A Dual-Prompting for Interpretable Mental Health Language Models

Hyolim Jeon[1,2*], **Dongje Yoo**[3*], **Daeun Lee**[1],
Sejung Son[1], **Seungbae Kim**[4], **Jinyoung Han**[1,2†]
[1]Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, South Korea
[2]Department of Human-AI Interaction, Sungkyunkwan University, Seoul, South Korea
[3]Department of Computer Engineering, Chung-Ang University, Seoul, South Korea
[4]Computer Science & Engineering Department, University of South Florida, Tampa, FL, USA
{gyfla1512, maze0717}@g.skku.edu {delee12, jinyounghan}@skku.edu
pass120cau.ac.kr, seungbae@usf.edu

Abstract

Despite the increasing demand for AI-based mental health monitoring tools, their practical utility for clinicians is limited by the lack of interpretability. The CLPsych 2024 Shared Task[1] aims to enhance the interpretability of Large Language Models (LLMs), particularly in mental health analysis, by providing evidence of suicidality through linguistic content. We propose a dual-prompting approach: (i) Knowledge-aware evidence extraction by leveraging the expert identity and a suicide dictionary with a mental health-specific LLM; and (ii) Evidence summarization by employing an LLM-based consistency evaluator. Comprehensive experiments demonstrate the effectiveness of combining domain-specific information, revealing performance improvements and the approach's potential to aid clinicians in assessing mental state progression.

1 Introduction

The global healthcare system faces significant challenges from mental health conditions such as depression and suicidal ideation (Darrudi et al., 2022), emphasizing the need for an advanced monitoring system for early intervention (Galea et al., 2020).

In response, NLP researchers have paid attention to identifying mental states, often leveraging social media data (Chen et al., 2023; Liu et al., 2023; Lee et al., 2023). Notably, the most recent development involves the application of Large Language Models (LLMs), which have demonstrated robust capabilities in general language processing in mental health analysis (Yang et al., 2023a,b; Xu et al., 2023b). Specifically, Amin et al. (2023) conducted a comparison of ChatGPT's zero-shot capability in identifying suicide and depression, contrasting it with previous methods that relied on previous Pre-trained Language Models (PLMs). Furthermore, Lamichhane (2023) evaluated ChatGPT's effectiveness in recognizing stress, depression, and suicide, emphasizing its strong grasp of language in texts related to mental health.

However, these studies have focused on identifying mental health status through a black box model, posing a challenge in interpreting the rationale behind their outcomes (Schoene et al., 2023; Zhang et al., 2022). Accordingly, efforts have been made to enhance the interpretability of mental health analysis, such as guiding LLMs to emphasize emotional cues (Yang et al., 2023a) and developing open-source LLMs by training them with data from mental health-related social media (Yang et al., 2023b). Nevertheless, a lack of reliability still remains; recent LLMs are often unreliable or inconsistent (Agrawal et al., 2022), potentially due to the lack of mental health-related knowledge (Yang et al., 2023a). This problem has significantly delayed the practical use of LLMs in clinical settings (Malhotra and Jindal, 2024).

To address this issue, the CLPsych 2024 Shared Task (Chim et al., 2024) introduces the challenge of utilizing open-source LLMs to enhance their interpretability in mental health analysis, specifically focusing on detecting suicidality through linguistic content in social media data. Particularly, the shared task includes two subtasks: (i) *Task A* requires finding key phrases from each post to support suicide risk, and (ii) *Task B* aims to provide a summary of evidence related to the user's suicide risk across multiple posts.

In this paper, we design an enhanced prompt for the extraction task (Task A) by assigning an expert identity, enabling LLMs to function as an expected agent (Xu et al., 2023a), and leveraging a suicide dictionary (Lee et al., 2022) to capture suicide-related context. Here, we utilize mental health-specific LLM, MentaLLaMA (Yang et al., 2023a). For the summarization task (Task B), we employ a consistency evaluator (Luo et al., 2023) to

[*] Equal contribution.
[†] Corresponding author.
[1] https://clpsych.org/shared-task-2024/

