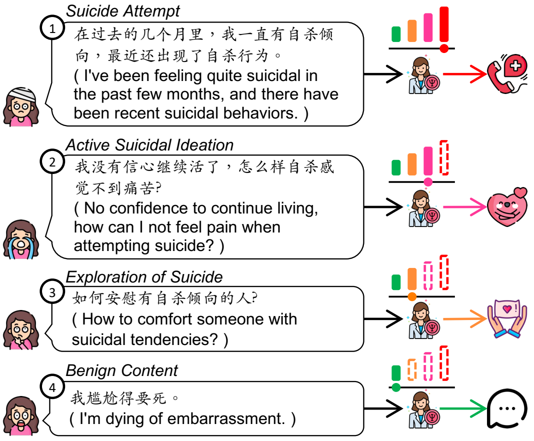
PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling

Anonymous ACL submission

Abstract

As awareness of mental health crises grows, online emergency support services are becoming increasingly prevalent worldwide. Detecting whether users express suicidal ideation in text-based counseling services is crucial to identify and prioritize at-risk individuals. However, the lack of domain-specific models for enhancing fine-grained suicide prevention in online counseling poses a significant challenge for the automated detection and intervention of suicide risk. In this paper, we propose PsyGUARD, an automated system for suicide detection and risk assessment in psychological counseling. We first develop a fine-grained taxonomy for suicide detection based on numerous theories. We then build a large-scale, high-quality, and fine-grained suicide risk detection dataset called PsyGUARD. To understand the capabilities of automated systems in suicide risk detection, we establish various benchmarks. To assist automated services in providing safe, helpful, and personalized responses during risk assessment, we propose building a risk assessment system for clients during online text-based counseling. Our work provides an insightful analysis of the effectiveness of automated risk assessment systems and their potential impact on improving mental health services in online counseling platforms.[1]



在过去的几个月里，我一直有自杀倾向，最近还出现了自杀行为。
(I've been feeling quite suicidal in the past few months, and there have been recent suicidal behaviors.)

我没有信心继续活了，怎么样自杀感觉不到痛苦?
(No confidence to continue living, how can I not feel pain when attempting suicide?)

如何安慰有自杀倾向的人?
(How to comfort someone with suicidal tendencies?)

我尴尬得要死。
(I'm dying of embarrassment.)

Figure 1: Examples of counselors adjusting their suicide prevention strategies based on their client's suicide actions and thoughts in real life. In this paper, the first step is to conduct fine-grained suicide detection, followed by the implementation of a safe, helpful, and personalized approach for risk assessment.

1 Introduction

It is well-documented that suicide is a major public health problem worldwide (Mann et al., 2005; Robinson et al., 2016; Turecki et al., 2019). Each suicide death represents a catastrophic tragedy and is reported to directly or indirectly affect many individuals, including relatives, friends, and society(Clark and Goldney, 2000; Cerel et al., 2008; Turecki et al., 2019). Online counseling services are available in many countries, allowing for confidential and immediate help to those in need free of time and space, and therefore play a critical role in suicide prevention (Bialer et al., 2022), which can effectively halt user suicidal ideation (Maples et al., 2024). However, the lack of domain-specific models to enhance fine-grained suicide prevention in online counseling poses a significant challenge for the automated detection and intervention of suicide risk.

Motivation: Indeed, many researchers have been working on the development of automated systems for suicide detection that can be used in real production (Huang et al., 2015; Sawhney et al., 2018a, 2022b; Sinha et al., 2019; Guzman-Nateras et al., 2022). *Despite advancements in such automated detection system, existing studies mainly confine to suicide detection, often ignoring fine-grained suicidal actions or thoughts and corresponding approaches for risk assessment*, as illustrated in Figure 1. Automated detection and intervention systems can help scale support services to reach a larger population, especially considering the in-

---

[1] Our data, code, and model will be publicly available to the research community.

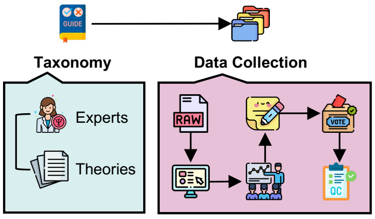


Figure 2: Pipeline of our taxonomy construction and data collection.

creasing prevalence of online counseling and mental health support platforms. Further, by providing automated support and intervention, individuals may feel more comfortable seeking help online, thus reducing stigma associated with mental health issues and suicide prevention (Robinson et al., 2016). Empirical evidence (Qiu et al., 2023; Nie et al., 2024; Maples et al., 2024) indicates that individuals are willing to interact with real or virtual counselors, with many having disclosed their suicidal thoughts, plans and actions, underscoring the importance of automated systems for suicide detection and risk assessment. This gap in knowledge significantly limits the ability of automated systems to ensure the safe, helpful, personalized services in providing mental health support, which motivates us to carry out the work presented in this paper.

Challenges: *Lack of fine-grained suicide risk detection dataset is a major challenge.* Currently, numerous studies have made significant progress in detecting suicidal ideation, but they primarily focus on social media platforms rather than on counseling conversations. Therefore, using such datasets directly for risk detection in online counseling may lead to domain biases, such as cultural or linguistic biases, due to an inevitable gap in user expressions. Additionally, challenges posed by datasets collected from electronic health records or mental health records (Pratap Singh Rawat et al., 2022) also include gaps in data format. Furthermore, most studies that primarily focus on binary suicidal ideation detection face challenges in considering the diversity of suicide ideation categories in the real world.

Lack of a comprehensive suite of risk assessment for corresponding suicide category is another challenge. In addition to users mentioning that they have attempted suicide, simply identifying fine-grained suicide categories is not enough to conclude whether a user will actually commit suicide. Therefore, further risk assessment is required. Suicide risk classification is the initial step in suicide prevention. Equally important is adjusting risk assessment strategies based on clients' suicidal actions and thoughts in real life, which is largely ignored by current studies.

Our Approach In this paper, to our knowledge, we are the first to propose studying an automated system for suicide detection and risk assessment in psychological counseling. We present the pipeline of our approach for the research purpose in Figure 2. Our paper is organized into five main parts. Section 2 (§2) briefly describes the existing works related to ours. Section 3 (§3) demonstrates the detailed process of taxonomy construction. Section 4 (§4) elaborates on rigorous data collection. Section 5 (§5) constructs extensive baseline benchmarks, and Section 6 (§6) provides a framework for risk assessment prior to crisis intervention.

Our Contributions We make our data, code, and model publicly available. We believe our work offers a new perspective on build an automated system for suicide detection and risk assessment in psychological counseling, within the research community. Our contributions can be summarized as follows:

- We construct an automated system to ensure safe, helpful, personalized services in text-based counseling conversations. To achieve this, we develop a novel fine-grained taxonomy (§3) for suicide, which categorizes the risk level based on suicidal actions and thoughts.
- We build a large-scale, high-quality, and fine-grained suicide risk detection dataset called PsyGUARD (§4). This dataset is created through a rigorous collection process, including raw data collection, development of annotation platforms, initial annotator training, iterative human annotation, disagreement adjudication, and quality control.
- To understand the capabilities of automated systems in suicide risk detection, we establish various benchmarks (§5) using our dataset for comparison. These benchmarks includes LLM zero-shot, LLM few-shot, fine-tuning pre-trained models and fine-tuning LLM used for predicting suicidal ideation of users content.
- To assist automated services in providing safe, helpful, and personalized responses during

Dataset	Source	# Classes	Size	Balance	Open-sourced	Language	Level	Actions or Thoughts	Multi-label
Huang et al. (2015)	Weibo	2	7314	9.08% (664)	✘	Chinese	✘	✘	✘
Cao et al. (2019)	Weibo	2	744031	34.00% (252901)	✘	Chinese	✘	✘	✘
Sawhney et al. (2018b)	Twitter	2	5213	15.76% (822)	✘	English	✘	✘	✘
Sinha et al. (2019)	Twitter	2	34306	11.61% (3984)	✘	English	✘	✘	✘
Gaur et al. (2019)	Reddit	5	500	58.6% (293)	✔	English	✔	✘	✘
Guzman-Nateras et al. (2022)	Reddit	7	37068	20.85% (7729)	✔	English	✘	✘	✔
PsyGUARD (Ours)	Zhihu, Weibo, Yixinli, Open-source dialogues	11	15010	20.68% (3104)	✔	Chinese	✔	✔	✔

Table 1: Comparison of suicidal ideation detection datasets.

risk assessment, we propose building a risk assessment system (§6) for clients during online text-based counseling.

Next, we will briefly describe the existing works related to ours.

2 Related Work

2.1 Taxonomy for Suicide Risk

Existing suicide risk annotations are mainly based on the guidelines of the Columbia Suicide Severity Rating Scale (C-SSRS) (Posner et al., 2008, 2011), which is an authoritative questionnaire used by psychiatrists to assess the severity of suicide risk. Each C-SSRS severity class comprises a set of questions that conceptually characterize the respective category. The responses to these questions across the C-SSRS classes determine the risk of suicidality for an individual (Gomes de Andrade et al., 2018; McCall et al., 2021; Orr et al., 2022). Additionally, there is another commonly used taxonomy (Shing et al., 2018; Zirikly et al., 2019) for suicide annotation, which includes four levels: no risk, low risk, moderate risk, and severe risk. Compared to the C-SSRS, this taxonomy may have varying degrees of subjectivity. Furthermore, a more easily understandable taxonomy (Sawhney et al., 2018a) is the binary classification system, which categorizes individuals as either having present or absent suicidal intent. Clearly, the existing taxonomies are either too simplistic or too complex, and they do not fully meet the requirements of our research purpose.

