Prompt-Based Modeling of Moments of Change and Change Summaries in Mental Health Timelines

Do Minh Duc and Pham Trung Tin and Vu Tran and Nguyen Le Minh
Japan Advanced Institute of Science and Technology (JAIST)
Ishikawa, Japan
{minhducdo,tinpham,vu-tran,nguyenml}@jaist.ac.jp

Abstract

This paper presents our prompt-based approach for modeling mental health timelines from Reddit user posts. We address two tasks: identifying moments of change and generating summaries of clinically meaningful changes across post sequences. Our framework uses large language models with in-context learning to analyze self-states and mental health indicators without task-specific fine-tuning. We build an inference pipeline with vLLM and Qwen2.5-72B-Instruct-GPTQ-Int8, and experiment with few-shot prompting, and balanced few-shot sampling. We also examine how the number of visible posts affects the model's ability to capture temporal changes. Our results suggest that prompt-based methods provide a practical and competitive baseline in low-resource and sensitive mental health settings, particularly for modeling self-state dynamics and generating summaries of psychological change over time.

1 Introduction

Mental health concerns are a pressing global issue (World Health Organization, 2022), motivating computational methods that can support scalable and longitudinal monitoring while accounting for the dynamic nature of psychological well-being. Over the past decade, social media platforms such as Reddit have become important spaces where individuals disclose emotional states, personal experiences, and mental health struggles over extended periods (Coppersmith et al., 2014; Shing et al., 2018; Tsakalidis et al., 2022b). These temporally rich user-generated data provide opportunities to study mental health trajectories over time, enabling earlier detection of meaningful changes and supporting more timely intervention.

Early NLP work in mental health primarily focused on static user-level or post-level classification. However, recent research has emphasized that mental health is better understood as a longitudinal process, where mood, behavior, cognition, and interpersonal context fluctuate across time. This shift has motivated shared tasks that move beyond static prediction toward modeling temporal dynamics, including the identification of Moments of Change and the generation of interpretable summaries of mental health trajectories (Tsakalidis et al., 2022a; Tseriotou et al., 2023). The CLPsych 2026 Shared Task (Ali et al., 2026) further extends this direction by focusing on adaptive and maladaptive self-state components, grounded in the MIND framework (Atzil-Slonim, 2025, 2026), and by requiring systems to identify changes and summarize the psychological processes leading up to them.

In this paper, we propose a prompt-based framework for modeling moments of change and generating change summaries in mental health timelines. Our main contributions are as follows:

- We develop a prompt-based in-context learning framework for analyzing longitudinal posts, focusing on self-states and mental health indicators without task-specific fine-tuning.
- We apply the framework to two CLPsych shared-task settings: detecting Moments of Change and generating summaries of clinically meaningful changes across post sequences.
- We investigate different prompting strategies, including few-shot prompting, and balanced few-shot sampling, to study their effectiveness in a low-resource and sensitive mental health domain.

2 Related work

Mental health assessments on social media have gained significant attention in recent years. Previous CLPsych shared tasks have explored various

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2026), pages 504–509

aspects of mental health analysis, including longitudinal modeling of mood changes (Tsakalidis et al., 2022a) and evidence generation for suicidality risk (Zirikly et al., 2019; Shing et al., 2018).

In-context learning (ICL) has emerged as a powerful technique for leveraging large language models without task-specific fine-tuning (Brown et al., 2020). By providing relevant examples within the prompt, ICL allows models to learn from demonstrations rather than parameter updates. Nevertheless, few-shot performance is highly sensitive to the selected examples and their label composition. Balanced sampling across labels can provide more informative demonstrations, reduce bias toward overrepresented classes, and improve the robustness of prompt-based predictions. This motivates our use of balanced few-shot prompting, in line with recent work showing that strategic example selection plays an important role in ICL performance (Pecher et al., 2025).

3 Problem Formalization

3.1 Task 2: Moments of Change

This task focuses on detecting clinically meaningful moments of change within a chronological user timeline. Given a chronologically ordered sequence of posts from a single individual, teams must identify two types of change:

- SWITCH: A sudden change in well-being between two consecutive posts, occurring when $|\text{Wellbeing}(t) - \text{Wellbeing}(t-1)| \geq 2$.
- ESCALATION: A gradual change over multiple consecutive posts, where well-being shifts from a neutral or mild state toward a more extreme positive or negative state.

