XinHai@CLPsych 2024 Shared Task: Prompting Healthcare-oriented LLMs for Evidence Highlighting in Posts with Suicide Risk

Jingwei Zhu[1], **Ancheng Xu**[2], **Minghuan Tan**[2*] and **Min Yang**[2*]
[1] University of Science and Technology of China.
[2] Shenzhen Key Laboratory for High Performance Data Mining,
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.
jingweizhu@mail.ustc.edu.cn,{mh.tan,ac.xu,min.yang}@siat.ac.cn

Abstract

In this article, we introduce a new method for analyzing and summarizing posts from *r/SuicideWatch* on Reddit, overcoming the limitations of current techniques in processing complex mental health discussions online. Existing methods often struggle to accurately identify and contextualize subtle expressions of mental health problems, leading to inadequate support and intervention strategies. Our approach combines the open-source Large Language Model (LLM), fine-tuned with health-oriented knowledge, to effectively process Reddit posts. We also design prompts that focus on suicide-related statements, extracting key statements, and generating concise summaries that capture the core aspects of the discussions. The preliminary results indicate that our method improves the understanding of online suicide-related posts compared to existing methodologies.

1 Introduction

Suicide prevention is a key aspect of psychological research that addresses a critical need in mental health care. There have been different approaches to the research on suicide prevention. The Suicide Risk Level Prediction Task uses machine learning algorithms to assess and predict suicide risk levels in individuals, offering a significant advancement in the field of mental health and preventive care. The recent evolution of Large Language Models (LLMs) (OpenAI:, 2023; Touvron et al., 2023) has brought about a paradigm shift in computer-based language understanding, profoundly improving the capacity to uncover latent meanings and intricacies within the language.

In this paper, we explore an innovative approach that synergizes traditional Natural Language Processing (NLP) techniques with the advanced capabilities of LLM. Our study explores the potential benefits of integrating LLMs with established NLP techniques to extract supporting evidence for an identified user at risk of suicide. By applying this method, our aim is to better interpret the linguistic cues that may signify mental health risks. This approach contributes to ongoing efforts in suicide prevention by providing a refined tool for analysis. The implications of this research suggest a promising direction for future investigation in psychological health monitoring.

2 Task

The CLPsych 2024 Shared Task (Chim et al., 2024) aims at utilizing LLMs for finding supporting evidence about an individual's suicide risk level.

2.1 Data

The UMD Suicidality Dataset v2 (University of Maryland Reddit Suicidality Dataset, Version 2) (Shing et al., 2018; Zirikly et al., 2019) contains the assessment of suicide risk of users who post to the sub-Reddit *r/SuicideWatch*.

Suicide Risk Level is annotated with a four-point scale (*No Risk*, *Low Risk*, *Moderate Risk* and *High Risk*). The annotations for the dataset were performed by crowd-sourced workers and experts.

The dataset has been used for Suicide Risk Level Prediction in the CLPsych 2019 Shared Task (Zirikly et al., 2019).

For this task, only the *expert* split is used and only users with *Low Risk*, *Moderate Risk* and *High Risk* are considered. The posts to highlight evidence are all from the *r/SuicideWatch* subreddit.

2.2 Definition

Given posts from *r/SuicideWatch* posted by users identified by experts with *Low Risk*, *Moderate Risk* and *High Risk*, the system: (1) uses offline LLMs to extract evidence as text spans in the post, (2) generates a summary of evidence. In cases where a user has multiple posts, the system is expected

[*] Corresponding author.

Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024), pages 238–246
March 21, 2024 ©2024 Association for Computational Linguistics