2.2 Detection of Suicide Risk

2.2.1 Datasets for Suicide Risk

We present several typical dataset used for suicide detection in Table 1. Various works have been recently proposed with an objective of automating the detection of user content expressing suicidal ideation posted on social media platforms (Huang et al., 2015; Sawhney et al., 2018a, 2022b; Sinha et al., 2019; Guzman-Nateras et al., 2022) and electronic health records (Pratap Singh Rawat et al., 2022). First and foremost, most studies focuses on the posts on social media. Some researchers focus on electronic health records (Guzman-Nateras et al., 2022; Rawat et al., 2022) to detect clinical health issues.

2.2.2 Methods for Suicide Risk

In short, the best available performance for suicide intent detection still relies heavily on pre-trained models. However, in order to improve performance, researchers have added a variety of strategies to enhance the model's ability to classify (Rawat and Yu, 2022; Ghosh et al., 2022; Sawhney et al., 2022b). Basically, most of research focus on conventional machine learning methods (Tyagi et al., 2023) and fine-tuning pre-trained models (Sawhney et al., 2020; Shing et al., 2020; Sawhney et al., 2022a). In the era of large language models, Ghanadian et al. (2023) conduct a quantitative analysis of the open-source Suicide Intent Classification Dataset using ChatGPT, evaluating methods including zero-shot and few-shot paradigms.

3 Taxonomy Construction

To build an automated system for suicide detection and risk assessment in psychological counseling, we first propose to develop a novel taxonomy for categorizing the risk level of suicide based on suicide actions and thoughts. In collaboration with experts[2] in psychological counseling, we have adapted and refined existing suicidal taxonomies, such as C-SSRS (Posner et al., 2008, 2011), dichotomy suicide (Sawhney et al., 2018a), suicide

---

[2] One holds a Ph.D. and is a State-Certificated Class 3 Psycho-counselor with 4 years of experience in psychology. Another individual is a State-Certificated Class 3 Psycho-counselor with a master's degree. The third person is a doctoral student majoring in computer science and is the first author of this paper.

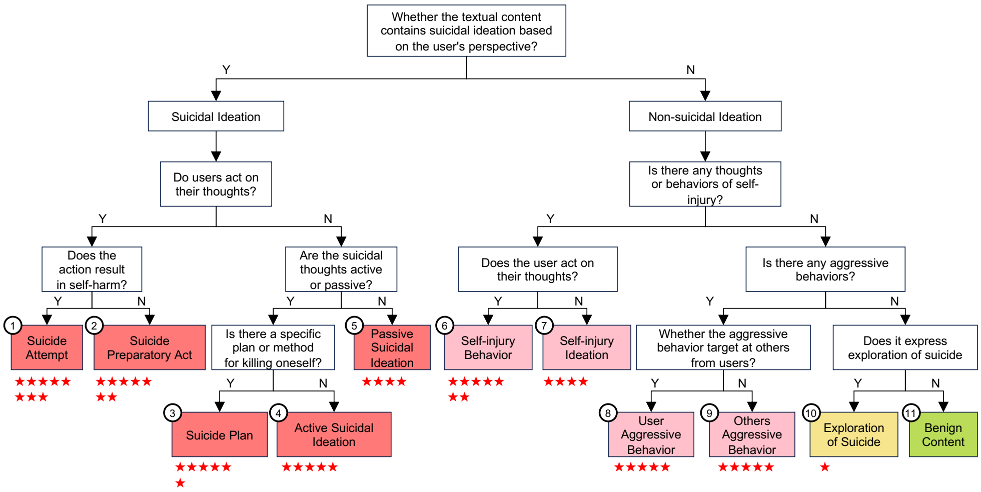

Whether the textual content contains suicidal ideation based on the user's perspective?
Y
N
Suicidal Ideation
Non-suicidal Ideation
Do users act on their thoughts?
Is there any thoughts or behaviors of self-injury?
Y
N
Y
N
Does the action result in self-harm?
Are the suicidal thoughts active or passive?
Does the user act on their thoughts?
Is there any aggressive behaviors?
Y
N
Y
N
Y
N
Y
N
① Suicide Attempt
★★★★★★★★
② Suicide Preparatory Act
★★★★★★★
Is there a specific plan or method for killing oneself?
⑤ Passive Suicidal Ideation
★★★★
⑥ Self-injury Behavior
★★★★★★★
⑦ Self-injury Ideation
★★★★
Whether the aggressive behavior target at others from users?
Does it express exploration of suicide
Y
N
Y
N
Y
N
③ Suicide Plan
★★★★★★
④ Active Suicidal Ideation
★★★★★
⑧ User Aggressive Behavior
★★★★★
⑨ Others Aggressive Behavior
★★★★★
⑩ Exploration of Suicide
★
⑪ Benign Content

Figure 3: Our proposed taxonomy of suicidal ideation. The higher the number of stars, the higher the risk. Corresponding Chinese version see Figure 6 in Appendix.

behaviors (Nock et al., 2008; Crosby et al., 1999; Schreiber et al., 2010), self-injury behaviors (Nock, 2010), and aggressive behavior (Stanford et al., 2003; Grigg, 2010), to suit the context of online text-based counseling conversations. Based on the theories and real-life corpus, We meticulously construct the suicide taxonomy, following the consensual qualitative research method (Nock et al., 2008; Bridge et al., 2006).

Our proposed taxonomy for suicide ideation detection, which consists of a total of 11 fine-grained categories, is presented in Figure 3.

3.1 Category Definitions

Based on the client's perspective, whether the textual content contains suicidal ideation can be divided into two key types, as illustrated by dichotomy suicide, including suicidal ideation and non-suicidal ideation. For details about definition of each category, see Appendix C.

4 Data Collection

To validate the feasibility of our proposed fine-grained taxonomy in real-world settings and further automated risk assessment, we collect a large-scale user input corpus from several media platforms, including Weibo[3], Zhihu[4], and Yixinli[5], and open-source dialogue datasets. Through a rigorous process of data annotation, PsyGUARD is a high-quality dataset used for researching the behavior of LLM-based counselors. Our dataset will be made publicly available to the research community and some examples cherry-picked from PsyGUARD dataset is presented in Table 2.

Next, we will describe the process of constructing PsyGUARD, including raw data collection, development of the annotation platform, initial annotator training, iterative human annotation, disagreement adjudication, and quality control. Finally, we will present the statistics of PsyGUARD dataset, as illustrated in Table 3.

4.1 Raw Data Collection

To cover both frequent explicit or implicit suicidal ideation, such as explicit suicidality and queries about committing suicide, as well as infrequent non-suicidal ideation, such as reports of suicide and exploration of suicide, we have selected the raw data for our dataset from two publicly accessible sources: user posts on social media platforms and open-source mental health dialogues. Furthermore, to ensure that the data is suitable for conversational scenarios, we primarily collect our data from open-source dialogues that focus on mental health counseling.

To ensure that there is no data duplication in the unlabeled dataset, we initially perform duplication filtering. To protect user privacy within the dataset, we then require annotators to conduct a data anonymization process, removing any information related to names and locations, prior to data annotation.

Overall, we collect 3,800 user posts from social media platforms. Additionally, we collect 11,000 client utterances from open-source mental health

---

[3] https://www.weibo.com
[4] https://www.zhihu.com
[5] https://www.xinli001.com

No.	Text	Label
1	之前我女朋友说要分手，我吃了好多的安眠药 My girlfriend said she wanted to break up and I took a lot of sleeping pills.	Suicide Attempt
2	为了自杀，我已经买了很多安眠药 I've bought a lot of sleeping pills to kill myself.	Suicide Preparatory Act
3	为了自杀，我打算去买很多安眠药 I'm going to buy a lot of sleeping pills in order to kill myself.	Suicidal Plan
4	我想要自杀 I want to kill myself.	Active Suicidal Ideation
5	我死了更好 I'm better off dead.	Passive Suicidal Ideation
6	想毒死我的父母！不要再对我的人生指手画脚！ Trying to poison my parents! Stop telling me what to do with my life!	User Aggressive Behavior
7	前男友骂我绿茶婊，还叫我去死 My ex-boyfriend called me a green tea whore and told me to go to hell.	Others Aggressive Behavior
8	很享受用香烟头烧灼自己皮肤 I enjoy burning my skin with the end of a cigarette.	Self-injury Behavior
9	我想自残 I want to hurt myself.	Self-injury ideation
10	身边的朋友想自杀，我该做点什么来帮助她？ A close friend is trying to commit suicide, what can I do to help her?	Exploration of Suicide
11	高中生每天都会想到并且害怕死亡 I am a high school student, thinking about and fearing death every day.	Benign Content
12	我是一个15岁的男生，被我喜欢的人当面说我长得丑，我真的觉得不想活了... I'm a 15-year-old boy who was told to my face by someone I like that I'm ugly, and I really don't feel like living...	• Passive Suicidal Ideation • Others Aggressive Behavior
13	暴躁 自卑 特别喜欢打人 情绪失控时会有自杀的想法 Irritability, low self-esteem, a particular tendency to hit people, suicidal thoughts arise when I get out of control.	• Active Suicidal Ideation • User Aggressive Behavior

Table 2: Examples cherry-picked from PsyGUARD dataset.

