# Team ISM at CLPsych 2024: Extracting Evidence of Suicide Risk from Reddit Posts with Knowledge Self-Generation and Output Refinement using A Large Language Model

**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

This paper presents our approach to the CLPsych 2024 shared task: utilizing large language models (LLMs) for finding supporting evidence about an individual's suicide risk level in Reddit posts. Our framework is constructed around an LLM with knowledge self-generation and output refinement. The knowledge self-generation process produces task-related knowledge which is generated by the LLM and leads to accurate risk predictions. The output refinement process, later, with the selected best set of LLM-generated knowledge, refines the outputs by prompting the LLM repeatedly with different knowledge instances interchangeably. We achieved highly competitive results comparing to the top-performance participants with our official recall of 93.5%, recall–precision harmonic-mean of 92.3%, and mean consistency of 96.1%.

## 1 Introduction

In the unprecedented rapid evolution of large language models (LLMs), the ninth workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024) introduced the shared task of utilizing LLMs for finding supporting evidence about an individual's suicide risk level in Reddit posts (Chim et al., 2024). It is evident that recent work on LLMs suggest their potential applications on clinical tasks such as information extraction (Agrawal et al., 2022) and question answering (Singhal et al., 2023).

The CLPsych 2024 shared task uses the same Reddit dataset as the CLPsych 2019 shared task (Shing et al., 2018; Zirikly et al., 2019) which consisted of Reddit posts and annotated users' suicide risk labels at 4 levels: no risk, low risk, moderate risk, and high (severe) risk. The annotations were performed by both experts and crowd-source workers. The CLPsych 2024 shared task focuses on the expert annotations of users and posts on the subreddit 'r/SuicideWatch', where users are annotated with risk labels at 3 levels: low risk, moderate risk or high risk. From the user posts and annotations, the task's goal is to find supporting evidence in the form of post highlights and a summarized evidence given a user. It is important to note that the ground-truth evidence is not available for the task participants, so no direct optimization using ground-truth evidence is possible.

**Task Definition.** Formally, given a user $i$ who was assessed with either low risk, moderate risk, or high risk of committing suicide, a set of their Reddit posts $V_i$, and their suicide risk assessment $A_i$ in 3 risk levels (low risk, moderate risk, high risk), the goal is to identify the evidence supporting the assessment $A_i$ in the form of post highlights and a summary: $\{H_i, S_i\}$.

**Overview of Our Approach.** Our framework is constructed around an LLM with knowledge self-generation and output refinement. Inspired by MedPrompt (Nori et al., 2023) where using LLM-generated knowledge of solving a task as a part of task prompting can help boost the performance of such task, we ask the LLM to respond with its knowledge of how to make suicide risk assessment and find supporting evidence from social media posts. The best set of LLM-generated knowledge is selected based on the prediction accuracy. Now we have output candidates from potential knowledge leading to accurate predictions. We design an output refinement process to aggregate and refine the output candidates to obtain the final output. In the refinement process, the LLM is prompted with interchanging knowledge on the same input repeatedly, so when finished, we obtained an output with more agreement among different knowledge instances. Similar to our idea of output refinement, Madaan et al. (2023) proposed a self-refinement process where the output is put back to the same LLM, albeit the same conversation, to get feedback

*Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 191–196
March 21, 2024 ©2024 Association for Computational Linguistics

and refined output. Our approach, instead of going in-depth with one conversation, collectively refines the output under various input knowledge.

## 2 Method

### 2.1 Framework

Our framework is constructed around an LLM and consists of two stages: 1) knowledge self-generation, and 2) output refinement. In stage 1 - knowledge self-generation, we ask the LLM to generate its knowledge of how to handle the task and use the LLM's generated responses as a part of the inputs for finding evidence, i.e. extracting highlights and generating summaries, and making suicide risk predictions. We, then, find the best set of the LLM's generated knowledge responses leading to accurate predictions. In stage 2 - output refinement, with the best set of knowledge, outputs are aggregated and refined by repeatedly prompting the LLM with each instance in the best set of the generated knowledge until no further change is observed. Our framework is illustrated in Figure 1.

