Universiteit
Leiden
The Netherlands

DreamerNLplus: interpretable modeling of mental health dynamics from social media timelines using hybrid rule-based and RAG methods

Zhyrko, M.; Lal, D.; Mulligen, E. van; Han, L.; Zirikly, A.; Bar, K.; ... ; Desmet, B.

Citation
Zhyrko, M., Lal, D., Mulligen, E. van, & Han, L. (2026). DreamerNLplus: interpretable modeling of mental health dynamics from social media timelines using hybrid rule-based and RAG methods. *Proceedings Of The 11Th Workshop On Computational Linguistics And Clinical Psychology (Clpsych 2026)*, 458-471. doi:10.18653/v1/2026.clpsych-1.36

Version:	Publisher's Version
License:	Creative Commons CC BY 4.0 license
Downloaded from:	https://hdl.handle.net/1887/4308936

Note: To cite this publication please use the final published version (if applicable).

DreamerNLplus: Interpretable Modeling of Mental Health Dynamics from Social Media Timelines using Hybrid Rule-Based and RAG Methods

Maryia Zhyrko[1†], **Daisy Monika Lal**[2†], **Erik van Mulligen**[3], **Lifeng Han**[1,4]
On behalf of the 4D PICTURE consortium
[1]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, NL
[2]School of Computing and Communications (SCC), Lancaster University, UK
[3]Department of Medical Informatics, Erasmus University Medical Center Rotterdam, NL
[4]Biomedical Data Sciences, Leiden University Medical Center, NL
[†] co-first, corresponding: {l.han} @ lumc.nl

Abstract

We present DreamerNLplus, a hybrid framework for modeling mental health dynamics from social media timelines in the CLPsych 2026 shared task. Our system addresses three tasks: psychological state modeling, temporal change detection, and sequence-level summarization. For Task 1, we combine LLM-based data augmentation, DeBERTa classification, and Random Forest regression for structured state prediction. For Task 2, we use few-shot prompting with a locally deployed Llama 3.1 model to detect Switch and Escalation events using short-term temporal context. For Task 3.1, we explore both a deterministic rule-based summarization pipeline and a few-shot LLM-based approach, ranking **2nd** officially. Our RAG-based method achieves strong performance in Task 3.2, ranking **1st** for Improvement and **3rd** for Deterioration, demonstrating its ability to capture recurrent psychological change patterns across timelines. Our analysis reveals key challenges, including the mismatch between classification and regression performance, the difficulty of modeling temporal transitions, and the disagreement between semantic and similarity-based evaluation metrics. These findings highlight the complexity of modeling mental health dynamics and motivate future work on unified evaluation frameworks. We share our code and prompts at https://github.com/4dpicture/CLPsych2026

1 Introduction and Background

We describe our system submissions to the CLPsych 2026 Shared Task (Ali et al., 2026): Capturing and Characterizing Mental Health Changes through Social Media Timeline Dynamics, where we have attended all the sub-tasks. Earlier reviews on NLP for mental health research (Le Glaz et al., 2021; Malgaroli et al., 2023) provide an overview of traditional NLP and ML methods including rule-based, statistical (TF-IDF, decision trees, SVMs, CRFs, random forests, NNs), pretrained LMs (BERT-like encoder models, sequence-to-sequence BART models, domain-specific MentalBERT), and early generative decoders (e.g., GPT2). They also pointed out the *limitations* of existing works such as low reproducibility issue. The resources used include electronic health records (EHRs), Psychological Evaluation reports, social media and interview data. For this study, we used social media post data from the shared task. To improve model *interpretability* as well as investigating newer open-source models, we applied a hybrid combination of a rule-based method, a random forest classifier, DeBERTa models, RAG, and open-sourced LLMs. Recent work related to ours include RAG (Kermani et al., 2025; Bogdanova et al., 2026), prompt engineering and PA-ISP (perspective-aware) (Chan et al., 2025; Romero et al., 2025; Ren et al., 2025), and CLPsych systems/dataset (Tseriotou et al., 2025; Atzil-Slonim, 2025; Tsakalidis et al., 2022; Atzil-Slonim, 2026).

The three main tasks and their sub-tasks:[1]

- **Task 1:** Predict adaptive and maladaptive ABCD element combinations: (1.1) Post-level identification of dominant ABCD subelements and self-state composition; (1.2) Self-state presence rating.
- **Task 2:** Identify moments of change.
- **Task 3:** Summary of change: (3.1) Summarizing sequences surrounding change events; (3.2) Identifying recurrent dynamic signatures of change across timelines.

2 DreamerNLplus Methods

Figure 1 summarizes the overall DreamerNLplus framework. Across the three shared tasks, our sys-

[1]Task 1: https://www.codabench.org/competitions/14057/ Task 2: https://www.codabench.org/competitions/14703/ Task 3: https://www.codabench.org/competitions/14669/

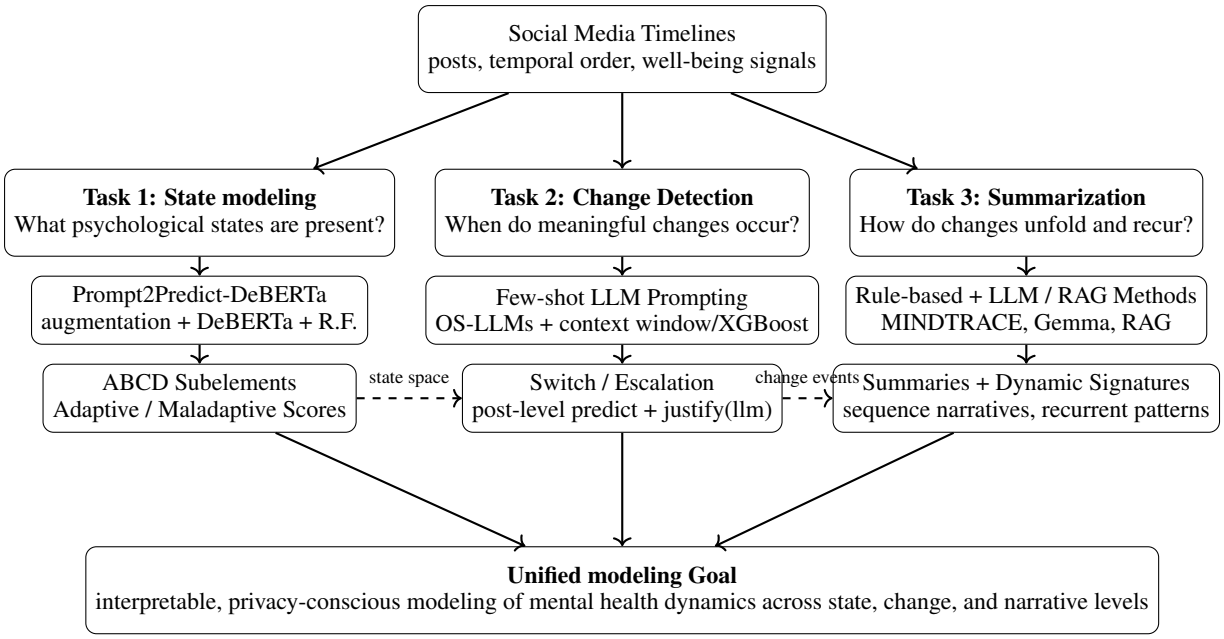


Figure 1: Overview of the DreamerNLplus system across the CLPsych 2026 shared tasks. Task 1 models psychological state representations, Task 2 detects temporal moments of change, and Task 3 summarizes and generalizes change dynamics across sequences.