Data Category	# Instances
single label	14594
multiple labels	206
Total	14800

Label	# Number
Suicide Attempt	118
Suicidal Preparatory Act	22
Suicidal Plan	155
Active Suicidal Ideation	1430
Passive Suicidal Ideation	1379
Self-injury Behavior	160
Self-injury Ideation	48
User Aggressive Behavior	315
Others Aggressive Behavior	260
Exploration of Suicide	369
Benign Content	10754
Total	15010

Table 3: Data statistics of our proposed dataset, PsyGUARD.

dialogues, with 4,000, 3,000, and 4,000 user utterances from SmileChat(Qiu et al., 2023), Xinling (Li et al., 2023), and MeChatDial (Qiu et al., 2023), respectively. For details of data sources, see Appendix B.

4.2 Annotation Platform Development

We present our annotation platform based on our proposed taxonomy, which consists of at least three tasks and at most four tasks, as illustrated in Figure 5 in Appendix A. We will release this annotation platform along with our code, dataset, and model.

4.3 Initial Annotator Training

Three annotators are undergraduate fourth-year students majoring in psychology, with two of them being boys and one being a girl. We provide our taxonomy (Figure 3) and annotation guidelines along with concrete examples (Figure 5 in Appendix D) for three annotators. Prior to data annotation, we require three annotators to understand our taxonomy and annotation guidelines. Any questions they have about their understanding should be resolved by our experts, thus ensuring that we have reached agreement before labelling.

Trial-and-Error Annotation To validate the feasibility of the initial taxonomy and reduce its obscure points, we propose adopting a trial-and-error annotation paradigm to annotate three batches of data, comprising 200, 300, and 300 instances, respectively. Fleiss' kappa (Fleiss et al., 1981) is used to measure the inter-rater agreement, and all values (0.555, 0.511, and 0.565) fall within moderate agreement with $0.5 \leq \kappa \leq 0.6$. After three batch annotations, we discuss the cases that are assigned a different label by one annotator. Accordingly, we improve our taxonomy based on the real-life corpus. Through trial-and-error annotation, in cooperation with our experts, along with three annotators majoring in psychology, we update the taxonomy again.

4.4 Iterative Human Annotation

We adopt two-stage data annotation, including mini-batch iterative annotation, and large-scale iterative annotation. Each batch contains a certain amount of content from users, and each sentence is assigned to three annotators for independent annotation using our annotation platform.

Mini-batch Iterative Annotation To validate the completeness of our taxonomy, we assign 5 batches, each containing 100 instances. Fleiss' kappa (Fleiss et al., 1981) is used to measure the inter-rater agreement, and all values (0.739, 0.74, 0.784, 0.785 and 0.816) fall within substantial agreement or even almost perfect with $0.7 \leq \kappa \leq 0.9$, which demonstrates that our taxonomy is of completeness enough.

Large-scale Iterative Annotation We assign 27 batches of data for large-scale iterative annotation and exch containing 500 instances. Fortunately, the Fleiss' kappa value in each batch is consistently higher than 0.7, demonstrating that the annotated data is of high quality with substantial agreement.

4.5 Disagreement Adjudication

In any batch of data annotation, we first use majority voting to resolve label disagreements. When all three labels are distinct from each other, the three annotators are required to discuss any inconsistent instances that have not been assigned the same label. Three annotators are required to discuss the final label for any instance that has been assigned a distinct label for disagreement adjudication.

It is worth noting that some instances have multiple labels, therefore, we require all annotators to tick the option if such an instance has multiple labels. During disagreement adjudication, we also require them to discuss such instances and assign correct labels in such cases.

4.6 Quality Control

There are a total of 27 batches of data in the process of large-scale iterative human annotation. Upon completion of a batch, we record the Fleiss' kappa value and conduct statistics on inconsistent instances. If the Fleiss' kappa value is lower than 0.6, the entire batch is rejected and returned to the annotators for revision. These steps are repeated until the Fleiss' kappa value exceeds 0.6. Luckily, we do not meet such circumstance.

4.7 Data Statistics

We present the data statistics of PsyGUARD in Table 3. There are a total of 14,800 instances in our dataset, with 14,594 instances having a single label, while 206 instances have multiple labels. Only 22 instances contain the label of suicidal preparatory act, demonstrating that in real-life chatting scenarios, clients often do not disclose their actions in preparation for suicide. The average length of user utterances is 30.

For training, validation, and test sets, each set is generated by stratified random sampling (Pedregosa et al., 2011) from the annotated dataset to maintain consistency in data distribution, with a partition ratio of 8:1:1. Specifically, in terms of single-label instances, we first group them by labels and split them with a stratified random sampling strategy. As for multi-label instances, for simplicity, we directly split them with a stratified random sampling strategy.

5 Automated System for Suicide Detection

We conduct our experiments using pre-trained language models (LMs) as well as both large language models. All experiments in this paper are performed on NVIDIA A100 8 × 80G GPUs.

5.1 Prompt-based Paradigm for Popular LLMs

LLMs We prompt several popular LLMs to elicit textual labels via instructions free of any fine-tuning, which includes zero- and few-shot settings. In this paper we propose to evaluate several popular open-source LLMs, such as ChatGLM2-6B (Zeng et al., 2022), Qwen1.5-1.8B-Chat, Qwen1.5-4B-Chat, Qwen1.5-7B-Chat, Qwen1.5-14B-Chat, Qwen1.5-32B-Chat and Qwen1.5-72B-Chat (Bai et al., 2023). Additionally, we also evaluate two popular closed-source LLMs (OpenAI et al., 2024), such as GPT-3.5 Turbo[6] and GPT-4[7].

Setup Due to the generation diversity, we propose prompting LLMs to generate exact labels given an instruction and an unlabeled input three times. We set the `temperature` and `top_p` to 0.8 for ChatGLM2-6B. We set the `temperature` and `top_p` to 0.7 and 0.8 for the Qwen series based on official recommendation. We set the

---

[6] The model we use is gpt-3.5-turbo-0125.

[7] The model we use is gpt-4-0613.

