Prompt Engineering for Capturing Dynamic Mental Health Self-States from Social Media Posts

Callum Chan[1], Sunveer Khunkhun[1],
Diana Inkpen[1], Juan Antonio Lossio-Ventura[2]
[1]School of Electrical Engineering and Computer Science, University of Ottawa, Canada
[2]Machine Learning Core, National Institute of Mental Health,
National Institutes of Health, USA
{cchan073, skhun073, dinkpen}@uottawa.ca, juan.lossio@nih.gov

Abstract

With the advent of modern Computational Linguistic techniques and the growing societal mental health crisis, we contribute to the field of Clinical Psychology by participating in the CLPsych 2025 shared task. This paper describes the methods and results obtained by the uOttawa team's submission (which included a researcher from the National Institutes of Health in the USA, in addition to three researchers from the University of Ottawa, Canada). The task consists of four subtasks focused on modeling longitudinal changes in social media users' mental states and generating accurate summaries of these dynamic self-states. Through prompt engineering of a modern large language model (Llama-3.3-70B-Instruct), the uOttawa team placed first, sixth, fifth, and second, respectively, for each subtask, amongst the other submissions. This work demonstrates the capacity of modern large language models to recognize nuances in the analysis of mental states and to generate summaries through carefully crafted prompting.

1 Introduction

Large Language Models (LLMs) have been explored in the mental health domain for tasks such as analyzing emotional states, sentiment, and depression (Xu et al., 2024a; Yang et al., 2024; Xin et al., 2024; Malgaroli et al., 2025). Their ability to process language at scale presents potential benefits for automating various processes in Clinical Psychology. Prompt engineering helps optimize LLM performance across domains (Liu et al., 2023), including biomedical and clinical applications (Hu et al., 2024; Sivarajkumar et al., 2024), where zero-shot and few-shot techniques have been investigated. These recent studies and our observations suggest that tailoring prompts to specific tasks can help improve performance in mental and clinical Natural Language Processing (NLP) tasks (Xu et al., 2024b), such as evaluating changes in individuals' mental well-being over time (Owen et al., 2023). Analyzing shifts in mental states over time provides valuable insights into overall mental health (Tsakalidis et al., 2022). This shared task builds on previous efforts by incorporating the generation of easily interpretable summaries, encouraging the Computational Linguistics community to further explore the dynamics of self-states. In this work, we use Llama-3.3-70B-Instruct (Grattafiori et al., 2024) to demonstrate how modern LLMs can identify textual indicators of adaptive and maladaptive self states, and generate summaries of social media posters' mental state as it changes over time. An adaptive self-state is characterized by aspects of an individual's mental state which facilitates the realization of basic needs and desires. Contrarily, a maladaptive self-state is characterized by aspects which impede the realization of basic needs and desires (Slonim, 2024).

2 Shared Task

The CLPsych 2025 shared task consists of four subtasks (Tseriotou et al., 2025). The first one (Task A.1) is to identify and extract textual evidence (spans) of adaptive and maladaptive self states of individual social media posts. The second (Task A.2) is to generate a well-being score from 1 to 10 of each post. The third (Task B) is to generate a summary for each post that describes the evidence contributing to the dominant self state, the non-dominant self state and the interplay between the two. The final task (Task C) is to generate a summary with similar requirements to Task B, but over a timeline of posts from a given user.

3 Dataset

The dataset was developed by the shared tasks organizers over the previous several years (Shing et al., 2018; Tsakalidis et al., 2022; Zirikly et al., 2019; Tseriotou et al., 2025). Thirty timelines of posts

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 256–267

were provided to the shared task participants this year to use as training data, totaling 343 posts in all. This training dataset is fully annotated with gold evidence of present self states, well-being scores, post-level self-state summaries, and timeline-level self-state summaries. Finally, ten additional timelines, totaling 94 posts in all, was provided to the participants as test data. These timelines were used in the evaluation of our proposed methods.

4 Methods

To address the four subtasks of the CLPsych 2025 shared task, we employed prompt engineering techniques on a local instance of the Llama-3.3-70B-Instruct model. For Task A.2, that expects numeric predictions, we also implemented linear and multinomial logistic regression classifiers for various types of embeddings.