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 247–255

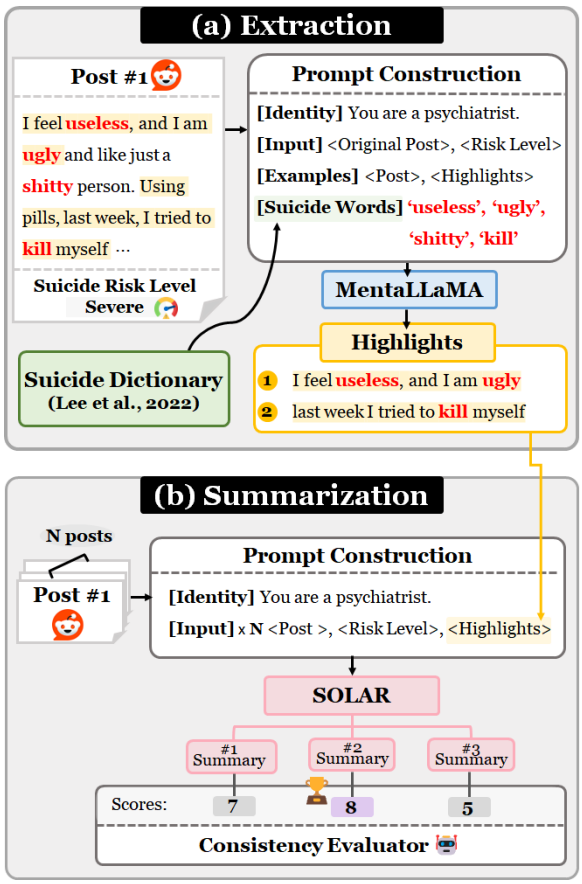

Figure 1: The overall architecture of the proposed approach: (a) Knowledge-aware Evidence Extraction (§2.1) and (b) Evidence Summarization with LLM-based Consistency Evaluator (§2.2)

improve the consistency of outcomes with multiple summaries.

The extensive experiments illustrate that combining domain-specific information with few-shot learning enhances the extraction of evidence, resulting in an improvement in recall from 91.0% to 92.2%. Additionally, our findings indicate that an LLM trained with general datasets is more effective in mitigating hallucination in summarization tasks than a domain-specific LLM. We believe our approach can support clinicians in assessing mental state progression.

2 Methodology

Our aim is (i) to extract evidence supporting the user's suicide risk from each post and (ii) to summarize all the evidence across multiple posts. To this end, we design two prompting strategies to instruct LLMs for trustworthy reasoning in mental health analysis. These strategies include Knowledge-aware Evidence Extraction (§2.1) and Evidence Summarization with an LLM-based Consistency Evaluator (§2.2). The overall proposed approach is depicted in Figure 1, and the full text of each prompt is available in Appendix A.

Table 1: Example of suicide words (Lee et al., 2022).

Suicide Risk	# of Words	Examples
Low	48	emptiness, overthink
Moderate	83	psychiatric, pain
Severe	111	cutting, die

2.1 Knowledge-aware Evidence Extraction

As shown in Figure 1(a), the prompt for the extraction task includes the original post and the assigned user's suicide risk level. For few-shot learning, we incorporate examples that are not included in the evaluation dataset. Moreover, we apply the three prompting strategies to address the unreliability issue of LLMs arising from the lack of mental health-related knowledge (Yang et al., 2023a).

A. Mental health-specific LLM. In order to tackle the zero-shot challenge in LLMs (Han et al., 2023), we utilize MentaLLaMA-chat-13B (Yang et al., 2023b), fine-tuned with 105K mental health-related social media data, demonstrating its efficacy in mental health-related tasks.

B. Assigning expert identity. As LLMs tend to provide insight into their cognitive processes when assigning predefined roles (Li et al., 2023; Xu et al., 2023a), we employ prompts to allocate the domain expert identity (e.g., *'You are a psychiatrist'*).

C. Utilizing a suicide dictionary. Since a domain-specific dictionary can aid LLMs in capturing relevant context (Yang et al., 2023a), we utilize a suicide dictionary (Lee et al., 2022), which has proven effective in identifying suicidal ideation on social media data. As shown in Table 1, the dictionary uses the UMD Reddit Suicidality Dataset (Shing et al., 2018), comprising 279 words validated by domain experts. If the given post includes words from the suicide dictionary (Lee et al., 2022), the model identifies and incorporates these words into our prompt, instructing the LLM to consider these words attentively.