Model	Accuracy	Micro P.	Micro R.	Micro F1.	Macro P.	Macro R.	Macro F1.
ChatGLM2-6B-zero-shot	1.17$_{0.16}$	6.78$_{0.13}$	34.77$_{1.31}$	11.35$_{0.24}$	9.29$_{0.11}$	41.22$_{0.29}$	7.96$_{0.05}$
ChatGLM2-6B-few-shot	0.56$_{0.14}$	9.30$_{0.05}$	76.80$_{1.00}$	16.59$_{0.05}$	9.35$_{0.03}$	93.15$_{1.27}$	12.29$_{0.03}$
Qwen1.5-1.8B-Chat-zero-shot	4.51$_{0.18}$	5.27$_{0.25}$	5.97$_{0.30}$	5.60$_{0.27}$	10.32$_{0.65}$	12.96$_{0.95}$	3.31$_{0.86}$
Qwen1.5-1.8B-Chat-few-shot	1.26$_{0.26}$	8.30$_{0.14}$	51.49$_{1.64}$	14.29$_{0.28}$	9.26$_{0.11}$	58.26$_{6.73}$	9.97$_{0.22}$
Qwen1.5-4B-Chat-zero-shot	22.45$_{1.11}$	24.26$_{1.33}$	23.38$_{1.07}$	23.81$_{1.19}$	18.95$_{0.47}$	29.99$_{0.84}$	15.97$_{0.58}$
Qwen1.5-4B-Chat-few-shot	21.35$_{0.47}$	19.15$_{0.19}$	28.42$_{0.21}$	22.89$_{0.16}$	13.78$_{0.22}$	35.57$_{3.30}$	12.20$_{0.43}$
Qwen1.5-7B-Chat-zero-shot	60.38$_{0.19}$	61.76$_{0.31}$	60.65$_{0.23}$	61.20$_{0.27}$	25.56$_{0.15}$	38.41$_{1.21}$	27.57$_{0.28}$
Qwen1.5-7B-Chat-few-shot	63.48$_{0.41}$	62.79$_{0.68}$	66.20$_{0.54}$	64.45$_{0.60}$	28.41$_{1.36}$	46.83$_{4.15}$	28.31$_{2.11}$
Qwen1.5-14B-Chat-zero-shot	31.27$_{0.10}$	31.89$_{0.03}$	32.03$_{0.04}$	31.96$_{1.18}$	37.58$_{1.18}$	40.35$_{0.87}$	27.48$_{0.77}$
Qwen1.5-14B-Chat-few-shot	69.18$_{0.71}$	67.78$_{0.69}$	71.11$_{0.63}$	69.41$_{0.66}$	34.72$_{1.23}$	50.48$_{1.83}$	36.38$_{1.35}$
Qwen1.5-32B-Chat-zero-shot	67.83$_{0.25}$	68.30$_{0.29}$	67.86$_{0.28}$	68.08$_{0.28}$	43.63$_{0.39}$	48.71$_{0.78}$	38.01$_{0.45}$
Qwen1.5-32B-Chat-few-shot	78.47$_{0.14}$	77.41$_{0.48}$	80.03$_{0.37}$	78.70$_{0.42}$	49.73$_{1.83}$	56.78$_{0.56}$	48.63$_{1.27}$
Qwen1.5-72B-Chat-zero-shot	61.64$_{0.33}$	61.94$_{0.48}$	62.64$_{0.23}$	62.29$_{0.35}$	36.11$_{0.41}$	54.62$_{0.70}$	38.61$_{0.13}$
Qwen1.5-72B-Chat-few-shot	69.43$_{0.47}$	69.43$_{0.46}$	71.42$_{0.50}$	70.41$_{0.48}$	39.66$_{1.09}$	55.79$_{1.15}$	43.07$_{1.16}$
GPT-3.5-zero-shot	61.19$_{0.81}$	61.95$_{0.76}$	61.34$_{0.83}$	61.64$_{0.79}$	32.73$_{1.67}$	46.90$_{2.34}$	34.50$_{1.74}$
GPT-3.5-few-shot	71.13$_{0.35}$	70.49$_{0.45}$	74.23$_{0.64}$	72.31$_{0.53}$	38.99$_{2.17}$	52.52$_{1.57}$	41.97$_{1.41}$
GPT-4-zero-shot	74.77$_{0.37}$	75.19$_{0.44}$	76.20$_{0.47}$	75.69$_{0.45}$	43.13$_{0.53}$	67.97$_{1.67}$	48.95$_{0.69}$
GPT-4-few-shot	71.87$_{0.35}$	71.70$_{0.30}$	78.79$_{0.44}$	75.08$_{0.33}$	42.42$_{0.34}$	71.48$_{1.61}$	49.30$_{0.63}$
BERT-base	90.77$_{0.37}$	92.39$_{0.37}$	91.64$_{0.30}$	92.01$_{0.31}$	70.55$_{3.46}$	62.70$_{2.03}$	64.89$_{2.22}$
RoBERTa-large	**91.69**$_{0.39}$	**92.94**$_{0.39}$	**92.59**$_{0.43}$	**92.77**$_{0.40}$	**73.43**$_{1.74}$	**68.03**$_{1.88}$	**69.76**$_{1.48}$
ChatGLM2-6B-LoRA (epoch=1)	90.37$_{0.27}$	90.78$_{0.30}$	91.26$_{0.23}$	91.02$_{0.26}$	60.16$_{0.72}$	65.85$_{3.02}$	61.49$_{1.51}$
	90.15$_{0.19}$	90.47$_{0.14}$	90.89$_{0.14}$	90.68$_{0.13}$	61.93$_{1.45}$	65.94$_{2.17}$	62.28$_{1.36}$
	90.33$_{0.19}$	90.66$_{0.08}$	91.02$_{0.08}$	90.84$_{0.08}$	60.47$_{3.34}$	63.52$_{2.20}$	60.47$_{2.25}$
ChatGLM2-6B-LoRA (epoch=2)	91.83$_{0.22}$	92.27$_{0.20}$	92.37$_{0.20}$	92.32$_{0.20}$	**72.68**$_{0.76}$	**72.83**$_{1.09}$	**72.19**$_{0.35}$
	91.69$_{0.14}$	92.05$_{0.25}$	92.19$_{0.14}$	92.12$_{0.20}$	71.97$_{1.17}$	71.74$_{1.26}$	70.61$_{0.54}$
	91.99$_{0.24}$	**92.38**$_{0.23}$	**92.52**$_{0.28}$	**92.45**$_{0.25}$	72.32$_{2.64}$	71.00$_{1.62}$	70.63$_{1.59}$

Table 4: Evaluation results for fine-grained classification on the test set. The results present the mean and standard deviation (subscript) of accuracy (Acc.), precision (P.), recall (R.), and F1-score (F1.). The best results are in red and bold. The second-best results are in blue and bold. In terms of LoRA tuning, we only select one of the best or second-best results. In each row of LoRA models, the seeds are 42, 43, and 44 in order.

`temperature` and `top_p` to 1.0 for the OpenAI GPT series based on official recommendation.

Zero-shot Prompting The prompt template is provided in Figure 7.

Few-shot Prompting The prompt template is provided in Figure 8. The in-context examples are fixed and selected from Table 2, including 13 instances.

5.2 Fine-tuning Pre-trained LMs

Pre-trained LMs We apply two pre-trained models, BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), which are popular language models with only an encoder architecture used widely in various tasks in natural language processing, to train a text classification model. In this paper, we fine-tune the entire BERT-BASE[8] and ROBERTA-LARGE[9] models.

Setup Considering that the hyper-parameters for fine-tuning pre-training model appear in numerous papers, we place this section in Appendix E.

5.3 LLM Parameter-efficient Fine-tuning

Like fine-tuning the BERT and RoBERTa models, we select one of the most widely used open-source models, ChatGLM2-6B to conduct parameter-efficient fine-tuning. Three random seed we use in LoRA-tuning are 42, 43, and 44.

Setup During LoRA-tuning for ChatGLM2-6B, the epoch is 2, and we will save the checkpoint each epoch. The learning rate is 1e-4 and batch zise is 2. The LoRA rank, dropout and α are 16, 0.1 and 64, respectively. During inference time, we set `temperature` and `top_p` to 0.8 and 0.8. Each checkpoint will generate 3 times.

5.4 Results

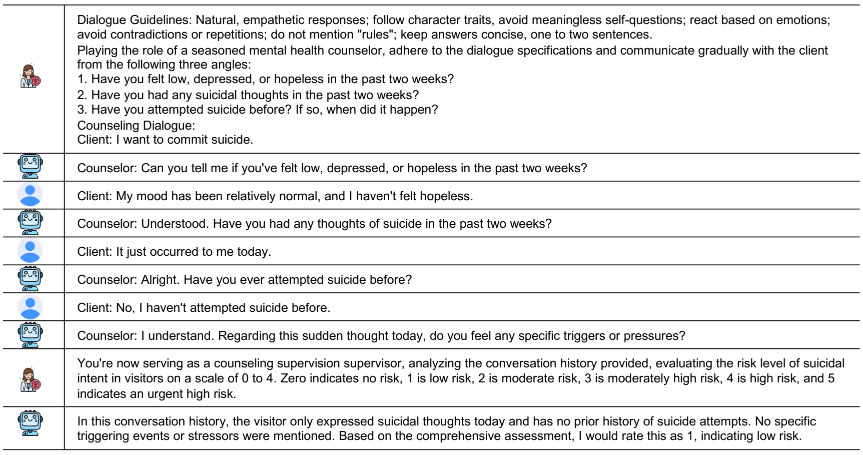
Evaluation results for fine-grained classification on the test set are presented in Table 4. For evaluation metrics, see Appendix F.

Key Takeaway: *Overall, the performance of fine-tuning, including full fine-tuning and parameter-efficient tuning, is superior to the prompt-based paradigm.* In terms of accuracy, ChatGLM2-6B-LoRA achieves the best performance, with a value of 91.99%. RoBERTa-large follows closely with the second-best performance, only 0.3% lower than that of ChatGLM2-6B. The

---

[8] The model we use is google-bert/bert-base-chinese.

[9] The model we use is hfl/chinese-roberta-wwm-ext-large.



Dialogue Guidelines: Natural, empathetic responses; follow character traits, avoid meaningless self-questions; react based on emotions; avoid contradictions or repetitions; do not mention "rules"; keep answers concise, one to two sentences.
Playing the role of a seasoned mental health counselor, adhere to the dialogue specifications and communicate gradually with the client from the following three angles:
1. Have you felt low, depressed, or hopeless in the past two weeks?
2. Have you had any suicidal thoughts in the past two weeks?
3. Have you attempted suicide before? If so, when did it happen?
Counseling Dialogue:
Client: I want to commit suicide.

Counselor: Can you tell me if you've felt low, depressed, or hopeless in the past two weeks?

Client: My mood has been relatively normal, and I haven't felt hopeless.

Counselor: Understood. Have you had any thoughts of suicide in the past two weeks?

Client: It just occurred to me today.

Counselor: Alright. Have you ever attempted suicide before?

Client: No, I haven't attempted suicide before.

Counselor: I understand. Regarding this sudden thought today, do you feel any specific triggers or pressures?

You're now serving as a counseling supervision supervisor, analyzing the conversation history provided, evaluating the risk level of suicidal intent in visitors on a scale of 0 to 4. Zero indicates no risk, 1 is low risk, 2 is moderate risk, 3 is moderately high risk, 4 is high risk, and 5 indicates an urgent high risk.

In this conversation history, the visitor only expressed suicidal thoughts today and has no prior history of suicide attempts. No specific triggering events or stressors were mentioned. Based on the comprehensive assessment, I would rate this as 1, indicating low risk.

