# Team MKC at CLPsych 2026: Capturing and Characterizing Mental Health Changes through Social Media Timeline Dynamics

**Kyomin Hwang*** **Hyeonjin Kim*** **Hyunho Lee*** **Nojun Kwak†**
Seoul National University, Seoul, Republic of Korea
{kyomin98, peaceful1, hhlee822, nojunk}@snu.ac.kr

## Abstract

Recent advances in Large Language Models (LLMs) have motivated their adoption across a wide range of domains, including Artificial Intelligence (AI) for mental health. Given the growing prevalence of mental health disorders worldwide and the limited accessibility of professional care, there is an increasing demand for scalable computational approaches that can assist in early detection and continuous monitoring of psychological well-being. In this area, ongoing efforts have focused on curating domain-specific datasets and leveraging them to develop LLMs capable of supporting holistic mental health analysis. In line with this direction, we propose an LLM-based pipeline for comprehensive mental health analysis over sequentially ordered user posts, as part of the CLPsych shared task. Our pipeline offers a unified framework that jointly enables post-level assessment and user-level temporal modeling.

## 1 Introduction

Recent advances in data collection and computational power have enabled the development of Large Language Models (LLMs) capable of performing a wide range of tasks as generalist AI systems (Brown et al., 2020). General-purpose LLMs such as GPT (Achiam et al., 2023) and Claude (Anthropic, 2025) demonstrate remarkable versatility across diverse domains, from natural language understanding to complex reasoning and code synthesis. Trained on vast corpora spanning web text, scientific literature, and source code, these models acquire broad world knowledge and generalize to novel tasks with minimal task-specific supervision (Kaplan et al., 2020; Kim et al., 2024; Radford et al., 2019). As the capabilities of LLMs continue to expand, there is growing interest in applying these models to specialized fields such as medical, law, and scientific discovery (Xie et al., 2023; Hwang and Kwak, 2026; Colombo et al., 2024).

As part of these advances, there have been efforts to apply LLMs to clinical psychology as well. However, early attempts have faced a critical bottleneck: the scarcity of training data, which stems largely from privacy concerns surrounding sensitive personal information. To address this limitation, Tsakalidis et al. (2022b) released a dataset while simultaneously introducing a novel task of identifying moments of change—temporal points at which shifts in a user's mood become observable. A growing body of subsequent work are leveraging this resource to formulate a range of tasks aimed at broader application of LLMs on mental health, with methodologies to tackle them (Ali et al., 2026; Tseriotou et al., 2025; Atzil-Slonim, 2026; Tsakalidis et al., 2022a).

Building on these advancements, we present an LLM-based pipeline for mental health analysis, developed as part of the CLPsych 2026 shared task (Ali et al., 2026), which leverages a subset of the dataset introduced by Tsakalidis et al. (2022b). Specifically, our pipeline addresses five interrelated tasks: 1) assessing a user's mental state at the post level; 2) estimating the presence rate of each mental state across a user's posting history; 3) identifying Moments of Change within a chronologically ordered sequence of a user's posts; 4) summarizing the change events observed in such sequences; and 5) detecting recurrent patterns and extracting the signature of improvement and deterioration over time. Our pipeline tackles these tasks in integration, spanning post-level assessment, user-level aggregation, and temporal modeling of mental state dynamics.

## 2 Method

In this section, we present a framework for detecting how the mental state and well-being of individ-

* Equal contribution.
† Corresponding author.



*Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026)*, pages 535–546

uals evolve over temporally ordered sequences of social media posts. The proposed framework consists of five components: 1) post-level identification of ABCD elements and self-state, 2) self-state presence rating, 3) identification of moments of change (MOC), 4) summarization of sequences surrounding change events, and 5) identification of recurrent dynamic signatures of change across timelines. In the following subsections, we describe in detail the method employed to address each component.