2.2 Evidence Summarization with LLM-based Consistency Evaluator

As shown in Figure 1(b), the prompt for the summarization incorporates multiple posts and the assigned user's suicide risk level. Additionally, an expert identity is assigned, similar to the previous step. However, despite the advancements in LLMs,

Table 2: Statistics of the evaluation dataset.

	Highlights	Summarization
Suicide Risk	**# posts (avg. # length)**	**# users (avg. # posts)**
Low	17 (1,149)	13 (1.31)
Moderate	91 (1,132)	75 (1.21)
Severe	54 (1,178)	37 (1.46)
Total	**162 posts**	**125 users**

Table 3: Comparison of performance between zero-shot and few-shot learning for the extraction task (Task A) using the evaluation dataset.

Model	**Highlights (Task A)**		
	Pre.↑	Rec.↑	F1↑
Ours w/ Zero-shot	**0.913**	0.910	0.911
Ours w/ Few-shot	0.912	**0.922**	**0.917**

hallucination and inconsistency still remain significant concerns (Tang et al., 2023a,b). To mitigate this issue, we apply the two following strategies that can enhance consistency.

A. Extract-then-Generate. Zhang et al. (2023) demonstrated the effectiveness of prompts that incorporate an extractive summary for abstractive summarization. Following this approach, the proposed prompt integrates the extracted phrases obtained from the preceding step (§2.1). The full text of each prompt is available in Appendix A.2.1.

B. Consistency Evaluator. We adopt a consistency evaluator proposed by Luo et al. (2023). Initially, multiple candidate answers are generated through the LLM. We then compute consistency scores (ranging from 1 to 10) for each candidate, assessing the extent to which the generated summary aligns with the original posts, utilizing the consistency evaluator. In the end, the answer with the highest score from multiple candidates is selected as the final result. Here, we adopt SOLAR (Kim et al., 2023) as the summarizer and evaluator, known for its recent outstanding performance[2]. Further details comparing summarizer and evaluator are provided in §4.2.

3 Experiments

3.1 Evaluation Dataset

The CLPsych 2024 shared task (Chim et al., 2024) provides the UMD Reddit Suicidality Dataset (Zirikly et al., 2019; Shing et al., 2018), consisting of 79,569 posts from 37,083 subreddits by 866 Reddit users who posted on r/SuicideWatch between 2008 and 2015. Each user in the dataset is assigned a label that indicates the severity of suicidality (i.e., No, Low, Moderate, or Severe), determined by crowdsourcers and domain experts.

The evaluation dataset comprises a subset of users labeled with *Low*, *Moderate*, and *Severe* risks validated by domain experts. It includes 162 posts distributed among 125 users and the statistics of the dataset are summarized in Table 2.

[2] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

3.2 Experimental Settings

All experiments are conducted on a GeForce RTX 3090 Ti GPU with 26GB of memory. To minimize the memory cost of 16-bit weights, we employ the bitsandbytes library (Dettmers et al., 2022a), converting them to int8 using vector-wise quantization (Dettmers et al., 2022b) without significant quality loss. Each prompt is processed independently to mitigate the impact of dialogue history.

3.3 Evaluation Metrics

Note that the ground truth dataset was not provided to the participants. Therefore, all the evaluation metrics and reported results are supplied by the organizers of the CLPsych 2024 Shared Task.

(1) Similarity: BERTScore (Zhang et al., 2019) is employed for the extraction task to assess token similarity using contextual embeddings.

(2) Consistency: For the summarization task, a natural language inference (NLI) model (Laurer et al., 2024) is applied to assess the consistency of individual sentences in the provided evidence summary. Specifically, the contradiction scores are calculated between the predicted outcomes and each ground truth summary sentence. The resulting sentence-level consistency score is then determined as 1 minus the probability of the contradiction prediction.

4 Results & Analysis

To demonstrate the effectiveness of the proposed method, we compare its performance with various approaches and conduct the case study where our proposed approach performs better. Note that due to the absence of ground truth from the organizer, quantitative analysis was limited, leading us to focus on qualitative analysis instead. Additionally, we manually paraphrase any examples from the data to preserve user anonymity.

4.1 Analysis on Knowledge-aware Evidence Extraction

Table 3 shows the results of our approach on the evaluation dataset, with precision, recall, and F1 scores, for the highlights task (Task A).