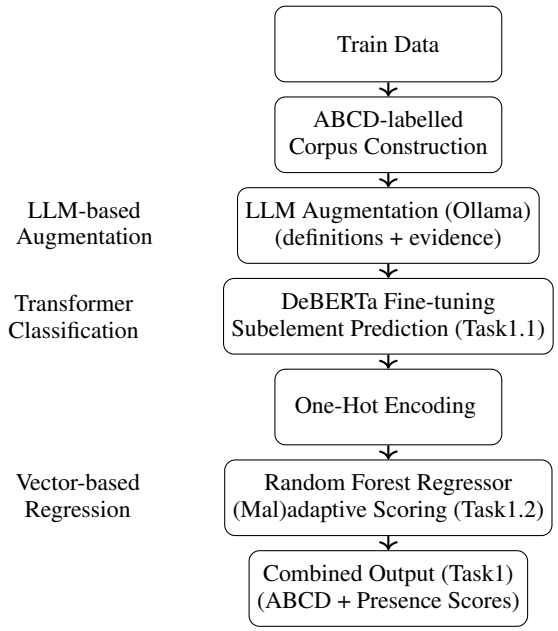


Figure 2: Prompt2Predict-DeBERTa pipeline including data preprocessing for Task 1 (Predict adaptive and maladaptive ABCD element combinations).

tems follow a unified modeling perspective: Task 1 identifies psychological state representations, Task 2 detects transitions between states, and Task 3 converts these dynamics into sequence-level summaries and recurrent signatures.

Tasks 1 and 2 on State Modeling and Change Detection. For **Task-1**, we propose **Prompt2Predict-DeBERTa**, a simple multi-stage framework for predicting psychological subelements and presence scores, as in Figure 2. First, we extract evidence for each label from the original training data and augment the evidence using synthetic data for model training purposes. To do this, we employ Ollama to expand the dataset by *generating new examples* through prompts that include label definitions and annotated evidence for every ABCD category. The augmented data is used for DeBERTa model fine-tuning on the task of subelement prediction (Task1.1). The prediction output (ABCD labels) will be encoded as one-hot vector and fed as input to the Random Forest Regressor to further predict the Adaptive / Maladaptive scoring (Task1.2). Finally, we combine both outputs to include two elements: ABCD labels and their presence ratings. We also deployed a rule-based approach for Task1, as in Figure 3.

This pipeline combines LLM-based augmentation (data preprocessing), transformer-based classification, and lightweight vector-based regression to produce structured and interpretable predictions, as well as comparing well-defined rules.

Task-2: Few-Shot LLM Prompting and XGBoost We first design a framework that supports multiple LLM backends (Ollama, HuggingFace) and runs locally for privacy, with Llama 3.1 8B as our submitted model (Figure 4). In this framework,

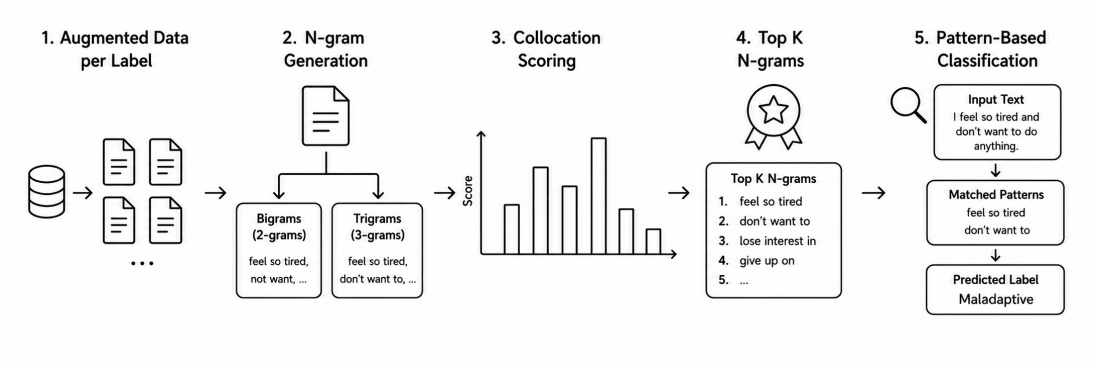


Figure 3: Task 1 rule-based pattern matching approach using n-gram collocations to classify post sentences into ABCD sub-categories.

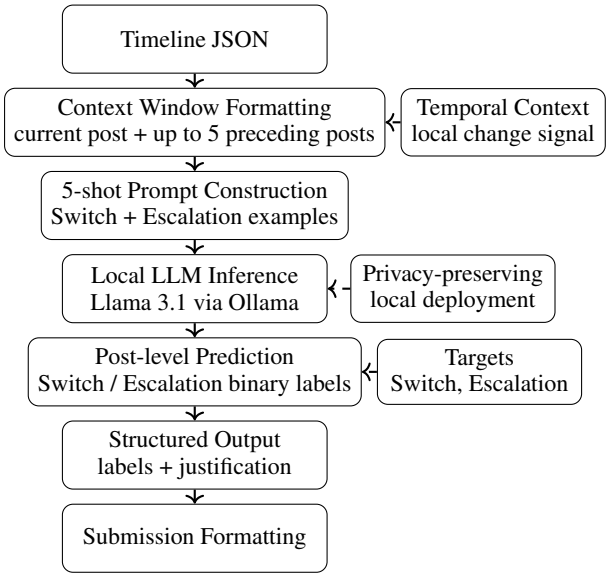


Figure 4: Task 2 few-shot prompting pipeline (Identify moments of change).

we create a local context window by keeping the preceding 5 timeline posts as context to the LLM prompts for predicting targets (Switch and Escalation). The prompt also defines Switch and Escalation, in addition to the context examples. Our examples include the combinations of Switch-only, Escalation-only, both, and neither, plus a first-post case where no change is possible (no preceding context). The LLM is asked to return its answer in a fixed JSON format with Switch, Escalation, and short justification, and the framework retries the prompt if the response does not match this format.

