difference. I
    usually begin binge eating around 1011pm. If anyone is willing to help me stay
    motivated about my health, Id truly appreciate it. We could support and
    encourage one another.


Example output self-state(s) composition:

Adaptive Self state:
A: (5) Content, happy, joy, hopeful
B-O: (1) Relating behavior
B-S: (1) Self care and improvement
C-O: (1) Perception of the other as related
D: (5) Competence, self esteem, self-care

Maladaptive Self state:
A: (4) Depressed, despair, hopeless
B-S: (2) Self harm, neglect and avoidance
D: (6) Expectation that competence needs will not be met
```

Listing 8: Example of Expected Input and Output for Task 1.1.

```
Example Input post:
Im searching for someone who could help keep me accountable with daily exercise and
    prevent late-night binge eating. I have a heart issue that involves low blood
    pressure (it doesnt increase with activity) along with a high heart rate. I need
     to do light activities, such as walking, every day to maintain my strength and
    energy. If I skip even one day, it becomes very difficult to get out of bed the
    following day, so staying consistent is essential for my health. Im not
    overweight, but I sometimes overeat at night and would really value support in
    avoiding that pattern. When I consume a large amount of carbs or calories at
    once, it seems to worsen my symptoms. I feel more lightheaded and notice my
    heart rate climbing. Because of that, I really need to prevent those episodes,
    but at times my depression and lack of motivation take over and I end up
    thinking "forget it" and doing it anyway. Having someone to talk to during those
     moments would make a big difference. I usually begin binge eating around 1011pm
    . If anyone is willing to help me stay motivated about my health, Id truly
    appreciate it. We could support and encourage one another.


Presence scores:

Adaptive Self state presence score: 4

Maladaptive Self state presence score: 2
```

Listing 9: Example of Expected Input and Output for Task 1.2 (Presence Scores).

```
Example Input post:

Post t: "Thanks so much for the advice on how to take care of my hair. Im very
    pleased with the result, and its made me feel so much better about myself".

Post t+1: "Its becoming more and more difficult to manage. Living with a chronic
    illness and trying to cope with everything that comes with it feels overwhelming
     lately. Im struggling a lot. I often feel worthless, and sometimes I convince
    myself that being sick means Im undeserving of love. I keep wondering what the
    point is if I cant have the kind of love and life I want. I know logically that
    other people with chronic illnesses still deserve love and care, so I dont
    understand why I have such a hard time believing that about myself. I dont
    really know how to deal with these feelings anymore. My thoughts have been
    getting darker, and that scares me, especially because Ive been in that place
    before. I know I need help, but Im terrified of being admitted again because my
    past experience at the hospital was awful. The last time I felt ignored and
    unsafe, and it made everything worse instead of better. So right now I feel
    stuck, like I need support, but I dont trust the place that is supposed to help.
     Maybe I just needed to get this out of me. I just feel so unworthy of love and
    even of life sometimes, and its getting harder and harder to keep coping".

Post t+2: "I plan to harm myself. I want to fall asleep and stay that way forever. I
     want to escape this reality. I dont want to feel so awful anymore. I feel as if
     love and kindness are just not meant for me. I want to apologize to anyone whom
     I might have hurt. Please forgive me. Ill try to sleep and hope it feels a
    little lighter later".


Example output of change label:

Post t: 0 (Escalation label), 0 (Switch label)

Post t+1: E (Escalation label), S (Switch label)

Post t+2: E (Escalation label), 0 (Switch label)
```

Listing 10: Example of Expected Input and Output for Task 2 (Change Detection).

```
Example Input: Timeline d0fb4b962e; Posts: 6, 7, 9, 10

Post index 6
MOC label: -
Wellbeing score = 7
Maladaptive presence: 1, Adaptive presence: 3
Maladaptive self-state composition: -
Adaptive self-state composition: B-O - (1) Relating behavior, D - (1) Relatedness
Post: "Does anyone here have a job? I know this isnt really related to the sub, but
    I was curious whether anyone here is employed. If so, what kind of work do you
    do?"