Analysis on few-shot learning. We find an improvement in recall from 91.0% to 92.2% by integrating few-shot learning. This suggests the importance of providing examples for few-shot learning in domain-specific tasks, particularly in clinical settings (Han et al., 2023).

Analysis on suicide-dictionary. We find integrating a suicide dictionary (Lee et al., 2022) also improves domain knowledge in extracting evidence. Specifically, it allows thorough consideration of suicide risk factors that might be overlooked due to their general meaning, such as *'family'* and *'credit'*, which have been validated by domain experts as suicide-related words. Examples of the results are provided below.

> ***Response w/ Suicide Dictionary:*** ["Fear of failing.","Fear of hurting."], ["working as of credit problems."], ["Don't want to be a burden or face my friends and family."]

Analysis on expert identity. We explore the performance of the LLM by employing different expert identities, such as psychology, counseling, and psychiatry. This analysis aims to understand how the model's behavior varies depending on the assigned role. For example, when the role is assigned as a psychologist, the LLM tends to prioritize the user's negative self-perception (e.g., *'ugly'* and *'hate'*) to a greater extent. Conversely, adopting the identity of a counselor enables the model to focus on the relationship (e.g., *'broke up'* and *'divorce'*), which may contribute to feelings of isolation. Additionally, we observe that assigning a psychiatrist role is likely to focus on clinical markers, such as emotional distress (e.g., *'anxiety'*) and history of abuse (e.g., *'assaulted'*), which can be connected to suicidal ideation. Hence, we suggest that selecting an appropriate identity aligned with the research objective can offer valuable insights.

> ***Response w/ Psychology Identity:*** ["I am ugly, I am annoying, I am unwanted"], ["I hate me"]
>
> ***Response w/ Counselor Identity :*** ["Fuck, we broke up three weeks ago"], ["disconnected from everybody"]
>
> ***Response w/ Psychiatry Identity:*** ["I will never go to school because of my depression."], ["I am feeling anxious/angry and constantly lonely"], ["When I was 4 years old, I was sexually abused"]

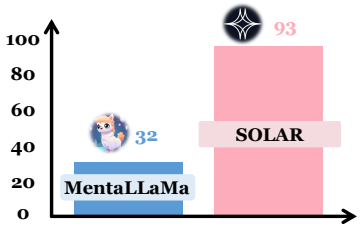

Figure 2: Winner count comparison for MentaLLaMA (Yang et al., 2023b) and SOLAR (Kim et al., 2023) in 125 evaluation dataset using evaluator.

4.2 Analysis on Evidence Summarization with LLM-based Consistency Evaluator

Analysis on Extract-then-Generate. We explore the efficacy of incorporating extractive summaries from Task A for the evidence summarization task. We observe that the hallucination issue frequently arises when extractive summaries are absent. This indicates that our approach enhances consistency by providing contextual information (Zhang et al., 2023). For a better understanding, we provide an example below. We notice that the LLM misinterprets the expression *'wishing to do it'* as a desire for success, resulting in generating *'self-distrust in achievements'* by the LLM.

> ***Posts:*** I was thinking about when I tried to hang myself, wishing to do it now.
> ***Response w/o Extract-then-Generate:*** They exhibit risk due to cognitions (self-distrust in achievements).

Comparison LLMs with Consistency Evaluator. Table 4 shows the results of our approach on the evaluation dataset, along with the mean consistency scores for the summarization task (Task B). We find that using only SOLAR (Kim et al., 2023) as a summarizer performed better than using both SOLAR (Kim et al., 2023) and MentaLLaMA (Yang et al., 2023b). This also can be found in Figure 2, when we use both summarizers, the evaluator selects 93 results from SOLAR (Kim et al., 2023) and 32 from MentaLLaMA (Yang et al., 2023b) as the final outputs from the 125 evaluation set. This implies that domain-specific models tend to perform worse than general LLMs, like

Table 4: Comparison of performance among different summarizers for the summarization task (Task B) using the evaluation dataset.

Summarizer	Summarization (Task B)
	Consistency↑
SOLAR & MentaLLaMa	0.970
SOLAR	**0.973**

ChatGPT (Luo et al., 2023) or SOLAR (Kim et al., 2023), on general linguistic tasks such as abstractive summarization (Wu et al., 2023). Moreover, MentaLLaMA (Yang et al., 2023b) exhibits biased hallucination issues by generating mental-health-related words like '*stuck*' or '*bother*' regardless of original contexts, leading to inconsistency. In future work, we plan to explore the comparison of evaluators and summarizers using a broader range of LLMs to gain additional insights.