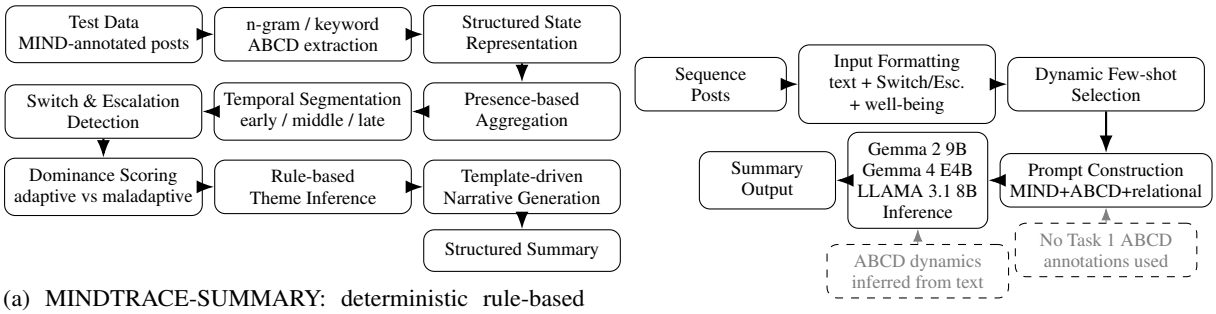
As a non-LLM comparison, we build a **classifier model** using the XGBoost library. In the preprocessing of this model, we use TF-IDF and sentence transformer embeddings (all-mpnet-base-v2) to represent the posts. To get the temporal difference features, we use 1) the embedding difference between the current and previous posts (i.e., current post embedding - previous post embedding $h_t - h_{t-1}$), 2) their element-wise product, and 3) timeline position. We also used additional feature embeddings, including 14 linguistic features related to sentiment, punctuation, and length (Table 1). Two binary classifiers then predict Switch and Escalation, with the rare positive examples given more weight during training so the model does not simply default to predicting no change everywhere.[2]

Task 3 on Change Summarization and Pattern Mining. For **Task 3.1** "summarizing sequences surrounding change events", we explore two distinct summary generation strategies (Figure 5).

MINDTRACE-SUMMARY is a fully deterministic, multi-stage framework for generating sequence-level psychological signatures from MIND-annotated posts. The model generates summaries by following a fixed structure and converting ABCD annotations into natural language. A full sample template is provided in Appendix D.3. It first determines whether adaptive or maladaptive states dominate, then builds a narrative with five parts: central theme, initial state, interaction dynamics, transition (switch or escalation), and outcome. ABCD labels are rewritten into fluent psychological descriptions, and the summary emphasises how states reinforce or shift over time. The method prioritizes consistency and structure over free-form generation, ensuring summaries clearly reflect change dynamics across the sequence.

For **Few-Shot LLM Prompting**, the prompt specifies the MIND framework, ABCD abbreviation

[2]Positive examples are up-weighted via XGBoost's `scale_pos_weight` parameter, set to $\min\left(\frac{n_{\text{neg}}}{n_{\text{pos}}}, 20\right)$ independently for each classifier.



(a) MINDTRACE-SUMMARY: deterministic rule-based pipeline.

(b) Few-shot LLM prompting pipeline using OS-LLMs.

Figure 5: Task 3.1 summary generation methods from DreamerNLplus – rule-based (left) vs OS-LLMs (right).

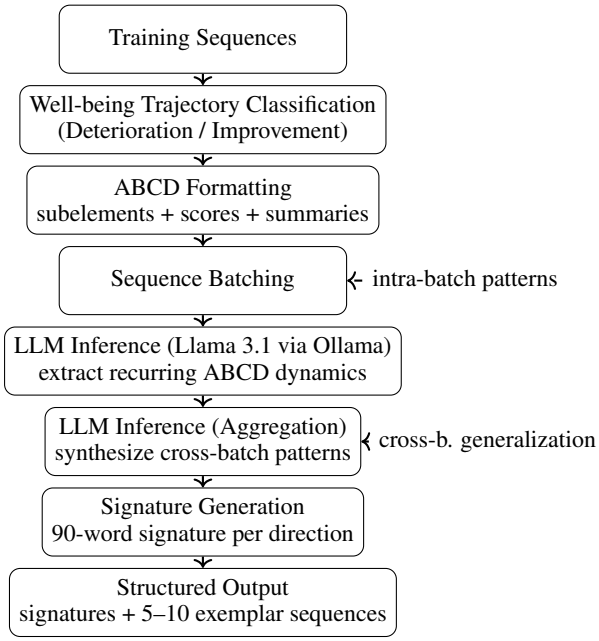


Figure 6: Overview of the RAG-LLM Signature Mining framework for Task 3.2.

conventions, relational dynamics vocabulary, and required summary structure covering pre-change phase, within/between-state dynamics, and explicit change event identification. Task 1 ABCD annotations are not used; the model infers ABCD dynamics from post text guided by the prompt.

For **Task 3.2** "identifying recurrent dynamic signatures of change across timelines", we propose **RAG-LLM Signature Mining**, a two-stage LLM-based framework for identifying recurrent dynamic signatures of psychological change. As illustrated in Figure 6, our framework separates intra-batch pattern extraction from cross-batch signature synthesis, enabling the identification of recurrent psychological dynamics across timelines. Sequences are batched and fed to an open-source LLM (Llama 3.1via Ollama) with per-post ABCD subelements, well-being scores, and gold summaries. Stage 1 extracts recurring ABCD dynamics per batch; Stage 2 synthesizes these into one 90-word signature per direction (deterioration/improvement), with 5–10 exemplar sequences as evidence.

3 Results and Analysis

Task 1: State modeling For Task 1, our system ranks 22nd in Task 1.1 (subelement classification) and 20th in Task 1.2 (presence estimation), as shown in Table 2. To further understand this discrepancy, we analyze the relationship between classification performance and regression performance in Figure 7. The results show a moderate negative correlation ($r = -0.486$, $p = 0.00354$), indicating that stronger subelement classification performance does not necessarily translate into better presence estimation. This observation reflects the fundamental difference between the two subtasks: Task 1.1 requires fine-grained categorical prediction of psychological subelements, while Task 1.2 evaluates continuous intensity estimation. Our system, which combines DeBERTa-based classification with Random Forest regression, appears more robust in the regression setting, suggesting that downstream aggregation mitigates upstream classification errors.

Task 2: Change Detection Our submitted LLM system reaches a combined F1 of 0.442, ranking 11th overall (Figure 10), while the XGBoost variant scores 0.327. The two approaches show opposite error patterns: the LLM has high recall but low precision (Switch 0.762/0.302, Escalation 0.917/0.393) meaning it picks up most change cues but tends to over-predict them, while XGBoost has high precision but low recall on rare positive classes(Switch precision 0.455, recall 0.238), reflecting how hard it is to learn these labels from only 30 gold timelines. This shows that accurately capturing subtle transitions remains challenging across paradigms,

and that errors on weak or ambiguous change signals can propagate across the sequence. Despite not relying on task-specific fine-tuning, the LLM method stays competitive performance while providing interpretable predictions with textual justifications. This highlights the effectiveness of few-shot prompting for modeling temporal dynamics in low-resource settings.

Task 3: Summarization and Pattern Mining For Task 3.1, in our processing, we kept other teams best average-ranked submission and our best-performing submission (ID 693964 prompting-w-ablation) ranks 4th overall based on the average ranking across multiple metrics (Figure 11). However, a notable observation is the strong disagreement between evaluation metrics.

Interestingly, shown in Figure 11, the rule-based submission (694142) achieves the best CT and strong CS performance, yet ranks worst overall due to poor ROUGE and BERT scores (a detailed discussion of the evaluation is provided in Appendix E.1.), highlighting a clear disagreement between semantic coherence metrics and similarity-based metrics: CS and CT emphasize semantic coherence and psychological consistency, whereas ROUGE and BERTScore prioritize surface-level similarity to reference summaries. As a result, optimizing for one set of metrics may degrade performance on the other. However, in the official Task 3.1 ranking, DreamerNLplus ranks **2nd** overall with submission 693964, based on the average of metric-specific ranks across CS, CT, ROUGE-L Recall, and BERTScore Recall.