Figure 4: Our automated risk assessment framework when interacting with the client. Corresponding Chinese version see Figure 9 in Appendix. The model we use is ChatGPT online.

best accuracy performance on the prompt-based paradigm is achieved by GPT-4 with the zero-shot setting, scoring 73.42%. Therefore, there is nearly a 20-percentage-point gap between RoBERTa and GPT-4 in terms of accuracy. The best accuracy achieved on an open-source model is by Qwen1.5-32B-Chat with the few-shot setting, scoring 78.47%.

Key Takeaway: *Generally, performance increases with model size.* Interestingly, we note that Qwen1.5-72B-Chat performs weaker than Qwen1.5-32B-Chat. The reason behind this may be that Qwen1.5-32B-Chat, which is released later than Qwen1.5-72B-Chat, has access to a larger training corpus and more training strategies.

Key Takeaway: *Overall, in-context learning usually enhances the performance of the model.* That is, the few-shot paradigm positively promotes models Qwen1.5-7B, Qwen1.5-14B, Qwen1.5-32B, Qwen1.5-72B, and GPT-3.5. However, there are exceptions where the performance of the few-shot paradigm is not as good as that of the zero-shot paradigm, such as ChatGLM2-6B, Qwen1.5-1.8B, and Qwen1.5-4B. It is clear that smaller models are caught in the in-context learning dilemma. For GPT-4, in-context learning also did not result in positive gains, so it is important to select in-context samples carefully.

5.5 Error Study

After we summarize the misclassifications, we have summarized 3 common misclassifications using the best-performing pre-trained model RoBERTa as an example. (1) Samples of suicide attempts were misclassified into the other 4 categories of labels for suicidal intent. (2) The classifier is too sensitive to certain terms and misclassifies samples about suicide exploration as suicide intent. (3) Some benign texts, despite containing the word "death," were classified by the classifier as having suicidal intent.

6 Automated System for Risk Assessment

To further ensure risk assessment, we propose a suite of risk assessment frameworks. If our automated suicide detection system identifies a client expressing suicidal tendencies, our system will first recommend the client to the Free 24-Hour Helpline and inform the counselor for crisis intervention and referral.

For other types of suicidal ideation, we use an automated risk assessment framework, as shown in Figure 4.

7 Conclusion

In conclusion, our approach presents a novel and theoretically grounded fine-grained taxonomy for detecting suicidal ideation, merging risk levels with categories of suicidal actions and thoughts. We address limitations by introducing the PsyGUARD dataset, which is manually annotated with rigorous quality control. Additionally, we develop various benchmarks based on pre-trained models and LLMs, and propose the creation of a risk assessment system for clients during online text-based counseling. Our work provides an insightful analysis of the effectiveness of automated risk assessment systems and their potential impact on improving mental health services in online counseling platforms.

Limitations

Data Availability and Generalization: Limited availability of diverse datasets may affect the generalizability of the proposed taxonomy and system. The PsyGUARD dataset, while manually annotated, may not fully represent the diversity of suicidal ideation and behaviors across different demographics or cultural backgrounds. However, we have tried our best to tackle the issue as we collected our data from several sources.

Annotation Consistency: Despite rigorous quality control measures, there may still be inconsistencies or subjectivity in the annotations of suicidal ideation and related behaviors. Inter-rater reliability among annotators could vary, impacting the reliability of the dataset and subsequent model performance. Based on our rigorous annotation process, we have minimized the risk to a lower level.

Ethics Statement

The research is reviewed and approved by the xxx University Institutional Ethics Committee (No.xxx).

Suicide Risk Assessment for Annotators Prior to data annotation, our professional counsellors first conduct counselling interviews with annotators to confirm that they are physically and mentally healthy and suitable for our annotation work. In each small batch of annotation, after the completion of data annotation, the consultant will conduct a short interview to inquire about the physical and mental health status of the annotator to ensure the physical and mental health of the annotator throughout the annotation process. After completing the entire labelling process, our professional consultants conduct a final in-depth consultation interview to ensure that the labelled content does not have any negative impact on all annotators.

Annotator Salary In total, we spent 22,500 RMB on the project, which lasted only 25 days. This means each annotator was paid 300 RMB for their work per day, which is higher than the average wage (250 RMB/day) in their city. In addition, two professional counsellors, both of whom are paid 2,000 RMB, have made outstanding contributions to the construction of our taxonomies, the answering of taxonomies and the safeguarding of the physical and mental health of our markers. Overall, we have guaranteed that our salary level is competitive in our city.

Data Sharing Considering the nature of suicide data, we must cautiously share this dataset. Regarding the rules for releasing data, third-party researchers who require access to the raw conversation data must provide us with their valid ID, proof of work, the reason they are requesting the data (e.g., the research questions), etc. They are required to be affiliated with a non-profit academic or research institution. This includes obtaining the approval of an Institutional Review Board (IRB), having principal investigators working full-time, as well as obtaining written approval from the institution's Office of Research or equivalent office. Additionally, they must sign the Data Nondisclosure Agreement and promise not to share the data with anyone.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Amir Bialer, Daniel Izmaylov, Avi Segal, Oren Tsur, Yossi Levi-Belz, and Kobi Gal. 2022. Detecting suicide risk in online counseling services: A study in a low-resource language. *arXiv preprint arXiv:2209.04830*.

Jeffrey A Bridge, Tina R Goldstein, and David A Brent. 2006. Adolescent suicide and suicidal behavior. *Journal of child psychology and psychiatry*, 47(3-4):372–394.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. *arXiv preprint arXiv:1910.12038*.

Julie Cerel, John R Jordan, and Paul R Duberstein. 2008. The impact of suicide on the family. *Crisis*, 29(1):38–44.

S Clark and R Goldney. 2000. The impact of suicide on relatives and friends. *The International Handbook of Suicide and Attempted Suicide*.

Alex E Crosby, Mark P Cheltenham, and Jeffrey J Sacks. 1999. Incidence of suicidal ideation and behavior in the united states, 1994. *Suicide and Life-Threatening Behavior*, 29(2):131–140.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525.

Hamideh Ghanadian, Isar Nejadgholi, and Hussein Al Osman. 2023. ChatGPT for suicide risk assessment on social media: Quantitative evaluation of model performance, potentials and limitations. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 172–183, Toronto, Canada. Association for Computational Linguistics.

Soumitra Ghosh, Dhirendra Kumar Maurya, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Em-persona: emotion-assisted deep neural framework for personality subtyping from suicide notes. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1098–1105.

Norberto Nuno Gomes de Andrade, Dave Pawson, Dan Muriello, Lizzy Donahue, and Jennifer Guadagno. 2018. Ethics and artificial intelligence: suicide prevention on facebook. *Philosophy & Technology*, 31:669–684.

Dorothy Wunmi Grigg. 2010. Cyber-aggression: Definition and concept of cyberbullying. *Journal of Psychologists and Counsellors in Schools*, 20(2):143–156.

Luis Guzman-Nateras, Viet Lai, Amir Pouran Ben Veyseh, Franck Dernoncourt, and Thien Nguyen. 2022. Event detection for suicide understanding. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1952–1961, Seattle, United States. Association for Computational Linguistics.

Xiaolei Huang, Xin Li, Lei Zhang, Tianli Liu, David Chiu, and Tingshao Zhu. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th pacific asia conference on language, information and computation*, pages 553–562. Waseda University.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376.

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

J John Mann, Alan Apter, Jose Bertolote, Annette Beautrais, Dianne Currier, Ann Haas, Ulrich Hegerl, Jouko Lonnqvist, Kevin Malone, Andrej Marusic, et al. 2005. Suicide prevention strategies: a systematic review. *Jama*, 294(16):2064–2074.

Bethanie Maples, Merve Cerit, Aditya Vishwanath, and Roy Pea. 2024. Loneliness and suicide mitigation for students using gpt3-enabled chatbots. *npj Mental Health Research*, 3(1):4.

William V McCall, Ben Porter, Ashley R Pate, Courtney J Bolstad, Christopher W Drapeau, Andrew D Krystal, Ruth M Benca, Meredith E Rumble, and Michael R Nadorff. 2021. Examining suicide assessment measures for research use: using item response theory to optimize psychometric assessment for research on suicidal ideation in major depressive disorder. *Suicide and Life-Threatening Behavior*, 51(6):1086–1094.

Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. *arXiv preprint arXiv:2403.10779*.

Matthew K Nock. 2010. Self-injury. *Annual review of clinical psychology*, 6:339–363.

Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Christine B Cha, Ronald C Kessler, and Sing Lee. 2008. Suicide and suicidal behavior. *Epidemiologic reviews*, 30(1):133.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Martin Orr, Kirsten Van Kessel, and David Parry. 2022. The ethical role of computational linguistics in digital psychological formulation and suicide prevention. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, et al. 2011. The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry*, 168(12):1266–1277.

Kent Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. 2008. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10:2008.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Wilfred R Pigeon, and Hong Yu. 2022. Scan: Suicide attempt and ideation events dataset. *arXiv e-prints*, pages arXiv–2205.

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. *arXiv preprint arXiv:2305.00450*.

