CUNY at CLPsych 2026: A Pipeline Approach to Classification and Summarization of Mental Health Changes

Amirmohammad Ziaei Bideh[†], **Shameed Charlomar Job**[*‡], **Ava Yahyapour**[*†], **Alla Rozovskaya**[†‡]
[†]Computer Science Department, CUNY Graduate Center
[‡]Linguistics Department, CUNY Graduate Center
amirmohammad.ziaeibideh69@cuny.edu

Abstract

We describe our submission to the CLPsych 2026 Shared Task on capturing and characterizing mental health changes through social media timeline dynamics. To infer the dominant self-states in posts (Tasks 1.1 and 1.2), we ensemble in-context learning of three open-weight large language models using majority voting. For predicting moments of change in a timeline (Task 2), we train supervised classifiers on features derived from Task 1.1 predictions. To summarize the patterns of mood dynamics and their progression over time within a timeline (Task 3.1), we augment in-context example labels predicted by upstream systems (Tasks 1.1, 1.2, and 2), yielding performance gains over zero-shot and unaugmented in-context learning baselines. Our submission ranked first on Task 1.1, fourth on Task 1.2, fourth on Task 2, and third on Task 3.1.[1]

1 Introduction

Mental health conditions affect a significant portion of the global population (World Health Organization, 2013; National Institute of Mental Health (NIMH), 2024), creating a need for scalable tools to monitor individuals' psychological states over time. Social media platforms offer longitudinal data for tracking how mental states evolve in response to life events and social interactions (Tsakalidis et al., 2022), and large language models (LLMs) have shown strong potential in supporting such analysis (Yang et al., 2024; Chan et al., 2025). Recent practice-oriented research further demonstrates how multimodal analysis and AI can uncover the intrapersonal and interpersonal dynamics underlying therapeutic change (Atzil-Slonim, 2026). The CLPsych 2026 shared task (Ali et al., 2026) addresses this need by asking participants to track and characterize how users' mental states evolve across longitudinal Reddit timelines.

This paper describes our submission to the CLPsych 2026 shared task (Ali et al., 2026). Our best submission for Task 1.1 uses an ensemble of seven LLM predictions across three model backbones with subelement-level In-Context Learning (ICL), where annotated training examples are included directly in the prompt to guide predictions. For Task 1.2, our best submission ensembles five predictions using ICL and Retrieval-Augmented Generation (RAG), which retrieves the most semantically similar training posts as in-context demonstrations. For Task 2, we train supervised classifiers on self-state features derived from upstream task predictions, using a Support Vector Machine (SVM) for Switch detection and a Random Forest (RF) for Escalation detection. For Task 3.1, our best submission uses label-augmented ICL, enriching prompts with predicted ABCD subelement and Moment of Change (MoC) labels from upstream tasks. For Task 3.2, we apply a batch-and-merge pipeline to identify recurrent dynamic signatures of improvement and deterioration across timelines. Our submission ranked **first** on Task 1.1, **fourth** on Task 1.2, **fourth** on Task 2, and **third** on Task 3.1.

The contributions of our work are as follows:

1. We show that LLM ensembling via majority voting substantially outperforms single-model baselines for subelement classification (Task 1.1), and that expanding the ensemble beyond five members further improves subelement classification but slightly hurts presence scoring (Task 1.2), suggesting diminishing returns for larger ensembles on ordinal prediction tasks.

2. We show that escalation is easier to predict than switch, and traditional supervised classifiers trained on LLM-derived presence scores are competitive for MoC detection (Task 2).

[*]Equal contribution.
[1]The source code for the experiments is available at https://github.com/amirzia/clpsych26-cuny.

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 441–457

3. We find that propagating predicted self-state labels from upstream tasks into downstream prompts with larger models yields consistent gains in summary quality over zero-shot prompting with smaller models and standard ICL baselines (Task 3.1).

2 Shared Task Description

The shared task is grounded in the Multimodal Intrapersonal and Interpersonal Dynamics (MIND) framework (Atzil-Slonim, 2025). The framework includes widely used psychotherapy constructs on the patient mental state referred to as self-state. A self-state is a dominant mode of experience and consists of a combination of Affect, Behavior, Cognition, and Desire (ABCD) components, their adaptivity level and their more fine-grained subcategorization into subelements (Atzil-Slonim, 2025).

Task 1.1 aims to identify which predefined ABCD subelements are expressed in a post and how they combine into *adaptive* and *maladaptive* self-states, referred to as the two *valences* of a self-state. Task 1.2 requires quantifying the degree to which each identified self-state is present in the post on a 1-5 scale (referred to as the presence score). Task 2 involves detecting clinically meaningful MoC within a user timeline,[2] identifying Switches (sudden change in well-being) and Escalations (gradual intensification of mood) (Tsakalidis et al., 2022). Task 3.1 involves generating a structured summary describing the progression of self-state dynamics within a sequence of posts surrounding an identified change event. Task 3.2 aims to identify recurrent dynamic signatures of improvement and deterioration that recur across multiple sequences and individuals.

Dataset. The training set contains Reddit timelines from 30 users, annotated with self-states according to the MIND framework. Figure 3 shows an anonymized excerpt of a sample timeline. The training set also provides gold summaries for sequences, describing the patterns of self-state dynamics. A sequence is a chronologically ordered list of posts within a timeline that culminates in an MoC. We randomly hold out 10 training timelines as our validation set for selecting the best models for submission. The 20 timelines are used for training and providing in-context examples.

[2] A *timeline* is a chronologically ordered collection of posts authored by a single user.

3 Related Work

Previous editions of the CLPsych shared task have attracted a wide range of system submissions. In CLPsych 2022 (Tsakalidis et al., 2022), team `BLUE` (Bucur et al., 2022) experimented with several text representation methods, and their best system consisted of an ensemble of machine learning (ML) classifiers. Team `WResearch` (Bayram and Benhiba, 2022) adopted a pipeline approach in which pre-trained BERT (Devlin et al., 2019) was used to compute emotion and sentiment scores that were passed as input features to downstream ML models. Team `UoS` (Azim et al., 2022) achieved competitive results using a bidirectional long short-term memory (Bi-LSTM) network for mood change prediction and suicide risk level assessment.

The following year's edition, CLPsych 2025 (Tseriotou et al., 2025), introduced more challenging subtasks such as evidence span detection and summarization. Team `uOttawa` (Chan et al., 2025) explored various prompt engineering strategies on top of a 70B-parameter LLM and obtained the best score on self-state identification. Team `BULUSI` (Ravenda et al., 2025) achieved strong results by combining an ensemble with an optimization step on top of LLM predictions. Finally, team `BLUE` (Sandu et al., 2025) obtained competitive performance on summarization through zero-shot prompting of open-weight LLMs.

