Exploring Instructive Prompts for Large Language Models in the Extraction of Evidence for Supporting Assigned Suicidal Risk Levels

Jiyu Chen[1] and **Vincent Nguyen**[1] and **Xiang Dai**[1]
and **Cécile Paris**[1,2] and **Sarvnaz Karimi**[1]
CSIRO's Data61[1]
Sydney, Australia
firstname.lastname@csiro.au

Diego Mollá[2,1]
Macquarie University[2]
Sydney, Australia
diego.molla-aliod@mq.edu.au

Abstract

Monitoring and predicting the expression of suicidal risk in individuals' social media posts is a central focus in clinical NLP. Yet, existing approaches frequently lack a crucial explainability component necessary for extracting evidence related to an individual's mental health state. We describe the CSIRO Data61 team's evidence extraction system submitted to the CLPsych 2024 shared task. The task aims to investigate the zero-shot capabilities of open-source LLM in extracting evidence regarding an individual's assigned suicide risk level from social media discourse. The results are assessed against ground truth evidence annotated by psychological experts, with an achieved recall-oriented BERTScore of 0.919. Our findings suggest that LLMs showcase strong feasibility in the extraction of information supporting the evaluation of suicidal risk in social media discourse. Opportunities for refinement exist, notably in crafting concise and effective instructions to guide the extraction process.

1 Introduction

The intersection between NLP and mental health research has provided valuable insights, uncovering the diagnostic potential inherent in language (Agrawal et al., 2022; Singhal et al., 2023). Previous research has primarily concentrated on static classifications of individuals' social media posts, with studies, for example, focusing on predicting the level of suicide risk within social media posts (O'dea et al., 2015; Shing et al., 2018; Zirikly et al., 2019) and tracking changes in emotion over time (Paris et al., 2015; Larsen et al., 2015; Tsakalidis et al., 2022b,a). Despite these advancements, the increasing reliance on computational models in mental health assessments unveils a prominent gap — the lack of an essential explainability component. This absence is critical for the nuanced extraction of evidence that explains an individual's mental health state. This deficiency assumes significance in supporting practitioners' decision-making as they navigate the intricacies of mental health diagnostics.

In response to this problem, a shared task is organised as part of the CLPsych 2024 workshop (Chim et al., 2024). In our participation, we investigate the application of an open-source Large Language Model (LLM), namely Llama-2 (Touvron et al., 2023) within a zero-shot learning framework. The principal objective is to systematically extract text spans that can be treated as evidence of an individual's assigned suicide risk level from their social media posts. Beyond the mere evaluation of LLM viability, we assume a proactive stance, aiming to formulate instructive prompts that guide the model in extracting accurate and semantically rich evidence. We use a sub-sample of the University of Maryland Reddit Suicidality Dataset, Version 2, which includes 125 randomly selected Reddit users and their $r/SuicideWatch$ posts (Shing et al., 2018; Zirikly et al., 2019), provided by the task organisers. The suicide risk levels of these users are annotated by psychologists.

The robustness and validity of our findings are ensured through evaluation against ground truth evidence annotated by domain experts, employing BERTScore (Zhang et al., 2020). Overall, we found instructing LLM with factor-oriented and risk-level-specific prompts achieved the best recall-oriented BERTScore of 0.919 among our experimented approaches.

2 Dataset

A sub-sample of 125 users and their posts on the $r/SuicideWatch$ subreddit was selected from the University of Maryland Reddit Suicidality Dataset (UMD Subset) (Shing et al., 2018; Zirikly et al., 2019). Each user in the subset creates 1.3 posts on average, with a maximum of three posts.

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 197–202

Psychology experts conducted annotations of the suicidal risk level for each user, classifying them as low, moderate, or high (or severe) risk through a comprehensive review of all posts associated with a particular user. Note that the risk level annotation is performed at the user level rather than the post level. Specifically, each user receives an annotation based on the highest risk level expressed throughout their entire collection of posts. To provide clarity, in instances where a user conveys high-risk suicidal thoughts in an initial post followed by expressing low risk in a subsequent post, the user's annotation reflects the highest risk level.