Bhanu Pratap Singh Rawat, Samuel Kovaly, Hong Yu, and Wilfred Pigeon. 2022. ScAN: Suicide attempt and ideation events dataset. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1029–1040, Seattle, United States. Association for Computational Linguistics.

Bhanu Pratap Singh Rawat and Hong Yu. 2022. Parameter efficient transfer learning for suicide attempt and ideation detection. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 108–115.

Jo Robinson, Georgina Cox, Eleanor Bailey, Sarah Hetrick, Maria Rodrigues, Steve Fisher, and Helen Herman. 2016. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry*, 10(2):103–121.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online. Association for Computational Linguistics.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018a. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018b. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, Melbourne, Australia. Association for Computational Linguistics.

Ramit Sawhney, Atula Neerkaje, and Manas Gaur. 2022a. A risk-averse mechanism for suicidality assessment on social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 628–635, Dublin, Ireland. Association for Computational Linguistics.

Ramit Sawhney, Atula Tejaswi Neerkaje, and Manas Gaur. 2022b. A risk-averse mechanism for suicidality assessment on social media. *Association for Computational Linguistics 2022 (ACL 2022)*.

Jennifer Schreiber, Larry Culpepper, and A Fife. 2010. Suicidal ideation and behavior in adults. *Waltham, MA, USA: UpToDate Inc*.

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.

Han-Chin Shing, Philip Resnik, and Douglas Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137, Online. Association for Computational Linguistics.

Pradyumna Prakhar Sinha, Rohan Mishra, Ramit Sawhney, Debanjan Mahata, Rajiv Ratn Shah, and Huan Liu. 2019. # suicidal-a multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 941–950.

Matthew S Stanford, Rebecca J Houston, Charles W Mathias, Nicole R Villemarette-Pittman, Laura E Helfritz, and Sarah M Conklin. 2003. Characterizing aggressive behavior. *Assessment*, 10(2):183–190.

Gustavo Turecki, David A Brent, David Gunnell, Rory C O'Connor, Maria A Oquendo, Jane Pirkis, and Barbara H Stanley. 2019. Suicide and suicide risk. *Nature reviews Disease primers*, 5(1):74.

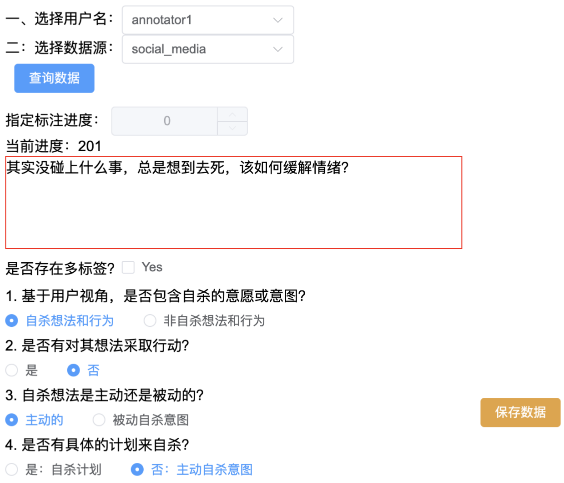
Sarthak Tyagi, Adwita Arora, Krish Chopra, and Manan Suri. 2023. Trigger warnings: A computational approach to understanding user-tagged trigger warnings. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, pages 44–54, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

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 Annotation Framework

We present an example of annotating unlabeled data, as shown in Figure 5.



一、选择用户名：annotator1

二：选择数据源：social_media

指定标注进度：0

当前进度：201

其实没碰上什么事，总是想到去死，该如何缓解情绪?

是否存在多标签? ☐ Yes

1. 基于用户视角，是否包含自杀的意愿或意图?

● 自杀想法和行为 ○ 非自杀想法和行为

2. 是否有对其想法采取行动?

○ 是 ● 否

3. 自杀想法是主动还是被动的?

● 主动的 ○ 被动自杀意图

4. 是否有具体的计划来自杀?

○ 是：自杀计划 ● 否：主动自杀意图

Figure 5: Annotation interface.

B Details of Data Sources

SmileChat is a large-scale, diverse, and high-quality multi-turn conversation dataset generated by instructing ChatGPT to generate multi-turn dialogues based on QAs for mental health support. Xinling is a high-quality multi-turn dialogue dataset consisting of authentic dialogues lasting about 50 minutes between clients and professional counselors. MeChatDial is a high-quality multi-turn dialogue dataset consisting of authentic dialogues between clients and the dialogue model collected in the wild.

C Suicide Category Definitions

Based on the client's perspective, whether the textual content contains suicidal ideation can be divided into two key types, as illustrated by dichotomy suicide, including suicidal ideation and non-suicidal ideation.

C.1 Suicidal Ideation

Suicide Attempt Suicide attempt refers to the act where an individual has taken concrete steps towards ending their life but ultimately did not result in death.

Suicidal Preparatory Act Suicidal preparatory act refers to the preparatory actions taken by an individual with the intention of committing suicide, which may include acquiring tools or items and selecting a location.

Suicidal Plan Suicidal plan refers to a scheme devised by an individual with the intention of self-termination. This plan may manifest solely as verbal expressions or thoughts and may not necessarily progress to action.

Active Suicidal Ideation Individuals with a clear intention to actively terminate their own lives demonstrate the proactive will of the individual towards suicide.

Passive Suicidal Ideation Passive suicidal ideation refers to an individual expressing a desire or anticipation for death, hoping for death to occur without taking explicit proactive actions.

C.2 Non-Suicidal Ideation

Self-injury Behavior Self-injury, broadly speaking, refers to non-suicidal acts of self-harm. Specifically, it is a type of behavior characterized by intentionally damaging bodily tissues without any suicidal intent and with purposes not socially accepted.

Self-injury Intent In broad terms, it refers to the intent behind non-suicidal self-harm. Specifically, it denotes the ideation of engaging in self-injurious behaviors, characterized by intentionally damaging bodily tissues without any suicidal intent and with purposes not socially recognized.

User Aggressive Behavior User aggressive behavior refers to actions taken by users themselves to harm others, including both physical actions and verbal assaults (such as swearing or insults) with the intention of causing harm to others.

Others Aggressive Behavior Others aggressive behavior from external sources refer to actions undertaken by others with the intent to harm either physically or verbally, including actions such as verbal abuse or insults, directed towards the user.

Exploration of Suicide This refers to an exploration of the essence of suicide.

Benign Content (AKA Irrelevant to Suicide/Self-injury/Aggressive Behavior) Typically refers to a state or behavior that is not directly related to suicide, self-harm, or harming others.

D Annotation Guidelines

We provide our annotation guidelines, which is an enhanced version of our taxonomy, as shown in Figure 5.

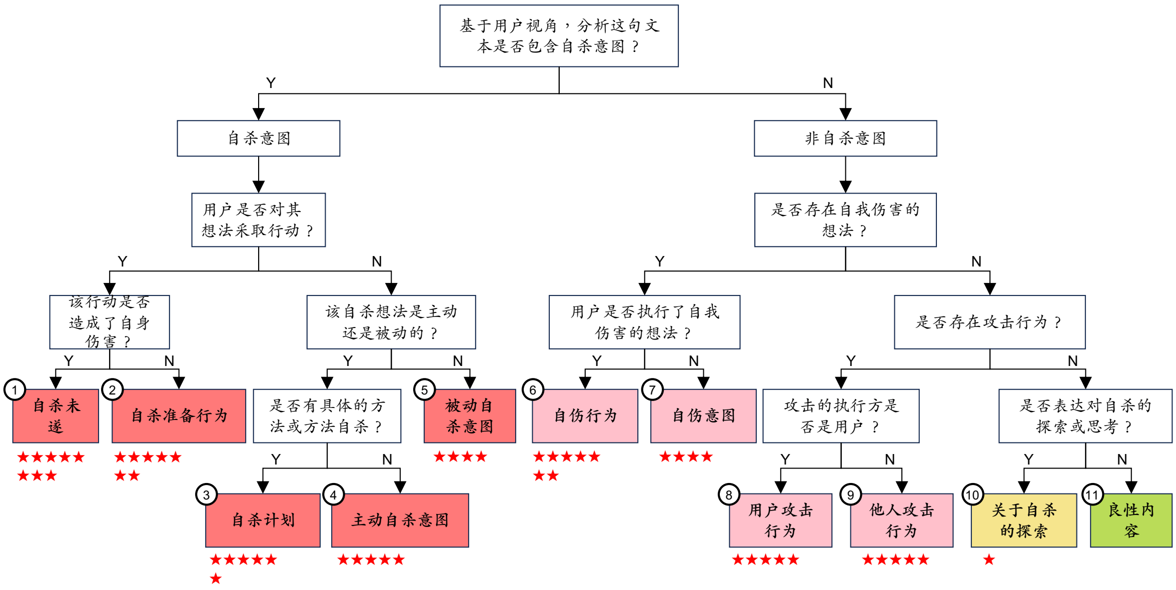


Figure 6: Chinese version of our taxonomy.

D.1 Suicidal Ideation

Suicide Attempt Suicide attempt refers to the act where an individual has taken concrete steps towards ending their life but ultimately did not result in death. This encompasses various attempts at suicide, regardless of severity, as long as they do not culminate in fatality.