### 2.1 Step 1

The first step is to classify ABCD subelements (Atzil-Slonim, 2025) and self-states at the post level. Self-states are categorized into two types, adaptive and maladaptive, each comprising six subdimensions. Each post is classified into one value from the subdimension-specific label set within each subdimension, with NONE always included as a valid option. This yields 12 independent classification targets per post, where both the number and the semantics of candidate labels vary across subdimensions, while the NONE option remains consistently available across all targets.

To address the multi-target classification problem, we employ Qwen3-4B-Embedding (Qwen Embed) (Zhang et al., 2025) as a feature extractor and fine-tune it via LoRA (Hu et al., 2022). The model is given a carefully designed prompt as in Figure 1. The resulting embeddings are passed to 12 independent classifiers, each trained with Cross-Entropy loss to handle its respective subdimension.

A central challenge is the severe class imbalance in the train dataset, which causes naive training to bias the model toward majority classes. We address this in two complementary ways. First, we apply inverse-frequency weighting to the Cross-Entropy loss, assigning each class a weight inversely proportional to its frequency. To prevent the extreme imbalance from inflating weights of rare classes and destabilizing training, we clip all weights to a maximum of 10. Second, since using only a subset of training data under such imbalance risks learning inadequate representations for minority classes, we adopt a 5-fold cross-validation strategy, training a separate model on each fold and ensembling the five models for final prediction.

### 2.2 Step 2

The second step builds on the outputs of Step 1 and focuses on predicting a numerical score for each self-state. Each of the two self-states, adaptive and maladaptive, is associated with a 5-point Likert scale score. The goal of Step 2 is to predict these two scores individually.

For this step, we introduce a dedicated predictor for each self-state, which regresses a single scalar value from the embeddings produced by the feature extractor in Step 1. Each predictor is trained using a weighted Huber loss (Huber, 1964), where per-sample weights are derived from the inverse frequency of each score bin (scores 1–5) in the training set, ensuring that underrepresented score levels receive proportionally greater emphasis during optimization. The model output is produced by passing the final linear layer's logit through a sigmoid activation and linearly rescaling it to the target range [1,5]. We adopt the same 5-fold cross-validation strategy as in Step 1, training a separate model per fold and ensembling them for the final prediction. The utilized prompt is in Figure 2.

### 2.3 Step 3

In the third step, we address the task of detecting two distinct types of transitions from a temporally ordered sequence of posts: 1) **Switch**, a substantial and sudden change in well-being between two consecutive posts, and 2) **Escalation**, a gradual intensification of mood over a sequence of consecutive posts. We formulated this step as binary classification problem for both types.

We fine-tune the Qwen Embed using LoRA, attaching two independent linear classifier for Switch and Escalation, each trained separately to specialize in its respective transition pattern. The model operates on a sliding window of size 2, taking as input the posts at time steps $t-1$ and $t$ to predict whether a Switch or Escalation occurs at time $t$. This design reflects the local temporal dependency inherent in each task, particularly for Switch, which is defined over consecutive post pairs.

To account for class imbalance, we employ Binary Cross-Entropy with adaptive class weighting, where each label's loss weight is set inversely proportional to its frequency in the training set. Given that training on a single fold under such imbalance risks learning inadequate representations for minority classes, we adopt a 5-fold cross-validation strategy, training a separate model per fold and ensembling the five predictions for the final output. The prompt used in this step is shown in Figure 3.

### 2.4 Step 4

In the fourth step, given a sequence of posts with their associated information (*e.g.*, sub-elements and self state), the objective is to generate a summary of the entire sequence. To this end, we perform Supervised Fine-Tuning on Qwen3-4B-Instruct-2507 (Yang et al., 2025) and Qwen3.5-4B (Qwen Team, 2026) with LoRA. During training, we provide the ground-truth annotations of each post as input, whereas at inference time we instead rely on the predictions produced by the models trained in Steps 1, 2, and 3, thereby faithfully reflecting the actual pipeline setting. The prompts used in this step are provided in Figure 4 and Figure 5.

### 2.5 Step 5

As the final step, building on the MIND (ABCD) framework and the self-state structure, we analyze how self-state components interact and evolve across sequences surrounding change events, with the aim of identifying and summarizing dynamic patterns of psychological deterioration and improvement that recur across individuals. To this end, we perform this analysis in a zero-shot manner by prompting Qwen3.5-9B (Qwen Team, 2026). The prompt is provided in Figure 6 and Figure 7. Concatenating every given summary at once as the model input resulted in Out-of-Memory error. We detoured this error by first splitting the given sequences into groups, extracting improvement and deterioration signature from each, and finally summarizing the signatures.

## 3 Experiment

### 3.1 Experimental Setting

In this section, we describe the experimental settings used across each step. For Steps 1, 2, and 3, we employed Qwen Embed as the backbone model and fine-tuned it using LoRA, with the rank $r = 16$ and scaling factor $\alpha = 32$. The learning rate was set to $1 \times 10^{-5}$, and a K-fold cross-validation strategy ($K = 5$) was applied consistently across all three steps. The maximum input token length was set to 786, the batch size to 16, and all models were trained for 30 epochs. All experiments were conducted on a single NVIDIA RTX A6000 GPU, and the random seed was fixed to 42 for all experiments. Training a single fold took approximately 42 min for Step 1 and 15 min for Step 2, with the K-fold ensemble scaling roughly linearly. For the SFT in Step 4, we employed LoRA with a rank of $r = 8$ and scaling factor $\alpha = 16$, attaching adapters to all linear layers. Training was performed with a learning rate of 2e-4, a batch size of 8, for 10 epochs.

| | | Step 1 | Step 2 |
|---|---|---|---|
| Adaptive | K-fold | Macro F1 (↑) | RMSE (↓) |
| × | None | 0.232 | 0.784 |
| | Mean | 0.210 | 0.677 |
| | Voting | 0.210 | **0.612** |
| ✓ | None | **0.333** | 0.777 |
| | Mean | 0.312 | 0.685 |
| | Voting | 0.332 | 0.700 |

Table 1: Results on the validation set for Step 1 and Step 2. *Adaptive* indicates whether class-wise adaptive weighting is applied to the classification loss. For K-fold, "None" denotes a single model trained without K-fold, "Mean" aggregates the logits of the ensembled models before selecting the final prediction, and "Voting" determines the final prediction by majority voting.

| Task | Adaptive | K-fold | Score |
|---|---|---|---|
| Step 1 (F1 ↑) | ✓ | None | 0.320 |
| | ✓ | Mean | **0.361** |
| Step 2 (RMSE ↓) | ✓ | None | 1.044 |
| | × | Mean | 1.010 |
| | × | Voting | **1.003** |

Table 2: Test-set results for Step 1 and Step 2. Other conventions follow Table 1.

### 3.2 Results

**Performance on Step 1 and Step 2:** Prior to evaluation on the test set, we randomly partition the full training data into training and validation splits at an 80/20 ratio. Table 1 reports the subelement-average macro F1 (Step 1) and RMSE (Step 2) on the held-out validation set. For both Step 1 and Step 2, we examine two design factors: i) whether adaptive weighting is applied to the classification loss (denoted by the *Adaptive* column), and ii) the K-fold aggregation strategy.

For **Step 1**, adaptive weighting consistently outperforms its non-weighted counterpart across all K-fold configurations, suggesting that adaptive weighting effectively mitigates the class imbalance inherent in the label distribution (Table 1). In contrast, once adaptive weighting is applied, the choice of K-fold aggregation strategy (mean or majority voting) yields only marginal differences relative to the gain obtained from adaptive weighting itself (Table 1). However, when these K-fold strategies are applied to the test set (Table 2), ensembling pro-

vides a clear improvement over the single-model baseline. We conjecture that this discrepancy stems from the difference in data utilization: the single-model setting (K-fold = None) is trained on only 80% of the training data, as the remaining 20% is held out for model selection, whereas the K-fold ensemble leverages the entire training set across its folds, generalizing better on the test set.

For **Step 2**, we observe that disabling adaptive weighting yields stronger validation performance (Table 1). This trend is preserved on the test set (Table 2), where the unweighted configuration again outperforms its weighted counterpart. A notable divergence from the validation results, however, is that K-fold aggregation produces consistent gains at test time across both aggregation strategies, where the *mean* strategy averages the continuous outputs of the per-fold models and the *voting* strategy rounds each fold's prediction to the nearest integer score and selects the mode as the final prediction. We attribute this improvement to the more complete utilization of the training data afforded by K-fold ensembling, which in turn translates into stronger generalization at test time.

**Performance on Step 3:** Table 3 presents the prediction performance on the Detection of Moments of Change (MOC). As shown in the table, the 4B model outperforms the 8B model, which is likely attributable to the limited size of the training data (approximately 500 samples). With such a small dataset, the larger model is more prone to overfitting and fails to fully leverage its capacity, resulting in degraded performance. Furthermore, ensembling the predictions of five models trained via K-fold cross-validation on top of Qwen3-4B yields additional performance gains. This suggests that maximally utilizing all available training data through K-fold ensemble is beneficial, particularly in few-shot learning settings. For the same reason, we kept the sliding window at size 2. Although a longer context could in principle capture Escalation more faithfully, models consuming wider temporal spans tend to overfit under our limited training regime, mirroring the overfitting pattern observed at the model-capacity level.

**Performance on Step 4 and Step 5:** Table 4 presents the performance on Step 4. As shown in the table, Qwen3.5 generally achieves higher performance than Qwen3 across most metrics. However, we observe that the Contradiction score of Qwen3.5 is worse than that of Qwen3. Nevertheless, both models underperform the Baseline, which relies on zero-shot prompting. We attribute this outcome to two potential factors: 1) noise propagated from the results of the preceding Steps 1–3 may have degraded the final performance, and 2) the training dataset may have been too small relative to the model's capacity, suggesting that a more carefully designed prompting could have been necessary. In such a low-resource regime, fine-tuning may also overwrite part of the LLM's pre-trained knowledge, whereas the zero-shot baseline preserves it in full. This is consistent with our Step 3 finding that the smaller Qwen3-4B outperforms the larger Qwen3-8B (Table 3). For Step 5, we adopted a zero-shot approach, which resulted in performance scores of 0.7266 on Improvement and 0.2301 on Deterioration.

| **Model** | **F1 Score** (↑) |
|---|---|
| Qwen3-4B | 0.53 |
| Qwen3-8B | 0.49 |
| Qwen3-4B w/ K-fold | **0.55** |

Table 3: Performance on the test set of Step 3.

| **Model** | **Consistency** (↑) | **Contradiction** (↓) | **Rouge-L Recall** (↑) |
|---|---|---|---|
| Baseline | 0.763 | 0.753 | 0.255 |
| Qwen3.5-4B | **0.669** | 0.857 | **0.290** |
| Qwen3-4B | 0.654 | **0.834** | 0.284 |

Table 4: Performance on the test set of Step 4.

## 4 Discussion

**Future Works** Our analysis of the dataset provided for the shared task revealed a significant class imbalance across the sub-dimensions used in the Step 1 post-level classification. More importantly, several sub-dimensions contained only a single training instance, and some contained none at all. This severe sparsity likely limited the model's ability to learn reliable decision boundaries for these categories, and may partly explain the lower performance observed on low-resource labels. To mitigate this issue, we adopted a K-fold ensemble strategy in the present work. However a more direct solution, which we leave for future research, is to use LLMs to generate synthetic user posts for the under-represented sub-dimensions. This approach could help reduce both the imbalance and the missing-

label problems, leading to a more balanced training signal across all labels.

## 5 Conclusion

In this paper, we presented an LLM-based pipeline for the CLPsych 2026 shared task on capturing mental health changes through social media timelines. To address the class imbalance that commonly appears in mental health data, we designed a loss function that reflects the label distribution and combined it with a K-fold ensemble strategy to improve robustness. We also enabled the model to describe users' mental states in natural language, either through fine-tuning or few-shot examples, making the outputs easier to interpret. Building on these components, our pipeline brings post-level and user-level temporal modeling together into a single unified framework.

## Limitations

In this paper, we proposed a holistic pipeline for Capturing and Characterizing Mental Health Changes through Social Media Timeline Dynamics. Nevertheless, several limitations remain to be addressed in future work. First, the scarcity of training data poses a significant challenge. The dataset employed for training exhibits a substantial class imbalance, with only approximately 500 samples available in total. Consequently, models with larger capacities tend to suffer from severe overfitting. To mitigate this issue, one promising direction is to leverage LLM-based data augmentation to generate additional samples, thereby constructing a more balanced dataset across all classes. Second, the design of sophisticated prompts tailored to the mental health domain warrants further investigation. The performance of LLMs is known to be highly sensitive to subtle variations in prompt formulation. Although this aspect was beyond the scope of the present study, we believe that incorporating domain-specific knowledge related to mental health into prompt engineering could further enhance the effectiveness of the proposed pipeline.

## Ethics

The dataset provided for the CLPsych 2026 shared task contains sensitive content related to users' mental health. Throughout this work, we strictly followed the data handling rules announced by the CLPsych organizers and we avoided any external API service that would have required transmitting the data outside our local environment. Our five-step pipeline is intended as a research artifact for studying mental health dynamics, and any use of it for screening or diagnosis would require oversight by trained mental health professionals. Automated inference of self-states and moments of change carries non-trivial clinical implications when errors occur: false positives may lead to unnecessary intervention, while false negatives could mean a missed opportunity for support.

Our proposed future direction of LLM-based synthetic data generation for underrepresented subdimensions (Section 4) also carries ethical risks that warrant careful consideration. LLMs trained on broad web data may misrepresent the lived experience of clinical populations and produce stereotypical or inaccurate depictions of distress, and any such distortions may be silently propagated once the generated samples are incorporated into training. Generating content in sensitive categories such as self-harm further raises content-safety concerns. Before using such generated data for training, future work should have clinical experts review the samples and check for fairness across different demographic and language groups.

## Acknowledgments

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343, RS-2025-25442338, 26-InnoCORE-01)

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Anthropic. 2025. Claude 3.7 Sonnet system card. Technical report, Anthropic.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, and 1 others. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Peter J. Huber. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, 35:492–518.

Kyomin Hwang and Nojun Kwak. 2026. Retrieval-augmented generation based nurse observation extraction. *arXiv preprint arXiv:2603.26046*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Taehoon Kim, Pyunghwan Ahn, Sangyun Kim, Sihaeng Lee, Mark Marsden, Alessandra Sala, Seung Hwan Kim, Bohyung Han, Kyoung Mu Lee, Honglak Lee, Kyounghoon Bae, Xiangyu Wu, Yi Gao, Hailiang Zhang, Yang Yang, Weili Guo, Jianfeng Lu, Youngtaek Oh, Jae Won Cho, and 23 others. 2024. Nice: Cvpr 2023 challenge on zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 7356–7365.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660, Dublin, Ireland. 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.

Tong Xie, Yuwei Wan, Wei Huang, Zhenyu Yin, Yixuan Liu, Shaozhou Wang, Qingyuan Linghu, Chunyu Kit, Clara Grazian, Wenjie Zhang, and 1 others. 2023. Darwin series: Domain specific large language models for natural science. *arXiv preprint arXiv:2308.13565*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

# A Appendix

## A.1 Related Works

### A.1.1 NLP Approaches in Psychological Assessment

The proliferation of large-scale datasets and the rapid advancement of computational resources have driven the emergence of foundation models capable of performing a wide range of tasks in a generalized manner, spanning both computer vision and natural language processing. These breakthroughs have subsequently inspired researchers to explore the applicability of such models in specialized domains, including psychology and mental healthcare (Tsakalidis et al., 2022a; Atzil-Slonim, 2026). MentalBERT (Ji et al., 2022) represents an early effort in this direction, extending the general-purpose BERT architecture (Devlin et al., 2019) through continued domain-adaptive pretraining on mental health-related corpora, enabling the extraction of psychologically grounded contextual embeddings. Building upon this line of work, MentalLLaMA (Yang et al., 2024) shifts from a purely representational paradigm to a generative one by leveraging LLaMA (Touvron et al., 2023), an instruction-tuned large language model, and further applying supervised fine-tuning on a curated collection of mental health-specific instruction datasets. This allows MentalLLaMA to perform a diverse array of mental health tasks in a conversational and interpretable manner, going beyond the embedding-focused capabilities of its predecessor. Motivated by these advances in NLP for mental health, this paper aims to leverage large language models (LLM) trained on large-scale data to capture and characterize longitudinal changes in mental health states within social media timelines.

### A.1.2 Collecting Psychological Data for NLP

Unlike general-purpose tasks, domains such as psychology present unique challenges in data collection, including privacy concerns and ethical constraints, which inevitably limit the scale and accessibility of psychological data. To address these limitations and advance NLP research in the psychological domain, Tsakalidis et al. curated a large-scale mental health counseling dataset sourced from social media. Building upon this dataset, subsequent works have explored a diverse range of mental health-related tasks, including mood change detection, suicide risk classification, and the identification and summarization of suicide risk evidence, thereby contributing to the broader development of AI for psychology. Motivated by these advances, this paper presents a framework proposed as part of the CLPsych shared task, aimed at capturing and characterizing mental health changes through social media timeline dynamics.

## A.2 Prompt Specifications

Figures 1, 2, 3, 4, and 5 illustrate the prompts employed in each respective task. For the few-shot examples in the Step 4 prompt, instances drawn from the provided dataset were utilized. Specifically, each example was constructed by incorporating the MoC label (*i.e.,* Switch and Escalation indicators), the Maladaptive presence score, the Adaptive presence score, and the composition of each corresponding state as input.

## Prompt used in Step 1

Instruct: Yor task is to identify evidence of adaptive and maladaptive self-states from a post (input text).
Each post can include either:
(1) a single self-state (adaptive or maladaptive),
(2) two complementary self-states (adaptive and maladaptive), or
(3) Evidence of neither an adaptive nor a maladaptive state.

Self-states constitute identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire/Need (ABCD dimensions).
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.

ABCD Dimensions:

1. Affect (A) – Adaptive: Calm/laid back, Sad/grieving, Content/hopeful, Vigorous, Justifiable anger, Proud, Feel loved/belong. Maladaptive: Anxious/fearful/tense, Depressed/hopeless, Mania, Apathic/blunted, Angry/contempt, Ashamed/guilty, Feel lonely.
2. Behavior toward other (B-O) — Adaptive: Relating behavior, Autonomous/adaptive control. Maladaptive: Fight or flight behavior, Over-controlled or controlling behavior.
3. Behavior toward self (B-S) — Adaptive: Self-care and improvement. Maladaptive: Self-harm, neglect and avoidance.
4. Cognition of Other (C-O) — Adaptive: Other as related, Other as facilitating autonomy. Maladaptive: Other as detached/over-attached, Other as blocking autonomy needs.
5. Cognition of Self (C-S) — Adaptive: Self-acceptance and compassion. Maladaptive: Self-criticism.
6. Desire/Need (D) — Adaptive: Relatedness, Autonomy and adaptive control, Competence/self-esteem/self-care. Maladaptive: Expectation that relatedness needs will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met.

Query:

Figure 1: Prompt used for Post-Level Identification of Dominant ABCD Sub-elements and Self-State Composition

## Prompt used in Step 2

Instruct: Your task is to estimate the degree to which each identified self-state is present in the post (input text) on a scale from 1 to 5.
Self-states constitute identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire/Need (ABCD dimensions).
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.

ABCD Dimensions:

1. Affect (A) – Adaptive: Calm/laid back, Sad/grieving, Content/hopeful, Vigorous, Justifiable anger, Proud, Feel loved/belong. Maladaptive: Anxious/fearful/tense, Depressed/hopeless, Mania, Apathic/blunted, Angry/contempt, Ashamed/guilty, Feel lonely.
2. Behavior toward other (B-O) — Adaptive: Relating behavior, Autonomous/adaptive control. Maladaptive: Fight or flight behavior, Over-controlled or controlling behavior.
3. Behavior toward self (B-S) — Adaptive: Self-care and improvement. Maladaptive: Self-harm, neglect and avoidance.
4. Cognition of Other (C-O) — Adaptive: Other as related, Other as facilitating autonomy. Maladaptive: Other as detached/over-attached, Other as blocking autonomy needs.
5. Cognition of Self (C-S) — Adaptive: Self-acceptance and compassion. Maladaptive: Self-criticism.
6. Desire/Need (D) — Adaptive: Relatedness, Autonomy and adaptive control, Competence/self-esteem/self-care. Maladaptive: Expectation that relatedness needs will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met.

Score Scale:
1=Not present,
2=Somewhat present (subtle, limited role),
3=Moderately present (clearly expressed, moderate contribution),
4=Much present (strongly influences the experience),
5=Highly present (strongly defines the overall experience).

Query:

Figure 2: Prompt used for Self State Presence Rating

## Prompt used in Step 3

Instruct: Given two consecutive posts from the same user timeline, determine whether a Switch (sudden and substantial change in wellbeing) or Escalation (gradual intensification of mood toward an extreme state) occurs at the second post.

Figure 3: Prompt used for Identify Moments of Change (MOC)

## Prompt used in Step 4 – (1)

### Task:
Your task is to generate a structured summary describing patterns of self-state dynamics and their progression over time within a sequence of posts surrounding a change (Switch / Escalation).
Each sequence includes:
(1) posts leading up to the change
(2) posts marking the change itself
For each sequence, the structued summary must describe:
(1) how psychological change processes evolve across the sequence
(2) how they culminate in (When it's a Switch), or unfold through (When it's an Escalation), the identified change event
(3) direction of the change (improvement / deterioration) must be indicated
(4) The identity of the change event (Switch / Escalation) must be indicated
(5) The Change Pattern using the ABCE elements and presence scores for the adaptive and maladaptive self-states.
### Definitions:
#### Self-state:
Self-states constitute identifiable units characterized by specific combinations of Affect, Behavior, Cognition, and Desire/Need(ABCD dimensions) that tend to be coactivated in a meaningful manner for limited periods of time.
- An **Adaptive self-state** pertains to aspects of Affect, Behavior (towards the self and others), Cognition (towards the self and others) that are conducive to the fulfillment of basic desires/needs.
- A **Maladaptive self-state** pertains to aspects of Affect, Behavior (towards the self and others), Cognition (towards the self and others) that hinder the fulfillment of basic desires/needs.
#### MOC (Moment of Change):
For each post, Moment of Change (MOC) is indicated. Moment of Change is categorized into three types: Switch, Escalation, and No Change.
- Switch: A switch reflects a substantial and sudden change in well-being between two consecutive posts.
- 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 (very negative or very positive) across a span of posts.
- No Change: A no change indicates that there is no significant shift in well-being between two consecutive posts.

Figure 4: Prompt used for Summarization

## Prompt used in Step 4 – (2)

#### ABCD Dimensions:
1. **Affect (A)**: Type of emotion expressed by a writer
- Adaptive: (1) Calm/ laid back (3) Sad, emotional pain, grieving (5) Content, happy, joy, hopeful (7) Vigor / energetic (9) Justifiable anger / assertive anger, justifiable outrage (11) Proud (13) Feel loved, belong
- Maladaptive: (2) Anxious/ fearful/ tense (4) Depressed, despair, hopeless (6) Mania (8) Apathic, don't care, blunted (10) Angry (aggression), disgust, contempt (12) Ashamed, guilty (14) Feel lonely
2. **Behavior of the self with the other (B-O)**: The writer's main behavior(s) toward the other
- Adaptive: (1) Relating behavior (3) Autonomous or adaptive control behavior
- Maladaptive: (2) Fight or flight behavior (4) Over controlled or controlling behavior
3. **Behavior toward the self (B-S)**: The writer's main behavior(s) toward the self
- Adaptive: (1) Self care and improvement
- Maladaptive: (2) Self harm, neglect and avoidance
4. **Cognition of Other (C-O)**: The writer's main perceptions of the other
- Adaptive: (1) Perception of the other as related (3) Perception of the other as facilitating autonomy needs
- Maladaptive: (2) Perception of the other as detached or over attached (4) Perception of the other as blocking autonomy needs
5. **Cognition of the self (C-S)**: The writer's main self-perceptions
- Adaptive: (1) Self-acceptance and compassion
- Maladaptive: (2) Self criticism
6. **Desire (D)**: The writer's main desire, expectation, need, intention, or fear
- Adaptive: (1) Relatedness (3) Autonomy and adaptive control (5) Competence, self esteem, self-care
- Maladaptive: (2) Expectation that relatedness needs will not be met (4) Expectation that autonomy needs will not be met (6) Expectation that competence needs will not be met
### Example:
SEQUENCE OF POSTS:
{Few-shot Example}
OUTPUT:
{Few-shot Example}
### SEQUENCE OF POSTS:
{}
OUTPUT:

Figure 5: Prompt used for Summarization

## Prompt used in Step 5 (Improvement)

The MIND (ABCD) model:
Affect (A): Type of emotion expressed by a writer.
Behavior of the self with the other (B-O): The writer's main behavior(s) toward the other.
Behavior toward the self (B-S): The writer's main behavior(s) toward the self.
Cognition of the other (C-O): The writer's main perceptions of the other.
Cognition of the self (C-S): The writer's main self perceptions.
Desire (D): The writer's main desire, expectation, need, intention, or fear.

Using the MIND (ABCD) framework and self-state structure, analyse how self-states and their elements interact and evolve in sequences surrounding identified change events.
Focus on high resolution patterns that recur across multiple individuals.

Extract Signature of Improvement.

The signature should describe a recurrent dynamic pattern observed across individuals that leads to and culminates as the change occurs.
Base all points on observable dynamic patterns in the timelines.
Signature should be under 90 words.

{Few-shot Example with Json Structure}

{Sequence of Posts with timeline and sequence id}

Figure 6: Prompt used for Step 5 (Improvement)

## Prompt used in Step 5 (Deterioration)

The MIND (ABCD) model:
Affect (A): Type of emotion expressed by a writer.
Behavior of the self with the other (B-O): The writer's main behavior(s) toward the other.
Behavior toward the self (B-S): The writer's main behavior(s) toward the self.
Cognition of the other (C-O): The writer's main perceptions of the other.
Cognition of the self (C-S): The writer's main self perceptions.
Desire (D): The writer's main desire, expectation, need, intention, or fear.

Using the MIND (ABCD) framework and self-state structure, analyse how self-states and their elements interact and evolve in sequences surrounding identified change events.
Focus on high resolution patterns that recur across multiple individuals.

Extract Signature of Deterioration.

The signature should describe a recurrent dynamic pattern observed across individuals that leads to and culminates as the change occurs.
Base all points on observable dynamic patterns in the timelines.
Signature should be under 90 words.

{Few-shot Example with Json Structure}

{Sequence of Posts with timeline and sequence id}

Figure 7: Prompt used for Step 5 (Deterioration)