We utilise the provided UMD Subset, consisting of 125 users, to investigate the application of LLMs for evidence extraction using zero-shot learning. For development, we randomly select nine users from the broader UMD dataset, where the suicide risk levels are annotated through crowd-sourcing, focusing on posts from $r/SuicideWatch$. This ensures their distinction from the 125 users in the provided UMD Subset.

3 Method

We design three approaches: (1) a baseline, (2) a factor-oriented, and (3) a risk level & factor-oriented approach. Each method varies in the design of the prompt in the zero-shot learning setting.

Baseline

The baseline employs a basic prompt (Listing 1) to instruct the LLM in extracting evidence supporting the annotation of a specific user's expression of suicidal thoughts. It is important to note that we do not explicitly indicate the risk level associated with users in specific posts. Two special linguistic markers are utilised in the pre-training stage of Llama-2 (Touvron et al., 2023). These linguistic markers, [INST] and <<SYS>>, are added during zero-shot learning to indicate the structure of the prompt. The [INST] token marks the boundary of the prompt instruction, while the <<SYS>> token marks the boundary of the system message used for setting the context for LLM.

Listing 1: Basic Prompt Template

```
[INST] <<SYS>> Here is a post containing
 suicidal ideation:{{post content}} <<
SYS>> Extract phrases as evidence that
support the suicide risk[/INST]
```

Factor-oriented Instruction

The factor-oriented approach depends on more instructive prompts, carefully designed with instructions that explicitly address the consideration of diverse psychological and socioeconomic factors when evaluating the risks of suicide. A study (Corbitt-Hall et al., 2016; Jones et al., 2003) indicates that humans tend to classify a post as having a high level of risk if it includes explicit expressions of self-harm, prolonged severe depression, and a lack of support from family or friends. Conversely, a user is less likely to have suicide risk if the post minimally contains overly dramatic complaints. A set of risk factors formed the foundation for creating these guidelines. We synthesised our compilation of risk factors and crafted prompts to instruct the LLM in extracting evidence related to specific factors. Table 1 shows the synthesised factors and their indication of suicide risk. Subsequently, We crafted a factor-oriented prompt (Listing 2) instructing the LLM in extracting evidence that supports each risk factor.

Listing 2: Factor-oriented Prompt Template

```
[INST] <<SYS>> Here is a post written
    by an individual: {{post content
    }} <</SYS>> Extract phrases if
    they covers any of the following
    aspects:

1. signs of fear, anger, or sadness
2. expression of thoughts or
    intention in self-harm or suicide
3. expression of difficulties in
    handling stress
4. expression of lacking support or
    connection from families or
    friends
5. diagnosis of chronic psychiatric
    disease, such as schizophrenia,
    bipolar, anxiety, eating
    disorder
6. signs of seeking public attention
[/INST]
```

Risk Level & Factor-oriented Instruction

Identifying evidence specific to various risk levels might present a challenge for LLMs. Hence, in the design of the baseline and factor-oriented approach, we did not explicitly specify the risk level associated with users in certain posts. Consequently, any text spans, irrespective of the expression of the risk level, will be extracted as evidence. To address this limitation, we propose a new approach that focuses on extracting evidence directly aligned with annotated risk levels in users' posts, providing a concise

Risk factors	Explanation
Emotion	Individual's emotional state, encompassing feelings such as fear, anger, or intense psychological distress.
Cognition	Individual's expression of the intention, the severity, and the frequency of self-harm or suicide thoughts.
Behavior	Individuals access to means or proposal of concrete plans to commit suicide
Motivation	The triggering events of individual's suicidal thoughts
Support	The unstable relationship and lack of support
Mental	The psychiatric diagnosis associated suicide risk, such as schizophrenia, bipolar, severe anxiety, or eating disorder
Environment	Exposure to suicide behaviour by others

Table 1: A collection of risk factors referred for the design of instructive factor-oriented prompt.

perspective for practitioners. To achieve this, we developed three prompt variations to guide the LLM. Specifically, our instruction emphasises extracting evidence indicative of acute situations that demand immediate interventions for users annotated with high risk. We incorporated selected risk factors to formulate risk level & factor-oriented prompts. For the formulation of risk factors, we referred to a previous study (Corbitt-Hall et al., 2016), in which researchers engaged college students in identifying socio-economic factors linked to various levels of suicide risk.

