From Posts to Timelines: Modeling Mental Health Dynamics from Social Media Timelines with Hybrid LLMs

Zimu Wang[1,2,*], **Hongbin Na**[3,*], **Rena Gao**[4], **Jiayuan Ma**[5], **Yining Hua**[6], **Ling Chen**[3], **Wei Wang**[1]
[1]School of Advanced Technology, Xi'an Jiaotong-Liverpool University
[2]University of Liverpool [3]University of Technology Sydney
[4]The University of Melbourne [5]The University of Sydney [6]Harvard University
Zimu.Wang19@student.xjtlu.edu.cn, Hongbin.Na@student.uts.edu.au

Abstract

Social media data is recognized for its usefulness in the early detection of mental disorders; however, there is a lack of research focused on modeling individuals' longitudinal mental health dynamics. Moreover, fine-tuning large language models (LLMs) on large-scale, annotated datasets presents challenges due to privacy concerns and the difficulties on data collection and annotation. In this paper, we propose a novel approach for modeling mental health dynamics using hybrid LLMs, where we first apply both classification-based and generation-based models to identify adaptive and maladaptive evidence from individual posts. This evidence is then used to predict well-being scores and generate post-level and timeline-level summaries. Experimental results on the CLPsych 2025 shared task demonstrate the effectiveness of our method, with the generative-based model showing a marked advantage in evidence identification.

1 Introduction

Mental disorders have emerged as a critical global challenge, being recognized as one of the leading contributors to illness and disability (Hua et al., 2024; Na et al., 2025). The World Health Organization[1] (WHO) reports that over 25% individuals will experience mental or neurological disorders in their lifetime. This phenomenon has been further exacerbated by COVID-19, leading to significant increases in anxiety and depression (Penninx et al., 2022), underscoring the urgent need for enhanced monitoring systems to facilitate early intervention.

Despite this phenomenon, mental health services remain undertreated and under-resourced, particularly in low- and middle-income countries. Social media platforms, such as X[2] and Reddit[3], offer significant potential for the early detection of mental disorders, as users regularly express their thoughts, emotions, and behaviors on these platforms. By leveraging machine learning algorithms, especially those utilizing large language models (LLMs), to analyze this data, it becomes possible to identify patterns indicative of disorders like depression or anxiety, facilitating earlier interventions (Shing et al., 2018; Tsakalidis et al., 2022a,b; Chim et al., 2024; Wang et al., 2024a,b; Qian et al., 2024). However, these methods are often limited to individual posts, with the longitudinal modeling of individuals' mental health dynamics largely overlooked in prior research. Moreover, due to privacy concerns and the challenges associated with collecting and annotating mental health data, fine-tuning LLMs on large-scale, curated annotated datasets remains challenging. As a result, prompt engineering are emerged as a promising and valuable line for mental health-related research (Peng et al., 2023; Na et al., 2024; Ma et al., 2025).

In this paper, we introduce a novel approach to modeling mental health dynamics from social media using hybrid LLMs, where the tasks explored include *Adaptive/Maladaptive Evidence Identification*, *Overall Well-being Rating*, and *Post-level and Timeline-level Summaries*. Specifically, in accordance to the prompts organized in Figure 3, we first leverage both classification-based and generation-based models with LLMs to identify adaptive and maladaptive evidence from individual posts (Figure 1). This evidence is then integrated to predict users' well-being scores and generate post-level and timeline-level summaries (Figure 2). The evidence identification and well-being rating tasks are performed using fine-tuned LLMs based on Qwen2.5-7B (Yang et al., 2025), while the summaries are generated using Qwen2.5-32B through in-context learning (ICL, Brown et al., 2020). In-context examples are selected from the training set based on the highest post similarity with the

[*]Equal contribution.
[1]https://www.who.int/
[2]https://x.com/
[3]https://www.reddit.com/