> ***Posts:*** If I couldn't return, I would jump on the train, or my dad wouldn't take me to the TV show ...
> ***Response w/ MentaLLaMa:*** The user shows a feeling of being stuck and bothered by others.

4.3 Error Analysis

While our proposed approach demonstrates outstanding performance, there are a few cases where the model fails to recognize crucial evidence supporting the suicide risk level and extracts sentences that are irrelevant to the potential suicide risk. Concerning practical utility, the lack of reliability has considerably impeded the implementation of LLMs in clinical settings (Malhotra and Jindal, 2024).

> ***Response w/ Expert Identity:*** ["This subreddit is a fantastic place."]
>
> ***Response w/ Suicide Dictionary:*** ["I love everyone in this subreddit."]

5 Conclusion

In this study, we introduced promising prompting strategies that can provide evidence supporting suicide risk levels on social media data. We enhanced the LLM interpretability by incorporating domain-specific elements like assigning a psychiatrist identity and combining a suicide word. Additionally, we improved the consistency in summarization by using a consistency evaluator with multiple candidates. The proposed dual-prompting approach provides reliable reasoning, making it suitable for monitoring mental health-related risks.

Limitations

Since ground truth is not provided, quantitative comparisons are limited. Therefore, we rely on qualitative comparisons, which may be subjective. Our experiments use only the smallest version of LLMs due to limited resources. Providing inferences about suicidality using social media data is inherently subjective, allowing for various interpretations among researchers (Keilp et al., 2012). Moreover, the experimental data may be sensitive to demographic and media-specific biases (Hovy and Spruit, 2016). While the effectiveness of leveraging social media data for mental health analysis may be constrained in specific clinical settings (Ernala et al., 2019), adopting a practical model promises the potential to discern diverse statistical patterns and biases across various objectives (Jacobson et al., 2020). Although the suicide dictionary (Lee et al., 2022) has demonstrated effectiveness in predicting suicide risk, its reliance on social media data for construction might restrict its generalizability. Furthermore, the dictionary was constructed using the same dataset as the one utilized in the Shared Task, which is anticipated to introduce a certain degree of bias.

Future Work. In future work, we plan to explore a wider range of prompt templates to enhance overall performance further. For instance, the prompts could be diversified by applying various LLM-based consistency evaluators, including ChatGPT and LLaMa2. Our ultimate objective is to expand the scope to cover diverse mental health domains, such as depression and bipolar disorder, and validate its effectiveness comprehensively. To achieve this, we plan to investigate domain-specific fine-tuning methods for LLMs in the mental health field, thereby extending the model to a more interpretable context.

Ethical Statement

The CLPsych 2024 shared task (Chim et al., 2024) prioritized responsible data utilization by providing exclusive access to the dataset for researchers aligned with ethical principles. Consequently, all task participants must adhere to data use agreements and ethical practices during the competition. Our research strongly emphasizes ethics, particularly in (i) protecting the privacy of Reddit users and (ii) preventing potential misuse of the dataset. We strictly adhered to Reddit's privacy

policy[3] to ensure user anonymity (Benton et al., 2017; Williams et al., 2017).

Acknowledgments

We commit to acknowledging the assistance of the American Association of Suicidology in making the dataset available, in any publications.

This research was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation (NRF) of Korea (NRF-2022S1A5A8054322) and the International Research & Development Program of the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (RS-2023-00265683).

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.

Mostafa M Amin, Erik Cambria, and Björn W Schuller. 2023. Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *IEEE Intelligent Systems*, 38:2.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102.

Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. 2023. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9071–9084, Singapore. 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.

Alireza Darrudi, Mohammad Hossein Ketabchi Khoonsari, and Maryam Tajvar. 2022. Challenges to achieving universal health coverage throughout the world: a systematic review. *Journal of preventive medicine and public health*, 55(2):125.

[3] https://www.reddit.com/policies/privacy-policy

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022a. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022b. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16.