4 Methodology

This section presents our approaches, which integrate findings from the CLPsych 2025 shared task, as we enhance and refine our strategies.

4.1 Tasks 1.1 and 1.2

We adopt a joint prediction setup in which an LLM is tasked with predicting both the subelements and their presence scores in a single pass. The default system prompt (Figure 4) contains a detailed description of the MIND framework, the definitions of self-states, the characteristics of each subelement, and the criteria for assigning presence scores. We compare several prompting techniques.

Zero-shot prompting. We use the default system prompt and pass the content of the post to be labeled in the user message.

Post-level In-Context Learning (ICL). Here, k full training posts selected uniformly at random, and their gold subelement annotations are included in the prompt. This exposes the LLM to the surrounding context of each subelement and to the

relationships between subelements within a post. A drawback is that coverage of all subelements is not guaranteed, and rare subelements may go unrepresented among the in-context examples.

Post-level ICL with RAG. Identical to post-level ICL, but the k in-context posts are retrieved by cosine similarity to the test post. Posts are encoded as the L2-normalized `CLS` embedding from `BAAI/bge-large-en-v1.5` (Xiao et al., 2023), truncated to 512 tokens.

Subelement-level ICL. In the subelement-level variant of ICL, we append k examples to the definition of each subelement in the system prompt. Each example is a relevant span from a training post that serves as evidence for the corresponding subelement. Note that, in this setting, we do not include the full post. Moreover, the k examples for a given subelement may come from different training posts, offering a more diverse view of how the subelement is expressed.

Ensemble. To reduce noise from any single prediction, we aggregate the outputs of multiple independent LLM runs via majority voting.

4.2 Task 2

We train separate supervised classifiers for switch and escalation changes using feature sets composed of the subelements and presence scores predicted in Tasks 1.1 and 1.2. Each classifier receives a fixed-size window of posts centered on a target post. The window includes both preceding and following posts; we denote the window that includes the following posts as having foresight. Each post in a sequence is labeled with predictions from the best submissions for Tasks 1.1 and 1.2. We experiment with the following features: the predicted presence for each valence, the absolute difference between presence per valence of the target and subsequent post, count of subelements per valence, and post index. We compare two machine learning algorithms: Support Vector Machine (SVM), and Random Forest. Classifiers are trained on 20 posts during validation (all 30 posts are used during test) and are tuned via grid search (Appendix B.2).

4.3 Task 3.1

We adopt an LLM prompting approach for Task 3.1. The system prompt that we use for the task (Figure 8) contains a detailed description of the MIND framework, the definitions of switch and escalation, and the required summarization aspects. We also experimented with a shortened version of this prompt (Figure 9) but observed negligible difference in performance, so all reported results use the longer prompt. The user prompt contains the chronologically ordered post contents of the test sequence. We compare the following approaches:

Zero-shot. The LLM produces the summary directly from the task description and the post contents, with no in-context examples.

ICL. The system prompt is augmented with k in-context examples. Each example consists of the post contents of a training sequence followed by the corresponding gold summary.

Label-augmented ICL. A pipeline-style extension of ICL in which subelement and change labels (switch or escalation) are included alongside the post contents. We consider two variants: (i) augmenting only the in-context examples with their gold labels, and (ii) additionally augmenting the test posts with predicted labels from our Task 1.1 and Task 2 systems. Figure 1 illustrates the full pipeline.

Summary of summaries. In this method, a single LLM first summarizes each post individually and then summarizes the sequence from those per-post summaries.

4.4 Task 3.2

Before prompting the LLM to identify recurrent signatures of improvement or deterioration, we filter the 74 gold training summaries via exact string matching, yielding 56 deterioration and 51 improvement sequences. These are passed to the LLM in batches of 10 to produce partial signatures, which are then merged in a final step that identifies patterns common across batches.

5 Results

This section provides the experimental results. We use the following open-weight LLMs for prompting: google/gemma-3-27b-it (`gemma`) (Gemma Team, 2025); Qwen/Qwen3.5-27B (`qwen`) (Qwen Team, 2026); and openai/gpt-oss-120b (`gpt`) (OpenAI, 2025). See Appendix A for the setup.

5.1 Tasks 1.1 and 1.2

Validation results across these dimensions are summarized in Table 1. On Task 1.1, ICL improves over zero-shot for `qwen` and `gemma` but not for `gpt`. For the ensemble, increasing k generally yields gains, unlike for individual models, and subelement-level ICL outperforms the other approaches. On Task 1.2, `qwen` attains substantially higher RMSE on presence scores than the other

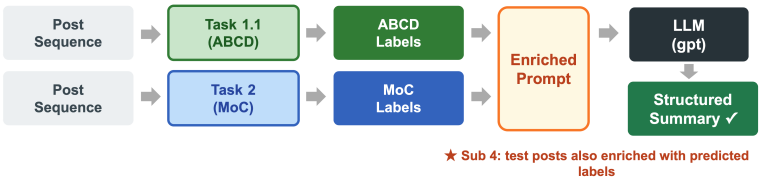


Figure 1: Label-augmented ICL pipeline for Task 3.1. Post sequences are processed through Task 1.1 to produce subelements and through Task 2 to produce MoC labels. Both label sets enrich the prompt fed to the LLM. Submission 3 enriches only the in-context examples with gold labels. Submission 4 additionally enriches the test posts with predicted labels from Tasks 1.1 and 2 at inference time.

Method	k	qwen		gemma		gpt		Ensemble	
		F1↑	RMSE↓	F1↑	RMSE↓	F1↑	RMSE↓	F1↑	RMSE↓
Zero-shot	–	0.309	1.124	0.324	0.926	0.340	0.908	0.330	0.913
Subelement ICL	1	0.328	1.027	0.335	0.860	**0.343**	0.948	0.347	0.852
	2	**0.362**	1.057	0.341	0.860	0.326	0.948	0.361	0.861
	3	0.333	**0.998**	0.354	**0.831**	0.339	**0.888**	**0.366**	**0.787**
Post ICL	1	0.331	1.072	0.330	0.877	0.311	0.939	0.335	0.867
	2	0.334	1.103	0.348	0.867	0.297	0.963	0.337	0.880
	3	0.330	1.047	0.336	0.860	0.307	0.971	0.332	0.859
RAG	1	0.337	1.048	**0.368**	0.920	0.309	0.914	0.330	0.898
	2	0.325	1.070	0.326	0.858	0.313	0.955	0.333	0.903
	3	0.332	1.028	0.327	0.853	0.319	0.953	0.342	0.876

Table 1: Validation results on Tasks 1.1 and 1.2, averaged over 5 runs. k is the number of in-context examples.

models. The LLM exhibits a clear tendency to over-predict subelements, resulting in a high false positive rate. The validation set contains 292 present subelements (adaptive and maladaptive combined), whereas one of the gemma predictions includes 555.