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 249–255

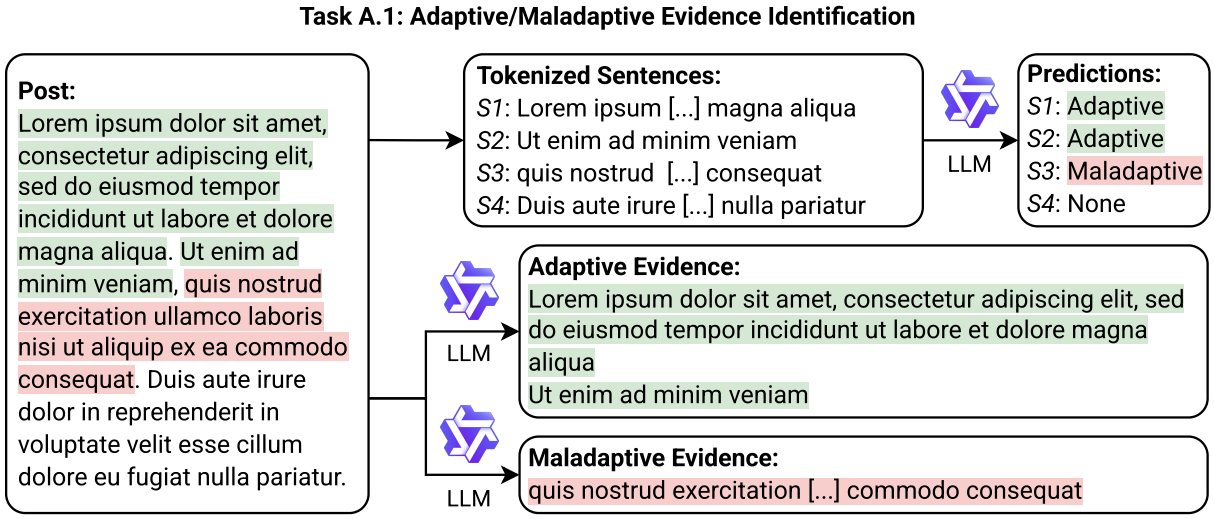

Figure 1: Framework for **Task A.1** using classification-based and generation-based models, with the post replaced by *lorem ipsum* for illustrative purposes.

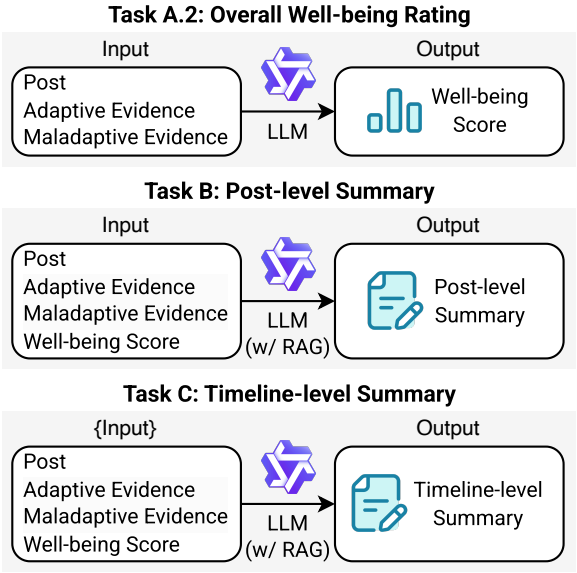

Figure 2: Frameworks for **Tasks A.2, B, and C**.

BGE-Large embedding model (Xiao et al., 2024). Experimental results on the CLPsych 2025 shared task (Tseriotou et al., 2025) highlight the effectiveness of our method, and the generative-based model demonstrates a significant advantage in evidence identification. Both the generative-based model and the classification-based model achieve similar performance when their extracted evidence is integrated into subsequent tasks.