**Stage 1.** We sample responses from the LLM to a knowledge prompt with our pre-defined knowledge generation instructions $I^g$ asking for the LLM's understanding of the task:

$$K^l = g_{\text{LLM}}(I^g) \tag{1}$$

and obtain a set of generated knowledge responses to be used as knowledge inputs: $\{K^l\}$.

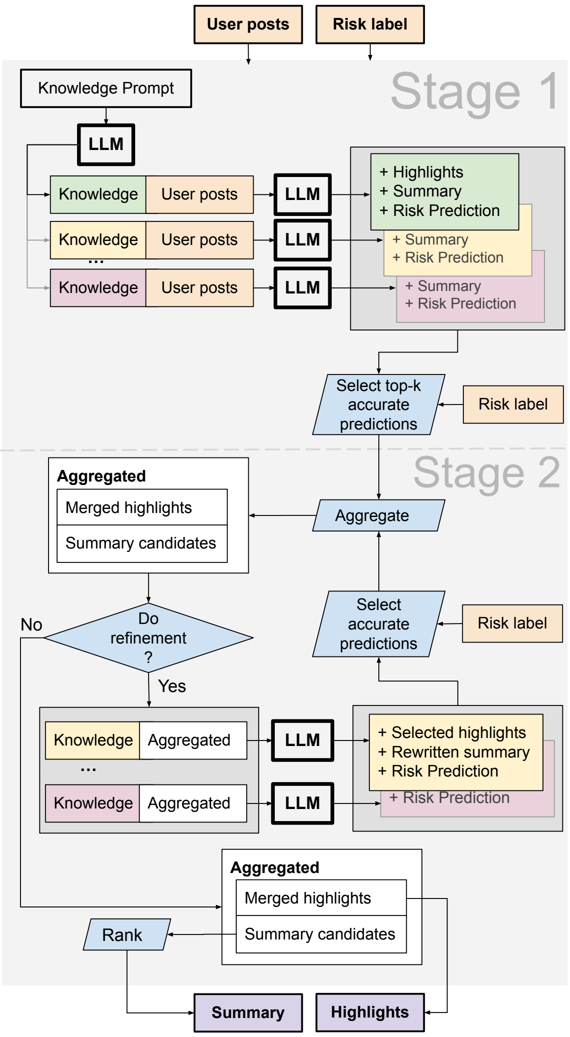
For each user $i$, we prompt the LLM with knowledge input $K^l$, the user's posts $V_i$, and our pre-defined instructions $I^h$ for extracting highlights $\hat{H}_i^l$, generating a summary $\hat{S}_i^l$ and making a risk-level prediction $\hat{A}_i^l$ as following:

$$\{\hat{H}_i^l, \hat{S}_i^l, \hat{A}_i^l\} = h_{\text{LLM}}(K^l, V_i, I^h) \tag{2}$$

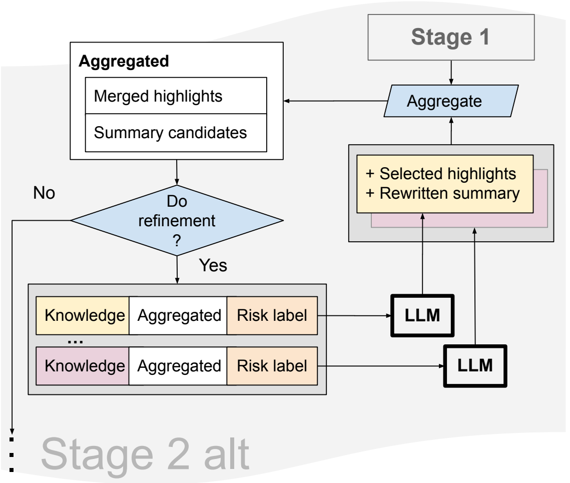
After that, for each user $i$, we select a subset of knowledge inputs $\{K^l\}$ leading to top-$k$ accurate predictions with the following scoring:

$$\text{score}_i(K^l) = (1_{\hat{A}_i^l = A_i} + \varepsilon) \times \sum_j 1_{\hat{A}_j^l = A_j} \tag{3}$$

where $\varepsilon$ is a very small positive number to avoid zero-scoring. It means that given a user, the selected knowledge inputs yield accurate predictions for that specific user and overall high accuracy for all users[1].



(a) Not using risk labels in stage 2 prompts.



(b) Using risk labels in stage 2 prompts.

Figure 1: Our framework with and without using ground-truth assessments, i.e. risk labels, in stage 2 prompts.

---

[1] Tie-breaks are decided by the earlier time order.

**Stage 2.** Instead of the user posts $V_i$ as in stage 1, we input the extracted highlights and generated summaries $\bigcup_l \{\hat{H}_i^l, \hat{S}_i^l\}$ aggregated over all $\{K^l\}$ selected in stage 1, and our pre-defined instructions $I^f$ to select new highlights, generate a new summary and make a new risk-level prediction. The procedure is repeated with newly extracted/generated highlights/summaries as inputs for further refinement as following:

$$\{\hat{H}_i^l, \hat{S}_i^l, \hat{A}_i^l\}_t = f_{\text{LLM}}(K^l, \bigcup_m \{H_i^m, S_i^m\}_{t-1}, I^f) \quad (4)$$

After each step $t$, we filter out knowledge inputs with inaccurate predictions $\hat{A}_i^l \neq A_i$.

We also investigate an alternative procedure of the refinement process. In addition to the highlights and summaries, we input the ground-truth risk assessment instead of predicting it as following:

$$\{\hat{H}_i^l, \hat{S}_i^l\}_t = f'_{\text{LLM}}(K^l, \bigcup_m \{\hat{H}_i^m, \hat{S}_i^m\}_{t-1}, A_i, I^{f'}) \quad (5)$$

where our pre-defined instructions $I^{f'}$ are for additionally using ground-truth risk assessment to select new highlights and generate a new summary.

We repeat the refinement process until $\bigcup_m \{\hat{H}_i^m\}_t = \bigcup_m \{\hat{H}_i^m\}_{t-1}$ for a maximum number of iterations $T$. Let $\tau \leq T$ be the stopped iteration, the final set of highlights is $\bigcup_m \{\hat{H}_i^m\}_\tau$, and the final summary is selected from the summary candidates $\{\hat{S}_i^m\}_\tau$ using TextRank (Mihalcea and Tarau, 2004) with BERTScore-F1 (Zhang et al., 2019) for measuring summary-pair similarity. As the results, the final set of highlights can be seen as a stable extraction across different knowledge inputs, and the final summary can be seen as the best summary over plausible summaries.

## 2.2 Experimental Settings

We used the LLM named Mixtral[2] with the specific version Mixtral-8x7B-Instruct-v0.1[3] trained to follow instructions. We used the original model weights and didn't further train the LLM. For efficient utilization of the LLM, we used the Huggingface transformers library[4] and loaded the LLM with 4-bit quantization (Dettmers et al., 2023). The temperature of the LLM is set to 1 for prompting knowledge (Equation 1) and is set to 0 for finding evidence and making predictions (Equations 2, 4, and 5). We set top-$k = 3$ for stage 1. For the final summary ranking in stage 2, we computed BERTScore-F1 using roberta-large (Liu et al., 2019).

In stage 1, to obtain the set of knowledge responses, we prompted the LLM with instructions $I^g$ as following:

1. *"Suppose you are a mental health care professional, describe in details steps to assess suicide risk of a person by reading their public posts on social media."*
2. *"Response"* from LLM, which describes abstractly about professional suicide risk analysis.
3. *"According to that, what are the cues to look for where the assessment is one of the 3 levels of risk (low, moderate, and high). Explain the cues for each of the level."*
4. *"Response"* from LLM to use as the generated knowledge response.

In addition to the generated knowledge responses, we also use manual provided by the shared task organizer, "aspects to consider" specifically, as an instruction of the aspects to focus on during finding evidence.