The test results for our three submissions are reported in Table 2. **Submission 1** uses subelement-level ICL with $k = 3$ and gemma as the backbone. **Submission 2** ensembles 5 predictions from qwen and gemma. **Submission 3** ensembles 7 predictions drawn from qwen, gemma, and gpt. Refer to Appendix B.1 for more details about the ensemble members.

Findings. Ensembling improves Task 1.1 macro F1 from 0.416 to 0.431 (+3.6% relative) and reduces Task 1.2 RMSE from 1.047 to 0.997 (−4.78% relative) between **Submissions 1** and **2**, consistent with the view that aggregating across models and runs reduces variance from sampling noise and from the choice of in-context examples (Yang et al., 2023). **Submission 3**, which adds two more members and a third backbone, achieves the **best** Task 1.1 F1 among all CLPsych 2026 participants. For Task 1.2, however, expanding the ensemble from 5 to 7 members slightly hurts performance, likely because the output space is restricted to integers in $\{1, \ldots, 5\}$, leaving little headroom for additional members to contribute useful signal.

	Task 1.1 Macro F1 ↑	Task 1.2 RMSE ↓
Submission 1	0.416	1.047
Submission 2	<u>0.431</u>	**0.997**
Submission 3	**0.442**	<u>1.007</u>
Official baseline	0.247	1.424

Table 2: Test results on Tasks 1.1 (subelement-level macro F1) and 1.2 (RMSE). Submission 3 achieves the best F1 among CLPsych 2026 participants, and Submission 2 achieves the 4th-best RMSE. The official baseline is a one-shot approach with `LLaMA-3.1-8B-Instruct`.

5.2 Task 2

Submission 1[3] uses Random Forest classifiers with predicted Task 1.1 presence scores as the feature, with window size 1 for switch and 2 for escalation.

Submission 2 uses SVM classifiers and expands the feature set with the absolute difference in presence between consecutive posts and the post index,

[3]The training data was not fully used in this submission. We provide the correct evaluation on the validation set.

	Average Macro F1 ↑
Submission 1	0.279
Submission 2	0.472
Submission 3	**0.572**
Official baseline 1	0.272
Official baseline 2	0.365
Official baseline 3	**0.572**

Table 3: Test results on Task 2. Baseline 1 adopts a zero-shot approach using `LLaMA-3.1-8B-Instruct`. The second baseline follows a pipeline approach where the predicted labels from Task 1.1 baseline are used to compute the well-being score deterministically to predict moments of change. Baseline 3 finetunes TempoFormer (Tseriotou et al., 2024).

	CS ↑	CT ↓	RL ↑	BSR ↑
Submission 1	0.797	0.696	0.283	0.249
Submission 2	0.722	0.808	0.244	0.218
Submission 3	0.789	0.714	0.292	**0.295**
Submission 4*	**0.818**	**0.621**	**0.305**	-
Official baseline 1	0.763	0.753	0.255	0.226
Official baseline 2	0.767	0.745	0.269	0.235

Table 4: Test results on Task 3.1. CS refers to consistency, CT refers to contradiction, RL refers to ROUGE recall, and BSR refers to BERTScore recall. The first baseline follows a zero-shot strategy using the model `LlamA-3.1-8B`. The second baseline adopts a pipeline approach where the predictions from the top performing baselines from Task 1 and 2 are used to enrich the prompt context. *Submission 4 was made during the analysis phase of the competition.

and increases window sizes: switch uses a window size 2 and escalation 3.

Submission 3 uses a window size 3 and a larger feature set than the previous submission. For both the switch and escalation model, we create a set of six features per post: presence, absolute presence difference, and subelement count for each valence. While the switch model has a window size 3, we remove foresight so that it does not include posts succeeding our target. This switch model also uses post index as an additional feature. Our validation results are reported in Table 12. The test results for our three submissions are reported in Table 3.

Findings. Our best system (Submission 3) achieved a combined F_1 of 0.572, placing us 4th in Task 2 ranking. Across submissions, escalation was consistently easier to predict than switch. Adding foresight for escalation and refining the window configuration in Submission 3 provided a further improvement, largely driven by a large gain in escalation F_1 (0.585 $\rightarrow$ 0.714 post-level).

5.3 Task 3.1

Validation results are reported in Table 10. ICL performance generally improves with k across all three backbones. Test results are reported in Table 4. **Submission 1** uses plain ICL with $k = 2$ and `gpt` as the backbone. **Submission 2** uses the summary-of-summaries approach with the same backbone, which degrades performance across all metrics: collapsing each post into an intermediate summary appears to discard the fine-grained signal needed to characterize cross-post dynamics. **Submission 3** is the label-augmented ICL variant, in which the in-context examples carry their gold subelements and change labels. **Submission 4** additionally augments the test posts with predicted labels from Tasks 1.1 and 2.

Findings. Submission 4 yields the best result on three of the four metrics among our submissions confirming that the pipeline signal is useful at inference time as well. Moreover, Submissions 1, 3, and 4 outperform both official baselines across all four metrics, demonstrating that larger models with an ICL approach consistently outperform smaller models in a zero-shot setting.

5.4 Task 3.2

We use `gpt` as the backbone LLM. The generated signatures are shown in Section C. Our submission achieved a fit score of 0, a recurrence score of 0, and a specificity score of 0.25. Since this task was evaluated manually by the organizers and gold annotations are not available, it is challenging to provide findings for this task.

6 Conclusion

We presented our submissions to the CLPsych 2026 Shared Task on analyzing self-state dynamics in longitudinal Reddit timelines. Our ensembling approach for Task 1.1 achieved the best performance in subelement classification among all participants. Our pipeline approach for Task 2, which used predictions from Task 1, yielded our strongest result on this task and ranked fourth. The label-augmented in-context learning approach for Task 3.1 ranked third in the competition. Overall, our results indicate that propagating predictions from upstream tasks leads to consistent gains on downstream tasks.

Limitations

Our work has several limitations. First, in our experiments, we used large LLMs that might not be accessible to everyone. This hinders the reproducibility and usage of the methods in resource-constrained environments. One solution is to use quantized models, which offer a reduced memory footprint at the cost of a slight degradation in performance. Second, the size of the datasets is relatively small and limited to 30 users in the training set and 10 users in the test set. Therefore, the conclusions from our paper might not be generalizable to the mental health domain in general. Third, the users in this study are Reddit users, who are not representative of the general population; to further extend the scope of the work, data from other social media platforms could be included. Finally, our submission to Task 3.2 received fit and recurrence scores of 0, suggesting that the batch-and-merge pipeline did not capture meaningful recurrent signatures; further investigation of the filtering and merging steps is needed.