For Task 3.2, as shown in Table 5, our RAG-based approach ranks **1st** on Improvement and **3rd** on Deterioration. In addition, our system achieves the highest score on Specificity and the second-highest on Recurrence for Improvement, and the second-highest Specificity for Deterioration. These results suggest that the proposed RAG-based framework is particularly effective at identifying precise and recurring dynamic patterns across timelines. By combining batch-level pattern extraction with cross-batch synthesis, the approach is able to capture higher-level psychological change signatures that generalize across individuals. However, the performance gap between Improvement and Deterioration also indicates that modeling deterioration patterns remains more challenging, potentially due to greater variability or ambiguity in negative self-state dynamics. This highlights the importance of modeling both intra-sequence consistency and inter-sequence generalization when identifying recurrent psychological dynamics.

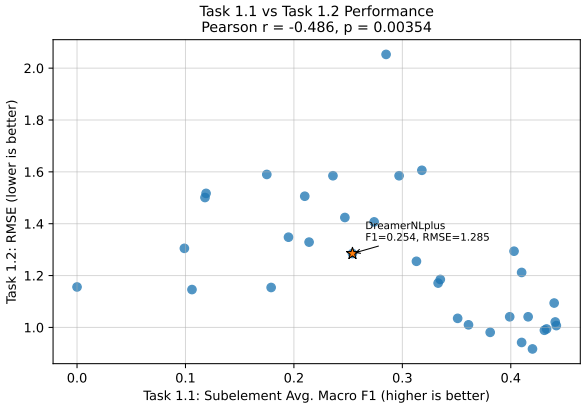


Figure 7: Task 1.1 vs 1.2 Relation. Each point = one team. X-axis = Task 1.1 (F1) → higher is better. Y-axis = Task 1.2 (RMSE) → lower is better.

Cross-Task Analysis Across all tasks, a consistent pattern emerges: different evaluation metrics capture distinct and sometimes conflicting aspects of performance. Task 1 reveals a trade-off between classification accuracy and regression stability, Task 2 highlights the difficulty of balancing local and temporal consistency, and Task 3 exposes a mismatch between semantic coherence and lexical similarity metrics. These findings suggest that modeling mental health dynamics requires balancing multiple objectives, including structured prediction, temporal reasoning, and semantic generation. Our hybrid approach demonstrates that combining interpretable representations with flexible LLM-based methods can provide robust performance across tasks, while also revealing important limitations in current evaluation frameworks.

4 Conclusions and Future Work

We presented **DreamerNLplus**, a hybrid framework for modeling mental health dynamics from social media timelines across three tasks: psychological state modeling, temporal change detection, and sequence-level summarization. By combining structured representations, few-shot prompting, and rule-based generation, our approach provides both interpretability and flexibility across different modeling paradigms. Future work should aim to design unified evaluation frameworks that reconcile semantic fidelity, temporal consistency, and textual similarity, enabling clinically meaningful assessment of mental health modeling systems.

Limitations

For data augmentation in Task 1, in this work, we extracted the evidence from original training data, and only asked LLMs to generate more evidence, subsequently, DeBERTa model is trained on the augmented evidence data to predict sub-element classes. This is similar to the dense prescription generation work in (Belkadi et al., 2025) for NER and engineering purposes only, without generating full clinical letters. In an ideal situation, we will further explore the generation of similar post-level, not only evidence.

Task 3 template-based summarization achieved high CS and low CT scores through predefined linguistic rules and structured feature-to-text mappings, enabling stable and interpretable summaries. However, the approach is task-specific and less flexible than human summarization, limiting its ability to capture nuanced contextual and emotional variations in narratives. Despite this, it is well-suited for formal or high-stakes settings where standardized and reproducible documentation is prioritized over linguistic diversity. Additionally, the framework relies on intermediate computations such as switch and escalation scores derived from upstream predictions, meaning errors in earlier stages may propagate into the final summaries.

Ethics

The shared task data we used in this paper is anonymized and annotated by CLPych2026 organizers. We only used secure methods and models to process the data, such as rule-based, locally hosted open-source LLMs, locally trained encoders, without releasing the data to any third parties with our best practice for privacy protection.

Acknowledgments

We thank David Lindevelt for the help on this project, including the automated prompt pipeline and framework adapted from another project (Han et al., 2026). We thank Prof. Suzan Verberne for editing the camera-ready version of this paper. We thank the reviewers for their valuable comments on our work. We thank the organizers preparing this shared task, and we are grateful for their communication during our registration and submissions, especially Talia Tseriotou and Iqra Ali from Queen Mary University of London. The 4D PICTURE consortium is funded by the European Union under Horizon Europe Work Programme 101057332. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. The UK team are funded under the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number: 10041120.

References

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

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

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

Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2025. Lt3: Generating medication prescriptions with conditional transformer. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 205–218.

Liliia Bogdanova, Shiran Sun, Lifeng Han, Natalia Amat Lefort, and Flor Miriam Plaza-del Arco. 2026. Flans at semeval-2026 task 7: Rag with open-sourced smaller llms for everyday knowledge across diverse languages and cultures. *arXiv preprint arXiv:2603.01910*.

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.

Lifeng Han, David Lindevelt, Sander Puts, Erik van Mulligen, and Suzan Verberne. 2026. Dutch

metaphor extraction from cancer patients' interviews and forum data using llms and human in the loop. *CL4Health WS at LREC2026, Palma, Spain*.

Arshia Kermani, Veronica Perez-Rosas, and Vangelis Metsis. 2025. A systematic evaluation of LLM strategies for mental health text analysis: Fine-tuning vs. prompt engineering vs. RAG. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 172–180, Albuquerque, New Mexico. Association for Computational Linguistics.

Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, and 1 others. 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, 23(5):e15708.

Matteo Malgaroli, Thomas D Hull, James M Zech, and Tim Althoff. 2023. Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.

Libo Ren, Yee Man Ng, and Lifeng Han. 2025. Malei at multiclinsum: Summarisation of clinical documents using perspective-aware iterative self-prompting with llms.

Pablo Romero, Libo Ren, Lifeng Han, and Goran Nenadic. 2025. The manchester bees at peranssumm 2025: Iterative self-prompting with claude and o1 for perspective-aware healthcare answer summarisation. In *Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health)*, pages 340–348.

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

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

A CLPsych 2026 Shared Tasks

CLPsych2026 is affiliated with The Workshop on Computational Linguistics and Clinical Psychology, a workshop series founded in 2014. CLPsych 2026 will be held at ACL in San Diego, July 4th, 2026.

A.1 Task Evaluations

Unified Evaluation Framework The CLPsych 2026 shared tasks evaluate complementary aspects of modeling mental health dynamics from social media timelines, spanning classification, regression, temporal change detection, and summarization. Across Tasks 1–3, the evaluation framework progressively moves from local, structured predictions to global, sequence-level reasoning and natural language generation.

Task 1 (State modeling). Task 1 evaluates the ability to model psychological self-states at the post level. Task 1.1 focuses on discrete classification of ABCD elements and subelements, using macro-averaged F1 scores across adaptive and maladaptive categories. Task 1.2 evaluates continuous presence estimation on a 1–5 scale, using regression metrics such as RMSE, with ranking based on the mean RMSE across valences. Together, these subtasks capture the challenge of jointly modeling structured categorical representations and continuous mental state intensity.