Post index 7
MOC label: -
Wellbeing score = 7
Maladaptive presence: 3, Adaptive presence: 2
Maladaptive self-state composition: A - (4) Depressed, despair, hopeless, B-S - (2)
    Self harm, neglect and avoidance, C-S - (2) Self criticism, D - (6) Expectation
    that competence needs will not be met
Adaptive self-state composition: A - (11) Proud, B-S - (1) Self care and improvement
Post: "After a week, Im back to square one. I was really proud that I managed to go
    seven days without drinking, but I ended up relapsing. And it was to Vodka. It
    makes me feel terrible, thinking that I might never be able to stop drinking
    that shitor alcohol in general."

Post index 9
MOC label: Escalation
Wellbeing score = 6
Maladaptive presence: 4, Adaptive presence: 2
Maladaptive self-state composition: A - (4) Depressed, despair, hopeless, B-S - (2)
    Self harm, neglect and avoidance, B-O - (2) Fight or flight behavior, C-S - (2)
    Self criticism, C-O - (2) Perception of the other as detached or over attached,
    D - (2) Expectation that relatedness needs will not be met
Adaptive self-state composition: B-O - (1) Relating behavior, D - (5) Competence,
    self esteem, self-care
Post: "When relatives or friends bring it up, I wish I could be honest with them...
    Has anyone here spoken with a psychologist? If so, what was your experience like
    , and did it help?"

Post index 10
MOC label: Switch, Escalation
Wellbeing score = 4
Maladaptive presence: 5, Adaptive presence: 2
Maladaptive self-state composition: A - (4) Depressed, despair, hopeless, B-O - (2)
    Fight or flight behavior, C-S - (2) Self criticism, C-O - (2) Perception of the
    other as detached or over attached, D - (6) Expectation that competence needs
    will not be met
Adaptive self-state composition: D - (5) Competence, self esteem, self-care
Post: "All of this feels like its going to affect me in my future responsibilities
    ... I sometimes think it would be better if I werent here causing stress for
    everyone."

Sequence Summary:
The central psychological theme revolves around the writers struggle with persistent
     self-doubt (C-S) and social anxiety (A), shifting from initial attempts at
    engagement (B-O) and self improvement (B-S) towards escalating despair (D),(A).
    Initially, the adaptive self-state is prominent, characterized by the writers
    desire to seek knowledge from others (D), exhibited in relating behavior (B-O).
    Then, the adaptive self-state loses presence while the maladaptive state gains
    dominance. The adaptive state is characterized by the writer exhibiting self-
    care behavior (B-S) which reinforced feelings of pride (A). But the maladaptive
    self-state dominates, overshadowing the adaptive state, instead featuring self-
    criticism (C-S) following engagement in the addictive behavior (B-S),
    intensifying despair (A) and the expectation that improvement is impossible (D)
    ...
```

Listing 11: Example of Expected Input and Output for Task 3.1 (Sequence Summary).

```
Examples of Recurrent Dynamic Signatures of change:
These are mock examples and are not based on real data. They illustrate
high-level dynamic patterns, including both within-self-state and
between-self-state interactions.

Example for signatures of deterioration:
Many sequences of deterioration show a progression from initial self-neglect
    behavior (B-S) into more severe self-harm behaviors (B-S), mediated by a gradual
     intensification of self-criticism (C-S) that is mutually reinforcing with
    anxious affect (A), reflecting the growing dominance of the maladaptive state.
    This intensification increasingly suppresses the adaptive state, where
    perceptions of others as facilitating needs (C-O), driven by the desire for
    competence (D), become overshadowed by the maladaptive dynamic.

Example for signatures of improvement:
Many sequences of improvement show a progression from an initial strong maladaptive
    feedback loop between detached perceptions of others (C-O) and depressive affect
     (A), which suppresses the adaptive state, into a reduced maladaptive presence.
    This is followed by a dominant adaptive feedback loop where self-acceptance (C-S
    ) fuels relating behavior (B-O), mutually reinforcing with the desire for
    autonomy (D).
```

Listing 12: Examples of Recurrent Dynamic Signatures of Change.