Other instructions are:

- $I^h$: *"Firstly, do a step-by-step analysis of the user posts. Secondly, give a list of extracted text spans from the TITLE and the BODY, which serve as evidence for your assessment. Thirdly, give a summary of the evidence in less than 100 words. Finally, give your assessment in just one of the three options: low risk, moderate risk, or high risk."*
- $I^f$: *"Firstly, select all important highlights linked to the suicide risk level. Secondly, in less than 100 words, write a summary given the selected highlights and the above summary candidates. Finally, give your assessment in just one of the three options: low risk, moderate risk, or high risk. "*
- $I^{f'}$: *"The mentioned user has been assessed with a suicide risk level of* ***{risk-level}****. Firstly,*

---

[2] https://mistral.ai/news/mixtral-of-experts/

[3] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

[4] https://huggingface.co/

*select all important highlights linked to the suicide risk level. Secondly, in less than 100 words, write a summary given the selected highlights and the above summary candidates.”* Where **{risk-level}** is filled with the user’s suicide risk level annotated.

For the final submission, we submitted 3 runs with the following options:

- Run 1: Skipping the refinement process. The highlights/summaries from stage 1 are merged/ranked to obtain the final outputs.
- Run 2: Not using ground-truth risk labels as input of the refinement process (Equation 4).
- Run 3: Using ground-truth risk labels as input of the refinement process (Equation 5).

## 3 Results & Discussions

The results are obtained with the metrics briefly described below.

- Highlights: recall, precision, and recall–precision harmonic-mean. Recall weighted by length-ratio of gold highlights vs. submitted highlights is also reported.
- Summary: mean consistency and max contradiction between submitted summary (hypothesis) and gold summary (premise) using a natural language inference model.

Please refer to the organizer’s paper (Chim et al., 2024) for the details of the evaluation metrics.

As the official results shown in Table 1, we achieved a recall of 0.935 (4th rank), a harmonic-mean of 0.923 (3rd rank) and a mean consistency of 0.961 (8th rank). The top results are pretty close with the best recall of 0.944, the best harmonic-mean of 0.929 and the best mean consistency of 0.979.

In stage 1, we acquired 320 knowledge responses, averaging 55.4% accuracy in risk prediction, with a peak accuracy of 72.8%. An example of the responses is shown in Appendix A.

In stage 2, we observed convergence with stopped iteration $\tau$ not going over 5 and having an average of 1.4 for run 2 and 1.9 for run 3.

The refinement process helps reduce the length of extracted highlights from 53% to 32–33%, which leads to a big improvement of weighted recall despite the cost of lower recall (Table 2). The process is, however, shown to reduce mean consistency and increase max contradiction. The problem could be because of the lack of context when using only highlights to generate a summary.

| Team Name | Rec | HM | MeC |
|---|---|---|---|
| CSIRO (baseline) | .919 | .917 | - |
| DONUTS Colaboratory | .872 | .907 | .942 |
| INF@UoS | .850 | .896 | .934 |
| LAMA | .577 | .888 | .964 |
| MHNLP | .910 | .909 | .873 |
| SBC | .907 | .909 | .976 |
| SCALAR-NITK | .886 | .889 | .901 |
| SKKU-DSAIL | .922 | .917 | .970 |
| sophiaADS | **.944** | .924 | .944 |
| SWELL | .915 | .903 | .973 |
| UniBuc Archaeology | .939 | .914 | .973 |
| UoS NLP | .943 | **.929** | .966 |
| UZH_CLyp | .910 | .913 | **.979** |
| Xinhai | .887 | .911 | .958 |
| ISM (Ours) | .935 | .923 | .961 |
| Our ranking | 4 | 3 | 8 |

Table 1: Official results. Rec: recall, HM: recall–precision harmonic-mean, MeC: mean consistency.

| Run | Rec | Prec | WR | HM | MeC | MaC |
|---|---|---|---|---|---|---|
| 1 | **.935** | .911 | .564 | **.923** | **.961** | **.125** |
| 2 | .910 | **.918** | .715 | .913 | .952 | .145 |
| 3 | .904 | .917 | **.744** | .910 | .957 | .127 |

Table 2: Our results for different runs. Rec: recall, Prec: precision, WR: weighted recall, HM: recall–precision harmonic-mean, MeC: mean consistency, MaC: maximum contradiction (lower is better).