Additionally, we established rules for choosing one of the three prompts based on the associated risk level. To illustrate the distinctions in prompt design for guiding evidence extraction concerning low and high risk, we present the covered risk factors in Listing 3.

Post-processing

We employ a set of Backus-Naur form grammars (Listing 4), which is the standard mechanism, to regulate the output of Llama, directing it to generate only the extracted content from the original text, without including descriptions or explanations. We observed that Llama can automatically correct spelling errors within the original text and may slightly rephrase the content. For instance, it rectifies "beleive" to its correct form "believe" or omits certain words, such as "just" in the extracted evidence of phrases like "I just feel so trapped". Nevertheless, the occurrences of auto-correction or rephrasing are intermittent and unpredictable, posing challenges in making strategies to revert the modified extracted text back to its original form. We propose a solution by instructing Llama to extract only concise phrases as evidence. In post-processing, we discard any extraction that does not match the content of the original post, ignoring capitalization.

Listing 3: Risk level & Factor-oriented Prompt Template

```
#low risk:
Extract phrases if they cover one or more of the
    following aspects:
1. expression of difficulties in handling stress
2. expression of lacking support or connection
   from families or friends
3. expression of emotion
4. action of overly dramatic reaction
5. seeking attentions
6. exposure to other people who commit suicide


#high risk:
Extract phrases if they cover one or more of the
    following aspects:
1. expression of self-harm or suicide plans
2. expression of serious warnings
3. calling for help
4. expression of emotional states, especially
   depression, anger, and fear
5. diagnosis of mental disorders, such as
   schizophrenia, bipolar, anxiety, eating
   disorder
6. expression of taking medicines or
   prescriptions for psychiatric treatment
```

Experiments

LLM	llama-2-70b-chat.Q4_0.gguf [1]
GPU	NVIDIA RTX 3500 Ada
context size	4096
batch size	4096
temperature	0

Table 2: The key environment setting and parameters for running the experiment.

[1] https://huggingface.co/TheBloke/Llama-2-70B-Chat-GGUF

We utilise Llama-2-70B-Chat (Touvron et al., 2023) as our LLM for the task and implement it using the Llama C++ framework [2] and 4-bits quantisation. A detailed parameters and hardware settings for running Llama is shown in Table 2.

4 Evaluation Metrics

The evaluation was conducted by the shared task organisers using BERTScore (Zhang et al., 2020). Assume G is a set of 4 gold highlights $G = \{g_1, g_2, g_3, g_4\}$ and H is a set of 2 submitted highlights $H = \{h_1, h_2\}$. Then, the evaluation metrics are:

- *Recall*: For a given user, take the average of the maximum BERTScore from each g_* to each h_*.
- *Precision*: For a given user, find the g_* with the maximum BERTScore to each each h_*, and then take the average over H.
- weighted-*Recall*: For a given user, sum the token count (tokenised by Zhuang et al.) of G as $len(G)$ and of H as $len(H)$. Weigh the user-level *Recall* by the $\frac{len(G)}{len(H)}$, if $len(H) > len(G)$.

The overall submission-level score is the mean across all test users.

5 Results

Table 3 demonstrates that the risk & factor-oriented (RF-oriented) approach is the most effective in extracting evidence associated with all three levels of pre-annotated risks when measured under recall-oriented BERTScore ($+0.015$ to baseline and $+0.007$ to factor-oriented approach). Specifically, we observed that the RF-oriented approach notably facilitates the extraction of evidence for user annotations with low risks (0.924). The extraction of this risk level presents a greater challenge, as the scores for high-risk tend to be higher than those for medium and low risks. This discrepancy is likely attributed to the fact that posts with lower risk levels tend to employ lexicons that express suicidal ideation less explicitly. In contrast, posts with a high risk level may explicitly include contents like "I cannot stop thinking of kill myself" or "I want to commit suicide". Shifting to precision-oriented BERTScore, its deficiency compared to the Baseline is minimal (0.01) and remains consistent with the factor-oriented approach, showcasing its robust nature. Nevertheless, the RF-oriented approach extracts longer context as evidence to support low-risk annotations (0.504 in weighted-*Recall*). Consequently, it yields worse weighted recall than the baseline and factor-oriented approach.

The baseline demonstrated excellent *Precision* (0.918) in extracting evidence. This observation suggests that while the LLM may not comprehensively grasp the causative factors for evaluating suicide risk levels in context, and may fail to cover all aspects, it has embedded enough knowledge to accurately identify relevant context. It also achieved the best weighted-*Recall* of 0.740 among the experimented approaches, indicating its extraction length is closer to the human annotation compared to the instruction that explicitly covers the risk factors as guidance.

Upon comparing the RF-oriented approach to the factor-oriented approach, we noticed that refining instructions to the LLM for conciseness led to improved performance ($+0.007$ in *Recall*; $+0.002$ in *Precision*; $+0.022$ in weighted-*Recall*) in evidence extraction. Specifically, when excluding the extraction of evidence for low-risk annotations, the RF-oriented approach, with instructions tailored for different risk levels, demonstrated the ability to extract shorter context and achieved better weighted-*Recall*.

6 Conclusions

We investigated three approaches with varying levels of instruction detail to guide LLMs in extracting evidence related to users exhibiting low, moderate, or high suicide risk levels. All approaches demonstrated strong effectiveness, with the baseline excelling in precision for shorter text pieces. However, the factor-oriented and RF-oriented approaches, equipped with detailed instructions covering diverse mental health factors tailored to different risk levels, proved more effective in capturing comprehensive evidence, with the RF-oriented approach performing the best. Our findings highlight the robust feasibility of LLMs in extracting information supporting the evaluation of suicidal risk in social media discourse. There is room for improvement by creating clear and effective instructions to steer the extraction process. This could involve adapting existing manual annotation guidelines for evidence extraction into instructive prompts. Ad-

[2] https://github.com/ggerganov/llama.cpp

	Recall			*Precision*			weighted-*Recall*		
	low	moderate	high	low	moderate	high	low	moderate	high
Median		0.910			0.906			0.617	
Baseline		0.904			**0.918**			**0.740**	
	0.910	0.904	0.902	0.900	0.904	0.903	0.686	0.753	0.736
Factor-oriented		0.912			0.915			0.679	
	0.906	0.910	0.918	0.899	0.904	0.919	0.602	0.680	0.708
RF-oriented		**0.919**			0.917			0.701	
	0.924	0.919	0.920	0.899	0.917	0.923	0.504	0.721	0.737

Table 3: Results of baseline, Factor-oriented, and RF-oriented (Risk Level & Factor) approaches on submission level. The median denotes the score of the 8-th ranked participant in the shared task from the total of 15 participants. The median by risk level is not disclosed by the task organisers. The top row of each cell denotes the overall submission-level score across all three risk levels, with the greatest value presented in bold; The bottom row of each cell denotes the overall submission-level score by risk level, with the greatest value marked by underline.

dressing the auto-correction behaviour of the generative LLM is crucial for further improving *Recall*. The model's generative settings occasionally auto-correct spelling errors or rephrase extracted text, posing challenges in recovering the originally expressed content and impacting the fidelity of evidence. This unpredictability introduces complexities in formulating strategies to revert the modified text to its original form, adding an additional layer of intricacy to the evidence extraction process.

In future, we will conduct a more comprehensive qualitative analysis. We aim to refine the instructional prompts given to the model, adapting existing manual annotation guidelines to ensure clearer and more effective guidance. We will explore the integration of contextual information, aiming to enhance the model's ability to capture broader situational cues for improved risk assessment. Addressing the auto-correction behavior, especially in terms of spelling errors, will be a priority, involving fine-tuning the model or implementing post-processing steps to preserve the original expressions in extracted text.

Limitations

The effectiveness of our approach heavily relies on the performance of the leveraged LLM in accurately processing mental health information. We noticed that when changing the Llama-2-70B model to Llama-2-7B, many text spans with the expression of evidence failed to be extracted.

Another limitation is the comprehensibility of the instructive prompts provided to the LLM. The design of prompts plays a crucial role in guiding the model's behaviour. However, achieving optimal prompt design is a challenging task, and variations in prompt comprehension could influence the accuracy and relevance of evidence extraction. We have noticed that slightly changing the order of the covered risk factors in the prompt may lead to a varied output. Due to the time constraints associated with this shared task, and the lack of labelled development and test data at the time of submission, we could not thoroughly analyze the impact of variations in the prompt text.

Besides, the extraction granularity cannot be systematically controlled. For some posts, the model tends to extract full sentences as evidence, while others may only extract single keywords. This inconsistency in extraction granularity poses challenges in achieving consistent and precise evidence granularity, requiring further exploration.

Lastly, our approach is based on zero-shot learning. This inherently limits the real-time adaptability of the model to evolving patterns in user behaviour or language expression. More advanced approaches, such as in-context learning, could be explored in the future.

Ethics Consideration

We affirm that the data utilised in this study is not shared with any external entities, including cloud services, third-party organizations, or companies. All data processing is conducted within our organization, ensuring a secure and protected environment. Our commitment includes presenting findings and insights responsibly, and avoiding potential harm. This involves careful interpretation of results and avoiding stigmatization based on extracted information.

Acknowledgments

Acknowledging the assistance of the American Association of Suicidology in making the UMD dataset available.

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.

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.

Darcy J Corbitt-Hall, Jami M Gauthier, Margaret T Davis, and Tracy K Witte. 2016. College students' responses to suicidal content on social networking sites: An examination using a simulated facebook newsfeed. *Suicide and Life-Threatening Behavior*, 46(5):609–624.

Jana E Jones, Bruce P Hermann, John J Barry, Frank G Gilliam, Andres M Kanner, and Kimford J Meador. 2003. Rates and risk factors for suicide, suicidal ideation, and suicide attempts in chronic epilepsy. *Epilepsy & Behavior*, 4:31–38.

Mark E Larsen, Tjeerd W Boonstra, Philip J Batterham, Bridianne O'Dea, Cecile Paris, and Helen Christensen. 2015. We feel: mapping emotion on twitter. *IEEE journal of biomedical and health informatics*, 19(4):1246–1252.

Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on twitter. *Internet Interventions*, 2(2):183–188.

Cécile Paris, Helen Christensen, Philip Batterham, and Bridianne O'Dea. 2015. Exploring emotions in social media. In *2015 IEEE Conference on Collaboration and Internet Computing (CIC)*, pages 54–61. IEEE.

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.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adam Tsakalidis, Jenny Chim, Iman Munire Bilal, Ayah Zirikly, Dana Atzil-Slonim, Federico Nanni, Philip Resnik, Manas Gaur, Kaushik Roy, Becky Inkster, et al. 2022a. Overview of the clpsych 2022 shared task: Capturing moments of change in longitudinal user posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 184–198.

Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. 2022b. Identifying moments of change from longitudinal user text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4647–4660. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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. Chinese Information Processing Society of China.

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.

A Appendix

Listing 4: Backus-Naur Form Grammars for Post-processing

```
root ::= Post
Post ::= "{" ws "\"highlights\":" ws
    stringlist "}"
Postlist ::= "[]" | "[" ws Post ("," 
    ws Post)* "]"
string ::= "\"" ([^"][^\t]*)␣"\""
boolean ::= "true" | "false"
ws ::= "\n\t"
number ::= [0-9]+ "."? [0-9]*
stringlist ::= "[" ws "]" | "[" ws
    string ("," ws string)* ws "]"
numberlist ::= "[" ws "]" | "[" ws
    string ("," ws number)* ws "]"
```