4.1 Prompt Engineering Strategies

We employed three distinct prompting strategies in our official submissions: zero-shot prompts, structured prompts with one example, and structured prompts with multiple examples. For each strategy, separate prompt templates were tailored for each of the four tasks. Thus, each strategy consisted of four prompt templates (totaling 12 templates for our official submissions). The templates included the following features: task requirements, definition of terms such as self state, adaptive and maladaptive, and guidelines to format responses. The prompts for Task A.2 included explanations of how to characterize well-being. The prompts for Task C included the responses from Task B, in order to generate the post timeline summary. Table 1 shows the key features of each strategy. Appendix A.3 presents examples of the structured contextual one-shot prompt used for tasks A.1, A.2, B, and C.

4.2 Regression Models on Embeddings

Our additional approach for Task A.2 consisted of two stages, which we describe below.

Embeddings: Embeddings were generated using various transformer-based LLMs. In the first stage, each post was passed to these models to generate embeddings, which are numerical vectors of hidden dimension d. Each LLM tokenizes the post and converts each token into vectors based on the context. The final vector for each post is obtained by averaging the token vectors. We used the base and large variants of BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021), as well as MentalBERT (Ji et al., 2022) and MentalRoBERTa (Ji et al., 2022), which are BERT and RoBERTa models pretrained on additional mental health-related data. We also incorporated SBERT (Sentence BERT) (Reimers and Gurevych, 2019) and LLaMA-3.3-70B-Instruct. For comparison, we included traditional techniques like TF-IDF and Bag-of-Words representations. Each embedding had a different context length and dimension depending on the model used. For more details, see Table 3 in Appendix A.1.

Regression: We trained linear (LR) and multinomial logistic regression (MLR) models on the embedding vectors of all posts. The models were trained and evaluated using 5-fold cross-validation, with hyperparameters optimized through grid search. For MLR, data was stratified by well-being score, and the loss function was adjusted for class imbalance by weighting errors. For LR, the output was rounded to the nearest integer.

4.3 Evaluation Metrics

Performance of each submission was evaluated using task-specific metrics specified and applied by the shared task organizers (Tseriotou et al., 2025). For Task A.1, recall and weighted recall was computed using BERTScore (Zhang et al., 2020). Incorrectly predicted empty span lists received a score of 0. Additionally, separate recall scores were provided for adaptive only spans and maladaptive only spans. Task A.2 was evaluated using the Mean Squared Error (MSE) and Macro-F1 across all posts. Additional MSE scores were also provided per well-being severity class (serious: 1–4, impaired: 5–6, minimal: 7–10), with incorrect null predictions being penalized by the maximum error. Tasks B and C assessed summary quality using a Natural Language Inference (NLI) model to measure mean consistency (absence of contradiction) and maximum contradiction between submitted and gold summaries; incorrectly predicted null summaries defaulted to 0. Task C applied these metrics at the timeline level.

5 Results

The results of our submissions to the CLPsych 2025 shared task demonstrate the effectiveness of prompt engineering in leveraging LLMs for mental health analysis. The performance of our methods across

	Contextual Zero-Shot Prompting (uOttawa 1)	**Structured Contextual One-Shot Prompting** (uOttawa 2)	**Structured Contextual Few-Shot Prompting** (uOttawa 3)
Prompt Structure	Prompts were loosely structured, with less explicit delineation between sections such as objectives, definitions, and output guidelines. Domain-specific terms (e.g., adaptive and maladaptive self states) were included into the prompts to provide context.	Prompts were finely structured, including clear delineations for task objectives, definitions of key terms, output guidelines, and one annotated example extracted from the training data.	Prompts were structured similarly to the second approach, with delineations for task objectives, definitions, and output guidelines. However, this submission included multiple examples: seven examples for Tasks A.1, A.2, and B, and three examples for Task C.
Approach	This strategy relied on the LLM's ability to infer task requirements and generate responses without explicit examples. The prompts were designed to guide the model in identifying textual evidence of self states (Task A.1), generating well-being scores (Task A.2), and creating summaries for individual posts (Task B), and timelines (Task C).	This strategy was designed to improve the LLM's understanding of the task by providing a single example for each subtask. The example served as a reference for the model to better align its responses with the desired format and content for more accurate outputs.	By providing multiple examples, this strategy aimed to further refine the LLM's ability to generate accurate and contextually appropriate responses. The examples were carefully selected to cover a range of self states, post lengths and timeline lengths in an effort to limit its reliance on existing knowledge and to prepare the LLM for a diverse set of possible test data.

Table 1: Prompting Strategies: Key Features.

the four subtasks is presented in Table 2.[1]

We also included scores for a baseline method using zero-shot learning with simple prompts and a smaller model (Llama-3.1-8B-Instruct), to assess the effectiveness of using the larger model (70B). Additionally, we included the results we obtained for Task A.2 using prompt engineering and linear regression.

Task A.1: Identification of Self States The uOttawa team achieved strong results in identifying self states, with uOttawa_2 (one-shot prompting) performing the best among our submissions and among all the submitted runs by the shared task participants. It achieved an overall recall of 0.637 (adaptive: 0.594, maladaptive: 0.681) and a weighted recall of 0.498 (adaptive: 0.542, maladaptive: 0.455). The results for the uOttawa_2 and uOttawa_3 submissions were better than those of uOttawa_1, highlighting the importance of structured prompts for this task. Surprisingly, the few-shot learning did not outperform the one-shot learning.

Task A.2: Well-Being Score For this task, uOttawa_3 (few-shot prompting) achieved the lowest overall MSE of 2.62, with strong performance across impairment levels (minimal: 2.91, impaired: 4.03, serious: 2.28). This demonstrates the effectiveness of providing multiple examples for accurate well-being prediction. The lowest MSE achieved by participating teams was 1.920. We ran additional experiments for this task (S4, S5, S6, and more), with better results as shown in Table 2. With S4 and S5, we obtained the lowest MSE, even compared to the official results obtained by other teams. Also, a linear regression classifier using Llama-3.3-70B embeddings achieved a competitive MSE score of 2.015. See Appendix A.2 for more details on the additional experiments.

Task B: Post-Level Summaries In post-level summary generation, uOttawa_2 (one-shot prompting) achieved the highest mean consistency (0.860),

[1]Lower scores are better for the metrics MSE and Max Contradiction.

			Official Submissions			Additional Submissions		
Task	**Metric**	Baseline	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**
A.1	Recall ↑	0.405	0.469	**0.637**	<u>0.550</u>	-	-	-
	Weighted Recall ↑	0.214	0.386	**0.498**	<u>0.455</u>	-	-	-
A.2	MSE ↓	4.682	2.830	3.430	2.620	**1.673**	<u>1.693</u>	2.015
	Macro F1 ↑	0.286	0.355	**0.378**	0.302	<u>0.361</u>	<u>0.361</u>	0.348
B	Mean Consistency ↑	0.780	0.773	**0.860**	<u>0.859</u>	-	-	-
	Max Contradiction ↓	0.815	**0.756**	0.832	<u>0.804</u>	-	-	-
C	Mean Consistency ↑	0.897	<u>0.926</u>	0.918	**0.943**	-	-	-
	Max Contradiction ↓	<u>0.747</u>	0.794	0.751	**0.714**	-	-	-

Table 2: Performance of the uOttawa team across the four subtasks. The best scores are **bolded**, and the runners-up are <u>underlined</u>. The baseline was based on zero-shot prompts with Llama-3.1-8B-Instruct (small model). S1–S6 represent the file submissions we made for evaluation. S1, S2, and S3 correspond to uOttawa_1, uOttawa_2, and uOttawa_3, respectively. S4 and S5 were based on prompt engineering slightly different from the initial (S1, S2, and S3), with S4 containing 4 examples per class of well-being and S5 without examples. S6 was based on linear regression with Llama-3.3-70B-Instruct embeddings.

indicating coherent and consistent summaries. The few-shot strategy (uOttawa_3) also performed well, with a mean consistency of 0.859. The highest mean consistency achieved by a team was 0.910.

Task C: Timeline-Level Summaries For timeline-level summaries, uOttawa_3 (few-shot prompting) achieved the highest mean consistency (0.943) and the lowest max contradiction score (0.714), demonstrating its ability to capture longitudinal dynamics effectively. The highest mean consistency achieved by a team was 0.946, only a little higher than ours.

Overall Performance The uOttawa team placed first in Task A.1, sixth in Task A.2, fifth in Task B, and second in Task C in the shared task. These results showcase the potential of prompt engineering to enhance LLM performance in nuanced mental health analysis tasks. The one-shot and few-shot strategies proved effective in guiding the model.

6 Discussion

The results of team uOttawa's methods demonstrate competitive performance when applied to the CLPsych 2025 shared task.

For task A.1, the one-shot strategy outperformed zero-shot and few-shot approaches, achieving the highest recall (0.637) and weighted recall (0.498). The one-shot strategy's superior performance likely stems from how span identification benefits from precise, unambiguous guidance. A single well-chosen example appears sufficient to demonstrate what constitutes adaptive/maladaptive evidence, while avoiding the potential confusion that multiple examples might introduce. Few-shot prompts risk including borderline cases that could confuse the model's judgment, whereas zero-shot lacks any concrete reference points altogether. This suggests that for evidence extraction tasks, one carefully selected example may serve as an ideal template, providing just enough context to ensure consistent span detection without overcomplicating the prompt structure.

For task A.2, the few-shot strategy achieved the lowest MSE (2.62), demonstrating the value of multiple examples for fine-grained predictions. However, the relatively high MSE scores across all submissions suggest that well-being scoring remains challenging for strategies that do not rely on fine-tuning on training data. Therefore, we conducted additional experiments based on prompt engineering and regression models on embeddings (S4, S5, and S6 in Table 2), which improved MSE (1.673, 1.693, and 2.015 respectively, see appendix A.2), outperforming our initial results as well as those of the other participants. The relative success of few-shot prompting in numerical scoring could be explained by its ability to demonstrate the contextual nature of well-being assessments through multiple examples. By showing how similar phrases (e.g., "I'm exhausted") might receive different scores depending on surrounding context, the few-shot approach may help the model develop a more nuanced scoring rubric. However, the even stronger performance of regression models suggests an important limitation of prompting strategies for numerical tasks - they may ultimately be less effective than approaches that can directly learn statistical patterns from embeddings. This could indicate that

numerical scoring depends more on quantitative feature recognition than qualitative example-based learning.

For task B, the minimal difference between one-shot (0.860 MSE) and few-shot (0.859) approaches suggests diminishing returns for additional examples in summary generation. While zero-shot (0.773) trailed both approaches, its relatively strong performance indicates that the base model already possesses substantial summarization capability. The pattern reveals a hierarchy of effectiveness: one-shot learning provides optimal guidance for most cases (balancing structure and simplicity), few-shot learning offers slight contradiction reduction (0.804 vs. 0.832) for complex posts, and zero-shot learning serves as a competent but less reliable baseline. This implies that while a single example sufficiently anchors the task, the choice between approaches could prioritize consistency over edge-case handling or the other way around.

For task C, the few-shot strategy excelled, achieving the highest mean consistency (0.943) and lowest max contradiction score (0.714). Few-shot prompting's strong performance in timeline analysis may reflect the fundamentally different cognitive demands of longitudinal reasoning compared to single-instance tasks. The multiple examples likely help the model recognize various temporal patterns and transitional relationships that would be difficult to convey with just one example. This could explain why the additional context proves so valuable: it may allow the model to build a more comprehensive understanding of how mental states evolve over time, including recognizing triggers (such as job loss) and their typical emotional consequences. The one-shot approach's limitation here suggests that temporal reasoning may require exposure to multiple case examples to be effective.

Overall, our results demonstrate the effectiveness of prompt engineering in guiding LLMs for mental health analysis. The one-shot strategy excels in tasks requiring precise identification and summarization, while the few-shot strategy is better for nuanced tasks. The zero-shot strategy, while competitive, consistently underperformed, highlighting the importance of examples and structured guidance.

7 Clinical Applications

Our work presents powerful state-of-the-art methods to the greater clinical community. Not only do these approaches achieve impressive results, they are also very accessible and can be easily implemented by those with little background in machine learning or artificial intelligence. One such example is the self state monitoring of consenting high-risk social media users. Using users' post history and new posts, social media administrators could use these strategies to automatically flag high-risk users showing signs of degrading well-being and an increasing dominant maladaptive self state. Summaries generated by our methods can then be used to guide more personalized intervention strategies instead of generic responses (for example, offering specific tailored advice to manage stress instead of merely suggesting to contact a mental health hotline).

8 Conclusion and Future Work

We have showcased how using structured prompts with one or a few examples can lead to very good results when detecting and summarizing mental states from social posts.

Future work includes the continuation of development for the few shot learning approach by exploring different numbers of examples and the careful selection of the most relevant examples embedded within the prompts. We should also experiment with different types and various sizes of LLMs. These could also be further pre-trained or fine-tuned on data from the mental health domain. Finally, for Task A.2 that outputs a numeric score, we plan to model a sequence of decisions and that uses features extracted from previous posts at each step.

Limitations

Our experiments are limited to the type of social media data available for the shared task, focusing exclusively on English-language posts.

Additionally, we tested only one type of LLM, Llama, using a small version for the baseline and a larger version for the main method. Additionally, we experimented with several regression models, each using different text embeddings for Task A.2. We scored the posts of each user in sequence, but did not condition the prediction for the current post on the previous prediction. Such a strategy would

be useful to detect extended periods of a user exhibiting a dominant adaptive or maladaptive self state.

Furthermore, due to the limited time frame during which our submissions could be scored, we could not perform detailed ablation studies to analyze the specific aspects of our prompts which contributed to performance, nor explore alternative methods such as hyperparameter tuning of temperature or top_p.

Finally, biases introduced during the process of prompt engineering may skew the responses of the LLM and our results. One source of bias stems from our previous experiences and projects using the Llama model and influenced the way in which we structured our prompts and tuned them. Another source comes from the way we presented the definitions of adaptive and maladaptive self state in our prompts. This contextual information could distort the LLM's understanding of these terms and the task it is presented with.

Ethics

The data was collected from public sources and anonymized. However, it is still sensitive, since mental health status labels were assigned. For accessing the data, we signed the data sharing agreement for the shared task and complied with all the clauses therein. This ensures proper use of the data, solely for research purposes, as well as secure storage of the data. As requested by the shared task organizers, we did not use ChatGPT or other closed-source models that could use this data for further training or model refinement.

To preserve the privacy of the social media posters used in this shared task, this work should not be replicated using the data referenced throughout this paper. This work's contributions are merely the ideas it presents. While modern computation linguistic tools provide powerful means of mental health monitoring and assessment, we encourage the greater Artificial Intelligence community to take a measured approach when dealing with sensitive user data to ensure its privacy.

References

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, and 1 others. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

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.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Matteo Malgaroli, Katharina Schultebraucks, Keris Jan Myrick, Alexandre Andrade Loch, Laura Ospina-Pinillos, Tanzeem Choudhury, Roman Kotov, Munmun De Choudhury, and John Torous. 2025. Large language models for the mental health community: framework for translating code to care. *The Lancet Digital Health*.

David Owen, Dimosthenis Antypas, Athanasios Hassoulas, Antonio F Pardiñas, Luis Espinosa-Anke, Jose Camacho Collados, and 1 others. 2023. Enabling early health care intervention by detecting depression in users of web-based forums using language models: longitudinal analysis and evaluation. *JMIR AI*, 2(1):e41205.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: Algorithm development and validation study. *JMIR Medical Informatics*, 12.

Dana Atzil Slonim. 2024. Self-Other Dynamics (SOD): A Transtheoretical Coding Manual.

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*. Association for Computational Linguistics.

Alison W Xin, Dylan M Nielson, Karolin Rose Krause, Guilherme Fiorini, Nick Midgley, Francisco Pereira, and Juan Antonio Lossio-Ventura. 2024. Using large language models to detect outcomes in qualitative studies of adolescent depression. *Journal of the American Medical Informatics Association*, page ocae298.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024a. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024b. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

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*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

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 Appendix

A.1 Embeddings for Task A.2

Table 3 provides an overview of the context length and hidden dimension sizes for the LLMs used to generate embeddings for the additional experiments for Task A2, as well as the configurations for traditional techniques such as BoW and TF-IDF. The listed models, including BERT, SBERT, RoBERTa, and Llama, employ varying context lengths and dimensionalities, which are important factors for their performance in subsequent regression analysis.

Model	Context Length	Dimension
BERT-base	512	768
BERT-large	512	1024
RoBERTa	512	768
MentalBERT	512	768
MentalRoBERTa	512	768
SBERT	384	768
Llama-3.3-70B-Instruct	128K	8192
Bag-of-Words	N/A	3000
TF-IDF	N/A	3000

Table 3: Context length and hidden dimension sizes for the LLMs used to generate embeddings, along with traditional techniques like TF-IDF and Bag-of-Words.

A.2 Additional Results for Task A.2

We present results for additional experiments for Task A.2 in Tables 4 and 2. They were scored by the organizers after the shared task submission date. In our first additional round, we generated text embeddings with several methods (Bag-of-Words, TF-IDF, BERT, SBERT, and Llama 3.3 70B). Then, we trained multinomial logistic regression (MLR) classifiers to produce a class (1, 2, ... 10) and linear regression (LR) classifiers to output a numeric value that was rounded up to the closest integer between 1 and 10. The scores for this additional submission are presented in Table 4. We observe improved results compared to our official submissions (for S6).

Our second effort to improve our scores for Task A.2 was to revisit our strategy of prompt engineering. The prompt templates from our official submissions were further refined, included additional information and guidance for the LLM to respond with only one token. This additional submission included results from a 4-Shot and 0-Shot variant of this prompt template. These scores are presented in Table 2 as S4 and S5. This approach proved very effective and ranked the best amongst all teams' official submissions for Task A.2.

Model	Type	MSE ↓	MacroF1 ↑
BOW	MLR	3.844	0.267
	LR	4.216	0.167
TF-IDF	MLR	4.426	0.248
	LR	3.812	0.226
BERT	MLR	4.379	0.270
SBERT	MLR	4.649	0.250
	LR	<u>3.229</u>	<u>0.302</u>
Llama-3.3-70B-Instr	MLR	4.111	0.236
	LR	**2.015**	**0.348**

Table 4: Results for the regression methods for Task A.2. The "Model" column names the language model from which embeddings were extracted to train the regression model. The best scores are **bolded**, runners-up are <u>underlined</u>.

A.3 Examples of Structured Prompt

We present examples of structured prompt that we used for Task A.1, A.2, B, and C in the uOttawa_2 submission (one-shot learning). These are shown in tables 5, 6, 7 and 8 respectively. The example post, its evidence of adaptive and maladaptive self states, well-being score, and post summaries have been redacted to preserve the privacy of the training data.

Task:
Your 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 or maladaptive state. For each self-state (adaptive or maladaptive), the evidence is a set of continuous spans of text from the post.
—
Definitions:
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, Behaviour, and Cognition towards the self or others, which is conducive to the fulfillment of basic desires/needs (D), such as relatedness, autonomy and competence.
- A maladaptive self-state pertains to aspects of Affect, Behaviour, and Cognition towards the self or others, that hinder the fulfillment of basic desires/needs (D).

ABCD dimensions:
1. Affect (A): The type of emotion expressed by the person.
- Adaptive Examples: Calm/Laid back, Emotional Pain/Grieving, Content/Happy, Vigor/Energetic, Justifiable, Anger/Assertive Anger, Proud.
- Maladaptive Examples: Anxious/Tense/Fearful, Depressed/Despair/Hopeless, Mania, Apathetic/Don't care/Blunted, Angry (Aggressive, Disgust, Contempt), Ashamed/Guilty.

2. Behavior of the self with the Other (BO) : The person's main behavior(s) toward the other
- Adaptive Examples: Relating behavior, Autonomous behavior
- Maladaptive Examples: Fight or fight behavior, Overcontrolled/controlling behavior

3. Behavior toward the Self (BS): The person's main behavior(s) toward the self
- Adaptive Examples: Self-care behavior
- Maladaptive Examples: Self-harm, Neglect, Avoidance behavior

4. Cognition of the Other (CO): The person's main perceptions of the other
- Adaptive Examples: Perception of the other as related, Perception of the other as facilitating autonomy needs
- Maladaptive Examples: Perception of the other as detached or over attached, Perception of the other as blocking autonomy needs

5. Cognition of the Self (CS): How the person perceives themselves.
- Adaptive Examples: Self-acceptance and self-compassion
- Maladaptive Examples: Self-criticism

6. Desire (D): The person's main desire, need, intention, fear or expectation
- Adaptive Examples: Relatedness, Autonomy and adaptive control, Competence, Self-esteem, Self-care
- Maladaptive Examples: Expectation that relatedness need will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met
—
Guidelines for Output:
- Responses Section: Provide answers for both self-states under the headings '### Section Adaptive' and '### Section Maladaptive'.
- Each section should list adaptive and maladaptive self-states, respectively, with supporting text spans.
- Begin each extracted text span with a dash ('-').
- If no adaptive or maladaptive self-state is found, create both sections but leave them empty; do not include any dashes ('-').
- Finally, do not include any additional information; only the text spans are needed.
—
Example:

Input text:
Redacted example post.

Output text:

Section Adaptive
- Redacted adaptive evidence #1 from example post.
- Redacted adaptive evidence #2 from example post.

Section Maladaptive
- Redacted maladaptive evidence #1 from example post.
- Redacted maladaptive evidence #2 from example post.
—
Analyze the following input text based on the given criteria.