Although we achieved high (>90%) recall, precision, and consistency in finding evidence, risk level classification accuracy is at most 72.8%, a recognizable discrepancy between finding evidence and matching it with a correct risk level.

## 4 Conclusion

In the concept of the CLPsych 2024 shared task, we have constructed a framework for extracting evidence of suicide risk from Reddit posts with knowledge-self generation and output refinement using an LLM. We achieved competitive results among the top participants. Our future work needs to focus on improving output refinement, and tackling consistency problems in matching supporting evidence with risk prediction.

## Limitations

- No guarantee of adequate domain understanding. Mixtral, 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.
- No guarantee of a strong consistency between finding evidence and making predictions. Our framework relies on the prediction accuracy measurement as a critical part of our evidence finding mechanism. However, our framework has no mechanism for checking the consistency between the prediction and the evidence found.

## 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).

## Acknowledgements

The authors are particularly grateful to the anonymous users of Reddit whose data feature in this year's shared task dataset, to the crowdsource annotators and to the clinical experts from BarIlan University for the data annotation, to the American Association of Suicidology for their assistance in making the dataset available, to the CLPsych organizers for holding this meaningful shared task, to the reviewers for their constructive comments. 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

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 Ninth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine.

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.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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 Self-Generated Knowledge Responses

Below is an illustration of knowledge responses in stage 1. Potentially sensitive contents are omitted.

```
  Sure, I'd be happy to explain the cues
to look for when assessing the level of
suicide risk (low, moderate, or high)
based on a social media post.  Here are
some possible cues to look for at each
level:
  1.  Low Risk:
  Cues to look for at this level may
include:
  * Expressions of vague or fleeting
suicidal thoughts, such as "I feel [OMITTED]"
or "I don't know [OMITTED]"
  * Indirect references to suicide, such
as "I wish [OMITTED]" or "I'm tired
[OMITTED]"
  * Ambivalent or inconsistent statements
about wanting to live or die.
  * Statements that suggest a willingness
to seek help or talk to someone
about their feelings.
  * A support system that is actively
engaged and responsive.
  2.  Moderate Risk:
  Cues to look for at this level may
include:
  * Specific plans or intentions to commit
suicide, such as "I've been thinking
[OMITTED]" or "I've been looking for
[OMITTED]"
  * Preparations for suicide, such
as giving away prized possessions or
putting affairs in order.
  * A history of previous suicide attempts
or self-harm.
  * Access to means of suicide, such as
firearms or medications.
  * A lack of social support or a sense
of isolation.
  * Statements that suggest a belief
that there is no other way out or no
hope for the future.
  3.  High Risk:
  Cues to look for at this level may
include:
  * A specific and imminent plan to
commit suicide, such as "I have a [OMITTED]"
or "I'm going to [OMITTED]"
  * Recent suicide attempts or self-harm
that required medical attention.
  * Active suicidal ideation with a
sense of calm or determination.
  * A lack of impulse control or a sense
of hopelessness or desperation.
  * Substance abuse or intoxication
that increases the risk of impulsive or
dangerous behavior.
  * A lack of social support or a sense
of burdensomeness on others.
  It is important to note that these
cues are not exhaustive and may vary
depending on the individual and their
unique circumstances.  If you are concerned
about an individual's suicide
risk, it is important to involve trained
professionals in the assessment and
intervention process.
```

This knowledge response led to the best risk prediction accuracy of 72.8%.

## B Organizer Manual: Aspects to Consider

In addition to the generated knowledge responses, we also use manual provided by the shared task organizer. Specifically, we input the contents of "aspects to consider" to the LLM as an instruction of what to focus on during finding supporting evidence. A part of the "aspects to consider" is:

1. `Emotions:  How does the individual feel?  ...`
2. `Cognitions:  What are the individual's thoughts and perceptions about suicide?  ...`
3. `Behaviour and Motivation:  What are the individual's acts or behavior related to suicide?  ...`
4. `Interpersonal and social support: Does the individual have social support/stable relationships?  ...`
5. `Metal health-related issues:  Consider psychiatric diagnoses associated with suicide ...`
6. `Context/additional risk factors: ...  socioeconomic and demographic factors ...`