Task 2 (Change Detection). Task 2 evaluates the detection of temporal change signals, specifically Switch (sudden change) and Escalation (gradual change). Performance is measured using F1 scores at both post-level and timeline-level, with final ranking based on the average of these two perspectives. This design rewards systems that can detect both local transitions and maintain consistency across sequences.

Task 3 (Change Summarization and Pattern Mining). Task 3 evaluates sequence-level understanding and generation. Task 3.1 assesses the quality of generated summaries using a multi-metric framework including semantic consistency (CS), contradiction (CT), and similarity-based metrics (ROUGE-L and BERTScore), with final ranking based on averaged metric ranks. Task 3.2 focuses on identifying recurrent dynamic patterns across timelines, emphasizing generalization and abstraction of psychological change signatures.

Cross-Task Perspective. Taken together, the evaluation framework reflects a hierarchy of modeling challenges: Task 1 focuses on *what* psychological states are present, Task 2 focuses on *when* meaningful changes occur, and Task 3 focuses on *how* these changes can be summarized and generalized. Notably, the metrics across tasks capture partially complementary qualities, including classification accuracy, regression stability, temporal consistency, and semantic coherence, highlighting the inherent trade-offs in modeling mental health dynamics.

This unified evaluation highlights the difficulty of aligning discrete classification, continuous estimation, temporal reasoning, and semantic generation within a single modeling framework.

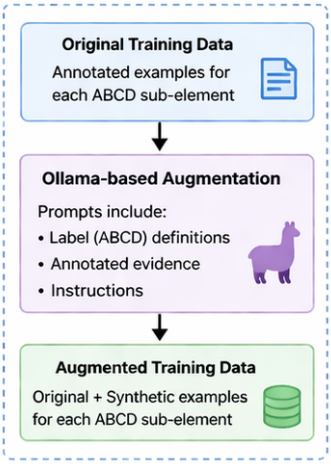


Figure 8: Targeted data augmentation strategy using Ollama.

B Methods and Experimental Development - Task1 (*More Details*)

We adopt a multi-paradigm modeling strategy combining rule-based, transformer-based, and LLM-assisted components for the overall shared task, to establish the relevance of the adopted approach for a given task.

B.1 Task 1.1: ABCD Element & Subelement Classification

Task 1.1 focuses on identifying fine-grained ABCD self-state elements and their corresponding subelements within each post. However, the task presents two key inherent challenges due to the fine-grained structure of the ABCD schema. First, the large number of subelements per category leads to a highly imbalanced and sparse label space, where certain subelements are significantly underrepresented. Second, there exists substantial seman-

tic overlap between closely related subelements, making boundary definition non-trivial even for pretrained language models.

Data Augmentation: To address data sparsity, we employ targeted data augmentation using Ollamas (see Fig. 8). Augmentation is guided by structured ABCD element definitions and subelement descriptions, enabling controlled expansion of training data while preserving label semantics (see Fig. 9). However, augmentation introduces additional challenges. While LLM-based generation improves data volume and diversity, it may also introduce distributional shifts, as synthetic samples often lack the contextual richness and temporal grounding of real social media posts. This is particularly critical in longitudinal mental health modeling, where self-state interpretation depends on subtle linguistic, emotional, and contextual cues.

Transformer-based Approach (DeBERTa): Initially, we considered a transformer-based sequence classifier, DeBERTa, for direct multi-label prediction of subelements. The model is fine-tuned to jointly learn element presence and subelement classification in a supervised setting, leveraging contextual embeddings to capture nuanced linguistic signals. However, due to extreme label granularity, semantic overlap between subelements (e.g., self-care vs. self-improvement, anxiety vs. despair), and limited training examples for several classes, purely supervised learning was found to be insufficient for robust generalisation.

Rule-based Approach (Augmented n-gram Modeling): Consequently, we introduce a rule-based pipeline that leverages label-conditioned augmented data for n-gram extraction and pattern-based classification (see Fig. 3). Augmented samples generated via Ollama are used to construct label-specific lexical signatures in the form of n-grams, which serve as interpretable indicators of subelement presence. Specifically, we compute top-k bigrams and trigrams from cleaned, stopword-removed text and rank them using likelihood ratio scores to identify statistically salient phrases. This approach is particularly effective in low-resource settings, where explicit lexical cues correlate strongly with specific psychological states.

B.2 Task 1.2: Presence Rating:

Task 1.2 focuses on estimating the overall presence rating (1–5) of adaptive and maladaptive states for each post, reflecting the psychological centrality of each state within the narrative. To address this task, we explore both regression-based and LLM-based approaches. We observed a clear relationship between the presence of specific subelements and the final presence ratings, but also noted that simple frequency-based counting of elements did not reliably correspond to the assigned scores. In particular, ratings were influenced not just by the number of subelements but by their specific types within the same ABCD category.

Regression-based Approach: Given the relationship observed between ABCD subelements and presence ratings, we adopt a regression framework using RandomForestRegressor. We construct structured binary feature vectors that serve as inputs to two separate regression models, with adaptive and maladaptive states modelled independently. This approach directly depends on the outputs of Task 1.1 during inference, where detected subelements are aggregated to form the input feature representation. Finally, the continuous outputs are rounded to discrete levels (1–5), aligning with the original annotation scheme.

B.3 Task 2 Features

C Cross-task Analysis

From our methods: Unlike Task 1, which learns explicit ABCD representations, Task 2 relies on contextual prompting to identify dynamic changes directly from timelines. Together, these approaches reflect a common emphasis on interpretability, temporal sensitivity, and privacy-preserving deployment. Task 1 models what psychological states are present, while Task 2 models when meaningful changes occur.

C.1 Stratified Sampling and K-fold

We have tried both Stratified Sampling and K-fold training data split for model development purposes. In the end, we adopted the K-fold approach for our data processing.

D Prompts and Rule-based Templates

We share our codes and full prompts used for the shared tasks at our Github page https://github.com/4dpicture/CLPsych2026.

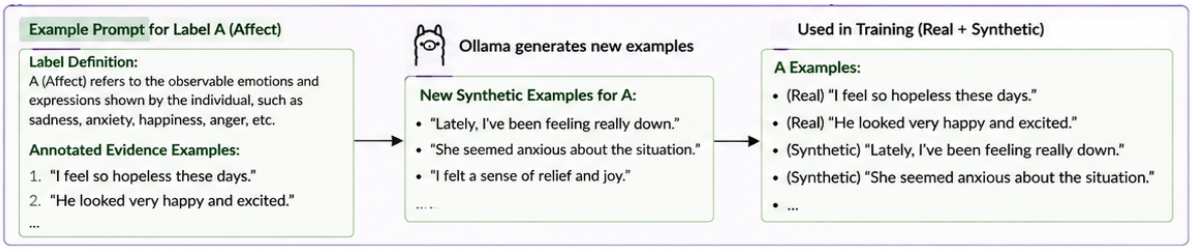


Figure 9: Data Augmentation Examples (paraphrased to preserve privacy in accordance with shared task guidelines).

Table 1: Hand-crafted linguistic features used in Task 2 (Switch/Escalation detection). Features are extracted per post and concatenated with the temporal embedding differences.