Ethical Considerations

Due to the sensitive nature of mental health data, we stored the dataset in secure, firewall-protected servers. We used only open-weight LLMs and served them locally. We did not use any commercial, closed-weight LLMs in order to preserve the confidentiality of the data. Furthermore, our systems are designed as a support tool to help professional mental health providers and should not be considered a replacement for certified professionals.

Acknowledgment

We thank Asmaa El Hansali for the help with the project and the anonymous reviewers for their insightful comments.

References

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

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

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

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

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.

Ana-Maria Bucur, Hyewon Jang, and Farhana Ferdousi Liza. 2022. Capturing changes in mood over time in longitudinal data using ensemble methodologies. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 205–212.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

National Institute of Mental Health (NIMH). 2024. Mental illness statistics. https://www.nimh.nih.gov/health/statistics/mental-illness. Accessed: 2026-04-27.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Federico Ravenda, Fawzia-Zehra Kara-Isitt, Stephen Swift, Antonietta Mira, and Andrea Raballo. 2025. From evidence mining to meta-prediction: a gradient of methodologies for task-specific challenges in psychological assessment. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 242–248, Albuquerque, New Mexico. Association for Computational Linguistics.

Anastasia Sandu, Teodor Mihailescu, Ana Sabina Uban, and Ana-Maria Bucur. 2025. Capturing the dynamics of mental well-being: Adaptive and maladaptive states in social media. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 225–234, Albuquerque, New Mexico. 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. 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.

Talia Tseriotou, Adam Tsakalidis, and Maria Liakata. 2024. 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.

World Health Organization. 2013. *Mental Health Action Plan 2013–2020*. World Health Organization, Geneva, Switzerland.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. One LLM is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

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*, pages 4489–4500, Singapore. ACM.

A Experimental Setup

All experiments and inference are conducted on secure internal servers to protect the privacy of the individuals represented in the dataset. Our system is equipped with 8 NVIDIA A40 GPUs (48 GB VRAM each) and 40 CPU cores at 2.30 GHz. We serve all LLMs locally with vLLM (Kwon et al., 2023), a high-throughput inference engine for large language models.

Figure 3 presents an example timeline in the training dataset.

B Additional Results

This section provides additional experiments and results.

B.1 Task 1.1 and 1.2 submission details

Tables 5 and 6 list the ensemble members for Submissions 2 and 3 in Task 1.1 and 1.2, respectively. All members use $k = 3$.

#	Approach	LLM
1	Post-level ICL with RAG	gemma
2	Post-level ICL	gemma
3	Subelement-level ICL	gemma
4	Post-level ICL	qwen
5	Subelement-level ICL	qwen

Table 5: Ensemble members for Submission 2 in Task 1.1 and 1.2. All members use $k = 3$.

B.2 Task 2 experiments on the validation set

Training While using the same training and validation set from Task 1, our model predicts F_1 scores generated by gemma and qwen across four prompting strategies—zero-shot (3 results from the gemma model and 2 results from qwen), post-level ICL (6 from each model), subelement-level ICL (6 from each model), and post-level RAG (6 from each model)— yielding 41 training results. We evaluated four feature sets of increasing complexity

#	Approach	LLM
1	Post-level ICL with RAG	`gemma`
2	Post-level ICL	`gemma`
3	Subelement-level ICL	`gemma`
4	Post-level ICL	`qwen`
5	Subelement-level ICL	`qwen`
6	Post-level ICL	`gpt`
7	Subelement-level ICL	`gpt`

Table 6: Ensemble members for Submission 3 in Task 1.1 and 1.2. All members use $k = 3$.

(see Figures 2b, 2d, and 2f): FS1 comprises adaptive and maladaptive presence scores with their absolute inter-post deltas; FS2 adds the post index; FS3 adds per-post element counts; and FS4 combines FS3 with the post index. For both switch conditions (with foresight and without foresight), F_1 improved monotonically with feature set complexity, with FS4 yielding averages of 0.463 and 0.472 respectively. For escalation, FS3 produced a substantially stronger average F_1 of 0.706 against 0.538 and 0.541 for FS1 and FS2, and was therefore selected.

We also evaluated window sizes w_0–w_3 (see Figures 2a, 2c, and 2e). Performance increased consistently with window size across all tasks, with w_3 achieving mean F_1 scores of 0.422, 0.443, and 0.607 for switch with foresight, switch without foresight, and escalation, respectively; w_3 was therefore adopted. For the switch task, removing foresight—restricting the window to prior posts only—produced stronger predictions (F_1 0.490 vs. 0.472 at FS4, w_3), and this formulation was retained for the final system. The best feature set configurations, FS4 for switch and FS3 for escalation, yield F_1 scores of 0.490 and 0.714, respectively.

Sub.	Task	Model	Hyperparam.
2	Escalation Switch	SVM	`C=1, kernel='rbf', gamma='scale'`
3	Escalation	RF	`n_est=200, depth=5, feat='log2', split=5`
3	Switch	SVM	`C=1, kernel='rbf', gamma='scale'`

Table 7: Task 2 hyperparameters by submission and change type. RF = RandomForestClassifier; n_est = n_estimators, feat = max_features; split = min_samples_split; depth = max_depth

B.3 Task 3.1: zero-shot vs. summary of summaries

We first tested a simple zero-shot prompt (Figure 6) in which the model receives the raw post texts and a brief instruction to generate a structured ABCD summary. This served as our baseline for both models.

For this section, we evaluate our approach on the Task 3.1 training set (74 sequences) using two language models: LLaMA 3.2 3B Instruct and Gemma 2 9B Instruct, both loaded with 4-bit quantization (NF4).

Sequential pipeline. Motivated by the winning system of Sandu et al. (2025) (BLUE team), we implement a two-step sequential pipeline. In the first step, the model generates a short post-level summary for each individual post describing the interplay between adaptive and maladaptive self-states. In the second step, these post-level summaries – rather than the raw post texts – are fed to the model to produce the final sequence summary. The prompt used in the second step is shown in Figure 7.

Results. Table 11 show the evaluation results across all configurations. The sequential pipeline consistently improved consistency over the zero-shot baseline for both models, with Gemma 2 9B achieving the best mean CS of 0.7382 ± 0.0068 across three runs.