For Task 2, let a user timeline be represented as:

$$T_u = (p_1, p_2, \ldots, p_n),$$

where p_t denotes the post at time step t. For each target post p_t, we formulate SWITCH detection over two consecutive posts:

$$\hat{y}_t^{\text{SWITCH}} = f_\theta\left(\mathcal{P}(p_{t-1}, p_t)\right).$$

We formulate ESCALATION detection over three consecutive posts:

$$\hat{y}_t^{\text{ESCALATION}} = f_\theta\left(\mathcal{P}(p_{t-2}, p_{t-1}, p_t)\right).$$

Here, f_θ denotes the LLM-based prediction function parameterized by θ. Both predictions are binary:

$$\hat{y}_t^{\text{SWITCH}} \in \{0, S\}, \hat{y}_t^{\text{ESCALATION}} \in \{0, E\}.$$

3.2 Task 3: Change Summaries

The goal of Task 3.1 is to generate a structured summary of self-state dynamics across a sequence of posts surrounding a change event. The summary should describe how psychological changes evolve over time, either culminating in a SWITCH or gradually unfolding as an ESCALATION. It should also indicate the direction of change and characterize the pattern using the MIND framework, including ABCD elements and adaptive or maladaptive self-state presence scores (Tseriotou et al., 2025).

For Task 3.1, the input is a sequence of posts associated with a change event:

$$X_i = (p_{a_i}, p_{a_i+1}, \ldots, p_{b_i}),$$

Where , X_i denotes the post sequence associated with the i-th change event, and a_i and b_i denote the start and end indices of that sequence. The output is a generated change summary:

$$\hat{s}_i = g_\theta(X_i),$$

where $\hat{s}_i$ describes the self-state dynamics and psychological change expressed across the sequence and g_θ denotes the LLM-based generation function used to produce the change summary.

4 Methodology

Our prompt-based modeling framework addresses Task 2 and Task 3.1 using in-context learning with local LLM inference through vLLM. We utilized Qwen2.5-72B-Instruct-GPTQ-Int8 as our primary language model, running it locally through vLLM to maintain data privacy and control over the inference process. Figure 1 presents an overview of our method.

Given a user timeline $T_u = \{p_1, p_2, \ldots, p_n\}$, we formulate each input instance x_i as a sequence of posts. Our prompt-based framework constructs an input prompt

$$\mathcal{P}(x_i) = [\mathcal{I}; \mathcal{D}_K; x_i],$$

where $\mathcal{I}$ denotes the task instruction and $\mathcal{D}_K$ denotes a set of K in-context demonstrations. In zero-shot prompting, $\mathcal{D}_K = \emptyset$. In few-shot prompting, $\mathcal{D}_K$ contains selected input-output examples.

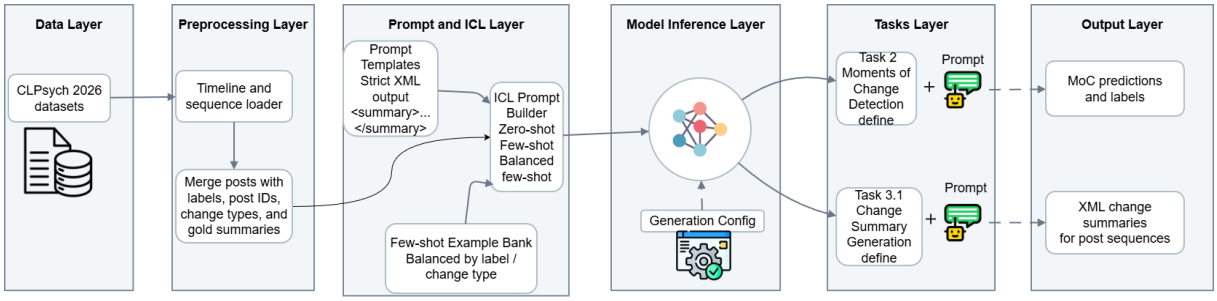

Figure 1: Overview of the prompt-based modeling framework.

For balanced few-shot prompting, demonstrations are sampled such that each label $c \in \mathcal{Y}$ appears with frequency: $|\mathcal{D}_c| \approx \frac{K}{|\mathcal{Y}|}$.

For Moment of Change detection, the model predicts:

$$\hat{y}_i = \arg\max_{y \in \mathcal{Y}} P_\theta(y \mid \mathcal{P}(x_i)), \tag{1}$$

whereas for Change Summary generation, the model produces:

$$\hat{s}_i = \text{Decode}_\theta(\mathcal{P}(x_i)) \tag{2}$$

We further define the input context with k visible posts as: $x_i^{(k)}$ following:

$$x_i^{(k)} = \{p_{i-k+1}, \ldots, p_i\},$$

where k denotes the number of visible posts provided to the model p_i is post in index i on timeline. To study the effect of context length, we vary the number of visible posts $k \in \{3, 4, 5\}$.

4.1 Prompt and ICL Layer

The Prompt and ICL layer is responsible for converting each processed timeline instance into a task-specific prompt. Each prompt consists of three components: a task definition, optional in-context demonstrations, and the target input sequence. The task definition describes the objective of the corresponding CLPsych 2026 task, while the demonstrations provide examples of input-output behavior for few-shot settings. In the zero-shot setting, the prompt contains only the task definition and the target instance. In the few-shot setting, we prepend selected examples from the training set. For balanced few-shot prompting, examples are sampled so that different labels or change types are represented with comparable frequency.