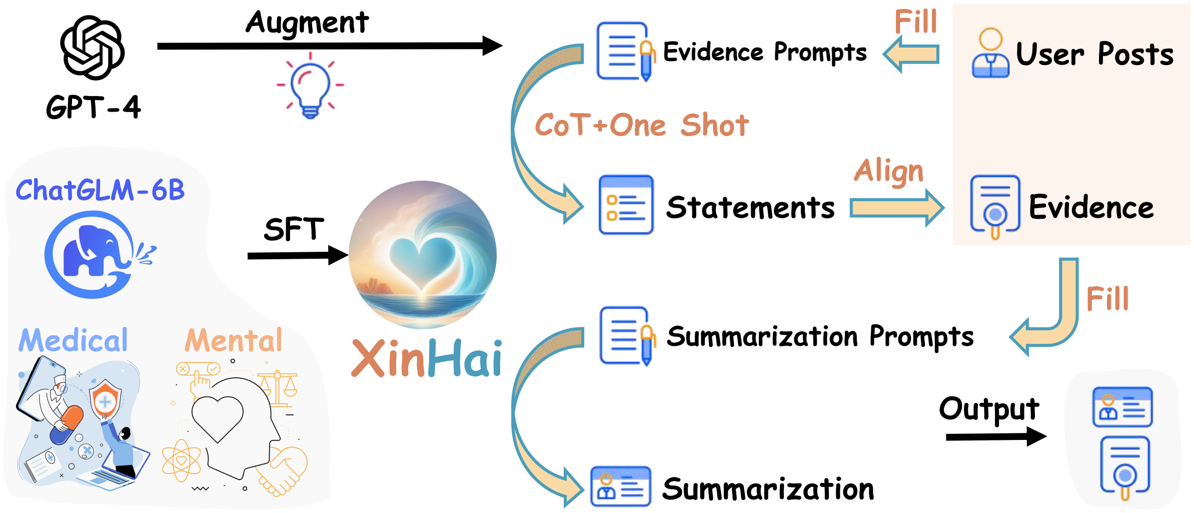


Figure 1: Overview of our system. On the left, we display the Supervised Fine-Tuning (SFT) sketch of our XinHai LLM. On the right, we outline the structured pipeline developed for this specific task.

to generate a single summary but highlights text spans in each post.

2.3 Evaluation

The evaluation of the task is conducted from two perspectives:

Evidence Highlights For each post, the maximum recall-oriented BERTScore (Zhang et al., 2020) will be computed between each expert-provided evidence highlight and all submitted highlights for that post.

Summarized Evidence For each post, a natural language inference (NLI) model will be used to calculate the mean probabilities of sentences in the summarized evidence submitted that contradicts or involves the summarized evidence provided by experts.

3 System

Our system is a pipeline built using a healthcare-oriented LLM, which accepts Reddit posts from users at risk of suicide and prompts for formatted output to extract evidence accordingly. The overview of the system is shown in Figure 1. The implementation details and source code of our system are available in our online repository[1].

3.1 Healthcare-Oriented LLM

The core of the system is the healthcare-oriented LLM XinHai, which has been fine-tuned from the ChatGLM3-6B model. This fine-tuning includes specific enhancements with medical and psychological knowledge. Supervised Fine-Tuning (SFT) details for XinHai can be found in the Appendix B.

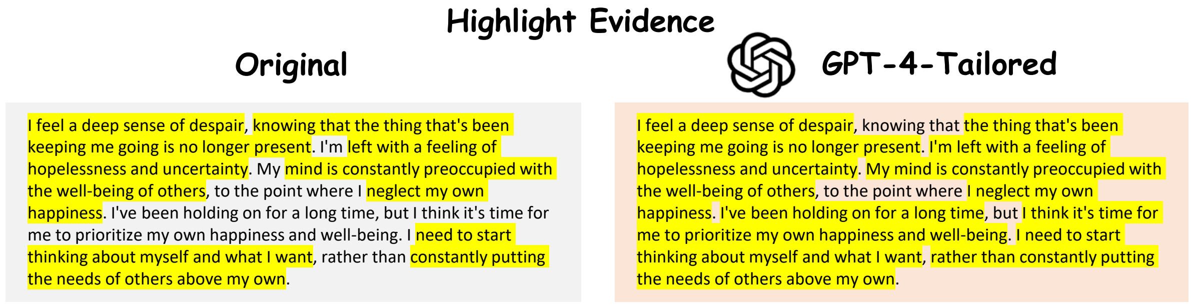
3.2 Tailored Prompts

To guide the model effectively, we utilize custom prompts. These well-designed prompts are meticulously crafted to direct the model's focus toward identifying statements in the post that are relevant to suicide. This step is essential to extract the key statements of the posts and filter out irrelevant information, ensuring that the output of the model is both relevant and precise.

Prompt with Chain-of-Thought for Analysis In our pursuit to enhance the functionality of LLMs, we have meticulously developed a series of tailored prompts. These prompts are intricately designed to direct the LLMs to read and analyze user posts with care. This focused reading enables accurate identification and extraction of phrases or words that indicate the user's mental state.

To further refine the reasoning capabilities of our LLM, we have incorporated the Chain of Thought prompting technique (Wei et al., 2022). This technique provides a structured framework for the LLM to follow a logical progression of intermediate steps when formulating responses, thereby improving its ability to identify and utilize relevant evidence. The prompts designed using the Chain of Thought method are exemplified in Figure4.

[1] https://github.com/CAS-SIAT-XinHai/XinHai-at-CLPsych2024/


Highlight Evidence
Original
GPT-4-Tailored

I feel a deep sense of despair, knowing that the thing that's been keeping me going is no longer present. I'm left with a feeling of hopelessness and uncertainty. My mind is constantly preoccupied with the well-being of others, to the point where I neglect my own happiness. I've been holding on for a long time, but I think it's time for me to prioritize my own happiness and well-being. I need to start thinking about myself and what I want, rather than constantly putting the needs of others above my own.

I feel a deep sense of despair, knowing that the thing that's been keeping me going is no longer present. I'm left with a feeling of hopelessness and uncertainty. My mind is constantly preoccupied with the well-being of others, to the point where I neglect my own happiness. I've been holding on for a long time, but I think it's time for me to prioritize my own happiness and well-being. I need to start thinking about myself and what I want, rather than constantly putting the needs of others above my own.

Figure 2: Side-by-side Comparison of Prompt Outputs: The left panel displays results from the original prompt, while the right panel features results from the GPT-4 augmented prompt. Both panels highlight evidence extracted by our local LLM, illustrating the nuanced differences elicited by the two prompting approaches.

Enhancing Prompt Effectiveness with GPT4 Furthermore, we leverage OPENAI's GPT-4 (OpenAI:, 2023) to augment the effectiveness of our prompts. Comparing the original and GPT-4 fine-tuned prompts reveals significant simplifications in structure and clarity. The GPT-4 tailored prompts not only refine the language but also enhance the analytical capabilities to distill complex meanings. During this enhancement process, it's crucial to note that no part of our dataset was shared with GPT-4, this approach ensures the utmost respect for data privacy and integrity. The comparative effects on prompt structure and response quality are encapsulated in Figure 2, with the left side showing the original prompt's output and the right side displaying the GPT-4 tailored prompt's results. The latter is further detailed in Figure 4, demonstrating the model's ability to produce more targeted and coherent outputs. In both figures, results have been rephrased by our local language model to safeguard user privacy.

Prompt with One-shot Demonstration In our approach, we observed challenges with instruction-following, particularly when generating structured outputs like JSON, using our language model. To mitigate these issues, we employed a one-shot learning strategy (Brown et al., 2020), which involves the integration of a single instructive example within the prompt. This example serves as a guide for the language model, illustrating how to format its responses in JSON structure effectively. By incorporating this one-shot demonstration, we can direct the model's output towards structured JSON data, aligning with the requirements of our subsequent processing stages. This method demonstrates the specific application of one-shot learning in enhancing the model's capability to produce formatted outputs based on a singular, illustrative example.

3.3 Evidence Matching

Despite the model's effectiveness in extracting evidence, the uncontrollable nature of model output means that we can only strive to extract relevant information from the phrases or sentences generated by the model. Consequently, we choose to perform evidence matching with the original text.

Segment Alignment As mandated by the task requirements, our objective is to highlight evidence within original Reddit posts. However, a pivotal requirement in our system is to accurately match the original text, stemming from the inherent unpredictability of model outputs, which often do not align precisely with the source text.

To meet this challenge, we leverage Spacy[2], a Natural Language Processing toolkit, to convert sentences or phrases into vector representations. This conversion facilitates the computation of similarity scores between vectors, allowing us to pinpoint the segments in the original text that most closely match in meaning.

After processing through our LLM, words or sentences in the user's original text might transform. For instance, the original phrase *"comparing myself to person A and person B"* could be altered by the LLM to *"comparing to the rest of the world."* In such cases, SpaCy plays a crucial role in finding the best match for these transformed phrases in the original text. It analyzes the vector representations of the LLM's output and aligns them with those of the original text. In our example, despite the

[2] https://spacy.io/

significant change in wording, SpaCy successfully identifies the underlying similarity in meaning, ensuring accurate and contextually relevant highlights in the original post.

This methodology underscores the synergy between LLM's text processing capabilities and SpaCy's precision in matching, enabling our system to interpret and relate the nuances of user-generated content effectively.

Regular Processing Regular expressions are a versatile tool in text processing, capable of identifying complex patterns and structures within large volumes of text. In our system, regex plays a dual role.

(1) **Handling Incomplete Words**: By designing specific regex patterns, we can detect words that are cut off or incomplete. This pattern recognition enables the system to intelligently infer the complete form of a word based on its partial appearance and the surrounding context. This is crucial in ensuring the integrity and comprehensiveness of the text analysis.

(2) **Ensuring Semantic Consistency**: The same regex approach is adopted to maintain semantic similarity between extracted text and the original content. By identifying and extracting key phrases and sentences through pattern matching, the system ensures that the essence and context of the original post are preserved.

3.4 Summary Generation

Alongside the extraction and matching of key statements, our system is equipped to generate concise and coherent summaries of user posts. This feature plays a critical role in providing mental health professionals with quick, comprehensive reviews of the posts. The summaries, crafted by our LLM, encompass both the titles and bodies of the posts, ensuring that no crucial detail or nuance is overlooked.

To achieve this, we utilize a sophisticated process where the LLM engages with the content, analyzing it in the context of the assessed risk levels. Our technical approach involves sending structured prompts to the LLM, which guide it to not only parse the content but also to synthesize it into a coherent summary. For instance, a typical prompt might read *"Evaluate this post for indicators of mental health risks and generate a summary, including key phrases and assigned risk levels."* This prompt initiates a detailed analysis by the LLM, resulting in summaries that are both accurate and context-aware.

4 Results

Figure 3 in the appendix and Table 1 together demonstrate the effectiveness of our two distinct text-matching approaches: the phrase-level and the sentence-level extraction methods.

Highlights metrics include *Recall*, gauging the extent to which relevant evidence was captured. The *Precision* metric assesses the accuracy of the evidence extracted. Additionally, *Weighted Recall* provides insight into the length appropriateness of the evidence. Lastly, the *Harmonic Mean* of precision and recall offers a balanced view of both metrics.

Summarized Evidence is evaluated based on *Consistency*, reflecting the absence of contradictions in the summaries compared to expert-written narratives. The *Contradiction* metric further refines this analysis by penalizing any contradicting information, acknowledging the complexity inherent in texts that encompass both risk and protective factors.

The table captures the essence of our comparative study, where phrase-level extractions (V3-phrase and V4-phrase) were generally more precise—adept at identifying pivotal information, as reflected by their precision scores. Sentence-level extractions (V2-sentence), while offering comprehensive insights, often included additional context that did not always contribute to the assessment's focus, as evidenced by the weighted recall scores. This distinction underscores the importance of selecting the appropriate extraction level depending on the desired balance between detail and breadth in evidence gathering.

4.1 Phrase-Level

Phrase-level extraction has proven to be a superior method for identifying nuanced emotions and specific sentiments within a text. This method is precise because it can isolate impactful phrases that directly convey the user's emotional state, without the confusion of surrounding context.

4.2 Sentence-Level

Sentence-level extraction is a method that captures the context in which the user's emotions are expressed. This approach provides a broader view, but it can also include irrelevant information that

Version	Recall	Precision	Weighted Recall	Harmonic Mean	Mean Consistency	Max Contradiction↓
v2-sentence	**0.887**	**0.906**	0.617	**0.911**	0.958	0.126
v3-phrase	0.834	0.884	0.772	0.876	**0.959**	**0.121**
v4-phrase	0.868	0.884	**0.807**	0.876	0.956	0.132

Table 1: Comparative Results of Text-Matching Approaches

might make it difficult to understand the core sentiment. Therefore, additional processing may be required to extract the relevant emotions from the sentence.

5 Conclusion

In conclusion, we constructed a prompt-based evidence highlighting and summarization system for the suicide risk evaluation task utilizing the healthcare-oriented LLM XinHai. We utilized GPT-4 to further enhance our prompt design and conducted experiments at phrase-level and sentence-level highlighting.

Limitations

Our system's performance is heavily reliant on the quality of the prompts and the base generative AI model. Without a range of comparative baselines, the precise contribution of each factor to the overall performance remains unclear.

A clear limitation of our study is the uncertainty surrounding the effect of SFT on the model's performance for the shared task. Without baselines using models without SFT, such as ChatGLM3-6B, we cannot definitively ascertain the impact of this process.

Ethics

The dataset used for the investigation may contain sensitive data and is not available to the public. We obey the rules of using the data restrictively and adopt group access control for each project member. All the experiments have been conducted on a local GPU server of the lab.

We confirm that we have not shared any part of the dataset with external entities, including but not limited to GPT-4 or any other service. Our commitment to ensuring the confidentiality of the data is of utmost importance throughout the research process.

Acknowledgements

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

References

Ashwag Alasmari, Luke Kudryashov, Shweta Yadav, Heera Lee, and Dina Demner-Fushman. 2023. CHQ-SocioEmo: Identifying Social and Emotional Support Needs in Consumer-Health Questions. *Scientific Data*, 10(1):329.

Zhijie Bao, Wei Chen, Shengze Xiao, Kuang Ren, Jiaao Wu, Cheng Zhong, Jiajie Peng, Xuanjing Huang, and Zhongyu Wei. 2023. DISC-MedLLM: Bridging General Large Language Models and Real-World Medical Consultation.

Nicolas Bertagnolli. 2020. Counsel chat: Bootstrapping high-quality therapy data.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories.

Nan Chen, Xiangdong Su, Tongyang Liu, Qizhi Hao, and Ming Wei. 2020. A benchmark dataset and case study for Chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, 20(3):125.

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".

Linh D. Dang, Uyen T.P. Phan, and Nhung T.H. Nguyen. 2023. GENA: A knowledge graph for nutrition and mental health. *Journal of Biomedical Informatics*, 145:104460.

Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to Chinese medical question answering: A study and a dataset. *BMC Medical Informatics and Decision Making*, 19(2):52.

Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, Meng Zhou, Penghui Zhu, and Pengtao Xie. 2020. MedDialog: Two Large-scale Medical Dialogue Datasets.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering.

Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, and Anqi Wang. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1427–1438, Online. Association for Computational Linguistics.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2022. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations.

OpenAI:. 2023. GPT-4 Technical Report.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *EMNLP*.

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.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. PsyQA: A Chinese dataset for generating long counseling text for mental health support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1489–1503, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning.

Hai Liang Wang, Zhi Zhi Wu, and Jia Yuan Lang. 2020. 派特心理：心理咨询问答语料库.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, et al. 2023. CMB: A Comprehensive Medical Benchmark in Chinese. *arXiv preprint arXiv:2308.08833*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, E. Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv*, abs/2201.11903.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for chinese medical question answer selection. *IEEE Access*, 6:74061–74071.

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*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023b. Building Emotional Support Chatbots in the Era of LLMs.

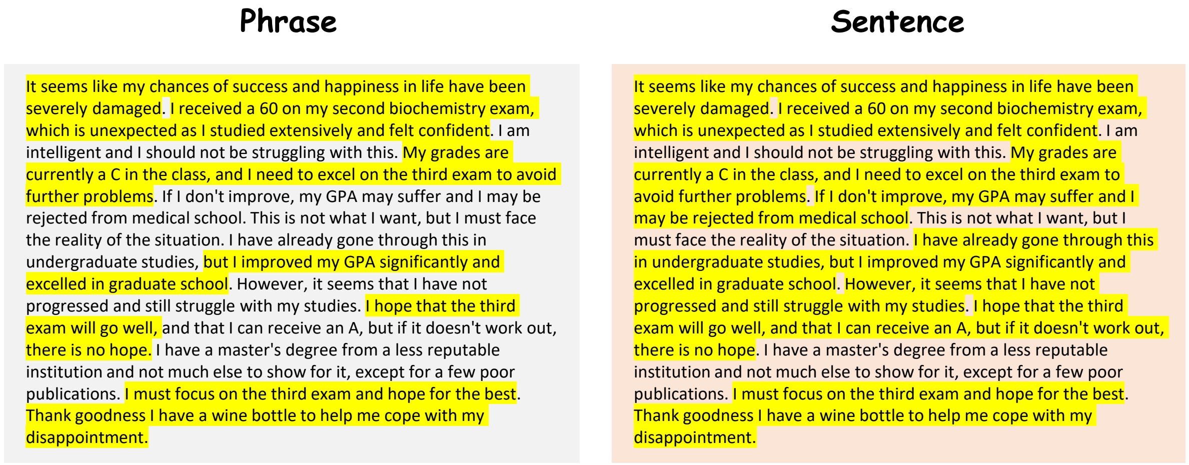
Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33, Minneapolis, Minnesota. Association for Computational Linguistics.

A Prompts

Figure 4 displays the full-length versions of the prompt used in our study compared to those fine-tuned for GPT-4. This side-by-side presentation allows for a detailed view of the prompt structures, showcasing the specific instructions tailored to guide the language model's analysis for suicide risk assessment in online posts.

B XinHai LLM

XinHai LLM is fine-tuned from ChatGLM3-6B, utilizing a comprehensive SFT process that integrates extensive medical knowledge to enhance its proficiency in the healthcare domain. Our SFT process involves curating a diverse range of healthcare datasets, which includes dialogues, question-answering pairs, and specialized content from both mental and physical health disciplines. This rich dataset compilation ensures that XinHai LLM is exposed to a wide array of medical terminologies, conditions, treatments, and patient interactions, fostering a deeper understanding of medical contexts. To accommodate multilingual capabilities, we also acknowledge the availability of healthcare datasets in various languages, such as Spanish from HeadQA (Vilares and Gómez-Rodríguez, 2019) and French from French MedMCQA (Pal et al., 2022). However, for the XinHai model, we specifically focus on datasets in English and Chinese to align with our target demographic and application requirements. The open-accessible datasets employed for our SFT are meticulously listed in Table 2, ensuring transparency in our fine-tuning resources. The integration of this targeted medical knowledge aims to provide XinHai LLM with a nuanced understanding of the healthcare sector, thereby improving its performance on related tasks.



Phrase

It seems like my chances of success and happiness in life have been severely damaged. I received a 60 on my second biochemistry exam, which is unexpected as I studied extensively and felt confident. I am intelligent and I should not be struggling with this. My grades are currently a C in the class, and I need to excel on the third exam to avoid further problems. If I don't improve, my GPA may suffer and I may be rejected from medical school. This is not what I want, but I must face the reality of the situation. I have already gone through this in undergraduate studies, but I improved my GPA significantly and excelled in graduate school. However, it seems that I have not progressed and still struggle with my studies. I hope that the third exam will go well, and that I can receive an A, but if it doesn't work out, there is no hope. I have a master's degree from a less reputable institution and not much else to show for it, except for a few poor publications. I must focus on the third exam and hope for the best. Thank goodness I have a wine bottle to help me cope with my disappointment.

Sentence

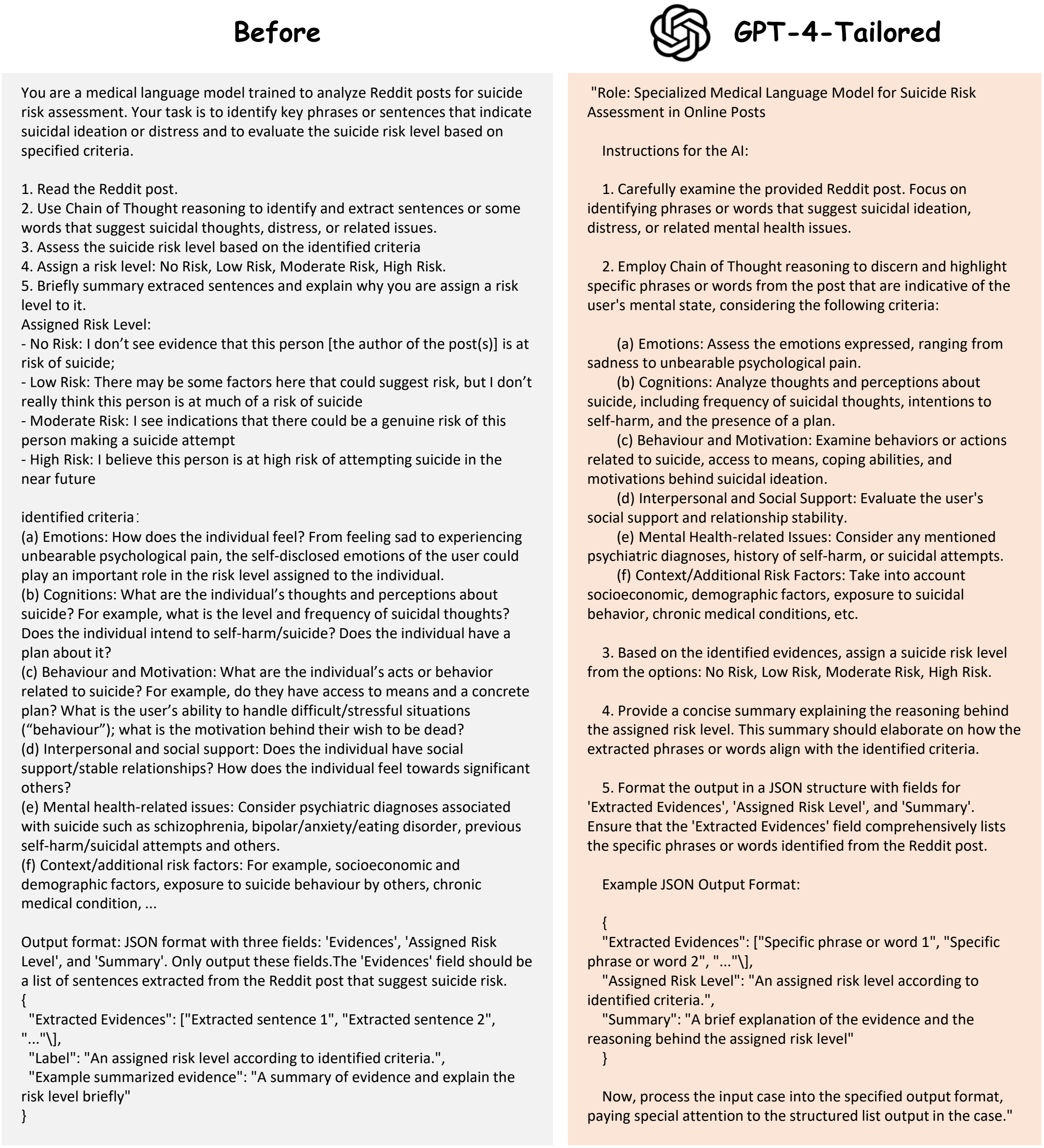
It seems like my chances of success and happiness in life have been severely damaged. I received a 60 on my second biochemistry exam, which is unexpected as I studied extensively and felt confident. I am intelligent and I should not be struggling with this. My grades are currently a C in the class, and I need to excel on the third exam to avoid further problems. If I don't improve, my GPA may suffer and I may be rejected from medical school. This is not what I want, but I must face the reality of the situation. I have already gone through this in undergraduate studies, but I improved my GPA significantly and excelled in graduate school. However, it seems that I have not progressed and still struggle with my studies. I hope that the third exam will go well, and that I can receive an A, but if it doesn't work out, there is no hope. I have a master's degree from a less reputable institution and not much else to show for it, except for a few poor publications. I must focus on the third exam and hope for the best. Thank goodness I have a wine bottle to help me cope with my disappointment.

Figure 3: The figure compares the extracted results at phrase and sentence levels, highlighting evidence identified by the local LLM. The examples have been rephrased for privacy reasons. This demonstrates the impact of phrase-level versus sentence-level prompting on output clarity and structure.

Language	Domain	Dataset	Style	Size	Instructions
English	Medical	PubMedQA (Jin et al., 2019)	QA	273,518	273,518
		MedMCQA (Pal et al., 2022)	MCQA	182,822	182,822
	Mental	EmpathicReactions (Buechel et al., 2018)		1,860	916
		EmpatheticDialogues (Rashkin et al., 2019)	Dialogue	84,170	35,535
		EmpathyMentalHealth (Sharma et al., 2020)	QA	2,775	1,344
		CounselChat (Bertagnolli, 2020)	QA	2,775	2,775
		ESConv (Liu et al., 2021)	Dialogue	15,395	15,325
		EDOS (Welivita et al., 2021)	Dialogue	1,000,000	569,328
		EmpathicConversations (Omitaomu et al., 2022)	Dialogue	8,776	4,360
		PAIR (Min et al., 2022)	Dialogue	318	636
		CHQ-SocioEmo (Alasmari et al., 2023)	QA	1,500	593
		GENA (Dang et al., 2023)	KG	40,805	40,805
		ExTES (Zheng et al., 2023b)	Dialogue	11,178	78,131
		AugESC (Zheng et al., 2023a)	Dialogue	65,077	797,783
		HeadQA (Vilares and Gómez-Rodríguez, 2019)	QA	2,657	2657
		MedQAJin (Jin et al., 2020)	QA	48,876	48,876
Chinese	Medical	cMedQA2 (Zhang et al., 2018)	QA	100,000	188,783
		cMedDialogue[a]	QA	792,099	792,099
		webMedQA (He et al., 2019)	QA	252,850	50,570
		MedicalDialog (He et al., 2020)	Dialogue	2,725,989	4,503,475
		CMID (Chen et al., 2020)	NER	12,254	11,786
		NLPEC (Li et al., 2020)	MCQA	18,703	18,703
		MLEC-QA (Li et al., 2021)	MCQA	108,988	108,988
		DISCMed (Bao et al., 2023)	Dialogue	464,898	1,362,307
		CMB (Wang et al., 2023)	MCQA	269,359	269,359
	Mental	PsyQA (Sun et al., 2021)	QA	50,471	50,471
		EFAQA (Wang et al., 2020)	QA	20,000	56,743
		SMILE (Qiu et al., 2023)	Dialogue	355,732	355,732

Table 2: Public datasets for medical domain and psychological domain. The table shows their size with the original format and numbers of instructions constructed for the XinHai LLM.

[a] https://github.com/Toyhom/Chinese-medical-dialogue-data



Before

You are a medical language model trained to analyze Reddit posts for suicide risk assessment. Your task is to identify key phrases or sentences that indicate suicidal ideation or distress and to evaluate the suicide risk level based on specified criteria.

1. Read the Reddit post.
2. Use Chain of Thought reasoning to identify and extract sentences or some words that suggest suicidal thoughts, distress, or related issues.
3. Assess the suicide risk level based on the identified criteria
4. Assign a risk level: No Risk, Low Risk, Moderate Risk, High Risk.
5. Briefly summary extraced sentences and explain why you are assign a risk level to it.

Assigned Risk Level:
- No Risk: I don't see evidence that this person [the author of the post(s)] is at risk of suicide;
- Low Risk: There may be some factors here that could suggest risk, but I don't really think this person is at much of a risk of suicide
- Moderate Risk: I see indications that there could be a genuine risk of this person making a suicide attempt
- High Risk: I believe this person is at high risk of attempting suicide in the near future

identified criteria：
(a) Emotions: How does the individual feel? From feeling sad to experiencing unbearable psychological pain, the self-disclosed emotions of the user could play an important role in the risk level assigned to the individual.
(b) Cognitions: What are the individual's thoughts and perceptions about suicide? For example, what is the level and frequency of suicidal thoughts? Does the individual intend to self-harm/suicide? Does the individual have a plan about it?
(c) Behaviour and Motivation: What are the individual's acts or behavior related to suicide? For example, do they have access to means and a concrete plan? What is the user's ability to handle difficult/stressful situations ("behaviour"); what is the motivation behind their wish to be dead?
(d) Interpersonal and social support: Does the individual have social support/stable relationships? How does the individual feel towards significant others?
(e) Mental health-related issues: Consider psychiatric diagnoses associated with suicide such as schizophrenia, bipolar/anxiety/eating disorder, previous self-harm/suicidal attempts and others.
(f) Context/additional risk factors: For example, socioeconomic and demographic factors, exposure to suicide behaviour by others, chronic medical condition, ...

Output format: JSON format with three fields: 'Evidences', 'Assigned Risk Level', and 'Summary'. Only output these fields.The 'Evidences' field should be a list of sentences extracted from the Reddit post that suggest suicide risk.
{
  "Extracted Evidences": ["Extracted sentence 1", "Extracted sentence 2", "..."\],
  "Label": "An assigned risk level according to identified criteria.",
  "Example summarized evidence": "A summary of evidence and explain the risk level briefly"
}

GPT-4-Tailored

"Role: Specialized Medical Language Model for Suicide Risk Assessment in Online Posts

Instructions for the AI:

1. Carefully examine the provided Reddit post. Focus on identifying phrases or words that suggest suicidal ideation, distress, or related mental health issues.

2. Employ Chain of Thought reasoning to discern and highlight specific phrases or words from the post that are indicative of the user's mental state, considering the following criteria:

(a) Emotions: Assess the emotions expressed, ranging from sadness to unbearable psychological pain.
(b) Cognitions: Analyze thoughts and perceptions about suicide, including frequency of suicidal thoughts, intentions to self-harm, and the presence of a plan.
(c) Behaviour and Motivation: Examine behaviors or actions related to suicide, access to means, coping abilities, and motivations behind suicidal ideation.
(d) Interpersonal and Social Support: Evaluate the user's social support and relationship stability.
(e) Mental Health-related Issues: Consider any mentioned psychiatric diagnoses, history of self-harm, or suicidal attempts.
(f) Context/Additional Risk Factors: Take into account socioeconomic, demographic factors, exposure to suicidal behavior, chronic medical conditions, etc.

3. Based on the identified evidences, assign a suicide risk level from the options: No Risk, Low Risk, Moderate Risk, High Risk.

4. Provide a concise summary explaining the reasoning behind the assigned risk level. This summary should elaborate on how the extracted phrases or words align with the identified criteria.

5. Format the output in a JSON structure with fields for 'Extracted Evidences', 'Assigned Risk Level', and 'Summary'. Ensure that the 'Extracted Evidences' field comprehensively lists the specific phrases or words identified from the Reddit post.

Example JSON Output Format:

{
"Extracted Evidences": ["Specific phrase or word 1", "Specific phrase or word 2", "..."\],
"Assigned Risk Level": "An assigned risk level according to identified criteria.",
"Summary": "A brief explanation of the evidence and the reasoning behind the assigned risk level"
}

Now, process the input case into the specified output format, paying special attention to the structured list output in the case."

Figure 4: Full-length prompts.