2 Methodology

2.1 Evidence Identification

We begin by identifying the adaptive and maladaptive evidence in each post, where each post may contain a single self-state, two complementary self-states, or neither, represented by continuous spans extracted from the post. To accomplish this, we employ classification-based and generation-based models for this process:

Classification-based Model. Given a post p consisting of multiple sentences, we first tokenize the text into individual sentences $\{t_1, \ldots, t_M\}$, where M represents the total number of sentences in the post. We then apply a fine-tuned LLM, denoted as $\mathcal{M}_C$, to classify each sentence into an estimated label $\hat{e} \in \mathcal{E}$ conditioned on an instruction I_C:

$$\hat{e} = \arg\max_{e} P(e|t_i, I_C, \mathcal{M}_C), \quad (1)$$

where $\mathcal{E} = \{\text{ADAPTIVE}, \text{MALADAPTIVE}, \text{NONE}\}$. t_i ($i \in [1, M]$) represents an input tokenized sentence. Intuitively, this method is highly dependent on the results of sentence tokenization, and may not be well-suited for cases involving complementary self-states, where both adaptive and maladaptive evidence might coexist.

Generation-based Models. In this approach, we leverage two LLMs, $\mathcal{M}_A$ and $\mathcal{M}_M$, each independently trained to identify adaptive and maladaptive evidence, respectively, enabling the direct generation of target evidence within each post p:

$$E_A = \mathcal{M}_A(p, I_A), \quad E_M = \mathcal{M}_M(p, I_M), \quad (2)$$

where I_A and I_M represent instructions for extracting adaptive and maladaptive evidence, while E_A and E_M denote the corresponding lists of sentences containing each type of evidence.

2.2 Overall Well-being Rating

The well-being score, derived from the Global Assessment of Functioning (GAF, American Psychiatric Association et al., 1994), measures an individual's overall functioning across three key domains: social functioning, occupational functioning, and psychological well-being. In this work, we utilize a fine-tuned LLM, denoted as $\mathcal{M}_W$, to predict the well-being score s based on a given post p and its corresponding adaptive and maladaptive evidence, E_A and E_M, respectively, as follows:

$$s = \mathcal{M}_W(p, E_A, E_M, I_W), \quad (3)$$

where I_W represents the instruction for predicting the well-being score.

2.3 Post-level and Timeline-level and Summaries

Next, we generate a post-level summary that captures the interaction between adaptive and maladaptive states identified in each post. Previous research has revealed that prompting larger LMs yields superior summarization performance than fine-tuning smaller models (Thulke et al., 2024); therefore, we leverage ICL (Brown et al., 2020) that conditions LLMs with few-shot demonstrations in producing more effective summaries. To identify the most suitable in-context example, given a candidate post p and a set of annotated posts $\mathcal{D} = \{p'_1, \ldots, p'_N\}$ belonging to multiple users, we first utilize an embedding model $Emb(\cdot)$ to generate the embeddings for the posts, and then locate the post p' that exhibits the highest semantic similarity to p:

$$\boldsymbol{v}_p = Emb(p), \quad (4)$$

$$\boldsymbol{v}_{p'_i} = Emb(p'_i) \quad \forall p'_i \in \mathcal{D}, \quad (5)$$

$$p' = \arg\max_{p'_i \in \mathcal{D}} \frac{\boldsymbol{v}_p \cdot \boldsymbol{v}_{p'_i}}{||\boldsymbol{v}_p|| \cdot ||\boldsymbol{v}_{p'_i}||}, \quad (6)$$

where $\boldsymbol{v}_{(\cdot)}$ denotes the embedding of a given post. Afterwards, we generate the summary m based on the post p and its corresponding evidence E_A and E_M, incorporating the retrieved $\{p', E'_A, E'_M, m'\}$ with an LLM $\mathcal{M}$ and an instruction I_{PS}:

$$m = \mathcal{M}(p, E_A, E_M, p', E'_A, E'_M, m', I_{PS}). \quad (7)$$

For the timeline-level summary, the generation process follows a similar approach to the post-level summary. However, instead of individual posts, all posts associated with each user are concatenated to identify the most relevant in-context example. Additionally, evidence from all posts is incorporated during the generation process.

3 Experiments

3.1 Dataset

The CLPsych 2025 shared task (Tseriotou et al., 2025) integrates longitudinal modeling of social media timelines with evidence generation, offering annotated evidence for both adaptive and maladaptive self-states, as well as a score representing the overall well-being reflected in each post. It also provides post-level summaries that capture the interaction between adaptive and maladaptive self-states within individual posts, and timeline-level summaries that offer clinical insights, along with a dynamic narrative of mental state fluctuations and trajectories over time. This task is organized around the MIND framework (Slonim, 2024), a pan-theoretical model that conceptualizes human experience as a series of self-states that evolve and fluctuate over time.

3.2 Baselines and Evaluation Metrics

Task A.1. The baselines for the task *Adaptive/ Maladaptive Evidence Identification* include a zero-shot Llama 3.1-8B (Grattafiori et al., 2024) and a fine-tuned BART-Large (Lewis et al., 2020) model. The input for both models consists of either a single post or a window of five consecutive posts. Experimental results were evaluated using recall-oriented BERTScore (Zhang et al., 2020) and weighted recall metrics computed over adaptive and maladaptive spans.

Task A.2. The baselines for *Overall Well-being Rating* include zero-shot Llama 3.1-8B and a fine-tuned BERT model (Devlin et al., 2019), where the input for both models consists of either a single post or a window of five consecutive posts. Metrics for this task include Mean Squared Error (MSE), computed for each post within a timeline and then averaged across all timelines. The MSE for posts that indicate serious impairment (1 to 4), impaired (5 to 6), or minimal impairment (7 to 10) to functioning were also calculated. Macro F1-scores were also evaluated based on the aforementioned classes and their corresponding ranges.

Tasks B and C. The baselines for the *Post-level and Timeline-level Summary* tasks include a zero-shot Llama 3.1-8B model, with an intermediate post-level summary also utilized to generate a self-state summary. The evaluation metrics encompass mean consistency, maximum contradiction, and maximum entailment.

Model	Overall		Adaptive		Maladaptive	
	R	W	R	W	R	W
Llama 3.1	35.8	33.7	30.6	29.3	38.2	41.1
w/ Win.	49.6	26.2	36.5	25.2	**62.7**	27.2
BART	40.4	38.2	47.3	46.4	33.6	29.9
w/ Win.	26.0	25.8	28.2	27.9	23.8	23.7
Ours (C.)	34.1	31.4	24.9	24.9	43.3	37.8
Ours (G.)	**50.7**	**45.6**	**49.9**	**46.5**	51.6	**44.6**

Table 1: Experimental results of our proposed method against baselines on Task A.1 (*Adaptive/Maladaptive Evidence Identification*). "R" and "W" denote recall and weighted recall; "C." and "G." denote classification-based and generation-based models; *w/ Win.* represents the incorporation of post windows.

Model	MSE↓	M-S	M-I	M-M	F1
Llama 3.1	4.22	4.67	3.66	3.20	25.5
w/ Windows	4.46	1.67	3.20	7.07	27.4
BERT	2.90	3.38	2.32	2.81	13.9
w/ Windows	4.56	5.68	**1.01**	5.34	13.5
Ours (w/ Class.)	**2.01**	1.25	3.11	**2.16**	**36.6**
Ours (w. Gen.)	2.17	**1.23**	3.60	2.31	34.3

Table 2: Experimental results of our proposed method against baselines on Task A.2 (*Overall Well-being Rating)*. "M-S", "'M-I", and "M-M" denote MSE across serious impairment, impaired, and minimal impairment.

3.3 Experiment Setup

We utilized two distinct LLMs for different tasks in our research. For Tasks A.1 and A.2, we fine-tuned Qwen2.5-7B (Yang et al., 2025) on the relevant datasets using LoRA (Hu et al., 2022). During the fine-tuning process, we configured the number of epochs to 10, the batch size to 2, and set the gradient accumulation steps to 8. For Tasks B and C, we used Qwen2.5-32B as the base model, and we leveraged BGE-Large (`EN-v1.5`, Xiao et al., 2024) as the embedding model to select the in-context example with the highest similarity to the target post. All experiments were conducted on 2 NVIDIA L20 graphics cards.

3.4 Experimental Results

Tables 1, 2, 3, and 4 present the performance of our approach compared to the baselines for Tasks A.1, A.2, B, and C, respectively. From the results, we observed that our method outperformed the majority of the baselines except on Task C. Among the two evidence identification models we proposed, the generative-based model proved to be significantly more effective due to its more accurate identification of evidence locations. However, we also found

Model	Mean Con.	Max Con.↓	Max Ent.
Llama 3.1	88.0	84.8	–
w/ Windows	**89.1**	83.6	–
Ours (w/ Class.)	82.9	80.8	**75.0**
Ours (w/ Gen.)	88.0	**78.1**	69.2

Table 3: Experimental results of our proposed method against baselines on Task B (*Post-level Summary)*. "Mean Con.", "Max Con.", and "Max Ent." denote mean consistency, maximum contradiction, and maximum entailment.

Model	Mean Con.	Max Con.↓
Llama 3.1	87.8	79.9
w/ Windows	**94.0**	**58.0**
Ours (w/ Class.)	91.4	78.5
Ours (w/ Gen.)	91.5	87.6

Table 4: Experimental results of our proposed method against baselines on Task C (*Timeline-level Summary)*.

that both approaches performed comparably when their extracted evidence was incorporated into subsequent tasks. For Task C, as indicated in Table 4, none of the methods outperformed the baselines. This underscores a significant limitation of current LLMs in long-term, timeline-level summarization under standard few-shot prompting, pointing to a promising avenue in future research, such as incorporating post windows into the summarization process, as evidenced by the baseline results.

4 Conclusions and Future Work

We introduced a novel approach to modeling mental health dynamics from social media using hybrid LLMs, where the tasks explored include *Adaptive/Maladaptive Evidence Identification*, *Overall Well-being Rating*, and *Post-level and Timeline-level Summary*. Specifically, we first leveraged both classification-based and generation-based models with LLMs to identify adaptive and maladaptive evidence from individual posts. This evidence was then integrated to predict users' well-being scores and generate post-level and timeline-level summaries. Experimental results on the CLPsych 2025 shared task highlighted the effectiveness of our method, and the generative-based model demonstrated a significant advantage in evidence identification. In the future, we will dedicate on proposing more advanced models for generating timeline-level summaries, such as incorporating post windows into the summarization process.

Limitations

Our study has two primary limitations: (1) Due to time constraints, we evaluated our approach using two state-of-the-art LLMs, Qwen2.5-7B and Qwen2.5-32B, while more established models such as Llama3.1-8B/70B were not included in our experiments; (2) Our method, regardless of whether the evidence was obtained through classification-based or generation-based models, did not outperform the baseline models when generating timeline-level summaries. Future work could address this limitation, potentially by incorporating post windows into the summarization process, as evidenced by the baseline results.

Ethical Considerations

We discuss the ethical considerations and broader impact of this work here: (1) **Intellectual Property:** Our approach is applied to the CLPsych 2025 shared task, adhering to the data access form and ensuring compliance with data protection protocolsensuring responsible data handling practices. All illustrative examples, including those in figures and prompts, have been replaced by *lorem ipsum* to respect data confidentiality. (2) **Intended Use.** This approach is designed for research purposes focused on understanding mental health patterns over time through social media timelines. It includes identifying adaptive and maladaptive evidence, predicting overall well-being scores, and summarizing posts and timelines. (3) **Misuse Risks**. This method is not intended for processing sensitive, personal, or non-consensually obtained data. Furthermore, the output generated is inherently dependent on the input text and should not be used to support financial, political, or clinical decision-making without appropriate human oversight and ethical approval.

References

A American Psychiatric Association, American Psychiatric Association, et al. 1994. *Diagnostic and statistical manual of mental disorders: DSM-IV*, volume 4. American psychiatric association Washington, DC.

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, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Jenny Chim, Adam Tsakalidis, Dimitris Gkoumas, Dana Atzil-Slonim, Yaakov Ophir, Ayah Zirikly, Philip Resnik, and Maria Liakata. 2024. Overview of the CLPsych 2024 shared task: Leveraging large language models to identify evidence of suicidality risk in online posts. In *Proceedings of CLPsych*, pages 177–190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *Proceedings of ICLR*.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi han Sheu, Peilin Zhou, Lauren V. Moran, Sophia Ananiadou, Andrew Beam, and John Torous. 2024. Large language models in mental health care: a scoping review. *Preprint*, arXiv:2401.02984.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

Jiayuan Ma, Hongbin Na, Zimu Wang, Yining Hua, Yue Liu, Wei Wang, and Ling Chen. 2025. Detecting conversational mental manipulation with intent-aware prompting. In *Proceedings of COLING*, pages 9176–9183.

Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. *Preprint*, arXiv:2502.11095.

Hongbin Na, Tao Shen, Shumao Yu, and Ling Chen. 2024. Multi-session client-centered treatment outcome evaluation in psychotherapy. *Preprint*, arXiv:2410.05824.

Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2023. When does in-context learning fall short and why? a study on specification-heavy tasks. *Preprint*, arXiv:2311.08993.

Brenda WJH Penninx, Michael E Benros, Robyn S Klein, and Christiaan H Vinkers. 2022. How covid-19 shaped mental health: from infection to pandemic effects. *Nature medicine*, 28(10):2027–2037.

Lu Qian, Yuqi Wang, Zimu Wang, Haiyang Zhang, Wei Wang, Ting Yu, and Anh Nguyen. 2024. Domain-specific guided summarization for mental health posts. *Preprint*, arXiv:2411.01485.

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 CLPsych*, pages 25–36.

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

David Thulke, Yingbo Gao, Rricha Jalota, Christian Dugast, and Hermann Ney. 2024. Prompting and fine-tuning of small llms for length-controllable telephone call summarization. In *Proceedings of FLLM*, pages 305–312.

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 CLPsych*, pages 184–198.

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 ACL*, pages 4647–4660.

Talia Tseriotou, Jenny Chim, Ayal Klein, Aya Shamir, Guy Dvir, Ali Iqra, 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 CLPsych*.

Yuqi Wang, Zimu Wang, Nijia Han, Wei Wang, Qi Chen, Haiyang Zhang, Yushan Pan, and Anh Nguyen. 2024a. Knowledge distillation from monolingual to multilingual models for intelligent and interpretable multilingual emotion detection. In *Proceedings of WASSA*, pages 470–475.

Zimu Wang, Wei Wang, Qi Chen, Qiufeng Wang, and Anh Nguyen. 2024b. Generating valid and natural adversarial examples with large language models. In *Proceedings of CSCWD*, pages 1716–1721.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of SIGIR*, page 641–649.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of ICLR*.

A Prompts

Classification-based Evidence Identification

```
Determine whether the following sentence demonstrates an adaptive state (1), a maladaptive state (2), or
 neither (0).
{Post Sentence}
```

Generation-based Evidence Identification

```
Identify all sentences or phrases from the following text that demonstrate adaptive states.
{Post Text}
```

Overall Well-being Rating

```
Based on the following text and the provided evidence of adaptive and maladaptive states, assess the
author's mental well-being. Provide a score from 1 to 10, where 1 indicates extremely poor and 10
indicates excellent.
Text: {Post Text}
Adaptive Evidence: {Adaptive Evidence}
Maladaptive Evidence: {Maladaptive Evidence}
```

Post-level Summary

```
Summarize the following post, considering the evidence of adaptive and maladaptive states and the well-
being score.

Post:
{Post Text for In-context Example}
Adaptive Evidence:
{Adaptive Evidence for In-context Example}
Maladaptive Evidence:
{Maladaptive Evidence for In-context Example}
Well-being Score:
{Well-being Score for In-context Example}
Summary:
{Summary for In-context Example}

Post:
{Post Text}
Adaptive Evidence:
{Adaptive Evidence}
Maladaptive Evidence:
{Maladaptive Evidence}
Well-being Score:
{Well-being Score}
Summary:
```

Figure 3: Prompts designed for each tasks. The prompt for timeline-level summary closely follows the structure of the one for post-level summary but integrates the information of multiple posts, with the incorporation of information from multiple posts to generate a summary of the entire timeline.