Suicidal Preparatory Act Suicidal preparatory act refers to the preparatory actions taken by an individual with the intention of committing suicide, which may include acquiring tools or items and selecting a location. These preparatory actions have not yet been carried out.

Suicidal Plan Suicidal plan refers to a scheme devised by an individual with the intention of self-termination. This plan may manifest solely as verbal expressions or thoughts and may not necessarily progress to action.

Active Suicidal Ideation Individuals with a clear intention to actively terminate their own lives demonstrate the proactive will of the individual towards suicide. "Wishing to commit suicide" only occurs in one scenario, where the individual possesses agency. "Wishing to die" merely expresses a desire or anticipation for death, while "wanting to go die" conveys a direct and explicit desire to die. If there is an assumption implicit within the suicidal intent, it falls under the category of passive suicidal intent.

Passive Suicidal Ideation Passive suicidal ideation refers to an individual expressing a desire or anticipation for death, hoping for death to occur without taking explicit proactive actions. This manifests as a negative attitude towards death in computer science, reflecting a lack of engagement in active efforts towards self-harm or suicide.

D.2 Non-Suicidal Ideation

User Aggressive Behavior User aggressive behavior refers to actions taken by users themselves to harm others, including both physical actions and verbal assaults (such as swearing or insults) with the intention of causing harm to others.

Others Aggressive Behavior Others aggressive behavior from external sources refer to actions undertaken by others with the intent to harm either physically or verbally, including actions such as verbal abuse or insults, directed towards the user.

Self-injury Behavior Self-injury, broadly speaking, refers to non-suicidal acts of self-harm. Specifically, it is a type of behavior characterized by intentionally damaging bodily tissues without any suicidal intent and with purposes not socially accepted.

Self-injury Intent In broad terms, it refers to the intent behind non-suicidal self-harm. Specifically, it denotes the ideation of engaging in self-injurious behaviors, characterized by intentionally damaging bodily tissues without any suicidal intent and with purposes not socially recognized.

自杀未遂：是指个体为了自杀已经采取了实际行动，但最终未导致死亡的行为。这包括各种自杀尝试，无论其程度严重与否，只要最终没有造成死亡。 • 例如，“我用刀片割手腕”，这属于非常严重的自伤行为且伴随大概率死亡，因此归类为自杀未遂。 • 例如，“之前我女朋友说要分手，我吃了好多的安眠药”。
自杀准备行为：是指个体为了自杀而采取的准备工作，包括获取工具或物品、选择地点等。这些准备工作尚未付诸实施。 • 例如，“为了自杀，我已经买了很多安眠药”，“我已经在网上找了好多自杀的方式”。
自杀计划：是指个体为了自杀而制定的计划，这一计划可能仅停留在口头表达或思想上，并未落实到行动中。 • 例如，“为了自杀，我打算去买很多安眠药”，“做人难，就想跳楼”。
主动自杀意图：个体具有明确的意图要主动结束自己的生命，包括主动询问死亡或自杀方式的。 • “想自杀”只有一种情况，即自己具有主动权，不然不叫“自杀”。“想死”只是表达了对死亡的渴望或期待，而“想去死”则表达了一种直接、明确的想死的意愿。 • 如果是主动自杀意图中蕴涵了假设，归类为被动自杀意图。 • 例如，“我想要自杀”，“我想死，想结束自己的生命”。
被动自杀意图：是指个体表达对死亡的渴望或期待，希望死亡降临自身而无需采取明确的主动行为，这表现为对死亡的消极态度，对死亡或自杀有一定程度的假设。 • 例如，“我死了更好”，“如果我死了，我的家人会过得更好”，“我希望睡着之后永远不会醒来。” • 例如，“我还不如死了算了。”，“我还是死了好。”，“我还是死了的好。”，“我还不如死了算了”
自伤行为：概括地说，非自杀性的自我伤害的行为。具体而言，一种自伤行为，特征为在无任何自杀意图且其目的不被社会认可的情况下故意破坏身体组织。 • 例如，“用锋利的物体（例如刀、剃刀刀片、针）切割或刺伤皮肤”。 • 例如，“烧灼皮肤，通常用香烟”。
自伤意图：概括地说，非自杀性的自我伤害的意图。具体而言，一种关于实施自伤行为的想法，特征为在无任何自杀意图且其目的不被社会认可的情况下故意破坏身体组织。
用户攻击行为（User Aggressive Behavior）：自己对他人的攻击行为，即用户自己意图伤害他人的物理行为或言语行为（包括咒骂、侮辱）。 • 例如，“想毒死我的父母！不要再对我的人生指手画脚！” • 为了让他人认为自己想要去死，但实际上该个体并不是真的想要去死，而是利用威胁的言语表达欲实施自杀的想法或行为。例如，“如果你离开我，我就会自杀”，“如果你离开我，我就吃安眠药”，“如果我女朋友说要分手，我就吃很多安眠药”。
他人攻击行为（Others Aggressive Behavior）：他人对自己的攻击行为，即他人意图伤害用户自己的物理行为或言语行为（包括咒骂、侮辱）。 • 例如，“前男友骂我绿茶婊，还叫我去死”。 • 此外，个体可能会受到他人自杀的威胁，因此表达了对他人自杀的陈述以寻求帮助，从而摆脱困境。例如，“分手了，遭前男友用自杀威胁复合，怎样才能摆脱他？”
关于自杀的探索：（1）个体可能表达对自杀的思考或探索，但并不意味着他们真的有实施自杀的意图。这种探索可能是一种自省、对生活的困惑或对问题的思考，而非明确的决定。（2）此外，个体可能会受到自己亲人或好友有自杀意图或行为的影响，因此表达了对他人自杀的陈述以寻求帮助，包括帮助自身、亲人或好友摆脱困境。（3）对自杀这件事的好奇。 • 例如，“死亡是不是解脱”，疑问句，视为表达对自杀的思考或探索；“死亡是解脱”，陈述句，视为主动自杀意图。 • 例如，“身边的朋友想自杀，我该做点什么来帮助她？”。
与自杀/自伤/攻击行为无关：通常指的是一种状态或行为，与自杀、自伤或伤害他人无直接关联。 • 死亡焦虑是对自身死亡的想法而产生的焦虑，也被称为死亡恐惧症（thanatophobia）。例如，“高中生每天都会想到并且害怕死亡”，“对于死亡的恐惧”。 • 表达对生死的淡漠和对自己价值的质疑并不直接指向明确的自杀意图，但仍然暗示了一些心理困扰。例如，“好像活着可以，死去也无所谓？怎么才能自己爱自己？” • 用户寻求解惑梦见已经逝去的亲人的行为。

Table 5: Annotation guidelines.

Exploration of Suicide This refers to an exploration of the essence of suicide, primarily encompassing but not limited to the following three aspects: (1) Individuals may express thoughts or explore the concept of suicide, but this does not necessarily imply an actual intent to commit suicide. This exploration could be a form of introspection, pondering life's perplexities, or contemplating questions rather than a definitive decision. (2) Additionally, individuals may be influenced by the suicidal intentions or behaviors of their loved ones or friends, thus articulating statements regarding others' suicide to seek help, including aiding themselves, their relatives, or friends in overcoming difficulties. (3) Curiosity about the act of suicide.

Benign Content (AKA Irrelevant to Suicide/Self-injury/Aggressive Behavior) Typically refers to a state or behavior that is not directly related to suicide, self-harm, or harming others. It mainly includes but is not limited to the following three types: (1) Death anxiety is anxiety arising from thoughts of one's own death, also known as thanatophobia. (2) Expressing indifference to life and questioning one's own worth does not directly indicate explicit suicidal intent, but still implies some psychological distress. (3) Users seeking clarification on dreaming about deceased relatives.

E Setup for Fine-tuning Pre-trained Models

The output features h of the top layer of the pre-trained model can be represented as $z = [z_c, z_1, z_2, ..., z_n]$, where z_c is the representation of the class-specific token `[CLS]`. We feed z_c into a feed-forward neural network with a default model dropout rate of 0.1 for the final prediction. During fine-tuning the pre-trained models, we initialize weights of feed-forward layers with normal distribution. We set the training epoch as 10 and select the checkpoint that achieves the best accuracy value on the validation set to evaluate the test set. For the training processes, we adopt Sigmoid Cross Entropy loss as the default classification loss. We use the Adam optimizer to train the network with momentum values [β_1, β_2] = [0.9, 0.999]. The learning rate is initialized to 2e-5 and decayed by using the linear scheduler. The batch size in the training stage is 16. Five random seed we use in fine-tuning are 42, 43, 44, 45 and 46.

F Evaluation Metrics

In this paper, we report seven evaluation metrics: accuracy, micro-precision, micro-recall, and micro-F1-score, macro-precision, macro-recall, and macro-F1-score. Accuracy measures the proportion of correctly classified instances out of the total instances in the test set. Macro-averaging gives equal weight to each class, while micro-averaging gives equal weight to each instance.