Sandro Galea, Raina M Merchant, and Nicole Lurie. 2020. The mental health consequences of covid-19 and physical distancing: the need for prevention and early intervention. *JAMA internal medicine*, 180(6):817–818.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Nicholas C Jacobson, Kate H Bentley, Ashley Walton, Shirley B Wang, Rebecca G Fortgang, Alexander J Millner, Garth Coombs III, Alexandra M Rodman, and Daniel DL Coppersmith. 2020. Ethical dilemmas posed by mobile health and machine learning in psychiatry research. *Bulletin of the World Health Organization*, 98(4):270.

John G Keilp, Michael F Grunebaum, Marianne Gorlyn, Simone LeBlanc, Ainsley K Burke, Hanga Galfalvy, Maria A Oquendo, and J John Mann. 2012. Suicidal ideation and the subjective aspects of depression. *Journal of affective disorders*, 140(1):75–81.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.

Moritz Laurer, Wouter Van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more

classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Daeun Lee, Migyeong Kang, Minji Kim, and Jinyoung Han. 2022. Detecting suicidality with a contextual graph neural network. In *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, pages 116–125.

Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim, and Jinyoung Han. 2023. Towards suicide prevention from bipolar disorder with temporal symptom-aware multitask learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4357–4369.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yujian Liu, Laura Biester, and Rada Mihalcea. 2023. Improving mental health classifier generalization with pre-diagnosis data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 566–577.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Anshu Malhotra and Rajni Jindal. 2024. Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *Cognitive Systems Research*, 84:101186.

Annika Marie Schoene, John Ortega, Silvio Amir, and Kenneth Church. 2023. An example of (too much) hyper-parameter tuning in suicide ideation detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1158–1162.

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.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023a. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023b. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Matthew L Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023a. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, Marzyeh Ghassemi, James Hendler, Anind K Dey, and Dakuo Wang. 2023b. Mental-llm: Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023a. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023b. Mentalllama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. Extractive summarization via ChatGPT for faithful summary generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3270–3278, Singapore. Association for Computational Linguistics.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

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, 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

A.1 Prompt of Knowledge-aware Evidence Extraction

Knowledge-aware Evidence Extraction

You are a psychiatrist.
<Examples>
Suicide Risk Level: [**Example's Label**]
Post: [**Example's Post**]
Highlights:
1. [**Highlights of Example**]
...

Referring to the <Examples> , Identify the original phrases in the post that express or reference suicide risk factors and list them without modification about <Question>.

<Question>
Suicide Risk Level: [**Label**]
Post : [**Post**]
- The answer must be in numbering format [examples of formatting]
- Phrases should be included in the given <Question>'s post.
- You MUST refer <Question>'s given post.
- Highlight only necessary phrases, not full sentences.
- Select the parts that haves the suicide-related words like [**Suicide Word List**]
- Select as many phrases as possible related to suicide even a little.

Highlights:

A.2 Evidence Summarization with LLM-based Consistency Evaluator

A.2.1 Prompt of Evidence Summarization

Extract-then-Generate

You are a psychiatrist.
You are willing to do an abstractive summary about the evidence that shows the user is at **[Label]** suicide risk.
There are suicide risk assessment aspects when seeing Reddit posts.

[**GROUND TRUTH**]
The suicide risk level of this user is [**Label**].
Here are Reddit posts and extractive evidence that supports the user is at **[Label]** suicide risk based on aspects.

[**Posts and Highlights**] Regarding the user's posts and extracted evidence and aspects of suicide risk assessments, Explain why the user is at [**Label**] suicide risk.
The revised summary should include the information in the extractive evidence and aspects.
The summary should be shorter than 300 letters.
The summary MUST be less than 300 letters.

Summarized evidence explain:

A.2.2 Prompt of Evaluating Consistency

Consistency Evaluator

You are an expertise psychiatry evaluator.
There are [**Label**] suicide risk user's posts and explain the reason for diagnosis based on posts.

[**Posts**]
Explain and summary of evidence: [**Summary**]
Score the following summary given the user posts concerning consistency from 1 to 10.
Note that consistency measures how much information the summary includes in the source posts. 10 points indicate that the summary contains only statements that are entailed by the source posts. 1 point indicates that the summary does not contain any word or statement that is entailed by the source posts.

Scores choices: from [1] to [10]

Give me a clear mark score and explain about it.
Keep the answer format
- Format: The score is [1]
to
- Format: The score is [10]

Scores: