Team ISM at CLPsych 2025: Capturing Mental Health Dynamics from Social Media Timelines using A Pretrained Large Language Model with In-Context Learning

Vu Tran
The Institute of Statistical Mathematics
Tokyo, Japan
vutran@ism.ac.jp

Tomoko Matsui
The Institute of Statistical Mathematics
Tokyo, Japan
tmatsui@ism.ac.jp

Abstract

We tackle the task by using a pretrained large language model (LLM) and in-context learning with template-based instructions to guide the LLM. To improve generation quality, we employ a two-step procedure: sampling and selection. For the sampling step, we randomly sample a subset of the provided training data for the context of LLM prompting. Next, for the selection step, we map the LLM generated outputs into a vector space and employ the Gaussian kernel density estimation to select the most likely output. The results show that the approach can achieve a certain degree of performance and there is still room for improvement.

1 Introduction

The CLPsych 2025 shared task (Tseriotou et al., 2025) combines longitudinal modeling in social media timelines with evidence generation (Chim et al., 2024), promoting the generation of humanly understandable rationales that support recognizing mental states as they dynamically change over time.

The task is structured around the MIND framework (Slonim, 2024), a pan-theoretical scheme for capturing self-states as combinations of Affect, Behavior, Cognition, and Desire (ABCD) components, and identifying mental fluctuations over time.

The shared task's provided dataset contains annotations of evidence aligned with the ABCD paradigm, well-being score and expert summaries at post-level and timeline-level (Shing et al., 2018; Zirikly et al., 2019; Tsakalidis et al., 2022).

Particularly, the shared task is organized into 4 tasks namely A.1, A.2, B, and C, focusing on different aspects of analyzing a given user's mental health state. Task A.1 focuses on extracting evidence of adaptive and maladaptive mental state from user posts. Task A.2 focuses on scoring the well-being of a user within the context of a given user post. Task B focuses on writing a summary of the user's mental health state within the context of a given user post. Task C focuses on writing a summary of the user's mental health state within the context of a given user timeline consisting of a series of posts.

We tackle the task by utilizing a pretrained large language model (LLM) and in-context learning (Dong et al., 2024) with template-based instructions to guide the LLM. Since we approach with a pretrained model without further fine-tuning and in-context learning is limited to the number of in-context examples, to improve generation quality, we employ a two-step procedure: sampling and selection. For the sampling step, we repeatedly randomly sample a subset of the provided training data for the context of LLM prompting. For the selection step, we map the LLM generated outputs into a vector space and employ the Gaussian kernel density estimation (Scott, 2015; Silverman, 2018) to select the most likely output. Details of our method is described in the next section.

2 Method

2.1 Overview

We design our framework consisting of an LLM and utilize in-context learning with a two-step procedure: sampling and selection.

Sampling We randomly sample a subset of the provided training data for the context of LLM prompting, and repeat for a number of rounds. We used meta-llama/Meta-Llama-3-8B-Instruct[1] as the LLM and set the sample size to 225. The temperature of LLM generation is set to 0.1.

Selection We map the LLM generated outputs into a vector space and employ the Gaussian kernel density estimation (Scott, 2015; Silverman, 2018) with the Scott's Rule for bandwidth selec-

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025), pages 287–291
May 3, 2025

```
You are a mental health expert and analyzing
a patient's social media post to determine
their well-being, their dominant self-state
of either adaptive or maladaptive.  The fol-
lowing is your past analysis.

Analysis 1:
<patient post contents>
Adaptive post segments:
* &lt;segment 1&gt;
* ...
Maladaptive post segments:
* &lt;segment 1&gt;
* ...
Well-being:  <well-being score>
Assessment:
<post summary>
...

Analysis i:  ...

Now analyze the following patent post.
<patient post>
Adaptive post segments:
<fill only post segments here, no analysis>
Maladaptive post segments:
<fill only post segments here, no analysis>
Well-being:  <give your score here>
Assessment:
<fill your assessment here>
```

Figure 1: Template for tasks A, B.

```
You are a mental health expert and analyzing
a patient's social media post to determine
their well-being, their dominant self-state
of either adaptive or maladaptive.  The fol-
lowing is your past analysis.
Past patient 1:
<patient post 1>
<patient post 2>
...
Final Assessment:

...

Past patient i:  ...

Now analyze the following patient.
<patient post 1>
<patient post 2>
...
Final Assessment:  <fill your assessment
here; it should be concise, and focus on
change of self-state in the beginning, middle,
and end of the post timeline; no need to men-
tion detailed post contents; must start with
Final Assessment:>
```

Figure 2: Template for tasks C.

tion (Turlach, 1993; Bashtannyk and Hyndman, 2001) to select the most likely output. We used sentence-transformers/all-MiniLM-L6-v2[2] as the sentence embedder model.

2.2 Tasks A & B

Since the evidence of adaptive and maladaptive states is the key for generating the summary of the given user post, we jointly tackle the two tasks A and B in one single flow. We design a prompting template (Figure 1) that instructs the LLM to extract evidence and summarize a given user post jointly. Specifically, we set the number of past analyses to 5, i.e. giving the LLM 5 past user posts with annotations as in-context learning examples.

After performing the sampling step, we collected a set of candidates for each post. We, then, proceed to the selection step. For each candidate, we map a triplet of ⟨adaptive-evidence, maladaptive-evidence, summary⟩ to a triplet of vectors ⟨vector(adaptive-evidence), vector(maladaptive-evidence), vector(summary)⟩. The concatenation of the 3 vectors in the triplet forms the representative vector of the candidate. The set of candidates' vectors are put through the Gaussian kernel density estimation, and the candidate whose vector has the highest density is selected as the final output for the given user post.

[2]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

2.3 Task C

Since a timeline may contain a lot of posts, and our resource is limited, even though we believe that the evidence and post-summary are valuable for making the timeline summary, we had to abandon the information and only use the timeline posts as the sole input. That leads to our designed prompting template shown in Figure 2. We set the number of past example timelines to 3. In our observation, a number of past timelines greater than 3 often resulted in junk responses, indicating that the selected LLM cannot handle such a long context.

The selection step is performed as described in Subsection 2.1, where each candidate is a summary generated.

3 Results

As shown in Table 1, our method achieved relatively good performance overall. Particularly, our system performs relatively better in evidence extraction than well-being scoring and summary generation.

For the results of Task A.1 (Table 2), our system, also similar to some other systems, did put more focus on extracting evidence related to maladaptive state than adaptive state. In one perspective, it is a sign that our system did put more alert on negative contents when doing analysis, which is understandable since many public LLMs, including the LLM used in this work, are aligned to recognize negative inputs for the purpose of safeguarding.

For the results of Task A.2 (Table 3), our system also did put more focus on problematic well-being

Team	**Task A1** Recall	**Task A2** MSE	**Task B** Mean Consistency	**Task C** Mean Consistency
Aquarius	0.507	2.010	0.880	0.915
BLUE	0.555	2.260	**0.910**	**0.946**
BULUSI	0.433	**1.920**	0.868	0.941
CIOL	0.246	3.990	0.612	0.610
CSIRO-LT	0.460	2.040	-	-
EAIonFlux	0.517	2.080	0.888	0.913
MMKA	0.602	6.610	-	-
NoviceTrio	-0.028	13.830	0.686	0.855
PsyMetric	0.168	3.230	0.698	0.926
ResBin	0.470	8.020	0.764	0.898
Seq2Psych	0.276	3.270	-	-
uOttawa	**0.637**	2.620	0.860	0.943
Zissou	0.579	3.140	0.846	-
ISM (ours)	0.561	2.760	0.859	0.852
our rank	4	7	6	9

Table 1: Official test results of participants.

	overall		**adaptive**		**maladaptive**	
Teams	Recall	Weighed Recall	Recall	Weighed Recall	Recall	Weighed Recall
Aquarius	0.507	0.456	0.499	0.465	0.516	0.446
BLUE	0.555	0.392	0.472	0.400	0.639	0.384
BULUSI	0.433	0.370	0.339	0.339	0.526	0.402
CIOL	0.246	0.174	0.230	0.151	0.262	0.198
CSIRO-LT	0.460	0.427	0.384	0.377	0.537	**0.478**
EAIonFlux	0.517	0.471	0.517	0.480	0.518	0.462
MMKA	0.602	0.343	0.522	0.374	0.681	0.313
NoviceTrio	-0.028	-0.028	-0.104	-0.104	0.047	0.047
PsyMetric	0.168	0.168	0.152	0.152	0.184	0.184
ResBin	0.470	0.302	0.258	0.255	0.682	0.350
Seq2Psych	0.276	0.236	0.245	0.238	0.308	0.235
uOttawa	**0.637**	**0.498**	**0.594**	**0.542**	0.681	0.455
Zissou	0.579	0.320	0.445	0.305	**0.713**	0.335
ISM (ours)	0.561	0.452	0.488	0.460	0.633	0.444
our rank	4	4	5	4	6	5

Table 2: Test results for task A.1.