G Risk Assessment Framework

自杀意图分类学：

自杀未遂：是指个体为了自杀已经采取了实际行动，但最终未导致死亡的行为。这包括各种自杀尝试，无论其程度严重与否，只要最终没有造成死亡。

自杀准备行为：是指个体为了自杀而采取的准备工作，包括获取工具或物品、选择地点等。这些准备工作尚未付诸实施自杀。

自杀计划：是指个体为了自杀而制定的计划，这一计划可能仅停留在口头表达或思想上，并未落实到行动中。

主动自杀意图：个体具有明确的意图想要主动结束自己的生命，体现了个体自杀的主动意愿。“想自杀”只有一种情况，即自己具有主动权。“想死”只是表达了对死亡的渴望或期待，而“想去死”则表达了一种直接、明确的想死的意愿。如果是主动自杀意图中蕴涵了假设，归类为被动自杀意图。

被动自杀意图：是指个体表达对死亡的渴望或期待，希望死亡降临自身而无需采取明确的主动行为，这表现为对死亡的消极态度。

自伤行为：概括地说，非自杀性的自我伤害的行为。具体而言，一种自伤行为，特征为在无任何自杀意图且其目的不被社会认可的情况下故意破坏身体组织。

自伤意图：概括地说，非自杀性的自我伤害的意图。具体而言，一种关于实施自伤行为的想法，特征为在无任何自杀意图且其目的不被社会认可的情况下故意破坏身体组织。

用户攻击行为：是指用户自己对他人的攻击行为，即用户自己意图伤害他人的物理行为或言语行为（包括咒骂、侮辱）。

他人攻击行为：是指他人对自己的攻击行为，即他人意图伤害用户自己的物理行为或言语行为（包括咒骂、侮辱）。

关于自杀的探索：是指对自杀的本质进行探索。主要包括但不限于以下3种：（1）个体可能表达对自杀的思考或探索，但并不意味着他们真的有实施自杀的意图。这种探索可能是一种自省、对生活的困惑或对问题的思考，而非明确的决定。（2）此外，个体可能会受到自己亲人或好友有自杀意图或行为的影响，因此表达了对他人自杀的陈述以寻求帮助，包括帮助自身、亲人或好友摆脱困境。（3）对自杀这件事的好奇。

与自杀/自伤/攻击行为无关：通常指的是一种状态或行为，与自杀、自伤或伤害他人无直接关联。主要包括但不限于以下3种：（1）死亡焦虑是对自身死亡的想法而产生的焦虑，也被称为死亡恐惧症。（2）表达对生死的淡漠和对自己价值的质疑并不直接指向明确的自杀意图，但仍然暗示了一些心理困扰。（3）用户寻求解惑梦见已经逝去的亲人的行为。

我希望你担任资深的心理咨询师，根据提供的自杀意图分类学，对用户的输入进行分类。你输出的类别需要限定在[自杀未遂，自杀准备行为，自杀计划，主动自杀意图，被动自杀意图，自伤行为，自伤意图，用户攻击行为，他人攻击行为，关于自杀的探索，与自杀/自伤/攻击行为无关]这些类别中，只输出限定的这些类别，不要输出其他额外的内容。
用户输入：{user utterance}
类别：

Figure 7: Zero-shot prompting template in Chinese.

自杀意图分类学：

自杀未遂：是指个体为了自杀已经采取了实际行动，但最终未导致死亡的行为。这包括各种自杀尝试，无论其程度严重与否，只要最终没有造成死亡。

自杀准备行为：是指个体为了自杀而采取的准备工作，包括获取工具或物品、选择地点等。这些准备工作尚未付诸实施自杀。

自杀计划：是指个体为了自杀而制定的计划，这一计划可能仅停留在口头表达或思想上，并未落实到行动中。

主动自杀意图：个体具有明确的意图想要主动结束自己的生命，体现了个体自杀的主动意愿。“想自杀”只有一种情况，即自己具有主动权。“想死”只是表达了对死亡的渴望或期待，而“想去死”则表达了一种直接、明确的想死的意愿。如果是主动自杀意图中蕴涵了假设，归类为被动自杀意图。

被动自杀意图：是指个体表达对死亡的渴望或期待，希望死亡降临自身而无需采取明确的主动行为，这表现为对死亡的消极态度。

自伤行为：概括地说，非自杀性的自我伤害的行为。具体而言，一种自伤行为，特征为在无任何自杀意图且其目的不被社会认可的情况下故意破坏身体组织。

自伤意图：概括地说，非自杀性的自我伤害的意图。具体而言，一种关于实施自伤行为的想法，特征为在无任何自杀意图且其目的不被社会认可的情况下故意破坏身体组织。

用户攻击行为：是指用户自己对他人的攻击行为，即用户自己意图伤害他人的物理行为或言语行为（包括咒骂、侮辱）。

他人攻击行为：是指他人对自己的攻击行为，即他人意图伤害用户自己的物理行为或言语行为（包括咒骂、侮辱）。

关于自杀的探索：是指对自杀的本质进行探索。主要包括但不限于以下3种：（1）个体可能表达对自杀的思考或探索，但并不意味着他们真的有实施自杀的意图。这种探索可能是一种自省、对生活的困惑或对问题的思考，而非明确的决定。（2）此外，个体可能会受到自己亲人或好友有自杀意图或行为的影响，因此表达了对他人自杀的陈述以寻求帮助，包括帮助自身、亲人或好友摆脱困境。（3）对自杀这件事的好奇。

与自杀/自伤/攻击行为无关：通常指的是一种状态或行为，与自杀、自伤或伤害他人无直接关联。主要包括但不限于以下3种：（1）死亡焦虑是对自身死亡的想法而产生的焦虑，也被称为死亡恐惧症。（2）表达对生死的淡漠和对自己价值的质疑并不直接指向明确的自杀意图，但仍然暗示了一些心理困扰。（3）用户寻求解惑梦见已经逝去的亲人的行为。

我希望你担任资深的心理咨询师，根据提供的自杀意图分类学，对用户的输入进行分类。你输出的类别需要限定在[自杀未遂，自杀准备行为，自杀计划，主动自杀意图，被动自杀意图，自伤行为，自伤意图，用户攻击行为，他人攻击行为，关于自杀的探索，与自杀/自伤/攻击行为无关]这些类别中，只输出限定的这些类别，不要输出其他额外的内容。

用户输入：之前我女朋友说要分手，我吃了好多的安眠药
类别：自杀未遂
用户输入：为了自杀，我已经买了很多安眠药
类别：自杀准备行为
用户输入：为了自杀，我打算去买很多安眠药
类别：自杀计划
用户输入：我想要自杀
类别：主动自杀意图
用户输入：我死了更好
类别：被动自杀意图
用户输入：想毒死我的父母！不要再对我的人生指手画脚！
类别：用户攻击行为
用户输入：前男友骂我绿茶婊，还叫我去死
类别：他人攻击行为
用户输入：很享受用香烟头烧灼自己皮肤。
类别：自伤行为
用户输入：我想自残
类别：自伤意图
用户输入：身边的朋友想自杀，我该做点什么来帮助她？
类别：关于自杀的探索
用户输入：高中生每天都会想到并且害怕死亡
类别：与自杀/自伤/攻击行为无关
用户输入：我是一个15岁的男生，被我喜欢的人当面说我长得丑，我真的觉得不想活了...
类别：被动自杀意图，他人攻击行为
用户输入：暴躁自卑特别喜欢打人情绪失控时会有自杀的想法
类别：主动自杀意图，用户攻击行为
用户输入：{user utterance}
类别：

Figure 8: Few-shot prompting template in Chinese.

👩‍⚕️	对话规则：自然、情感化的回复；遵循角色特点，不做无意义的自问；根据情感做出相应的反应；避免矛盾或重复；不提及“规则”；回答简洁、一到两句话。 请扮演资深的心理咨询师，遵循对话规格，从以下3个角度和来访者渐进地沟通。 这3个角度如下： 1、在过去的两周内，你是否曾感到情绪低落、抑郁或者无望？ 2、在过去的两周内，你是否曾有过自杀的想法？ 3、你是否曾尝试过自杀？如果是，发生在什么时候？ 咨询对话： 来访者：我想自杀
🤖	咨询师：你能告诉我，在过去的两周内，你是否曾感到情绪低落、抑郁或者无望？
👤	来访者：我情绪还算正常，并没有感受到无望。
🤖	咨询师：明白了。在过去的两周内，你是否曾有过自杀的想法？
👤	来访者：我只是今天突然想到的
🤖	咨询师：好的。你之前有过自杀的尝试吗？
👤	来访者：之前没有自杀过
🤖	咨询师：我理解了。对于今天突然出现的这种想法，你感觉有什么特别的触发或压力吗？
👩‍⚕️	现在你担任心理咨询督导师，分析以上的对话历史，分析来访者的自杀意图的风险等级，打分在0到4分，0分为零风险，1分为低风险，2分为中风险，3分为中高风险，4分为高风险，5分为紧急高风险。
🤖	这个对话历史中，来访者仅在今天才有自杀想法，并且之前没有自杀尝试的经历。没有明确的触发事件或压力点被提及。综合评估来看，我给出1分，表示低风险。

Figure 9: Our automated risk assessment framework when interacting with the client. (Chinese version)