For Task 2, the prompt defines the objective as identifying Moments of Change in a user's mental health trajectory based on longitudinal signals from a sequence of posts. The model is instructed to compare the target post or segment with the surrounding context and determine whether it reflects a meaningful change in the user's self-state or mental health condition. The output is constrained to the predefined task label format.

For Task 3.1, the prompt defines the objective as generating a concise and interpretable summary of clinically meaningful change across a sequence of posts. The model is instructed to summarize the change trajectory rather than producing independent post-level summaries. In particular, the prompt encourages the model to capture the user's self-state dynamics, including affective, behavioral, cognitive, and desire-related signals when they are present in the input. To make the output easier to parse, we require the model to return the generated summary in a strict XML format.

4.2 Tasks Layer

The Tasks layer separates the inference process into two CLPsych 2026 shared-task objectives. The first objective, Task 2, focuses on Moment of Change detection. Given a timeline context, the model predicts whether the target point corresponds to a meaningful transition in the user's mental health trajectory. This task is formulated as a label prediction problem, where the LLM produces a task label conditioned on the constructed prompt.

The second objective, Task 3.1, focuses on Change Summary generation. Given a sequence of posts associated with a change event, the model generates a short summary describing the psychological change expressed across the sequence. Unlike Task 2, this task is formulated as a conditional generation problem. The generated summary is expected to describe the direction and nature of the change, including the user's self-state before, during, and after the transition when such information is available.

Table 1: Task 2 results ranked by average Macro F1. Our team is highlighted in bold.

Team Name	Task 2 - Avg Macro F1
USAI	0.600
CtbuY	0.588
CUNY	0.572
MKC	0.554
Codezone Research Group	0.553
Aurevia	0.484
Meronym Labs	0.466
debju	0.447
JNLP few-shot	**0.566**
JNLP few-shot balance	**0.580**

Although both tasks share the same prompt-based inference backbone, they differ in their output structure. Task 2 produces discrete Moment of Change predictions, whereas Task 3.1 produces natural-language summaries wrapped in XML tags. This separation allows the same LLM inference pipeline to support both classification-style and generation-style tasks.

5 Experiment

We evaluate our prompt-based framework on two CLPsych 2026 shared-task settings: Task 2, which focuses on Moment of Change detection, and Task 3.1, which focuses on change-summary generation. For Task 2, we submit two configurations: a few-shot and a balanced few-shot in-context learning setting. This allows us to examine whether adding balanced demonstrations improves the model's ability to identify change patterns in mental health timelines.

Table 1 presents the Task 2 results in terms of average Macro F1. Our balanced few-shot configuration achieves an average Macro F1 of 0.580, improving over the few-shot standard configuration, which obtains 0.566. This suggests that balanced in-context examples help the model better distinguish between different types of change, especially in a low-resource and sensitive domain where task-specific fine-tuning is not applied.

We analyze different context window sizes $k \in \{3, 4, 5\}$ to examine how much local history is needed for detecting changes in mental health timelines in Table 4. This design reflects the difference between the two labels: SWITCH can often be detected from local contrast between consecutive posts, whereas ESCALATION may require a longer context to capture gradual changes over time. Based on this analysis and the official shared-task evaluation, we submitted two final configurations: standard few-shot prompting and balanced few-shot prompting.

Table 2: Task 3.1 results. Our team is highlighted in bold.

Team Name	Consistency	Contradiction	ROUGE-L
Aurevia	0.866	0.625	0.185
psytechlab	0.857	0.571	0.078
DreamerNLplus	0.845	0.439	0.189
Meronym Labs	0.801	0.659	0.266
CUNY	0.797	0.696	0.283
Meronym Labs	0.787	0.673	0.282
McMasterNLP	0.770	0.761	0.208
JNLP few-shot	**0.791**	**0.666**	**0.117**

For Task 3.1, we evaluate the ability of our system to generate summaries that describe self-state dynamics and psychological changes across post sequences. Unlike Task 2, which is formulated as a prediction task, Task 3.1 requires the model to generate structured natural-language summaries. We report results using three evaluation metrics: Consistency, Contradiction, and ROUGE-L.

Table 2 shows the Task 3.1 results. Our system achieves a Consistency score of 0.791, which is comparable to several higher-ranked systems. This indicates that the generated summaries are generally aligned with the input post sequences. However, the lower ROUGE-L score suggests that our generated summaries may differ lexically from the reference summaries, even when they capture similar change patterns. This is expected in open-ended generation tasks, where multiple valid summaries can describe the same psychological transition using different wording.

6 Conclusion

This paper presented a prompt-based framework for detecting Moments of Change and generating change summaries in mental health timelines. Using Qwen2.5-72B-Instruct-GPTQ-Int8 with vLLM, we explored zero-shot, few-shot, and balanced few-shot prompting strategies without task-specific fine-tuning. Our results show that balanced few-shot prompting improves Task 2 performance over standard few-shot prompting, while the generated summaries achieve reasonable consistency with the input timelines despite lower lexical overlap with reference summaries. These findings suggest that prompt-based LLMs can serve as a practical baseline for low-resource and sensitive mental health timeline modeling. Future work should explore

more robust prompt optimization, domain adaptation, and human-centered evaluation of generated change summaries.

7 Limitations

First, the LLM used in this paper was pretrained mainly on large-scale open general data, meaning there is no guarantee that it contains sufficient high-quality clinical or professional knowledge for understanding mental health-related content; fine-tuning or further adapting the model with expert-annotated mental health data may help reduce this limitation. **Second**, although in-context learning can guide LLMs to perform new tasks through task instructions and demonstrations, the model may still fail to fully understand complex mental health contexts, which is especially challenging for longitudinal timelines where psychological changes may be subtle, gradual, or expressed indirectly across multiple posts. **Third**, since our method relies on prompt-based inference, its performance can be affected by the wording of task instructions, the number of visible posts, and the selected few-shot examples; although balanced few-shot sampling helps control label distribution, different prompt formulations or demonstration choices may lead to different predictions.

8 Ethics

Secure access to the shared task dataset was provided with IRB approval under University of Maryland, College Park protocol 1642625 and approval by the Biomedical and Scientific Research Ethics Committee (BSREC) at the University of Warwick (ethical application reference BSREC 40/19-20). We do not include raw user posts or personally identifiable information in the paper. Any examples, prompt illustrations, or descriptions are paraphrased or abstracted to comply with the data access agreement and protect user privacy.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K16954. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the author(s)' organization, JSPS or MEXT.

References

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

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Branislav Pecher, Ivan Srba, Maria Bielikova, and Joaquin Vanschoren. 2025. Automatic combination of sample selection strategies for few-shot learning.

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

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

World Health Organization. 2022. *World Mental Health Report: Transforming Mental Health for All*. World Health Organization, Geneva.

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

A Prompt Templates

A.1 Prompt Template

The few-shot balance examples include representative cases for four label combinations: `Switch=0, Escalation=0`, `Switch=S, Escalation=0`, `Switch=0, Escalation=E`, and `Switch=S, Escalation=E`.

Prompt Instruction: Switch and Escalation Detection

 # Instruction
You are given a timeline of Reddit posts written by the same user. Your task is to determine whether the **last post** shows a mental health change.
Define the switch and Esscalation follow CLPsysch **# User Prompt**
Timeline:
{USER_TIMELINE}
 # Few-shot Examples
Examples
 # Answer
XML format: <Esscalation>...</Esscalation>, <Switch>...</Switch>

For Task 3.1, we use few-shot prompting for change summary generation. Few-shot demonstrations are randomly sampled from the training set and prepended to the target input. Each demonstration contains a post sequence and its reference change summary, helping the model learn the expected summary style, including temporal progression, psychological dynamics, and change direction.

Prompt Instruction: Change Summary Generation

 # Instruction
You are given a chronologically ordered sequence of Reddit posts surrounding a clinically meaningful mental health change event. Your task is to summarize the user's mental health changes across the sequence. Focus on psychological dynamics, not on summarizing each post separately.
Mention the main mental health theme, how it changes over time, the change type as Switch or Escalation, and the direction as improvement or deterioration.
 # User Prompt Posts in chronological order:
{POSTS}
 # Few-shot Examples
Examples
 # Answer XML format: <summary>...</summary>

A.2 Model Settings

Table 3 shows the model and decoding settings used in our experiments.

Table 3: Model and decoding settings used in our experiments.

Setting	Value
Model	Qwen2.5-72B-Instruct-GPTQ-Int8
Inference engine	vLLM
Prompting strategy	Few-shot prompting
Max new tokens	8192
Temperature	0.0
Top-p	0.95
Top-k	70
Decoding strategy	Greedy decoding

Table 4 reports the F1-scores for Switch and Escalation detection under different prompting strategies and context window sizes.

Table 4: F1-scores for Switch and Escalation detection under different prompting strategies and context window sizes on the training data.

Context Window	Switch F1	Escalation F1
Zero-shot		
$k = 3$	0.453	0.407
$k = 4$	0.420	0.378
$k = 5$	0.410	0.634
Few-shot		
$k = 3$	0.435	0.513
$k = 4$	0.427	0.566
$k = 5$	0.447	0.568
Few-shot balance		
$k = 5$	0.410	0.634