#	Feature	Description
Lexical		
1	`log_len`	$\log(1 + \text{word count})$; compresses post length onto a continuous scale
2	`n_sentences`	Number of sentences, split on `[.!?]+`
3	`avg_word_len`	Mean character length across words
4	`frac_upper`	Fraction of characters that are uppercase
5	`frac_punct`	Fraction of characters that are punctuation (`!?.,;:`)
Punctuation / prosodic signals		
6	`n_exclaim`	Raw count of exclamation marks (`!`)
7	`n_question`	Raw count of question marks (`?`)
8	`n_ellipsis`	Raw count of ellipses (`...`)
9	`emo_punct`	$\min(\texttt{n_exclaim} + \texttt{n_question},\ 10)$; capped emotional punctuation
Sentiment lexicon		
10	`frac_neg`	Fraction of words matching a negative lexicon (*hate, depressed, pain, alone*, ...)
11	`frac_pos`	Fraction of words matching a positive lexicon (*happy, hope, proud, grateful*, ...)
12	`sentiment_balance`	`frac_neg` − `frac_pos`; net sentiment polarity
Structural		
13	`words_per_sent`	word count/sentence count; average sentence length
14	`has_removed`	Binary flag for `[removed]` or `[deleted]` posts

D.1 Task2 Prompts

The full Task 2 prompt (Switch and Escalation detection) is provided in Appendix G.

D.2 Task3.1 Prompts

The full Task 3.1 summary-generation prompt is provided in Appendix G.

D.3 Task3.1 Summary Generation: Rule-based Template

We generate the final narrative using a rule-based template conditioned on extracted discourse features. Let M denote the total maladaptive score and A the total adaptive score. Let Δ denote structural change type and D denote trajectory direction.

Feature Sets Adaptive ($\mathcal{F}_a$) and Maladaptive ($\mathcal{F}_m$) features labels.

Initial Phase If $M \geq A$, maladaptive processes are dominant:

> Initially, maladaptive self-state processes are more dominant, characterized by elements such as $\mathcal{F}_m$, while adaptive processes remain less prominent.

Otherwise, adaptive processes are dominant:

> Initially, adaptive self-state processes are more dominant, characterized by elements such as $\mathcal{F}_a$, buffering against maladaptive tendencies.

Temporal Dynamics If $M > A$, maladaptive dynamics intensify over time:

> Maladaptive dynamics intensify over time through reinforcing cycles of negative affect, self-critical cognition, and behavioral withdrawal, suppressing adaptive functioning.

Otherwise, adaptive processes strengthen over time:

> Adaptive processes strengthen over time through increasing self-compassion, relational engagement, and constructive coping that counter maladaptive tendencies.

Structural Transition If $\Delta =$ switch:

> A transition point emerges within the sequence, reflecting a shift in the balance between adaptive and maladaptive self-states.

If $\Delta =$ escalation:

> An escalation unfolds across the sequence, reflecting progressive intensification of emotional, cognitive, and behavioural processes over time.

Trajectory Direction If $D =$ deterioration:

> In the later phase, maladaptive self-state dynamics dominate, reinforcing sustained distress and hopelessness.

If $D =$ improvement:

> In the later phase, adaptive self-state dynamics become dominant, supporting resilience and psychological recovery.

If $D =$ fluctuation:

> In the later phase, adaptive and maladaptive self-states remain in tension, reflecting ongoing fluctuation between distress and coping.

Global Template (Always Included) Across all sequences, the following integrative statements are included:

> The central psychological theme across the sequence reflects an evolving interaction between maladaptive distress and adaptive coping processes expressed through affect, cognition, behavior, and desire.
>
> As the sequence progresses, adaptive and maladaptive self-states increasingly interact, creating periods of internal conflict, reflection, and shifting psychological balance.
>
> Across the sequence, adaptive and maladaptive self-states alternate in dominance and suppression, shaping the overall trajectory of psychological change.

D.4 Task3.2 Prompts

The full Task 3.2 prompts (Stage 1 and Stage 2) are provided in Appendix G.

E Discussion

In this section, we discuss the performance characteristics, advantages, and limitations of the proposed methods.

E.1 Consistency vs. Lexical Similarity in Template-Based Summarization

The template-based summarization method achieved the highest CS and lowest CT scores, highlighting its strength in producing stable and logically coherent outputs. This behavior is expected, as the generation process is strictly governed by predefined linguistic rules and structured feature-to-text mappings, which reduce variability and minimize the risk of semantic drift or internally inconsistent statements. Such control is particularly important in sensitive domains such as palliative care narrative analysis, where reliability and interpretability of generated summaries are critical.

However, this same constrained generation process leads to lower ROUGE and BERTScore performance compared to more flexible neural baselines. Both metrics reward lexical overlap and semantic similarity to reference summaries, which are typically written in a more natural and varied style. Template-based outputs, while semantically faithful and structurally consistent, do not exhibit the paraphrastic richness or lexical alignment required to maximize these scores. As a result, there is an inherent trade-off between consistency-oriented generation and similarity-based evaluation metrics.

This trade-off suggests that template-based summarization is particularly well-suited for applications where factual stability, interpretability, and controllability are prioritized over surface-level

Task2	Team Name	Avg Macro F1
1	USAI	0.6
2	CtbuY	0.588
3	JNLP	0.58
4	CLPsych2026 - Baseline 3 (Tempoformer)	0.572
5	CUNY	0.572
6	MKC	0.554
7	Codezone Research Group	0.553
8	Aurevia	0.484
9	Meronym Labs	0.466
10	debju	0.447
11	**DreamerNLplus**	**0.442**
12	McMasterNLP	0.412
13	BLUE	0.403
14	NoviceTrio	0.385
15	DrosophilAI	0.383
16	psytechlab	0.372
17	CLPsych2026 - Baseline 2	0.365
18	CSE_IIT_Ropar	0.357
19	CLPsych2026 - Baseline 1	0.272
20	Afrilan	0.268
21	Lin Tan	0.26

Figure 10: Task 2 Ranking

similarity to reference texts. In the context of psychological trajectory modeling from patient and caregiver narratives, such methods are advantageous for producing reproducible summaries of adaptive and maladaptive self-state dynamics. Conversely, neural summarization models may be preferred in settings where linguistic expressiveness and alignment with human-written references are more important than strict structural consistency.

F Details on rankings and evaluation scores

We list our official ranking scores with other teams in this section.

Table 2: Task 1.1 and Task 1.2 rankings. Task 1.1 is ranked by Subelement Average Macro F1, while Task 1.2 is ranked by RMSE, where lower is better.