Teams	MSE	MSE serious	MSE impaired	MSE minimal	F1 Macro
Aquarius	2.010	2.160	3.110	1.250	0.366
BLUE	2.260	**1.410**	3.690	2.060	**0.393**
BULUSI	**1.920**	3.040	1.190	**0.650**	0.351
CIOL	3.990	7.310	**0.490**	2.890	0.119
CSIRO-LT	2.040	1.820	3.680	1.080	0.344
EAIonFlux	2.080	1.770	3.710	2.110	0.321
MMKA	6.610	4.220	11.760	4.950	0.257
NoviceTrio	13.830	3.160	11.590	18.620	0.135
PsyMetric	3.230	2.520	6.630	3.280	0.300
ResBin	8.020	20.260	3.710	1.890	0.192
Seq2Psych	3.270	4.980	1.380	2.630	0.191
uOttawa	2.620	2.280	4.030	2.910	0.302
Zissou	3.140	2.910	4.320	3.090	0.344
ISM (ours)	2.760	1.930	5.000	2.740	0.319
our rank	7	4	11	8	8

Table 3: Test results for task A.2.

Teams	Mean Consistency	Max Contradiction
Aquarius	0.880	0.781
BLUE	**0.910**	**0.533**
BULUSI	0.868	0.805
CIOL	0.612	0.966
CSIRO-LT	-	-
EAIonFlux	0.888	0.782
MMKA	-	-
NoviceTrio	0.686	0.885
PsyMetric	0.698	0.563
ResBin	0.764	0.835
Seq2Psych	-	-
uOttawa	0.860	0.832
Zissou	0.846	0.772
ISM (ours)	0.859	0.777
our rank	6	4

Table 4: Test results for task B.

Teams	Mean Consistency	Max contradiction
Aquarius	0.915	0.876
BLUE	**0.946**	0.540
BULUSI	0.941	0.714
CIOL	0.610	1.000
CSIRO-LT	-	-
EAIonFlux	0.913	0.760
MMKA	-	-
NoviceTrio	0.855	0.596
PsyMetric	0.926	**0.354**
ResBin	0.898	0.816
Seq2Psych	-	-
uOttawa	0.943	0.714
Zissou	-	-
ISM (ours)	0.852	0.833
our rank	9	8

Table 5: Test results for task C.

state as can be seen that MSE serious is relatively better than other categories.

For the results of Tasks B, and C (Tables 4, and 5), our system can generate relatively good summaries highly consistent with the expert annotated summaries. However, max contradiction metric results show that our system added contradictory analysis in the output summaries, which raises the concern of hallucination, a critical problem often found with LLMs (Huang et al., 2025).

4 Conclusion

We have presented our approach for the task by using a pretrained large language model (LLM) and in-context learning with template-based instructions to guide the LLM and designing a two-step procedure, namely sampling and selection, to improve system response quality. We achieved promising results even though the method is simple and requires manageable resources for processing. There is still room for improvement in several directions including choosing stronger LLMs, or fine-tuning with domain knowledge.

Limitations

- No guarantee of adequate domain knowledge. The LLM used in this paper was pre-trained on data extracted from the open Web, which means the model is not guaranteed to be trained on high-quality professional data needed to understand the domain data in this task. Finetuning the model with high-quality professional data may improve the limitation.

- No guarantee of adequate domain context understanding. Though in-context learning is an effective method for guiding an LLM to deal with a new task, the LLM may not understand fully the context, especially since there is no guarantee of adequate domain knowledge in the pre-trained model.

Ethics Statement

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).

Acknowledgments

This work was supported by "Strategic Research Projects" grant from ROIS (Research Organization of Information and Systems), Japan and 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

David M Bashtannyk and Rob J Hyndman. 2001. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298.

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 the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 177–190.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

David W Scott. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

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.

Bernard W Silverman. 2018. *Density estimation for statistics and data analysis*. Routledge.

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, and 1 others. 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.

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.

Berwin A Turlach. 1993. Bandwidth selection in kernel density estimation: a rewiew. Technical report, Humboldt Universitaet Berlin.

Ayah Zirikly, Philip Resnik, Ozlem 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.