Sub.	Label	Prec	Rec	F_1
2	Switch	0.500	0.381	0.432
	Escalation	0.591	0.542	0.565
	Macro F_1			*0.499*
3	Switch	0.326	0.714	0.448
	Escalation	0.625	0.833	0.714
	Macro F_1			***0.581***

Table 8: Task 2: Post-level precision, recall, and F_1 by submission.

Sub.	Label	Prec	Rec	F_1
2	Switch	0.450	0.342	0.349
	Escalation	0.675	0.508	0.542
	Macro F_1			*0.446*
3	Switch	0.353	0.592	0.416
	Escalation	0.714	0.721	0.709
	Macro F_1			***0.563***

Table 9: Task 2: Timeline-level precision, recall, and F_1 (macro-averaged over 10 timelines) by submission.

B.4 Task 3.1 results on validation set

In addition to the approaches we discussed in the main text, we implement the following approaches:

LLM as a judge. Three independent ICL summaries are generated by `qwen`, `gemma`, and `gpt`. A separate LLM (`gpt`) then selects the best of the three.

LLM as an aggregator. Same setup as the judge, except that the aggregator LLM is asked to produce a new summary that draws on the three candidates rather than selecting one of them.

The results of the all approaches are reported in Table 10.

C Task 3.2 Recurrent Signatures

Here is the detected recurrent signatures of change by our method:

Signature of Deterioration. The recurrent signature of deterioration is characterized by a shift from adaptive self-states marked by (C-S) and (D) to maladaptive self-states dominated by (A), (C-S), (C-O), and (D). Initially, adaptive self-states are present, often characterized by (A) and (D), but these are gradually overshadowed by maladaptive states. The maladaptive states intensify (A), (C-S), (C-O), and (D) mutually reinforcing each other, culminating in a sense of hopelessness and despair.

Signature of Improvement. The recurrent signature of improvement is characterized by a shift from maladaptive self-criticism (C-S) and depressive affect (A) to adaptive self-compassion (C-S) and content affect (A). This shift involves maladaptive self-neglect behaviors (B-S) being overshadowed by relating behaviors (B-O) and a strengthened desire for connection (D) mutually reinforcing adaptive self-compassion (C-S).

D Dataset

Figure 3 presents an example timeline in the training dataset.

E Prompts

Figures 4, 6, 7, 8, and 9 show the system prompts we used in our approaches.

Model	k	CS ↑	CT ↓	RL ↑	BSR ↑
qwen	1	75.40	76.96	25.47	28.63
	2	75.95	75.93	26.15	30.11
	3	76.20	75.33	27.45	31.49
gpt	1	77.88	70.60	27.90	27.43
	2	79.68	**66.14**	28.56	26.59
	3	**80.56**	67.71	28.59	27.85
gemma	1	75.23	75.32	26.40	31.59
	2	75.83	73.09	28.51	33.41
	3	74.78	75.16	**28.84**	**34.10**
aggregate	1	79.96	70.26	27.82	26.59
	2	78.61	69.89	28.56	26.57
	3	78.38	71.01	28.36	27.39
judge	1	76.01	72.50	27.19	29.34
	2	76.58	73.10	27.23	28.37
	3	77.74	71.86	27.47	27.22
simple-prompt	1	77.99	69.75	28.09	27.37
	2	77.73	71.70	28.03	26.60
	3	79.43	68.45	28.50	27.77

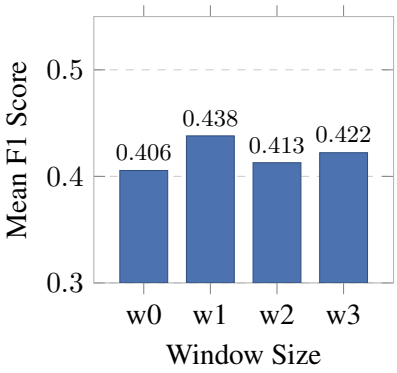
Table 10: Average CS (consistency), CT (contradiction), RL (ROUGE recall), and BSR (BERTScore recall) over 4 independent runs on the validation data. All the numbers are in percentages. The first three models represent ICL methods with different LLMs. In `aggregate` (LLM as an aggregator) and `judge` (LLM as a judge), `gpt` uses the predicitons form the first three models. The `short-prompt` model is ICL with `gpt` as the backbone but with a shorter prompt (Figure 9). Ensemble methods for summary generation degrade the performance of ICL methods.

Configuration	**Model**	**CS ↑**	**CT ↓**	**ROUGE-L ↑**
Zero-shot baseline	LLaMA 3.2 3B	0.7152	0.6233	0.1502
	Gemma 2 9B	0.7086	0.7920	0.1837
Sequential pipeline	LLaMA 3.2 3B – Run 1	0.7275	0.7187	0.2135
	LLaMA 3.2 3B – Run 2	0.7279	0.6979	0.2117
	LLaMA 3.2 3B – Run 3	0.7202	0.7106	0.2076
	LLaMA mean ± std	$\mathbf{0.7252 \pm 0.0040}$	–	–
Sequential pipeline	Gemma 2 9B – Run 1	0.7408	0.7644	0.2006
	Gemma 2 9B – Run 2	0.7304	0.7463	0.1900
	Gemma 2 9B – Run 3	0.7433	0.7674	0.1906
	Gemma mean ± std	$\mathbf{0.7382 \pm 0.0068}$	–	–

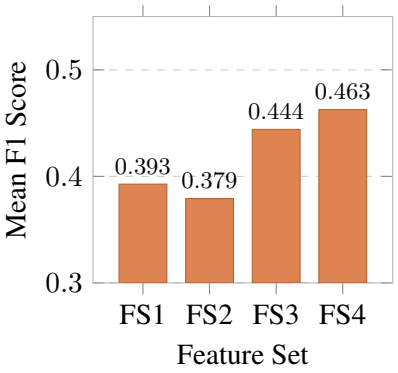
Table 11: Task 3.1 Evaluation Results on Training Set (74 sequences).

	Post-Level			**Timeline-Level**			
	Switch F1	Esc. F1	Macro F1	Switch F1	Esc. F1	Macro F1	Combined F1 ↑
Submission 1	0.304	0.588	0.446	0.239	0.480	0.359	0.403
Submission 2	0.346	0.507	0.426	0.344	0.319	0.331	0.379
Submission 3	**0.510**	**0.791**	**0.650**	**0.559**	**0.697**	**0.628**	**0.639**

Table 12: Validation results on Task 2. Post-level and timeline-level scores are reported per label and as macro F1; the combined F1 is the ranking metric.

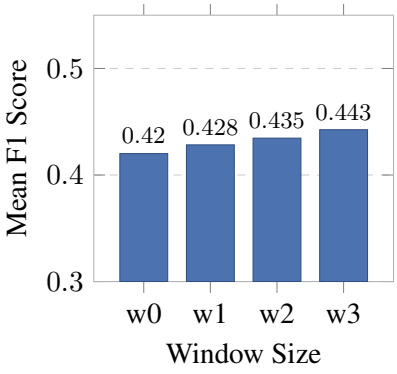


(a) Window size effect.

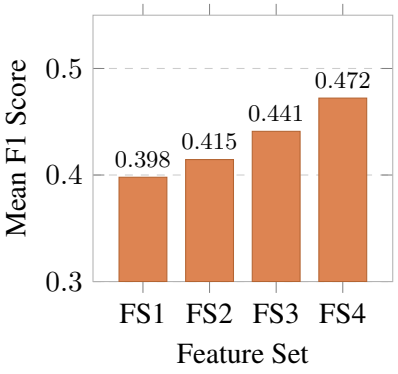


(b) Feature set effect.

Switch with Foresight (AS)

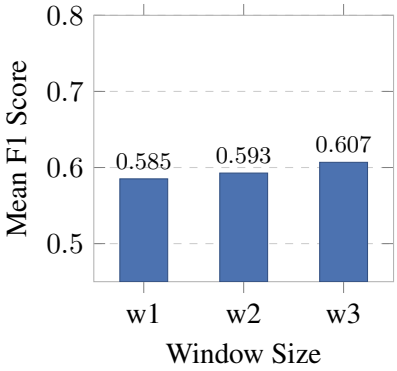


(c) Window size effect.

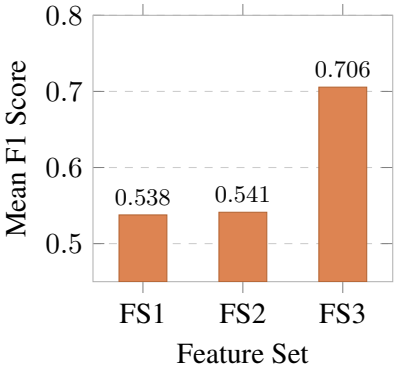


(d) Feature set effect.

Switch without Foresight (BS)



(e) Window size effect (w0 not evaluated).



(f) Feature set effect (FS4 not evaluated).

Escalation (ESC)

Figure 2: Mean F1 scores per window size (left column) and feature set (right column) for each task. FS1: presence + absolute deltas; FS2: FS1 + Post Index; FS3: FS1 + count features; FS4: FS3 + Post Index.

```
{
  "timeline_id": "7d9d2e0e0a",
  "posts": [
    {
      "post_index": 1,
      "post_id": "a33649e870",
      "date": "01-01-2020, 01:02:03",
      "Switch": "0",
      "Escalation": "E",
      "post": "I'm tired. [REMOVED]",
      "Well-being": 4,
      "evidence": {
        "adaptive-state": {
          "B-O": {
            "Category": "(1) Relating behavior",
            "highlighted_evidence": "[REMOVED]"
          },
          "Presence": 2
        },
        "maladaptive-state": {
          "A": {
            "Category": "(4) Depressed, despair, hopeless",
            "highlighted_evidence": "[REMOVED]"
          },
          "B-S": {
            "Category": "(2) Self harm, neglect and avoidance",
            "highlighted_evidence": "[REMOVED]"
          },
          "Presence": 5
        }
      }
    },
    {
      "post_index": 2,
      "post_id": "7f22000b69",
      "date": "01-01-2021, 01:02:03",
      "Switch": "S",
      "Escalation": "E",
      "post": "I'm tired of being alone [REMOVED]",
      "Well-being": 1,
      "evidence": {
        "adaptive-state": {
          "C-O": {
            "Category": "(1) Perception of the other as related",
            "highlighted_evidence": "[REMOVED]"
          },
          "Presence": 2
        },
        "maladaptive-state": {
          "A": {
            "Category": "(4) Depressed, despair, hopeless",
            "highlighted_evidence": "[REMOVED]"
          },
          "Presence": 5
        }
      }
    }
  ]
}
```


Figure 3: A sample timeline data.

System Prompt

```
## Task:
Your task is to identify adaptive and maladaptive self-states from a reddit post. A post can contain zero or more adaptive and maladaptive self-
↪ states.

### Definitions
Self-states are conceptualized as structured combinations of 6 elements: Affect (A), Behavior of the self with the others (B-O), Behavior of the
↪ self toward the self (B-S), Cognition of the others (C-O), Cognition of the self (C-S), and Desire (D). Each present element has exactly one
↪ subelement. The subelements are the specific manifestations of the elements in the post. Here are the definitions of the elements and their
↪ subelements. The percentage of subelements across all posts in the training data is provided in parentheses:
- *Affect* (A): Emotional tone or mood.
    - Adaptive subelements:
        1. Calm / laid back (0.60%)
        2. Sad, emotional pain, grieving (5.36%)
        3. Content, happy, joy, hopeful (8.33%)
        4. Vigor / energetic (0.60%)
        5. Justifiable anger/ assertive anger, justifiable outrage (1.19%)
        6. Proud (4.17%)
        7. Feeling loved, belong (0.00%)
    - Maladaptive subelements:
        1. Anxious/ fearful/ tense (16.67%)
        2. Depressed, despair, hopeless (34.52%)
        3. Mania (0.60%)
        4. Apathic, don't care, blunted (0.00%)
        5. Angry (aggression), disgust, contempt (4.76%)
        6. Ashamed, guilty (4.76%)
        7. Feel lonely (4.76%)
- *Behavior of the self with the others* (B-O): The writer's main behavior(s) toward the others.
    - Adaptive subelements:
        1. Relating behavior (48.21%)
        2. Autonomous or adaptive control behavior
    - Maladaptive subelements:
        1. Fight or flight behavior (13.69%)
        2. Over-controlled or controlling behavior
- *Behavior toward the self* (B-S): The writer's main behavior(s) toward the self.
    - Adaptive subelements:
        1. Self care and improvement (34.52%)
    - Maladaptive subelements:
        1. Self harm, neglect and avoidance (27.98%)
- *Cognition of the others* (C-O): The writer's main perceptions of the other.
    - Adaptive subelements:
        1. Perception of the other as related (19.05%)
        2. Perception of the other as facilitating autonomy needs (1.19%)
    - Maladaptive subelements:
        1. Perception of the other as detached or over attached (45.83%)
        2. Perception of the other as blocking autonomy needs (6.55%)
- *Cognition of the self* (C-S): The writer's main self-perceptions.
    - Adaptive subelements:
        1. Self-acceptance and compassion (25.00%)
    - Maladaptive subelements:
        1. Self criticism (57.14%)
- *Desire* (D): The writer's main desire, expectation, need, intention, or fear.
    - Adaptive subelements:
        1. Relatedness (24.40%)
        2. Autonomy and adaptive control (7.14%)
        3. Competence, self esteem, self-care (21.43%)
    - Maladaptive subelements:
        1. Expectation that relatedness needs will not be met (13.10%)
        2. Expectation that autonomy needs will not be met (4.76%)
        3. Expectation that competence needs will not be met (26.19%)

Self-state rating is the degree to which each identified self state is present in the post. It is an integer between 1 and 5 with the following
↪ definitions:
- 1 (Not present): The self state is not expressed in the post.
- 2 (Somewhat present): The self state is expressed, but plays a subtle, limited role in shaping the person's overall experience.
- 3 (Moderately present): The self state is clearly expressed and moderately contributes to the person's experience.
- 4 (Much present): The self state strongly influences and shapes the experience described in the post.
- 5 (Highly present): The self state strongly shapes and clearly defines the overall experience described in the post.
...
```

Figure 4: The system prompt for the Task 1.1 and 1.2.

System Prompt

```
...
Percentage of subelements across all posts in the training data is provided in parentheses:
- Affect (A):
    - Adaptive subelements: 1 (0.60%), 2 (5.36%), 3 (8.33%), 4 (0.60%), 5 (1.19%), 6 (4.17%), 7 (0.00%)
    - Maladaptive subelements: 1 (16.67%), 2 (34.52%), 3 (0.60%), 4 (0.00%), 5 (4.76%), 6 (4.76%), 7 (4.76%)
- Behavior of the self with the others (B-O):
    - Adaptive subelements: 1 (48.21%), 2 (4.76%)
    - Maladaptive subelements: 1 (13.69%), 2 (0.00%)
- Behavior toward the self (B-S):
    - Adaptive subelements: 1 (34.52%)
    - Maladaptive subelements: 1 (27.98%)
- Cognition of the others (C-O):
    - Adaptive subelements: 1 (19.05%), 2 (1.19%)
    - Maladaptive subelements: 1 (45.83%), 2 (6.55%)
- Cognition of the self (C-S):
    - Adaptive subelements: 1 (25.00%)
    - Maladaptive subelements: 1 (57.14%)
- Desire (D):
    - Adaptive subelements: 1 (24.40%), 2 (7.14%), 3 (21.43%)
    - Maladaptive subelements: 1 (13.10%), 2 (4.76%), 3 (26.19%)


### Output format
You need to output the presence and rating of the adaptive and maladaptive self-states as well as the subelements of the self-states. The
↪ subelement of non-existent self-states should be 0. Write your output in the following JSON format:
```json
{
    "adaptive_states": {
        "A": int (integer between 0 and 7),
        "B-O": int (integer between 0 and 2),
        "B-S": int (integer between 0 and 1),
        "C-O": int (integer between 0 and 2),
        "C-S": int (integer between 0 and 1),
        "D": int (integer between 0 and 3),
        "rating": int (integer between 1 and 5)
    },
    "maladaptive_states": {
        "A": int (integer between 0 and 7),
        "B-O": int (integer between 0 and 2),
        "B-S": int (integer between 0 and 1),
        "C-O": int (integer between 0 and 2),
        "C-S": int (integer between 0 and 1),
        "D": int (integer between 0 and 3),
        "rating": int (integer between 1 and 5)
    }
}
```
```

Figure 5: The system prompt for the Task 1.1 and 1.2. (Cont.)

System Prompt

```
...
SYSTEM_PROMPT = (
    'You are a clinical psychologist specialising in psychodynamic self-state analysis.\n'
    'Your task is to write a structured sequence summary for social media posts '
    'surrounding a mental health change event, grounded in the MIND (ABCD) framework.\n\n'
    'FRAMEWORK:\n'
    'Self-states combine: Affect (A), Behavior-Self (B-S), Behavior-Other (B-O), '
    'Cognition-Self (C-S), Cognition-Other (C-O), Desire (D).\n'
    'Each self-state is Adaptive or Maladaptive. Always abbreviate ABCD in parentheses.\n\n'
    'OUTPUT REQUIREMENTS (in this order, up to 350 words):\n'
    '1. CENTRAL THEME: Dominant ABCD pattern; direction of change '
    '(improvement/deterioration); change event type (Switch/Escalation/both); when it occurs.\n'
    '2. ADAPTIVE DYNAMICS: Presence trajectory and internal ABCD relational dynamics.\n'
    '3. MALADAPTIVE DYNAMICS: Same for the maladaptive state.\n'
    '4. CROSS-STATE DYNAMICS: Dominance, suppression, or dialogue between states.\n\n'
    'CONSTRAINTS: Do NOT print numeric presence scores. Max 350 words.'
)
...
```

Figure 6: Simple zero-shot system prompt for Task3.1 used as baseline for both LLaMA 3.2 3B and Gemma 2 9B.

System Prompt

```
...
{Step 1 -- Post-level prompt (applied to each post individually):}

{``Summarise the interplay between adaptive and maladaptive
self-states in this single post. Identify the dominant self-state
and describe how the core ABCD elements interact.
Write 2--3 sentences only.''}

{Step 2 -- Sequence-level prompt (applied to all post summaries):}

You are a clinical psychologist specialising in psychodynamic
self-state analysis. Your task is to write a structured sequence
summary grounded in the MIND (ABCD) framework.

{Change event definitions:}
SWITCH: sudden change in well-being between two consecutive posts.
ESCALATION: gradual intensification of mood across consecutive posts.

'OUTPUT REQUIREMENTS (up to 350 words, one paragraph):\n'
    '1. CENTRAL THEME: dominant ABCD pattern; direction
    (improvement/deterioration); change event type (Switch/Escalation)
    '2. ADAPTIVE DYNAMICS: presence trajectory; relational dynamics
    between ABCD subelements (MUST be described)
    '3. MALADAPTIVE DYNAMICS: same for the maladaptive state
    '4. CROSS-STATE DYNAMICS: dominance, suppression, reflective
    dialogue (MUST be described if present)
    'CONSTRAINTS: Do NOT print numeric presence scores. Max 350 words.'

{Constraints:} Use (A),(B-S),(B-O),(C-S),(C-O),(D) abbreviations.
No numeric scores. Start with:
{``The central psychological theme revolves around''}
...
```

Figure 7: Sequential pipeline prompt for Task3.1 following Sandu et al. (2025). Step 1 generates post-level summaries; Step 2 generates the sequence summary from those summaries rather than raw post texts.

System Prompt

```
You are a clinical psychologist specialising in psychodynamic self-state analysis.
Your task is to write a structured sequence summary grounded in the MIND (ABCD) framework.

# Framework
In the MIND framework, a self-state is defined as an identifiable unit characterized by specific combinations of Affect, Behavior (towards the
↪ self and others), Cognition (towards the self and others),and Desire (ABCD). An Adaptive self-state pertains to aspects of ABCD that are
↪ conducive to the fulfillment of basic desires/needs. A Maladaptive self-state pertains to aspects of ABCD that hinder the fulfillment of basic
↪ desires/needs. Each of the ABCD elements is operationalized through a set of subelements, representing distinct psychological expressions
↪ within each element:

1. *Affect* (A): Emotional tone or mood
    - Adaptive subelements: Calm/ laid back; Sad, emotional pain, grieving; Content, happy, joy, hopeful; Vigor / energetic; Justifiable anger/
↪ assertive anger, justifiable outrage; Proud; Feeling loved, belong
    - Maladaptive subelements: Anxious/ fearful/ tense; Depressed, despair, hopeless; Mania; Apathic, don't care, blunted; Angry (aggression),
↪ disgust, contempt; Ashamed, guilt; Feel lonely
2. *Behavior of the self with the others* (B-O): The writer's main behavior(s) toward the others
    - Adaptive subelements: Relating behavior; Autonomous or adaptive control behavior
    - Maladaptive subelements: Fight or flight behavior; Over-controlled or controlling behavior
3. *Behavior toward the self* (B-S): The writer's main behavior(s) toward the self
    - Adaptive subelements: Self care and improvement
    - Maladaptive subelements: Self harm, neglect and avoidance
4. *Cognition of the others* (C-O): The writer's main perceptions of the other
    - Adaptive subelements: Perception of the other as related; Perception of the other as facilitating autonomy needs
    - Maladaptive subelements: Perception of the other as detached or over attached; Perception of the other as blocking autonomy needs
5. *Cognition of the self* (C-S): The writer's main self-perceptions
    - Adaptive subelements: Self-acceptance and compassion
    - Maladaptive subelements: Self criticism
6. *Desire* (D): The writer's main desire, expectation, need, intention, or fear.
    - Adaptive subelements: Relatedness; Autonomy and adaptive control; Competence, self esteem, self-care
    - Maladaptive subelements: Expectation that relatedness needs will not be met; Expectation that autonomy needs will not be met; Expectation
↪ that competence needs will not be met

Given a chronologically ordered sequence of posts from a single individual (a timeline), there are two possible clinically meaningful moments of
↪ change:

- SWITCH:  A switch reflects a substantial and sudden change in well-being between two consecutive posts. 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 individual's mood
↪  progressively shifts from neutral or mildly valenced, toward a more extreme state. An escalation may reflect either improvement or
↪ deterioration.

# 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 or Escalation). The summary must describe how psychological change processes evolve across the sequence, and
↪  how they culminate in (when it's a Switch), or unfold through (when it's an Escalation), the identified change event. The direction of the
↪ change (improvement / deterioration) as well as the identity of the change event (Switch/Escalation) should be explicitly stated in the summary
↪ . Describe the change pattern using the MIND framework (ABCD elements).

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. Explain how this theme evolves across the sequence. 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 or when it
↪ culminates.
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 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.
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. examine whether dynamics occur between ABCD subelements across opposite self-
↪ states (suppression/attenuation, reflective dialogue, dominance competition, resilience or other structured interactions). If such cross-self-
↪ state dynamics are present in the sequence, they MUST be described.

# Output requirement
Each reference to an ABCD element should include its abbreviation in parentheses. Use the following mapping: (A) for Affect; (B-S) for Behavior-
↪ self; (B-O) for Behavior-other; (C-S) for Cognition-self; (C-O) for Cognition-other; (D) for Desire. Keep your summary below 350 words and
↪ write it ONE paragraph. Don not mention the post numbers in your summary.

Follow the structre of the examples and write your summary in the same format and start with the term "The central psychological theme revolves
↪ around..."

Think step by step and analyze the posts. Finally, write your summary in the following format:
```json
{
        "summary": "<YOUR SUMMARY>"
}
```
```

Figure 8: The system prompt for the Task 3.1.

System Prompt

```
You are a clinical psychologist specializing in psychodynamic self-state analysis. Write a structured sequence summary grounded in the MIND (ABCD)
↪  framework.

# Framework
A self-state is characterized by combinations of Affect (A), Behavior toward others (B-O), Behavior toward self (B-S), Cognition of others (C-O),
↪ Cognition of self (C-S), and Desire (D). Self-states are either Adaptive (ABCD elements conducive to fulfilling basic needs) or Maladaptive (
↪ ABCD elements that hinder need fulfillment).

Subelements:
- A - Adaptive: calm, sad/grieving, content/hopeful, vigor, justifiable anger, proud, feeling loved. Maladaptive: anxious, depressed/hopeless,
↪ manic, apathetic, aggression/contempt, ashamed, lonely.
- B-O - Adaptive: relating, autonomous/adaptive control. Maladaptive: fight/flight, over-controlling.
- B-S - Adaptive: self-care/improvement. Maladaptive: self-harm/neglect/avoidance.
- C-O - Adaptive: other as related, other as autonomy-facilitating. Maladaptive: other as detached/over-attached, other as blocking autonomy.
- C-S - Adaptive: self-acceptance/compassion. Maladaptive: self-criticism.
- D - Adaptive: relatedness, autonomy, competence/self-esteem. Maladaptive: expectation that relatedness, autonomy, or competence needs won't be
↪ met.

Change events:
- SWITCH: Sudden, substantial change in well-being between two consecutive posts.
- ESCALATION: Gradual intensification of mood across consecutive posts toward a more extreme state.
Both can reflect improvement or deterioration.

# Task
Write a structured summary describing self-state dynamics and their progression across a chronological post sequence surrounding a change event.
↪ Cover:
1. Central recurring theme - how it evolves before and through/culminating in the change.
2. Adaptive and maladaptive self-state dynamics - their relative presence and ABCD subelement interactions (co-activation, reinforcement,
↪ amplification, etc.) across the sequence.
3. Relationship between self-states - how dominance, suppression, coexistence, or cross-state dynamics shift across the sequence.

Explicitly state the change direction (improvement/deterioration) and type (Switch/Escalation). Include ABCD abbreviations inline. Stay under 350
↪ words, one paragraph, starting with: "The central psychological theme revolves around..."

Output format:
{
  "summary": "<YOUR SUMMARY>"
}
```

Figure 9: The shorter version of system prompt for the Task 3.1.