Team	Task 1.1 Rank	Task 1.2 Rank	Task 1.1 Macro F1	Task 1.2 RMSE
CUNY	1	6	0.442	1.007
StateOfMIND	2	8	0.441	1.021
StateOfMIND	3	12	0.440	1.094
StateOfMIND	4	5	0.433	0.994
CUNY	5	4	0.431	0.989
Meronym Labs	6	1	0.420	0.917
CUNY	7	10	0.416	1.041
Meronym Labs	8	18	0.410	1.212
USAI	9	2	0.410	0.942
USAI	10	21	0.403	1.294
Meronym Labs	11	11	0.399	1.041
Aurevia	12	3	0.381	0.981
MKC	13	7	0.361	1.010
McMasterNLP	14	9	0.351	1.035
ull	15	17	0.335	1.185
ull	16	16	0.333	1.171
BLUE	17	33	0.318	1.606
NoviceTrio	18	19	0.313	1.255
Afrilan	19	31	0.297	1.585
Afrilan	20	34	0.285	2.053
psytechlab	21	25	0.274	1.407
DreamerNLplus	**22**	**20**	**0.254**	**1.285**
CLPsych2026 - Baseline	23	26	0.247	1.424
CtbuY	24	30	0.236	1.585
Afrilan	25	23	0.214	1.329
CtbuY	26	28	0.210	1.506
debju	27	24	0.195	1.348
DrosophilAI	28	14	0.179	1.154
CSE_IIT_Ropar	29	32	0.175	1.590
ull	30	29	0.119	1.517
CtbuY	31	27	0.118	1.501
CSE_IIT_Ropar	32	13	0.106	1.146
DrosophilAI	33	22	0.099	1.305
debjy	34	15	0.000	1.156

A	B	C	D	E	F	G	H	I	J	K
Team	Consistency (CS)	CS_rank	Contradiction (CT)	CT_rank	ROUGE-L Recall	R-L_rank	BERTscore Recall	BERTscore_rank	Score Average	core Rank Average
MERONYM_LABS	0.80073322	4	0.659288617	4	0.265764584	12	0.34491533	7	0.5176754378	6.75
CUNY	0.789027221	7	0.714318694	8	0.29207881	5	0.294779657	18	0.5225510955	9.5
NoviceTrio	0.704759255	16	0.770804092	13	0.317894427	2	0.340504646	8	0.533490605	9.75
DreamerNLplus	*0.735013197*	*14*	*0.76715816*	*12*	*0.28453229*	*8*	*0.345456439*	*6*	*0.5330400215*	*10*
USAI	0.680537067	20	0.849299731	25	0.332895421	**1**	0.365483143	**2**	0.5570538405	12
DreamerNLplus	*0.739321004*	*13*	*0.739891788*	*10*	*0.225534182*	*21*	*0.328495991*	*10*	*0.5083107413*	*13.5*
Aurevia	0.865719654	1	0.624931034	3	0.184924076	26	0.226479176	25	0.475513485	13.75
MKC	0.668688559	21	0.856618807	27	0.29046832	6	0.362011761	3	0.5444468618	14.25
DreamerNLplus	*0.845004471*	***3***	*0.439257234*	***1***	*0.189083659*	*25*	*0.14373662*	*29*	*0.404270496*	*14.5*
Baseline	0.767319058	11	0.744681011	11	0.26853177	12	0.235086789	25	0.503904657	14.75
psytechlab	0.856826357	2	0.570653373	2	0.077716539	29	0.146949642	28	0.4130364778	15.25
JNLP	0.790974544	6	0.665859524	5	0.116772542	28	0.163692788	27	0.4343248495	16.5
McMaster NLP	0.770177034	10	0.761182828	11	0.208128021	23	0.255358784	23	0.4987116668	16.75
CSE_IIT_Ropar	0.688347356	18	0.812002344	18	0.24240965	17	0.306391245	16	0.5122876488	17.25
ULL	0.585453856	27	0.84554644	23	0.261929928	13	0.319793656	14	0.50318097	19.25
CtbuY	0.615451482	26	0.848440523	24	0.232144785	19	0.317244728	15	0.5033203795	21

Figure 11: Task 3.1 eval on test set - all teams (our filtering: by using the best averaged rank submission from other teams).

Table 3: **The Official rank** of Task 3.1 based on selected submissions of all teams. CS, ROUGE-L Recall, and BERTScore Recall are higher-is-better; CT is lower-is-better. Final rank is based on the average of metric-specific ranks. We rank the **2nd best** overall.

Rank	Team	Sub. ID	CS	CS Rank	CT	CT Rank	ROUGE-L	R-L Rank	BERT	BERT Rank	Avg. Rank
1	MERONYM_LABS	694229	0.801	3	0.659	3	0.266	6	0.345	4	4.00
2	**DreamerNLplus**	**693964**	**0.735**	**7**	**0.767**	**7**	**0.285**	**4**	**0.345**	**3**	**5.25**
3	CUNY	694216	0.789	5	0.714	5	0.292	3	0.295	9	5.50
4	NoviceTrio	693913	0.705	8	0.771	8	0.318	2	0.341	5	5.75
5	USAI	693912	0.681	10	0.849	13	0.333	1	0.365	1	6.25
5	Aurevia	693454	0.866	1	0.625	2	0.185	11	0.226	11	6.25
7	MKC	687777	0.654	11	0.834	10	0.284	5	0.359	2	7.00
8	psytechlab	689973	0.857	2	0.571	1	0.078	13	0.147	13	7.25
9	JNLP	691315	0.791	4	0.666	4	0.117	12	0.164	12	8.00
9	McMaster NLP	694189	0.770	6	0.761	6	0.208	10	0.255	10	8.00
11	CSE_IIT_Ropar	694311	0.688	9	0.812	9	0.242	8	0.306	8	8.50
12	ULL	694669	0.585	13	0.846	11	0.262	7	0.320	6	9.25
13	CtbuY	691446	0.615	12	0.848	12	0.232	9	0.317	7	10.00

Table 4: Summary of DreamerNLplus Submissions on Task 3.1 Across Evaluation Metrics (↑ higher is better, ↓ lower is better). Detail score refers to Fig. 11

Submission	CS ↑	CT ↓	ROUGE	BERT	Score Avg
693964	medium-high	high	decent	decent	**0.533 (best)**
694011	medium-high	medium	lower	medium	0.508
694142	**very high/3rd**	**very low/best**	low	low	0.404 (worst)

Table 5: Task 3.2 official rankings for recurrent dynamic signatures of Improvement and Deterioration. Best scores in each metric column are shown in bold, and second-best scores are underlined. We **won Improvement** category, and *2nd best* on Specificity for Deterioration category.

Direction	Team	Rank	Fit	Recurrence	Specificity	Overall
Improvement	*DreamerNLplus*	1	0.6250	<u>0.8125</u>	**1.0000**	**0.7608**
	CSE_IIT_Ropar	2	**1.0000**	0.6875	0.3750	<u>0.7426</u>
	MKC	3	<u>0.7500</u>	0.5625	<u>0.9375</u>	0.7266
	McMasterNLP	4	0.6875	0.3750	0.7500	0.5938
	psytechlab	5	0.6875	**1.0000**	0.2500	0.5437
	Aurevia	6	0.3750	0.6250	0.5000	0.4653
	MeronymLabs	7	0.2500	0.2500	0.5625	0.2981
	CtbuY	8	0.2500	0.2500	0.0000	0.1250
	CUNY	9	0.0000	0.0000	0.2500	0.0000
Deterioration	CSE_IIT_Ropar	1	<u>0.8750</u>	0.5625	**0.9375**	**0.7891**
	Aurevia	2	0.6875	<u>0.8125</u>	0.5625	<u>0.6761</u>
	DreamerNLplus	3	0.4375	0.6875	<u>0.8750</u>	0.6038
	MeronymLabs	4	0.4375	0.5625	0.8125	0.5511
	CtbuY	5	**1.0000**	**1.0000**	0.0000	0.5000
	psytechlab	6	0.6250	0.6250	0.2500	0.4911
	MKC	7	0.1875	0.1875	0.5000	0.2301
	McMasterNLP	8	0.1875	0.1875	0.3125	0.2109
	CUNY	9	0.0000	0.0000	0.3125	0.0000

G Full Prompts

This appendix collects the complete prompts used across all tasks.

G.1 Task 2: Switch and Escalation Detection (Llama 3.1 8B, Ollama, 5-shot)

```
You are an expert psychologist trained in the MIND framework for detecting moments of change in mental health trajectories from
      social media timelines.

## Your Task

Determine whether the target post represents a **switch** or an **escalation** (or neither) compared to the preceding posts.

## Definitions

**Switch**: A sudden, distinct shift in well-being between the target post and the preceding post.
- Moving from adaptive to maladaptive state or vice versa
- A qualitative change in dominant affect, behavior, or cognition
- The change is abrupt, not gradual

**Escalation**: A gradual intensification of the current mental health state across consecutive posts.
- The same trajectory continues but deepens in severity
- Progressive worsening or improvement over multiple posts
- The change unfolds over time, not suddenly

## Key Distinctions

- A **switch** is sudden and qualitative (shift in kind)
- An **escalation** is gradual and quantitative (shift in degree)
- A post can be BOTH or NEITHER
- Default to NEITHER unless there is clear evidence of change
- The FIRST post in a timeline cannot be a switch or escalation

## Examples

### Example 1: Switch only

Context posts:
- Post 1: "I'm really struggling. I wish that everyone wasn't so cruel. I wish that I didn't wish I could disappear." (strongly
      maladaptive: despair, loneliness, suicidal ideation)

Target post:
- "Calling All Garden Volunteers! I have recently secured 2 community plots. I want them to grow beautiful vegetables. I'm a
      broke high school student but I want to try my best to give this neighborhood a better space. Would anyone join me?"

Assessment: Switch = YES, Escalation = NO.
Justification: The preceding post was dominated by hopelessness and suicidal ideation (strongly maladaptive). The target post
      shifts abruptly to proactive helping behavior and hopefulness  a qualitative shift toward an adaptive state. This is sudden
      , not gradual, so it is a switch but not an escalation.

### Example 2: Both switch and escalation

Context posts:
- Post 1: "I'm not your personal career coach. I am definitely not going to give you job advice if YOU NEVER EVEN LISTEN TO IT."
       (anger, frustration directed outward)

Target post:
- "When the therapist said that facing my trauma would make my anxiety spike, they weren't joking. I feel like everyone around
      me is full of nothing but toxicity and fake smiles. I have zero future ahead of me. What is even the point. Maybe I'm
      better off isolating myself."

Assessment: Switch = YES, Escalation = YES.
Justification: The context showed anger directed outward. The target shifts to deep hopelessness and withdrawal - a
      qualitatively different maladaptive state (switch). The overall severity has also intensified significantly compared to
      prior posts (escalation).

### Example 3: Escalation only

Context posts:
- Post 1: "Im logging off now. I cant see myself being here next month. Should I just quit my job because I know I wont make it
      anyways." (depressed, hopeless, avoidant)

Target post:
- "Every single task is a massive mountain right now. I'm paralyzed by terror and the pressure just makes me want to end it. So
      please stop nagging about my career just for a minute. I'm sorry for being an utter failure."

Assessment: Switch = NO, Escalation = YES.
Justification: The trajectory remains consistently maladaptive  depression and hopelessness continue. However, severity has
      deepened: self-criticism ("an utter failure"), explicit desire to die, perception of others as pressuring. The same
      direction intensifies, but there is no qualitative shift in kind.

### Example 4: Neither

Context posts:
- Post 1: "I went to a meetup and really enjoyed it. I felt totally at ease." (mixed: mostly adaptive with some anxiety)

Target post:
- "Are there any Cook County dog rescue places willing to help find these pups a shelter? Please?"

Assessment: Switch = NO, Escalation = NO.
```

```
Justification: The post shows continued mixed adaptive/maladaptive elements consistent with the prior trajectory. There is no
      sudden shift and no progressive intensification. Normal variation.

### Example 5: First post  neither

Context posts: (none  this is the first post)

Target post:
- "How can anyone experience genuine joy? I literally don't get it."

Assessment: Switch = NO, Escalation = NO.
Justification: This is the first post in the timeline. Without prior context, no change can be detected.

## Instructions

1. Read the preceding context posts to understand the trajectory.
2. Read the target post carefully.
3. Compare the target post to the preceding posts.
4. Default to NO switch and NO escalation unless there is clear evidence.
5. Provide a brief justification.

## Post to Assess

{post_text}

{format_instructions}
```

G.2 Task 3.1: Summary Generation (Gemma 2 9B, Ollama)

```
STAGE1_SYSTEM = (
    "You are an expert clinical psychologist analyzing patterns in mental health "
    "change sequences using the MIND framework (ABCD elements)."
)

STAGE1_PROMPT = """Below are {n} sequences of social media posts that all show {direction}.
Each sequence includes per-post ABCD annotations and a gold summary.

Your task: Identify the 2-3 most recurrent ABCD dynamic patterns across these sequences.

For each pattern, describe:
1. Which maladaptive and/or adaptive ABCD subelements are involved
2. The relational dynamics between them (mutual reinforcement, co-activation, suppression, etc.)
3. How the balance between adaptive and maladaptive states shifts
4. Which specific sequences exhibit this pattern (list their timeline_id/sequence_id)

Use ABCD abbreviations: (A) Affect, (B-S) Behavior-self, (B-O) Behavior-other, (C-S) Cognition-self, (C-O) Cognition-other, (D)
      Desire.

Be specific  name exact subelement categories (e.g., "(2) Self criticism" not just "C-S").

=== SEQUENCES ===

{sequences}

=== YOUR ANALYSIS ===
"""
```

G.3 Task 3.2 Stage 1: Batch Pattern Extraction (Llama 3.1 8B, Ollama, RAG)

```
STAGE2_PROMPT = """You previously analyzed multiple batches of {direction} sequences from social media timelines.
Below are your batch-level analyses identifying recurrent ABCD dynamic patterns.

Your task: Synthesize ONE concise signature of {direction} that captures the single most
recurrent pattern across ALL batches.

REQUIREMENTS:
- The signature MUST be 90 words or fewer
- Use ABCD abbreviations in parentheses: (A), (B-S), (B-O), (C-S), (C-O), (D)
- Name specific subelement categories (e.g., "(2) Self criticism", "(4) Depressed, despair, hopeless")
- Describe BOTH within-self-state dynamics (how ABCD elements interact within maladaptive or adaptive state) AND between-self-
      state dynamics (how one state suppresses/overshadows the other)
- Describe how the pattern progresses over time (what happens early vs late in the sequence)
- Start with "Many {direction} sequences show..."

Also select the 5-10 sequences that BEST exemplify this signature.

Format your response EXACTLY as:

SIGNATURE:
[your signature text, 90 words max]

EVIDENCE:
[timeline_id_1]/[sequence_id_1]
[timeline_id_2]/[sequence_id_2]
...

=== BATCH ANALYSES ===
```

```
{batch_analyses}

=== AVAILABLE SEQUENCES ===
{available_sequences}

=== YOUR FINAL SIGNATURE ===
"""
```