Input Text:
{INCLUDE_TEXT (*POST*) }

Output Text:

Table 5: One-Shot Structured Prompt Template for Task A.1.

Task:
Your task is to rate the overall well-being present in the post on a scale from 1 (low well-being) to 10 (high well-being). The score is based on GAF (American Psychiatric Association, 2000), and reflects how well an individual has been doing based on three key domains: Social functioning (school,friendships), occupational functioning (work) and an individual's overall psychological functioning. The clinical cutoff score is 6, meaning that individuals scoring below 6 may be experiencing significant distress
—
Definitions:
Here is an example of the wellbeing scale:
10: No symptoms and superior functioning in a wide range of activities
9: Absent or minimal symptoms (eg., mild anxiety before an exam), good functioning in all areas, interested and involved in a wide range of activities.
8: If symptoms are present, they are temporary and expected reactions to psychosocial stressors (eg., difficulty concentrating after family argument). Slight impairment in social, occupational or school functioning.
7: Mild symptoms (eg., depressed mood and mild insomnia) or some difficulty in social, occupational, or school functioning, but generally functioning well, has some meaningful interpersonal relationships.
6: Moderate symptoms (eg., panic attacks) or moderate difficulty in social, occupational or school functioning.
5: Serious symptoms (e.g., suicidal thoughts, severe compulsions) or serious impairment in social, occupational, or school functioning (eg., no friends, inability to keep a job).
4: Some impairment in reality testing or communication, or major impairment in multiple areas (withdrawal from social ties, inability to work, neglecting family, severe mood/thought impairment).
3: A person experiences delusions or hallucinations or serious impairment in communication or judgment or is unable to function in almost all areas (eg., no job, home, or friends).
2: In danger of hurting self or others (eg., suicide attempts; frequently violent; manic excitement) or may fail to maintain minimal personal hygiene or significant impairment in communication (e.g., incoherent or mute)
1: The person is in persistent danger of severely hurting self or others or persistent inability to maintain minimal personal hygiene or has attempted a serious suicidal act with a clear expectation of death.
0: Unable to assess well-being.
—
Guidelines for Output:
- Provide each answer as "Well-being score: <number>", where '<number>' is the well-being score you assign.
- If a well-being score cannot be provided, answer as "Well-being score: 0".
- Do not include any additional information or explanations - only the score is needed.
—
Example:

Input text:
Redacted example post.

Output text:

Well-being score: Redacted
—
Analyze the following input text based on the given criteria.

Input Text:
{INCLUDE_TEXT (*POST*) }

Output text:

Table 6: One-Shot Structured Prompt Template for Task A.2.

Task:
Your task is to summarize self-states for the social media post below. Specifically, generate a summary of the interplay between adaptive and maladaptive states identified in the post. Begin by determining which self-state is dominant (adaptive/maladaptive) and describe it first. For each self-state, identify the central organizing aspect (A, B, C, or D) that drives the state and structure the summary around it. Describe how this central aspect influences the rest, emphasizing potential causal relationships between them. Then, proceed to the second self-state and follow the same approach. If the post contains only one self-state (either adaptive or maladaptive), summarize only that state. Note that the summary does not need to explicitly highlight A, B, C, or D, but should aim to naturally integrate these elements into the description.
—
Definitions:
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, Behaviour, and Cognition towards the self or others, which is conducive to the fulfillment of basic desires/needs (D), such as relatedness, autonomy and competence.
- A maladaptive self-state pertains to aspects of Affect, Behaviour, and Cognition towards the self or others, that hinder the fulfillment of basic desires/needs (D).

ABCD dimensions:
1. Affect (A): The type of emotion expressed by the person.
- Adaptive Examples: Calm/Laid back, Emotional Pain/Grieving, Content/Happy, Vigor/Energetic, Justifiable, Anger/Assertive Anger, Proud.
- Maladaptive Examples: Anxious/Tense/Fearful, Depressed/Despair/Hopeless, Mania, Apathetic/Don't care/Blunted, Angry (Aggressive, Disgust, Contempt), Ashamed/Guilty.

2. Behavior of the self with the Other (BO) : The person's main behavior(s) toward the other
- Adaptive Examples: Relating behavior, Autonomous behavior
- Maladaptive Examples: Fight or fight behavior, Overcontrolled/controlling behavior

3. Behavior toward the Self (BS): The person's main behavior(s) toward the self
- Adaptive Examples: Self-care behavior
- Maladaptive Examples: Self-harm, Neglect, Avoidance behavior

4. Cognition of the Other (CO): The person's main perceptions of the other
- Adaptive Examples: Perception of the other as related, Perception of the other as facilitating autonomy needs
- Maladaptive Examples: Perception of the other as detached or over attached, Perception of the other as blocking autonomy needs

5. Cognition of the Self (CS): How the person perceives themselves.
- Adaptive Examples: Self-acceptance and self-compassion
- Maladaptive Examples: Self-criticism

6. Desire (D): The person's main desire, need, intention, fear or expectation
- Adaptive Examples: Relatedness, Autonomy and adaptive control, Competence, Self-esteem, Self-care
- Maladaptive Examples: Expectation that relatedness need will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met
—
Guidelines for Output:
- Response Section: Provide an answer under the headings '### Summary:'.
- Format the answer as a single paragraph, making it clear and consise.
- The summary should be no more than 6 sentences. - Ensure the summary captures the main points without unnecessary details.
—
Example:

Input text:
Redacted example post.

Output text:

Summary:
Redacted summary.
—
Analyze the following input text based on the given criteria.

Input Text:
{INCLUDE_TEXT (*POST*) }

Output Text:

Table 7: One-Shot Structured Prompt Template for Task B.

Task:
Your task is to summarize self-states for each timeline, given the summaries for each post on the timeline. Specifically, generate a summary focusing on the Interplay between adaptive and maladaptive self-states along the timeline. Emphasize temporal dynamics focusing on concepts such as flexibility, rigidity, improvement, and deterioration. If applicable, describe the extent to which the dominance of the self-states changes over time and how changes in aspects (Affect, Behavior, Cognition, and Desire) contribute to these transitions.
—
Definitions:
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, Behaviour, and Cognition towards the self or others, which is conducive to the fulfillment of basic desires/needs (D), such as relatedness, autonomy and competence.
- A maladaptive self-state pertains to aspects of Affect, Behaviour, and Cognition towards the self or others, that hinder the fulfillment of basic desires/needs (D).

ABCD dimensions:
1. Affect (A): The type of emotion expressed by the person.
- Adaptive Examples: Calm/Laid back, Emotional Pain/Grieving, Content/Happy, Vigor/Energetic, Justifiable, Anger/Assertive Anger, Proud.
- Maladaptive Examples: Anxious/Tense/Fearful, Depressed/Despair/Hopeless, Mania, Apathetic/Don't care/Blunted, Angry (Aggressive, Disgust, Contempt), Ashamed/Guilty.

2. Behavior of the self with the Other (BO) : The person's main behavior(s) toward the other
- Adaptive Examples: Relating behavior, Autonomous behavior
- Maladaptive Examples: Fight or fight behavior, Overcontrolled/controlling behavior

3. Behavior toward the Self (BS): The person's main behavior(s) toward the self
- Adaptive Examples: Self-care behavior
- Maladaptive Examples: Self-harm, Neglect, Avoidance behavior

4. Cognition of the Other (CO): The person's main perceptions of the other
- Adaptive Examples: Perception of the other as related, Perception of the other as facilitating autonomy needs
- Maladaptive Examples: Perception of the other as detached or over attached, Perception of the other as blocking autonomy needs

5. Cognition of the Self (CS): How the person perceives themselves.
- Adaptive Examples: Self-acceptance and self-compassion
- Maladaptive Examples: Self-criticism

6. Desire (D): The person's main desire, need, intention, fear or expectation
- Adaptive Examples: Relatedness, Autonomy and adaptive control, Competence, Self-esteem, Self-care
- Maladaptive Examples: Expectation that relatedness need will not be met, Expectation that autonomy needs will not be met, Expectation that competence needs will not be met
—
Guidelines for Output:
- Response Section: Provide an answer under the headings '### Timeline Summary:'.
- Format the answer as a single paragraph, making it clear and consise.
- The summary should be no more than 6 sentences. - Ensure the timeline summary captures the main points without unnecessary details.
—
Example:

Input text:
A chronologically ordered sequence of summarized posts from timeline

Output text:

Timeline Summary:
A timeline summary.
—
Analyze the following input text based on the given criteria.

Input Text:
{INCLUDE_TEXT (*Summaries of posts on specific timeline from Task B*) }

Output Text:

Table 8: One-Shot Structured Prompt Template for Task C.