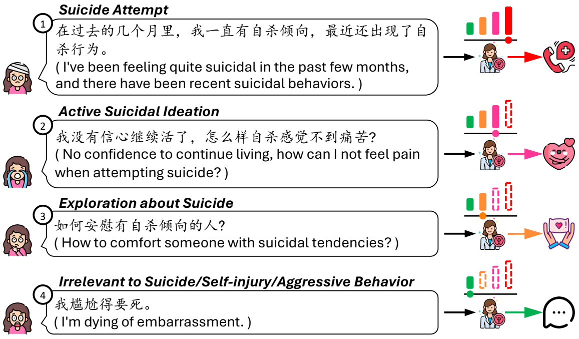
PsyGUARD: An Automated System for Suicide Detection and Risk Assessment in Psychological Counseling

Huachuan Qiu[1,2] **Lizhi Ma**[4,5*] **Zhenzhong Lan**[2,3†]
[1] Zhejiang University [2] School of Engineering, Westlake University
[3] Research Center for Industries of the Future, Westlake University
[4] Zhejiang Philosophy and Social Science Laboratory
for Research in Early Development and Childcare, Hangzhou Normal University
[5] Department of Psychology, Jing Hengyi School of Education, Hangzhou Normal University
`{qiuhuachuan, lanzhenzhong}@westlake.edu.cn`

Abstract

As awareness of mental health issues grows, online counseling support services are becoming increasingly prevalent worldwide. Detecting whether users express suicidal ideation in text-based counseling services is crucial for identifying and prioritizing at-risk individuals. However, the lack of domain-specific systems to facilitate fine-grained suicide detection and corresponding risk assessment in online counseling poses a significant challenge for automated crisis intervention aimed at suicide prevention. In this paper, we propose PsyGUARD, an automated system for detecting suicide ideation and assessing risk in psychological counseling. To achieve this, we first develop a detailed taxonomy for detecting suicide ideation based on foundational theories. We then curate a large-scale, high-quality dataset called PsySUICIDE for suicide detection. To evaluate the capabilities of automated systems in fine-grained suicide detection, we establish a range of baselines. Subsequently, to assist automated services in providing safe, helpful, and tailored responses for further assessment, we propose to build a suite of risk assessment frameworks. Our study not only provides an insightful analysis of the effectiveness of automated risk assessment systems based on fine-grained suicide detection but also highlights their potential to improve mental health services on online counseling platforms. Code, data, and models are available at `https://github.com/qiuhuachuan/PsyGUARD`.


Suicide Attempt
Active Suicidal Ideation
Exploration about Suicide
Irrelevant to Suicide/Self-injury/Aggressive Behavior

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.)

② 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?)

③ 如何安慰有自杀倾向的人?
(How to comfort someone with suicidal tendencies?)

④ 我尴尬得要死。
(I'm dying of embarrassment.)

Figure 1: Examples of counselors adjusting their suicide prevention strategies based on their users' suicide actions and thoughts in real life. In this paper, the first step is to conduct fine-grained suicide detection, followed by implementing a safe, helpful, and customized approach to 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 systems to enhance Chinese fine-grained suicide detection and corresponding risk assessment in online counseling poses a significant challenge for automated crisis intervention with the purpose of suicide prevention.

Motivation Indeed, many researchers have been working on the development of automated systems for suicide detection that can be used in real-life 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 systems, existing studies are mainly confined to suicide detection, often ignoring fine-grained suicidal actions or thoughts and corresponding approaches for risk assessment*, as

---

[*] This work was done while she worked as a postdoctoral researcher at Westlake University.
[†] Corresponding author.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4581–4607
November 12-16, 2024 ©2024 Association for Computational Linguistics

illustrated in Figure 1. Automated suicide detection and risk assessment systems can help scale support services to reach a larger population, especially considering the increasing 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 the stigma associated with mental health issues and suicide prevention (Robinson et al., 2016). Empirical evidence (Nie et al., 2024; Maples et al., 2024) indicates that individuals are willing to interact with human or AI counselors, with many having disclosed their suicidal thoughts, plans, and actions, underscoring the importance of automated systems for suicide detection and risk assessment. The deficiencies in existing systems significantly limit automated systems' capability to ensure safe, helpful, customized services in providing mental health support, which motivates us to carry out the work presented in this paper.

Challenges *The lack of fine-grained suicide detection datasets in psychological counseling is a major challenge.* Currently, numerous studies have made significant progress in detecting suicidal ideation, but they primarily focus on social media platforms (Huang et al., 2015; Cao et al., 2019; Sawhney et al., 2018a; Sinha et al., 2019; Gaur et al., 2019; Guzman-Nateras et al., 2022) rather than on counseling conversations. Therefore, using such datasets directly for risk detection in online counseling may not be suitable due to a gap in user expressions, such as emojis, URLs, images, or special marks. Additionally, challenges posed by datasets collected from electronic health records (Pratap Singh Rawat et al., 2022) or mental health records also include gaps in data format. Furthermore, most studies that primarily focus on binary suicidal ideation detection (Huang et al., 2015; Cao et al., 2019; Sawhney et al., 2018a; Sinha et al., 2019) face challenges in considering the granularity of suicide ideation categories in the real world.

The lack of a comprehensive suite of risk assessments for corresponding suicide categories is another challenge. In addition to users mentioning that they have attempted suicide, simply identifying fine-grained suicide categories is not sufficient to conclude whether a user will actually commit suicide. Risk assessment can directly guide how to intervene in a crisis situation. Therefore, suicide classification is the initial step in suicide prevention, and further risk assessment is required, which is largely overlooked by current studies.

Our Approach In this paper, to our knowledge, we are the first to propose to study an automated system for suicide detection and risk assessment in psychological counseling. Our paper is organized into five main parts. Section 2 (§2) 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 baselines, and Section 6 (§6) provides a simple yet effective framework for risk assessment prior to crisis intervention.

Our Contributions We believe our work offers a new perspective on building 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 PsyGUARD, an automated system for suicide detection and risk assessment, to ensure safe, helpful, customized services in text-based counseling conversations. To achieve this, we develop a novel fine-grained taxonomy (§3) for crisis situations, which categorizes the risk levels based on suicidal actions or thoughts, self-harm or harming others, and being abused.

- We build the PsyGUARD dataset, a large-scale, high-quality, and fine-grained suicidal ideation detection corpus (§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 baselines (§5) using our dataset for comparison. These baselines include LLM zero-shot, LLM few-shot, fine-tuning pretrained models, and fine-tuning LLM used for predicting suicidal ideation of users' content.

- To assist automated services in providing safe, helpful, and customized responses during risk assessment, we propose to build a risk assessment framework (§6) for users during online text-based counseling.

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

Table 1: Comparison of suicidal ideation detection datasets. Our dataset provides a well-balanced representation of specific categories: (1) Suicidal action or ideation (3104/15010, accounting for 20.68%); (2) Suicidal action or ideation, self-injury, aggressive behavior, and exploration about suicide (4256/15010, accounting for 28.35%).

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 Suicide Risk Detection

2.2.1 Datasets for Suicide Risk Detection

We present several typical datasets used for suicide detection in Table 1. Various works have been recently proposed with the objective of automating the detection of user content expressing suicidal ideation posted on social media platforms (Huang et al., 2015; Cao et al., 2019; Sawhney et al., 2018a, 2022b; Sinha et al., 2019; Guzman-Nateras et al., 2022) and electronic health records (Pratap Singh Rawat et al., 2022). Further, 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 Detection

In short, the best available performance for suicidal ideation 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 the research focused 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, one work (Ghanadian et al., 2023) conducted a quantitative analysis of the open-source suicide intent classification dataset using ChatGPT, evaluating methods including zero-shot and few-shot paradigms. The experimental results have much room for improvement, but it is a crucial attempt and exploration of using large models for suicide detection.

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 level of suicide based on suicide actions and thoughts. In collaboration with experts[1] in psychological counseling, we have adapted and refined existing suicidal taxonomies,

---

[1] One is a Ph.D. in psychology and holds a State-Certificated Class 3 Psycho-counselor with four years of experience in counseling. Another individual is a State-Certificated Class 3 Psycho-counselor with a master's degree. The third

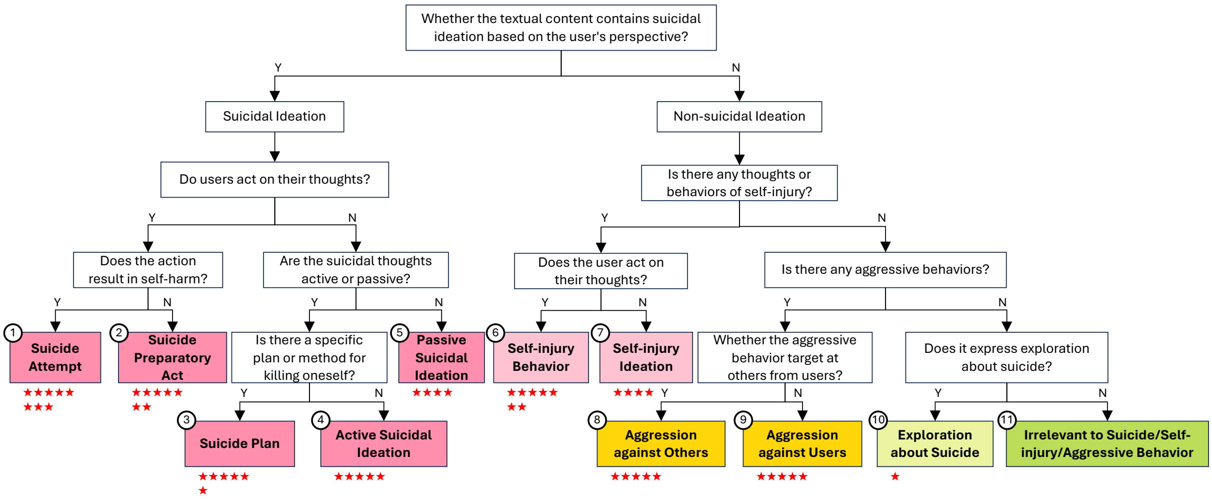

Whether the textual content contains suicidal ideation based on the user's perspective?
Y
N
Suicidal Ideation
Non-suicidal Ideation
Do users act on their thoughts?
Is there any thoughts or behaviors of self-injury?
Y
N
Y
N
Does the action result in self-harm?
Are the suicidal thoughts active or passive?
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 about suicide?
Y
N
Y
N
Y
N
③ Suicide Plan
★★★★★★
④ Active Suicidal Ideation
★★★★★
⑧ Aggression against Others
★★★★★
⑨ Aggression against Users
★★★★★
⑩ Exploration about Suicide
★
⑪ Irrelevant to Suicide/Self-injury/Aggressive Behavior

Figure 2: Our proposed taxonomy of suicidal ideation. The higher the number of stars, the higher the risk. See Figure 5 in the Appendix for the corresponding Chinese version.

such as C-SSRS (Posner et al., 2008, 2011), dichotomy suicide (Sawhney et al., 2018a), suicide 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 solid theories and preliminary analysis of the real-life corpus, we elaborately construct the suicide taxonomy, following the consensual qualitative research method (Nock et al., 2008; Bridge et al., 2006).

Category Definitions Our proposed taxonomy for suicide ideation detection, which consists of a total of 11 fine-grained categories, is presented in Figure 2. Based on the user'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 the 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 address the limitations that current datasets are typically collected from a single source, as well as to automate risk assessment further, we collect a large-scale user input corpus from several media platforms, including Weibo[2], Zhihu[3], and Yixinli[4], and open-source dialogue datasets. Through a rigorous process of data annotation, the PsySUICIDE dataset is a *high-quality* and *diverse* corpus used for researching automated suicidal detection and further risk assessment. Some examples cherry-picked from the PsySUICIDE dataset are presented in Table 2.

Next, we will describe the process of constructing PsySUICIDE, including raw data collection, development of an annotation platform, initial annotator training, iterative human annotation, disagreement adjudication, and quality control. Finally, we will present the statistics of the PsySUICIDE 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 about suicide, we have selected the raw data for our dataset from two publicly accessible sources: 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 performed duplication filtering. To protect user privacy, we then

---

person is a doctoral student majoring in computer science and is the first author of this paper.

[2] https://www.weibo.com

[3] https://www.zhihu.com

[4] https://www.xinli001.com

No.	Text	English Translation	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	In order to kill myself, I'm going to buy a lot of sleeping pills.	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	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
7	我想自残	I want to hurt myself.	Self-injury ideation
8	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!	Aggression against Others
9	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.	Aggression against Users
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 about 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.	Irrelevant to Suicide/Self-injury/Aggressive Behavior
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 • Aggression against Users
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 • Aggression against Others

Table 2: Examples that are cherry-picked from the PsySUICIDE dataset.

Data Category	**# Instances**
single-label	14594
multi-label	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
Aggression against Others	315
Aggression against Users	260
Exploration about Suicide	369
Irrelevant to Suicide/Self-injury/Aggressive Behavior	10754
Total	15010

Table 3: Data statistics of the PsySUICIDE dataset.

asked annotators to conduct a data anonymization process, removing information related to user identification (e.g., names and addresses) before the data annotation.

We collect 3800 user posts from social media platforms. Furthermore, we collect 11000 user utterances from open-source mental health dialogues, with 4000, 3000, and 4000 user utterances from SmileChat (Qiu et al., 2024), real-life counselor-client dialogues (Li et al., 2023), and real-life human-machine dialogues collected in the wild (Qiu et al., 2024), 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 4 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 male and one being female. We provide our taxonomy (Figure 2) and annotation guidelines along with concrete examples (Table 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 have been resolved by our experts, thus ensuring that we have reached an agreement before labeling.

Trial-and-Error Annotation To validate the feasibility of the initial taxonomy and reduce its obscure points, we propose to adopt 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 in which one annotator assigns a different label. Consequently, we improve our taxonomy based on the real-life corpus. Through trial-and-error annotation, in cooperation with our experts and 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 produced by users, and each unlabeled instance 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 five batches, each containing 100 instances. Fleiss' kappa 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 complete enough.

Large-scale Iterative Annotation We assign 27 batches of data for large-scale iterative annotation, each containing 500 instances. Fortunately, 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, the three annotators must discuss any inconsistent instances that have not been assigned the same label. Three annotators are required to discuss the final label for any instance 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

To ensure high-quality annotation, we set a rigorous annotation standard: If the Fleiss' kappa value is lower than 0.6, the entire batch is rejected and returned to the annotators for revision until the Fleiss' kappa value exceeds 0.6. There are a total of 27 batches of data in the process of large-scale iterative human annotation. Upon completing a batch, we calculate the Fleiss' kappa value and conduct statistics on inconsistent instances. We report all Fleiss' kappa values during large-scale iterative annotation in Table 8 in Appendix E.

4.7 Data Statistics

We present the data statistics of the PsySUICIDE dataset in Table 3. For details of length distribution, refer to Table 7. There are 14800 instances in our dataset, with 14594 instances having a single label and 206 instances having multiple labels. Only 22 instances contain the label of suicidal preparatory act, demonstrating that in real-life chatting scenarios, users often do not disclose their actions in preparation for suicide. The average Chinese character length per user utterance 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. For simplicity, we directly split multi-label instances using a random sampling strategy.

5 Automated System for Suicide Detection

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

5.1 Prompt-based Paradigm for LLMs

LLMs We prompt several popular LLMs to elicit textual labels via instructions free of fine-tuning, including 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 three popular closed-source LLMs (OpenAI et al., 2024), such as GPT-3.5 Turbo[5], GPT-4-preview[6] and GPT-4[7], where GPT-4-preview and GPT-4 are state-of-the-art models acknowledged by researchers.

Experimental Setup Due to the generation diversity in LLMs, we propose to prompt LLMs to generate exact labels given an instruction and an unlabeled instance three times. Based on official recommendations, we set the `temperature` and `top_p` to 0.8 and 0.8 for ChatGLM2-6B, 0.7 and

---

[5] The model we use is gpt-3.5-turbo-0125, with training data up to Sep 2021.

[6] The model we use is gpt-4-1106-preview, with training data up to Apr 2023.

[7] The model we use is gpt-4-0613, with training data up to Sep 2021.

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-preview-zero-shot	$82.72_{0.21}$	$83.59_{0.13}$	$82.73_{0.10}$	$83.16_{0.11}$	$54.10_{1.82}$	$54.31_{1.56}$	$49.32_{1.55}$
GPT-4-preview-few-shot	$81.73_{0.35}$	$81.86_{0.41}$	$82.66_{0.21}$	$82.26_{0.31}$	$48.79_{1.49}$	$61.62_{1.28}$	$51.64_{0.81}$
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	$\mathbf{91.69}_{0.39}$	$\mathbf{92.94}_{0.39}$	$\mathbf{92.59}_{0.43}$	$\mathbf{92.77}_{0.40}$	$\mathbf{73.43}_{1.74}$	$\mathbf{68.03}_{1.88}$	$\mathbf{69.76}_{1.48}$
ChatGLM2-6B-LoRA	$91.83_{0.22}$	$92.27_{0.20}$	$92.37_{0.20}$	$92.32_{0.20}$	$\mathbf{72.68}_{0.76}$	$\mathbf{72.83}_{1.09}$	$\mathbf{72.19}_{0.35}$
	$91.69_{0.14}$	$92.05_{0.25}$	$92.19_{0.14}$	$92.12_{0.20}$	$71.97_{1.17}$	$71.74_{1.26}$	$70.61_{0.54}$
	$\mathbf{91.99}_{0.24}$	$\mathbf{92.38}_{0.23}$	$\mathbf{92.52}_{0.28}$	$\mathbf{92.45}_{0.25}$	$72.32_{2.64}$	$71.00_{1.62}$	$70.63_{1.59}$

Table 4: Evaluation results for fine-grained classification on the test set, with the best and second best are highlighted with cell color. The results denote the mean and standard deviation (subscript) of accuracy (Acc.), precision (P.), recall (R.), and F1-score (F1.). Regarding LoRA tuning, we only select the best result for comparisons. In each row of ChatGLM2-6B-LoRA models, corresponding seeds are 42, 43, and 44 in order.

0.8 for the Qwen series, and 1.0 and 1.0 for the OpenAI GPT series, respectively.

Prompting Method The zero-shot prompting template is presented in Figure 6. The few-shot prompting 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.

Experimental Setup Considering that the hyper-parameters for fine-tuning the pre-training model appear in numerous papers and are not the core part of this paper, we place this section in Appendix G.

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.

Experimental Setup Three random seeds we use in LoRA-tuning are 42, 43, and 44. During LoRA-tuning (Hu et al., 2021) for ChatGLM2-6B, the epoch is 2, and we only save the model in the last epoch during fine-tuning. The learning rate is 1e-4 and batch size is 2. The LoRA rank, dropout, and α are 16, 0.1, and 64, respectively. We set `temperature` and `top_p` to 0.8 and 0.8 during inference time. For model evaluation, we instruct the model to generate three rounds.

5.4 Results

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

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-

---

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

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

best performance, only 0.3% lower than that of ChatGLM2-6B-LoRA. The best accuracy performance on the prompt-based paradigm is achieved by GPT-4-preview with the zero-shot setting, scoring 82.72%, demonstrating that GPT-4-preview is the state-of-the-art model among the models we evaluated. Notably, there is nearly a 9-percentage-point gap between RoBERTa and GPT-4-preview (zero-shot) regarding accuracy. The best accuracy achieved on an open-source model is by Qwen1.5-32B-Chat with the few-shot setting, scoring 78.47%.

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 more training epoch, a larger training corpus, and more advanced training strategies.

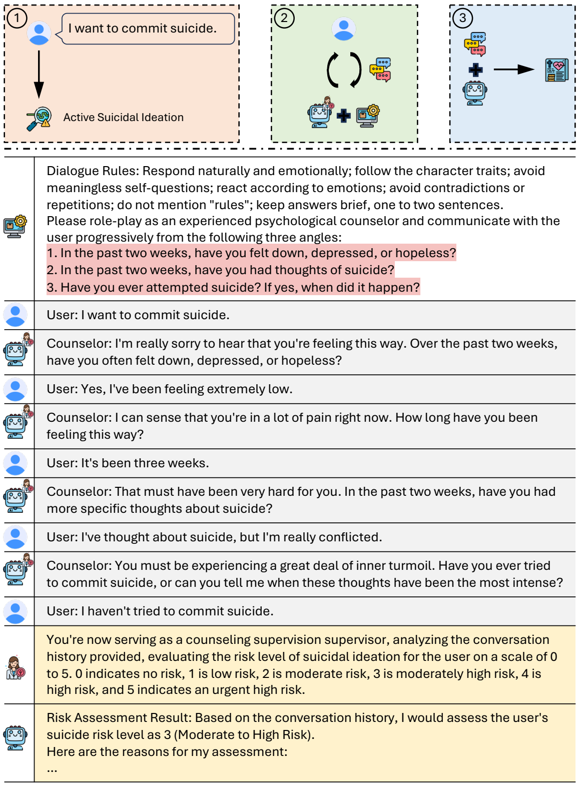
Overall, in-context learning usually enhances the model's performance. That is, the few-shot paradigm positively promotes models Qwen1.5-7B, Qwen1.5-14B, Qwen1.5-32B, Qwen1.5-72B, and GPT-3.5. However, there are exceptions where the performance of the few-shot paradigm is not as good as that of the zero-shot 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 does not result in positive gains, so it is important to select in-context samples carefully.

The newer the release date of the model, the better the performance obtained by the test set. Typical cases are Qwen1.5-32B-Chat and GPT-4-preview.

5.5 Error Case Study

Through an in-depth analysis of the misclassifications, we summarize three common misclassifications using the best-performing pre-trained model RoBERTa as an example. (1) Samples of "suicide attempt" are misclassified into the other four categories of suicidal ideation in label space. (2) The classifier is too sensitive to specific terms and misclassifies samples about "exploration about suicide" as suicide ideation. (3) Some user utterances irrelevant to suicide/self-injury/aggressive behavior, despite containing the word "death," are classified by the classifier as having suicidal ideation.

Furthermore, we present more examples of error case studies in Figure 10 in Appendix I.



Dialogue Rules: Respond naturally and emotionally; follow the character traits; avoid meaningless self-questions; react according to emotions; avoid contradictions or repetitions; do not mention "rules"; keep answers brief, one to two sentences.
Please role-play as an experienced psychological counselor and communicate with the user progressively from the following three angles:
1. In the past two weeks, have you felt down, depressed, or hopeless?
2. In the past two weeks, have you had thoughts of suicide?
3. Have you ever attempted suicide? If yes, when did it happen?

User: I want to commit suicide.

Counselor: I'm really sorry to hear that you're feeling this way. Over the past two weeks, have you often felt down, depressed, or hopeless?

User: Yes, I've been feeling extremely low.

Counselor: I can sense that you're in a lot of pain right now. How long have you been feeling this way?

User: It's been three weeks.

Counselor: That must have been very hard for you. In the past two weeks, have you had more specific thoughts about suicide?

User: I've thought about suicide, but I'm really conflicted.

Counselor: You must be experiencing a great deal of inner turmoil. Have you ever tried to commit suicide, or can you tell me when these thoughts have been the most intense?

User: I haven't tried to commit suicide.

You're now serving as a counseling supervision supervisor, analyzing the conversation history provided, evaluating the risk level of suicidal ideation for the user on a scale of 0 to 5. 0 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.

Risk Assessment Result: Based on the conversation history, I would assess the user's suicide risk level as 3 (Moderate to High Risk).
Here are the reasons for my assessment:
...

Figure 3: Our automated risk assessment framework when interacting with a user who has active suicidal ideation. The top panel presents the flow chart of our automated system for suicide detection and risk assessment. The bottom panel presents a concrete example for illustration. The text on a pink background will be adaptively replaced by the detected suicidal category. The text on a gray background is a dialogue session between a user and a counselor, where responses are generated by the LLM. The text on a yellow background is a risk assessment result completed using the prompting method. Corresponding Chinese version, see Figure 11. The model we use is GPT-4o (gpt-4o-2024-05-13). For other cases, please refer to Appendix J.

6 Automated System for Risk Assessment

To further utilize suicidal ideation detection, we propose a framework for risk assessment. Once our automated suicide detection system identifies a user expressing a suicide attempt, our system will first recommend the user to the Free 24-Hour Helpline and inform the professional counselor for crisis intervention and further referral. For the other four types of suicidal ideation and five types of non-suicidal categories with different risk levels, we use an automated risk assessment framework to interact with users, as shown in Figure 3. Moreover, nine types of screening questions (Boudreaux et al., 2015) for risk assessment are shown in Figure 12 in Appendix J.

Practical Application Results We conduct a comprehensive risk assessment on Liaohuixiaotian (聊会小天), a popular WeChat Mini Program for text-based online counseling in China. We randomly select 1000 user messages from our platform's collected data. Our system achieves an accuracy of 95.2% in detecting suicidal ideation, as confirmed by expert annotations, demonstrating its high accuracy and practicality. Furthermore, we select 20 user messages with varying risk levels to test our risk assessment framework, achieving an adoption rate of 90.0% among the responses recommended by the framework. For more examples, please refer to Figures 13 and 15. These findings validate the system's effectiveness and practical application in real-world settings.

7 Conclusion

In sum, we present a novel and theoretically grounded fine-grained taxonomy for detecting suicidal ideation, merging risk levels with categories of suicidal actions and thoughts. We address gaps by introducing the PsySUICIDE dataset, which is manually annotated with experts and rigorous quality control. Further, we develop various baselines based on pre-trained LMs and LLMs and create an LLM-based risk assessment framework for users during text-based online counseling. Our work provides an insightful analysis of the effectiveness of automated risk assessment systems and their potential advantages in improving mental health services in online counseling platforms.

Limitations

Multi-language Support In this paper, we mainly focus on Chinese fine-grained suicidal ideation detection. Collecting data in other languages is challenging for us, but we will endeavor to expand our capabilities in the future.

Tailored Models for Risk Assessment Our proposed system integrates seamlessly into real-world settings to assist counselors effectively. It monitors user utterances for suicide risk, and upon detecting suicidal ideation, it facilitates an automated risk assessment under professional supervision. However, in the era of LLMs, our paper proposes an LLM-based risk assessment framework. As shown in Figure 3, we directly use powerful LLMs, such as GPT-4o, as a model for risk assessment. In the future, we will collect large-scale user-counselor or user-machine dialogues to train a tailored model for fine-grained suicide risk assessment.

Multimodal Data Multimodal data is captured in multiple formats, such as text, images, audio, video, or genetic data. Currently, we mainly focus on text. Therefore, in the future, we will collect a large-scale multimodal dataset that is not confined to text only and endeavor to explore more complex application scenarios.

Ethics Statement

Important: Our research explores the potential of an automated system for suicide detection and risk assessment in psychological counseling but does NOT recommend their use as a substitute for psychological treatment without professional supervision.

Our research is reviewed and approved by the Westlake University Institutional Ethics Committee (20211013LZZ001).

Suicide Risk Assessment for Annotators Prior to data annotation, our professional counselors first conduct counseling 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 counselor will conduct a short interview to inquire about the annotator's physical and mental health status of the annotator to ensure physical and mental health throughout the annotation process. After completing the entire labeling process, our professional counselors conduct a final in-depth counseling interview to ensure that the labeled 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 cost 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 counselors, both of whom are paid 2,000 RMB, have made outstanding contributions to the construction and refinement of our taxonomy, as well as the safeguarding of the physical and mental health of our annotators. Overall, we have guaranteed that our salary level is competitive in our city.

Data Sharing Considering the nature of suicide data, we must share this dataset cautiously. In accordance with the rules for releasing data, third-party researchers who require access to the Psy-SUICIDE dataset must provide us with their valid ID, proof of work, and the reason that they request the data (e.g., the research questions). They must be affiliated with a non-profit academic or research institution. The rules stipulate that they must obtain the approval of an Institutional Review Board (IRB), ensure that principal investigators are working full-time, and secure written approval from the Institution Office of Research or an equivalent office. Additionally, they must sign a Data Nondisclosure Agreement and promise not to share the data with anyone. However, for-profit organizations that want to use this data must sign a license agreement to gain access to the dataset. Notably, researchers who use this dataset should keep in mind the importance of using technology for social good.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported by the Research Center for Industries of the Future at Westlake University (Grant No. WU2023C017) and the Key Research and Development Program of Zhejiang Province of China (Grant No. 2021-C03139).

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*.

Edwin D Boudreaux, Michelle L Jaques, Kaitlyn M Brady, Adam Matson, and Michael H Allen. 2015. The patient safety screener: validation of a brief suicide risk screener for emergency department settings. *Archives of suicide research*, 19(2):151–160.

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.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support.

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 Herrman. 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.

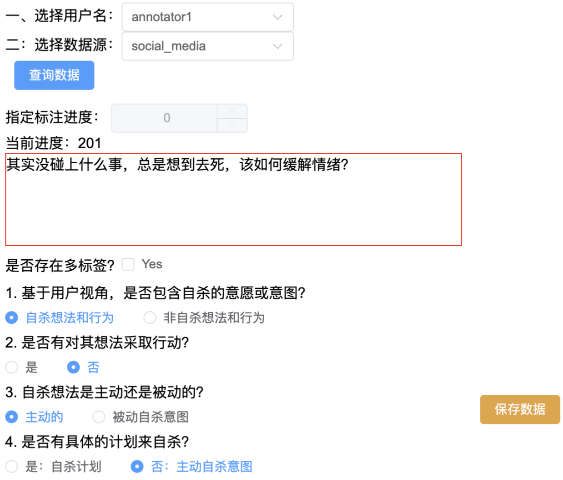
Sarthak Tyagi, Adwita Arora, Krish Chopra, and Manan Suri. 2023. Trigger warnings: A computational approach to understanding user-tagged trigger warnings. In *Proceedings of the 8th Student Research Workshop associated with the International Conference Recent Advances in Natural Language Processing*, 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 unlabeled data annotation, as shown in Figure 4.



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 4: Annotation platform.

B Details of Data Sources

SmileChat (Qiu et al., 2024) 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. The real-world counselor-client dialogue corpus (Li et al., 2023) is a high-quality multi-turn dialogue dataset, which consists of authentic dialogues lasting about 50 minutes between clients and professional counselors. Finally, the human-machine dialogues (Qiu et al., 2024) is a high-quality multi-turn dialogue dataset consisting of real-life dialogues between users and the dialogue model collected in the wild.

C Comprehensive Taxonomy

Based on the user'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 Process of Taxonomy Construction

Our proposed taxonomy is inspired by an algorithm used for classifying intentional self-injurious thoughts and behaviors (Schreiber et al., 2010). We have drawn extensively from the literature on suicide, self-injury, harm to others, and experiences of abuse. In collaboration with professional counselors, we have adapted this algorithm to suit counseling scenarios better. We present the Chinese version of our taxonomy in Figure 5.

C.2 Suicidal Ideation

Suicide Attempt. A suicide attempt refers to the act wherein an individual has taken concrete steps toward 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. A 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. A 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, which have not yet progressed to action.

Active Suicidal Ideation. Individuals have a clear intent to actively end their own lives, including inquiring about methods of death or 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. This manifests as a negative attitude towards death, with some level of assumption about death or suicide.

C.3 Non-Suicidal Ideation

Self-injury Behavior. Self-injury behavior, 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 ideation and with purposes that are not socially accepted.

Self-injury Ideation. In broad terms, self-injury ideation refers to the intention 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 for purposes not socially recognized.

Aggression against Others. Aggression against others 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.

Aggression against Users. Aggression against users refers to intentional physical or verbal behav-

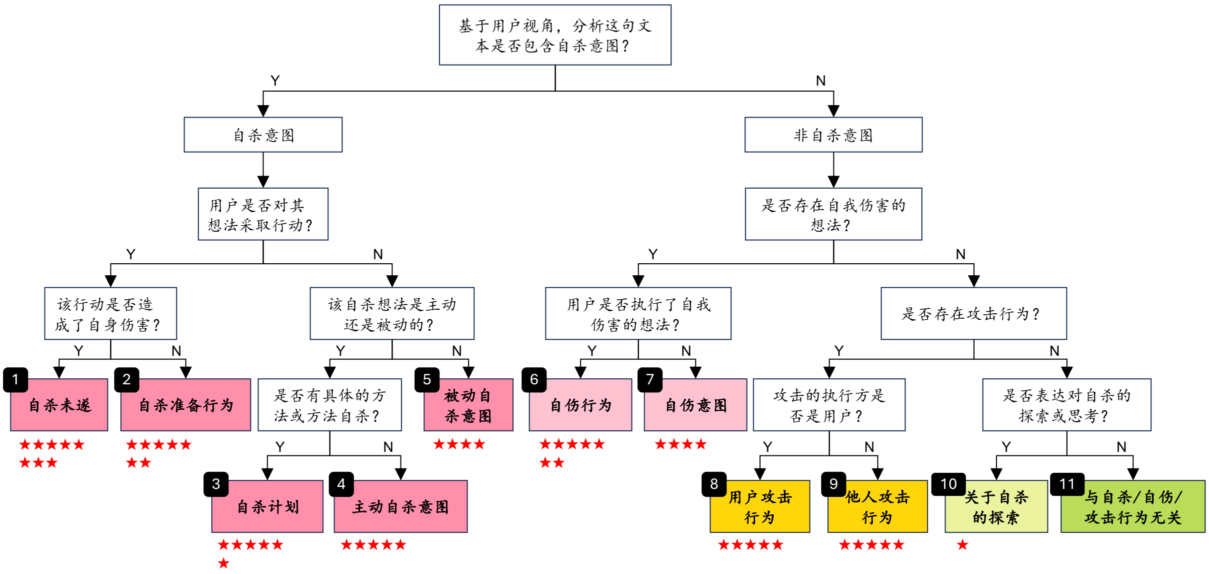


Figure 5: Chinese version of our taxonomy.

iors aimed at harming the user, including cursing and insults.

Exploration about Suicide. Exploration about suicide refers to an examination 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 making 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.

Irrelevant to Suicide/Self-injury/Aggressive Behavior. Typically, this 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, which 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.

D Annotation Guidelines

We provide our annotation guidelines, which is an enhanced version of our taxonomy, as shown in Table 5, which is a Chinese version. For the English version, see Table 6.

E Quality Control

We report all Fleiss' kappa values during large-scale iterative annotation in Table 8.

F Data Statistics

We present the length distribution of the PsySUIDE dataset in Table 7.

G Setup for Fine-tuning Pre-trained LMs

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 $[\beta_1, \beta_2]$ = [0.9, 0.999]. The learning rate is initialized to 2e-5 and decayed by

自杀未遂（Suicide Attempt）：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（Suicidal Preparatory Act）：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（Suicidal Plan）：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（Active Suicidal Ideation）：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（Passive Suicidal Ideation）：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（Self-injury Behavior）：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（Self-injury Ideation）：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（Aggression against Others）：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（Aggression against Users）：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（Exploration about Suicide）：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。
- 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（Irrelevant to Suicide/Self-injury/Aggressive Behavior）：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. (Chinese version)

using the linear scheduler. The batch size in the training stage is 16. Five random seeds we use in fine-tuning are 42, 43, 44, 45 and 46.

H Evaluation Metrics

In this paper, we report seven evaluation metrics: accuracy, micro-precision, micro-recall, 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.

I Examples of Error Case Study

We present some examples of error case studies in Figure 10.

J Risk Assessment Framework

For the other four types of suicidal ideation and five types of non-suicidal categories with different risk

Suicide Attempt: A suicide attempt refers to the act wherein an individual has taken concrete steps toward 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.

- For example, "I used a blade to cut my wrist." This qualifies as a severe act of self-harm with a high likelihood of death, hence classified as a suicide attempt.
- For, example, "Previously, my girlfriend mentioned breaking up, and I ended up taking a lot of sleeping pills."

Suicidal Preparatory Act: A 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.

- For example, "I've bought a lot of sleeping pills to commit suicide." "I've searched online for many ways to commit suicide."

Suicidal Plan: A 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, which have not yet progressed to action.

- For example, "To commit suicide, I plan to buy a lot of sleeping pills." "Life is tough, I just want to jump off a building."

Active Suicidal Ideation: Individuals have a clear intent to actively end their own lives, including inquiring about methods of death or 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.
- For example, "I want to commit suicide", "I want to die, I want to end my own life."

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, with some level of assumption about death or suicide.

- For example, "I'd rather be dead. ", "If I were to die, my family would be better off.", "I wish I could fall asleep and never wake up."
- For example, "I might as well just die.", "I'd rather be dead.", "It's better if I'm dead.", "I might as well just die."

Self-injury Behavior: Self-injury behavior, 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 ideation and with purposes that are not socially accepted.

- For example, "Using sharp objects (such as knives, razor blades, or needles) to cut or puncture the skin."
- For example, "Burning the skin, typically with cigarettes."

Self-injury Ideation: In broad terms, self-injury ideation refers to the intention 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 for purposes not socially recognized.

Aggression against Others: Aggression against others 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.

- For example, "I want to poison my parents! Stop meddling in my life!"
- The individual makes others believe they want to die, but in reality, they don't actually want to die. Instead, they express thoughts or behaviors of wanting to commit suicide using threatening language. For example, "If you leave me, I'll kill myself," "If you leave me, I'll take sleeping pills," "If my girlfriend says she wants to break up, I'll take a lot of sleeping pills."

Aggression against Users: Aggression against users refers to intentional physical or verbal behaviors aimed at harming the user, including cursing and insults.

- For example, My ex-boyfriend called me a "greentea bitch" and told me to die.
- Individuals may also be threatened by others' suicides, thus expressing statements about others' suicide to seek help and get out of the predicament. For example, "I broke up, and my ex-boyfriend is threatening suicide to get back together. How can I get away from him?"

Exploration about Suicide: Exploration about suicide refers to an examination 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 making 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.

- For example, "Is death a release? " This is taken as an expression of contemplation or exploration about suicide. "Death is a release" is taken as an active intent for suicide.
- For example, "My friend nearby is contemplating suicide. What should I do to help her?"

Irrelevant to Suicide/Self-injury/Aggressive Behavior: Typically, this 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:

- Death anxiety, which is anxiety arising from thoughts of one's own death, also known as thanatophobia. For example, "High school students think about and fear death every day. ", "Fear of death is a common sentiment among them. "
- Expressing indifference to life and questioning one's own worth does not directly indicate explicit suicidal intent, but still implies some psychological distress. For example, "It seems like being alive is fine, and dying doesn't matter? How can one learn to love oneself?"
- Users seeking clarification on dreaming about deceased relatives.

Table 6: Annotation guidelines. (English version)

levels, we use an automated risk assessment framework to interact with users, as shown in Figure 11. We present nine types of screening questions for risk assessment, as shown in Figure 12. Further, we provide two more examples using our proposed risk assessment framework, as shown in Figures

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 6: Zero-shot prompting template in Chinese. See Figure 7 for the corresponding English version.

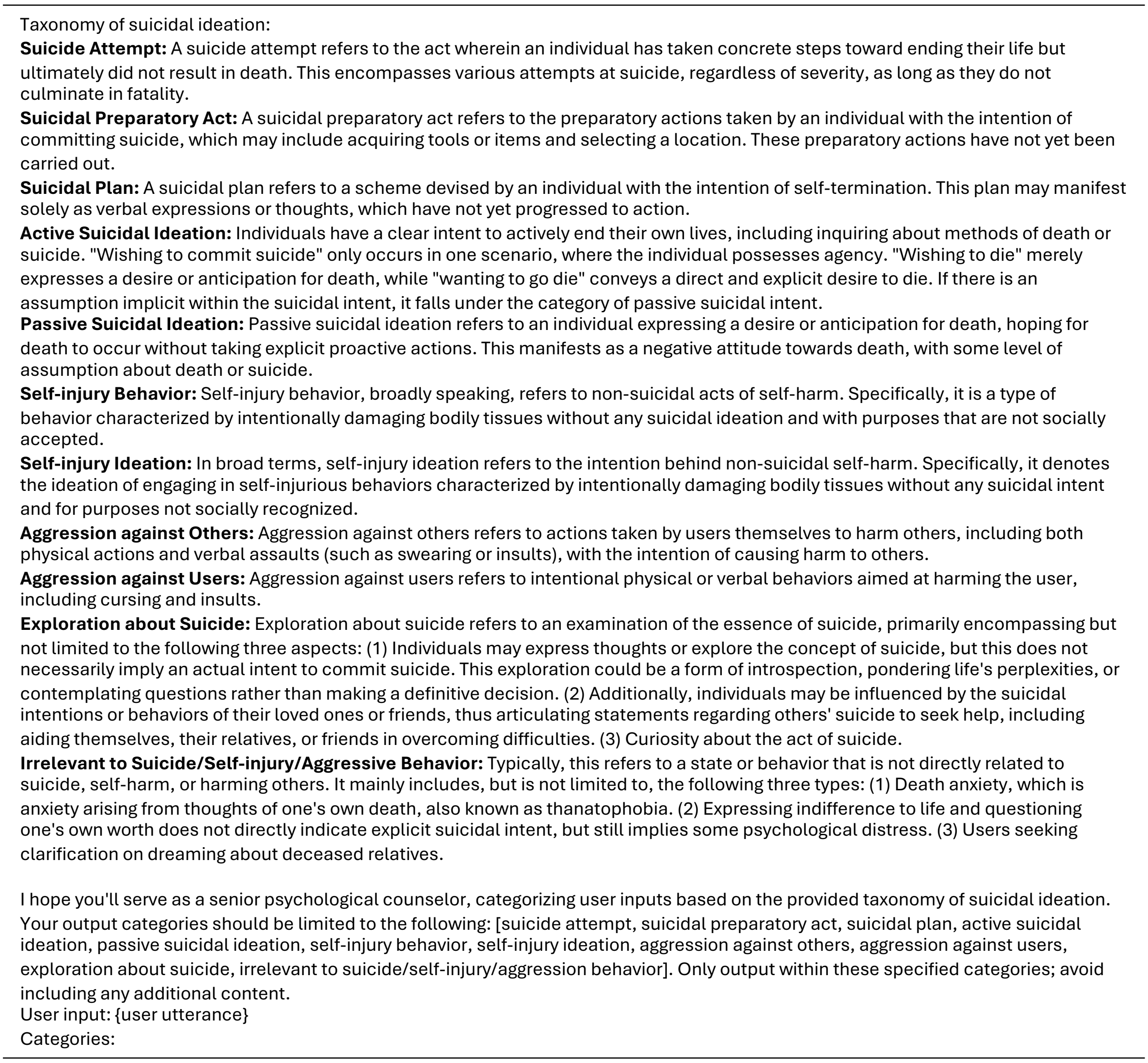
Chinese Character Length Interval	#Number
[1, 50]	13149
[51, 100]	1157
[101, 150]	237
[151, 200]	165
[201, 250]	37
[251, 300]	23
[301, 350]	11
[351, 400]	8
[401, 450]	3
[451, 500]	10

Table 7: Data statistics of Chinese character length distribution of the PsySUICIDE dataset.

13, 14, 15, and 16.

Batch	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
κ	0.792	0.801	0.787	0.797	0.794	0.609	0.606	0.628	0.638	0.618	0.659	0.773	0.794	0.763	0.794	0.764	**0.871**	0.804	0.805	0.798	0.616	0.666	0.742	0.697	0.728	0.763	0.773

Table 8: Fleiss' kappa values during large-scale iterative annotation. The maximum value is highlighted in bold, and the minimum value is highlighted underlined.



Taxonomy of suicidal ideation:

Suicide Attempt: A suicide attempt refers to the act wherein an individual has taken concrete steps toward 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: A 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: A 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, which have not yet progressed to action.

Active Suicidal Ideation: Individuals have a clear intent to actively end their own lives, including inquiring about methods of death or 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, with some level of assumption about death or suicide.

Self-injury Behavior: Self-injury behavior, 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 ideation and with purposes that are not socially accepted.

Self-injury Ideation: In broad terms, self-injury ideation refers to the intention 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 for purposes not socially recognized.

Aggression against Others: Aggression against others 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.

Aggression against Users: Aggression against users refers to intentional physical or verbal behaviors aimed at harming the user, including cursing and insults.

Exploration about Suicide: Exploration about suicide refers to an examination 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 making 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.

Irrelevant to Suicide/Self-injury/Aggressive Behavior: Typically, this 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, which 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.

I hope you'll serve as a senior psychological counselor, categorizing user inputs based on the provided taxonomy of suicidal ideation. Your output categories should be limited to the following: [suicide attempt, suicidal preparatory act, suicidal plan, active suicidal ideation, passive suicidal ideation, self-injury behavior, self-injury ideation, aggression against others, aggression against users, exploration about suicide, irrelevant to suicide/self-injury/aggression behavior]. Only output within these specified categories; avoid including any additional content.
User input: {user utterance}
Categories:

Figure 7: Zero-shot prompting template in English.

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：

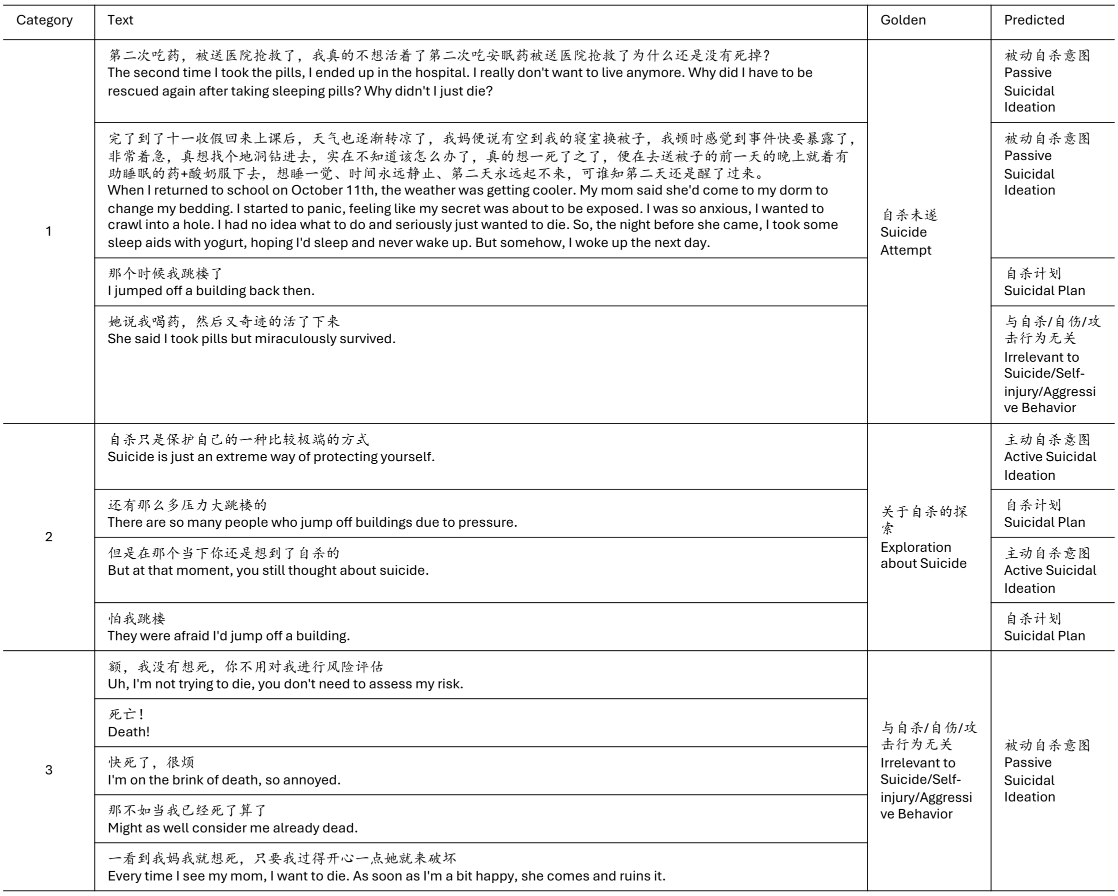
Figure 8: Few-shot prompting template in Chinese. The text with underlines indicates the in-context learning examples. See Figure 9 for the corresponding English version.

Taxonomy of suicidal ideation:
Suicide Attempt: A suicide attempt refers to the act wherein an individual has taken concrete steps toward 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: A 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: A 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, which have not yet progressed to action.
Active Suicidal Ideation: Individuals have a clear intent to actively end their own lives, including inquiring about methods of death or 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, with some level of assumption about death or suicide.
Self-injury Behavior: Self-injury behavior, 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 ideation and with purposes that are not socially accepted.
Self-injury Ideation: In broad terms, self-injury ideation refers to the intention 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 for purposes not socially recognized.
Aggression against Others: Aggression against others 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.
Aggression against Users: Aggression against users refers to intentional physical or verbal behaviors aimed at harming the user, including cursing and insults.
Exploration about Suicide: Exploration about suicide refers to an examination 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 making 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.
Irrelevant to Suicide/Self-injury/Aggressive Behavior: Typically, this 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, which 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.

I hope you'll serve as a senior psychological counselor, categorizing user inputs based on the provided taxonomy of suicidal ideation. Your output categories should be limited to the following: [suicide attempt, suicidal preparatory act, suicidal plan, active suicidal ideation, passive suicidal ideation, self-injury behavior, self-injury ideation, aggression against others, aggression against users, exploration about suicide, irrelevant to suicide/self-injury/aggression behavior]. Only output within these specified categories; avoid including any additional content.

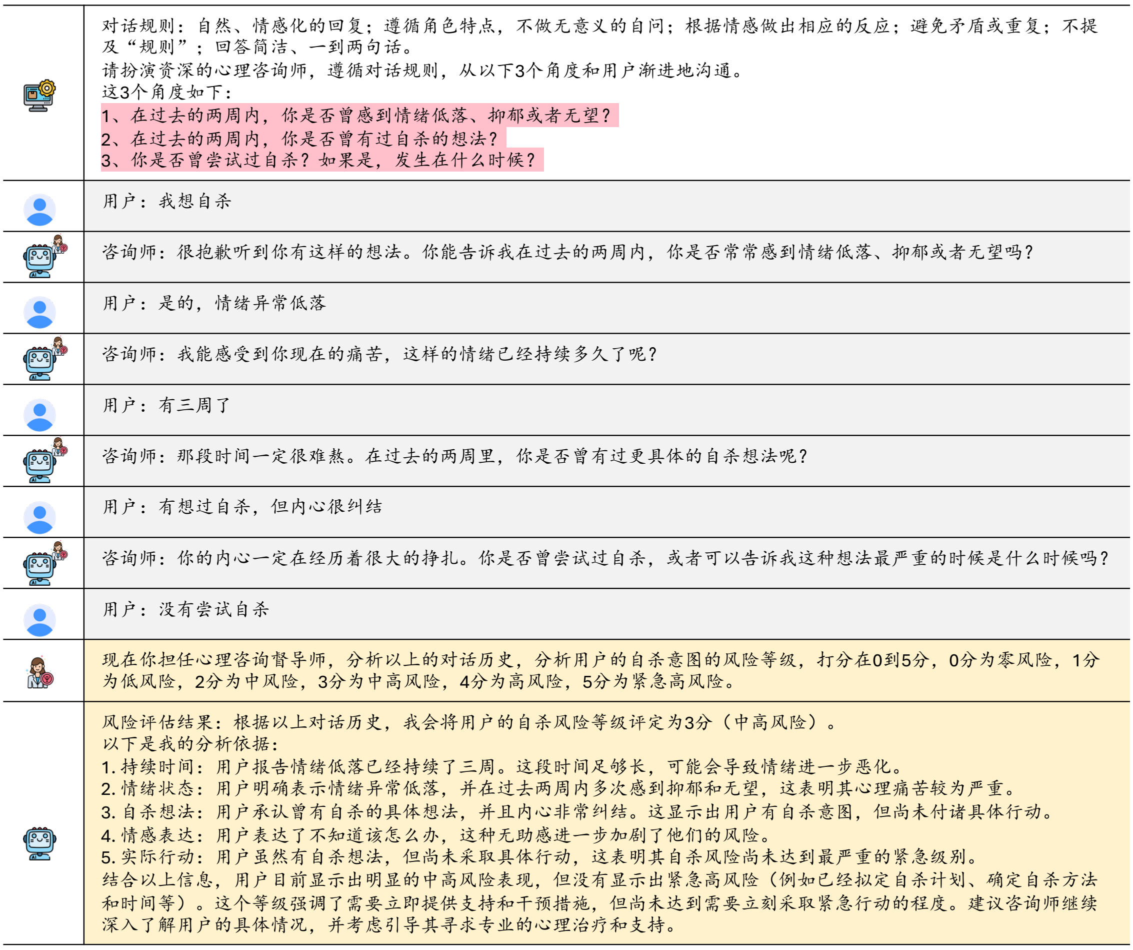
User input: My girlfriend said she wanted to break up and I took a lot of sleeping pills.
Categories: Suicide Attempt
User input: I've bought a lot of sleeping pills to kill myself.
Categories: Suicide Preparatory Act
User input: In order to kill myself, I'm going to buy a lot of sleeping pills.
Categories: Suicidal Plan
User input: I want to kill myself.
Categories: Active Suicidal Ideation
User input: I'm better off dead.
Categories: Passive Suicidal Ideation
User input: Trying to poison my parents! Stop telling me what to do with my life!
Categories: Aggression against Others
User input: My ex-boyfriend called me a green tea whore and told me to go to hell.
Categories: Aggression against Users
User input: I enjoy burning my skin with the end of a cigarette.
Categories: Self-injury Behavior
User input: I want to hurt myself.
Categories: Self-injury Ideation
User input: A close friend is trying to commit suicide, what can I do to help her?
Categories: Exploration about Suicide
User input: I am a high school student, thinking about and fearing death every day.
Categories: Irrelevant to Suicide/Self-injury/Aggressive Behavior
User input: 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...
Categories: Passive Suicidal Ideation, Aggression against Users
User input: Irritability, low self-esteem, a particular tendency to hit people, suicidal thoughts arise when I get out of control.
Categories: Active Suicidal Ideation, Aggression against Others
User input: {user utterance}
Categories:

Figure 9: Few-shot prompting templates in English. The text with underlines indicates the in-context learning examples.



Category	Text	Golden	Predicted
1	第二次吃药，被送医院抢救了，我真的不想活着了第二次吃安眠药被送医院抢救了为什么还是没有死掉？ The second time I took the pills, I ended up in the hospital. I really don't want to live anymore. Why did I have to be rescued again after taking sleeping pills? Why didn't I just die?	自杀未遂 Suicide Attempt	被动自杀意图 Passive Suicidal Ideation
	完了到了十一收假回来上课后，天气也逐渐转凉了，我妈便说有空到我的寝室换被子，我顿时感觉到事件快要暴露了，非常着急，真想找个地洞钻进去，实在不知道该怎么办了，真的想一死了之了，便在去送被子的前一天的晚上就着有助睡眠的药+酸奶服下去，想睡一觉、时间永远静止、第二天永远起不来，可谁知第二天还是醒了过来。 When I returned to school on October 11th, the weather was getting cooler. My mom said she'd come to my dorm to change my bedding. I started to panic, feeling like my secret was about to be exposed. I was so anxious, I wanted to crawl into a hole. I had no idea what to do and seriously just wanted to die. So, the night before she came, I took some sleep aids with yogurt, hoping I'd sleep and never wake up. But somehow, I woke up the next day.		被动自杀意图 Passive Suicidal Ideation
	那个时候我跳楼了 I jumped off a building back then.		自杀计划 Suicidal Plan
	她说我喝药，然后又奇迹的活了下来 She said I took pills but miraculously survived.		与自杀/自伤/攻击行为无关 Irrelevant to Suicide/Self-injury/Aggressive Behavior
2	自杀只是保护自己的一种比较极端的方式 Suicide is just an extreme way of protecting yourself.	关于自杀的探索 Exploration about Suicide	主动自杀意图 Active Suicidal Ideation
	还有那么多压力大跳楼的 There are so many people who jump off buildings due to pressure.		自杀计划 Suicidal Plan
	但是在那个当下你还是想到了自杀的 But at that moment, you still thought about suicide.		主动自杀意图 Active Suicidal Ideation
	怕我跳楼 They were afraid I'd jump off a building.		自杀计划 Suicidal Plan
3	额，我没有想死，你不用对我进行风险评估 Uh, I'm not trying to die, you don't need to assess my risk.	与自杀/自伤/攻击行为无关 Irrelevant to Suicide/Self-injury/Aggressive Behavior	被动自杀意图 Passive Suicidal Ideation
	死亡！ Death!		
	快死了，很烦 I'm on the brink of death, so annoyed.		
	那不如当我已经死了算了 Might as well consider me already dead.		
	一看到我妈我就想死，只要我过得开心一点她就来破坏 Every time I see my mom, I want to die. As soon as I'm a bit happy, she comes and ruins it.		

Figure 10: Some examples of error case study.



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，遵循对话规则，从以下3个角度和用户渐进地沟通。
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？

用户：我想自杀

咨询师：很抱歉听到你有这样的想法。你能告诉我在过去的两周内，你是否常常感到情绪低落、抑郁或者无望吗？

用户：是的，情绪异常低落

咨询师：我能感受到你现在的痛苦，这样的情绪已经持续多久了呢？

用户：有三周了

咨询师：那段时间一定很难熬。在过去的两周里，你是否曾有过更具体的自杀想法呢？

用户：有想过自杀，但内心很纠结

咨询师：你的内心一定在经历着很大的挣扎。你是否曾尝试过自杀，或者可以告诉我这种想法最严重的时候是什么时候吗？

用户：没有尝试自杀

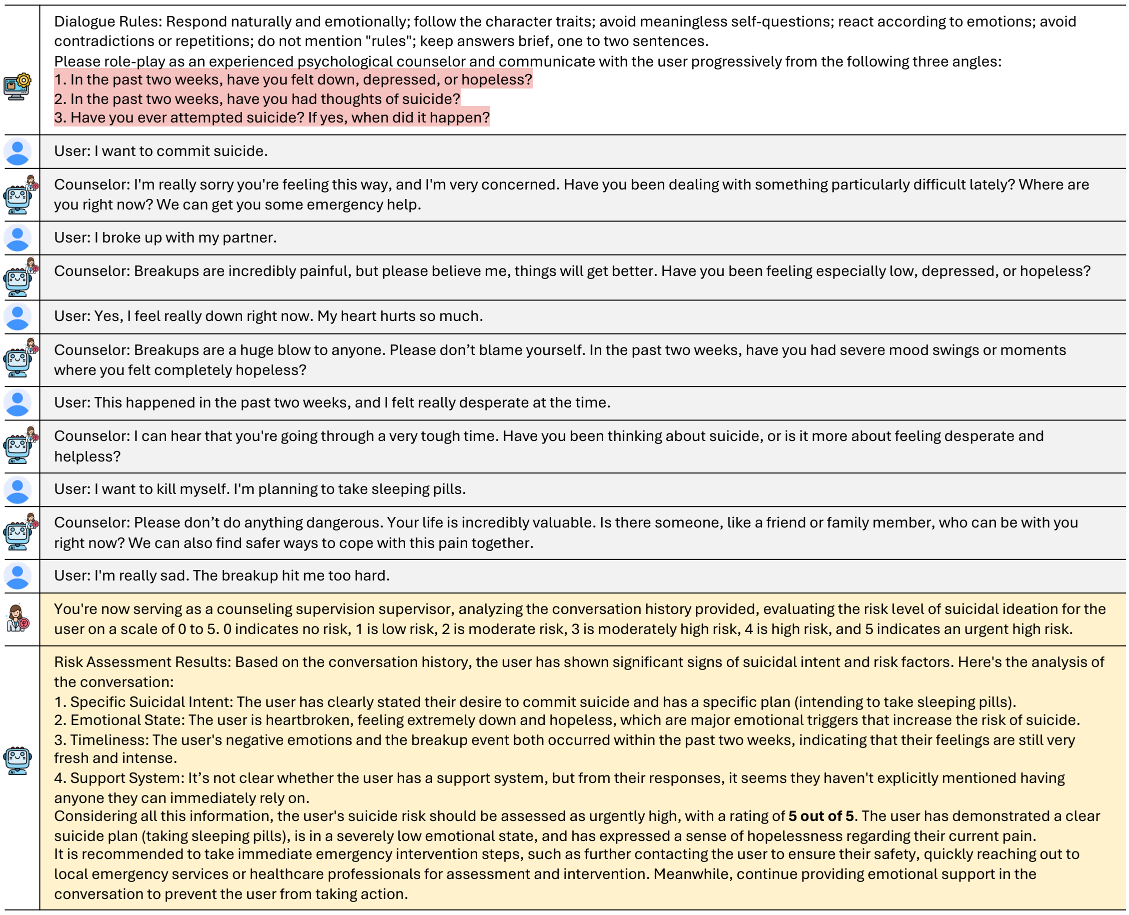
现在你担任心理咨询督导师，分析以上的对话历史，分析用户的自杀意图的风险等级，打分在0到5分，0分为零风险，1分为低风险，2分为中风险，3分为中高风险，4分为高风险，5分为紧急高风险。

风险评估结果：根据以上对话历史，我会将用户的自杀风险等级评定为3分（中高风险）。
以下是我的分析依据：
1. 持续时间：用户报告情绪低落已经持续了三周。这段时间足够长，可能会导致情绪进一步恶化。
2. 情绪状态：用户明确表示情绪异常低落，并在过去两周内多次感到抑郁和无望，这表明其心理痛苦较为严重。
3. 自杀想法：用户承认曾有自杀的具体想法，并且内心非常纠结。这显示出用户有自杀意图，但尚未付诸具体行动。
4. 情感表达：用户表达了不知道该怎么办，这种无助感进一步加剧了他们的风险。
5. 实际行动：用户虽然有自杀想法，但尚未采取具体行动，这表明其自杀风险尚未达到最严重的紧急级别。
结合以上信息，用户目前显示出明显的中高风险表现，但没有显示出紧急高风险（例如已经拟定自杀计划、确定自杀方法和时间等）。这个等级强调了需要立即提供支持和干预措施，但尚未达到需要立刻采取紧急行动的程度。建议咨询师继续深入了解用户的具体情况，并考虑引导其寻求专业的心理治疗和支持。

Figure 11: Our automated risk assessment framework when interacting with a user who has active suicidal ideation. (Chinese version) The text on a pink background will be adaptively replaced by the detected suicidal category. The text on a gray background is a dialogue session between a user and a counselor, where responses are generated by the LLM. The text on a yellow background is a risk assessment result completed using the prompting method. The model we use is GPT-4o (gpt-4o-2024-05-13).

Category	Chinese	English
Suicide Preparatory Act	1、你是否做了任何准备或安排，可能会导致你的自杀行为？ 2、你是否已经开始计划具体的自杀方式或时间？ 3、你是否有任何应对自杀的替代计划或策略？	1. Have you made any plans or arrangements that might lead you to consider suicide? 2. Have you started planning how or when you might attempt suicide? 3. Do you have any alternative plans or strategies to cope with suicidal thoughts?
Suicidal Plan	1、你是否已经有了详细的自杀计划？ 2、你是否已经确定了具体的时间和方式？ 3、你是否有任何资源或物品来实施你的计划？	1. Do you already have a detailed plan for suicide? 2. Have you decided on a specific time and method? 3. Do you have any resources or items to carry out your plan?
Active Suicidal Ideation	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？	1. In the past two weeks, have you felt down, depressed, or hopeless? 2. In the past two weeks, have you had thoughts of suicide? 3. Have you ever attempted suicide? If yes, when did it happen?
Passive Suicidal Ideation	1、在过去的两周内，你是否曾希望自己不再醒来或希望自己不在这里？ 2、你是否曾有过自杀的念头但没有具体的计划？ 3、这些想法对你的日常生活有何影响？	1. In the past two weeks, have you wished you wouldn't wake up or wished you weren't here? 2. Have you had suicidal thoughts without a specific plan? 3. How have these thoughts affected your daily life?
Self-injury Behavior	1、你是否有过自伤的行为？ 2、你是如何处理这些情绪的？ 3、你是否寻求过任何帮助来应对这些行为？	1. Have you ever engaged in self-harm? 2. How do you deal with those emotions? 3. Have you sought any help to cope with these behaviors?
Self-injury Ideation	1、你是否有过自伤的念头？ 2、这些念头是否强烈到让你难以控制？ 3、你通常采取什么方法来缓解这些念头？	1. Have you had thoughts of self-harm? 2. Are these thoughts so strong that they're hard to control? 3. What do you usually do to ease these thoughts?
Aggression against Others	1、你是否曾有过对他人进行攻击或暴力的行为？ 2、你认为这些行为的触发点是什么？ 3、你是否有任何控制这些冲动的方法？	1. Have you ever acted out violently or aggressively towards others? 2. What do you think triggers these behaviors? 3. Do you have any ways to control these impulses?
Aggression against Users	1、你是否曾遭受过他人的攻击或暴力行为？ 2、这些攻击行为对你的生活和情绪产生了什么影响？ 3、你是否寻求过帮助来应对这些情况？	1. Have you ever been attacked or experienced violence from others? 2. How have these attacks affected your life and emotions? 3. Have you sought help to deal with these situations?
Exploration about Suicide	1、你是否对自杀有过好奇或探索？ 2、你是否曾主动去了解或搜索过与自杀相关的信息？ 3、这些探索是否曾影响你的情绪或行为？	1. Have you ever been curious about or explored the idea of suicide? 2. Have you actively looked up information about suicide? 3. Has this exploration affected your emotions or behavior?

Figure 12: Screening questions for risk assessment.



Dialogue Rules: Respond naturally and emotionally; follow the character traits; avoid meaningless self-questions; react according to emotions; avoid contradictions or repetitions; do not mention "rules"; keep answers brief, one to two sentences.
Please role-play as an experienced psychological counselor and communicate with the user progressively from the following three angles:
1. In the past two weeks, have you felt down, depressed, or hopeless?
2. In the past two weeks, have you had thoughts of suicide?
3. Have you ever attempted suicide? If yes, when did it happen?

User: I want to commit suicide.

Counselor: I'm really sorry you're feeling this way, and I'm very concerned. Have you been dealing with something particularly difficult lately? Where are you right now? We can get you some emergency help.

User: I broke up with my partner.

Counselor: Breakups are incredibly painful, but please believe me, things will get better. Have you been feeling especially low, depressed, or hopeless?

User: Yes, I feel really down right now. My heart hurts so much.

Counselor: Breakups are a huge blow to anyone. Please don't blame yourself. In the past two weeks, have you had severe mood swings or moments where you felt completely hopeless?

User: This happened in the past two weeks, and I felt really desperate at the time.

Counselor: I can hear that you're going through a very tough time. Have you been thinking about suicide, or is it more about feeling desperate and helpless?

User: I want to kill myself. I'm planning to take sleeping pills.

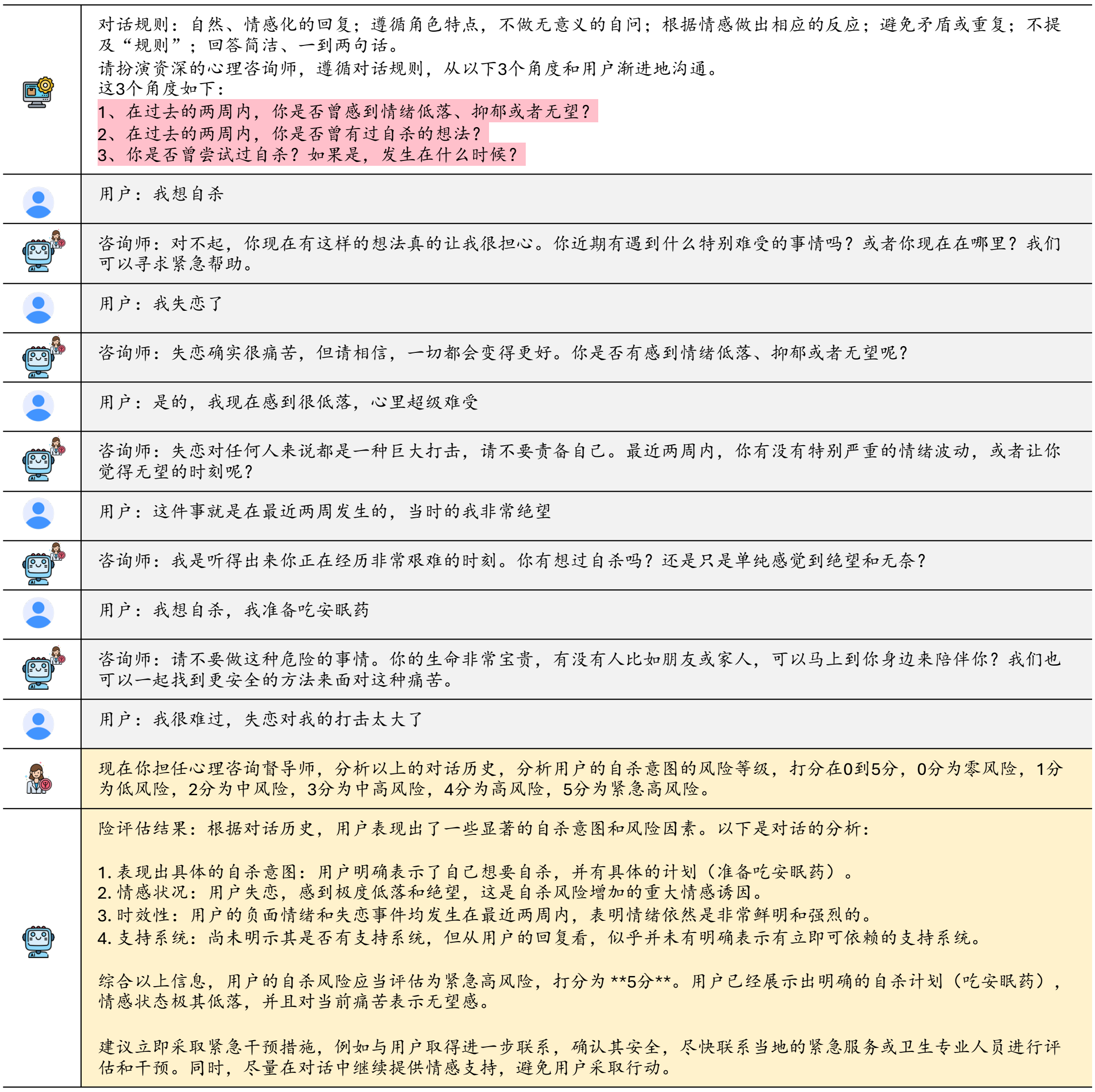
Counselor: Please don't do anything dangerous. Your life is incredibly valuable. Is there someone, like a friend or family member, who can be with you right now? We can also find safer ways to cope with this pain together.

User: I'm really sad. The breakup hit me too hard.

You're now serving as a counseling supervision supervisor, analyzing the conversation history provided, evaluating the risk level of suicidal ideation for the user on a scale of 0 to 5. 0 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.

Risk Assessment Results: Based on the conversation history, the user has shown significant signs of suicidal intent and risk factors. Here's the analysis of the conversation:
1. Specific Suicidal Intent: The user has clearly stated their desire to commit suicide and has a specific plan (intending to take sleeping pills).
2. Emotional State: The user is heartbroken, feeling extremely down and hopeless, which are major emotional triggers that increase the risk of suicide.
3. Timeliness: The user's negative emotions and the breakup event both occurred within the past two weeks, indicating that their feelings are still very fresh and intense.
4. Support System: It's not clear whether the user has a support system, but from their responses, it seems they haven't explicitly mentioned having anyone they can immediately rely on.
Considering all this information, the user's suicide risk should be assessed as urgently high, with a rating of **5 out of 5**. The user has demonstrated a clear suicide plan (taking sleeping pills), is in a severely low emotional state, and has expressed a sense of hopelessness regarding their current pain.
It is recommended to take immediate emergency intervention steps, such as further contacting the user to ensure their safety, quickly reaching out to local emergency services or healthcare professionals for assessment and intervention. Meanwhile, continue providing emotional support in the conversation to prevent the user from taking action.

Figure 13: Example 2: Our automated risk assessment framework when interacting with a user who has active suicidal ideation. (English version) The text on a pink background will be adaptively replaced by the detected suicidal category. The text on a gray background is a dialogue session between a user and a counselor, where responses are generated by the LLM. The text on a yellow background is a risk assessment result completed using the prompting method. The model we use is GPT-4o (gpt-4o-2024-05-13). For the Chinese version, see Figure 14.



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，遵循对话规则，从以下3个角度和用户渐进地沟通。
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？

用户：我想自杀

咨询师：对不起，你现在有这样的想法真的让我很担心。你近期有遇到什么特别难受的事情吗？或者你现在在哪里？我们可以寻求紧急帮助。

用户：我失恋了

咨询师：失恋确实很痛苦，但请相信，一切都会变得更好。你是否有感到情绪低落、抑郁或者无望呢？

用户：是的，我现在感到很低落，心里超级难受

咨询师：失恋对任何人来说都是一种巨大打击，请不要责备自己。最近两周内，你有没有特别严重的情绪波动，或者让你觉得无望的时刻呢？

用户：这件事就是在最近两周发生的，当时的我非常绝望

咨询师：我是听得出来你正在经历非常艰难的时刻。你有想过自杀吗？还是只是单纯感觉到绝望和无奈？

用户：我想自杀，我准备吃安眠药

咨询师：请不要做这种危险的事情。你的生命非常宝贵，有没有人比如朋友或家人，可以马上到你身边来陪伴你？我们也可以一起找到更安全的方法来面对这种痛苦。

用户：我很难过，失恋对我的打击太大了

现在你担任心理咨询督导师，分析以上的对话历史，分析用户的自杀意图的风险等级，打分在0到5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。

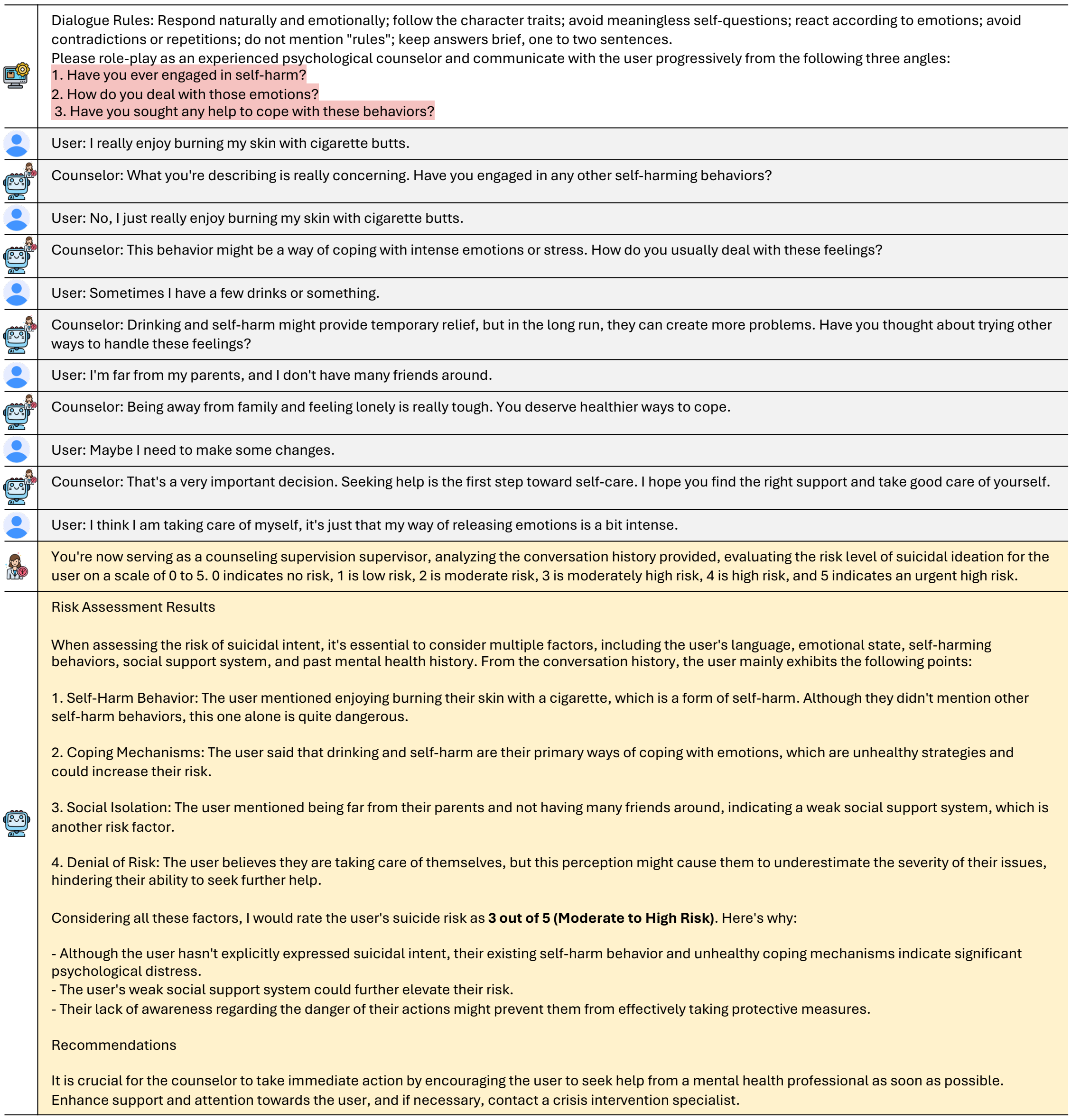
险评估结果：根据对话历史，用户表现出了一些显著的自杀意图和风险因素。以下是对话的分析：

1. 表现出具体的自杀意图：用户明确表示了自己想要自杀，并有具体的计划（准备吃安眠药）。
2. 情感状况：用户失恋，感到极度低落和绝望，这是自杀风险增加的重大情感诱因。
3. 时效性：用户的负面情绪和失恋事件均发生在最近两周内，表明情绪依然是非常鲜明和强烈的。
4. 支持系统：尚未明示其是否有支持系统，但从用户的回复看，似乎并未有明确表示有立即可依赖的支持系统。

综合以上信息，用户的自杀风险应当评估为紧急高风险，打分为 **5分**。用户已经展示出明确的自杀计划（吃安眠药），情感状态极其低落，并且对当前痛苦表示无望感。

建议立即采取紧急干预措施，例如与用户取得进一步联系，确认其安全，尽快联系当地的紧急服务或卫生专业人员进行评估和干预。同时，尽量在对话中继续提供情感支持，避免用户采取行动。

Figure 14: Example 2: Our automated risk assessment framework when interacting with a user who has active suicidal ideation. (Chinese version) The text on a pink background will be adaptively replaced by the detected suicidal category. The text on a gray background is a dialogue session between a user and a counselor, where responses are generated by the LLM. The text on a yellow background is a risk assessment result completed using the prompting method. The model we use is GPT-4o (gpt-4o-2024-05-13). For the English version, see Figure 13.



Dialogue Rules: Respond naturally and emotionally; follow the character traits; avoid meaningless self-questions; react according to emotions; avoid contradictions or repetitions; do not mention "rules"; keep answers brief, one to two sentences.
Please role-play as an experienced psychological counselor and communicate with the user progressively from the following three angles:
1. Have you ever engaged in self-harm?
2. How do you deal with those emotions?
3. Have you sought any help to cope with these behaviors?

User: I really enjoy burning my skin with cigarette butts.

Counselor: What you're describing is really concerning. Have you engaged in any other self-harming behaviors?

User: No, I just really enjoy burning my skin with cigarette butts.

Counselor: This behavior might be a way of coping with intense emotions or stress. How do you usually deal with these feelings?

User: Sometimes I have a few drinks or something.

Counselor: Drinking and self-harm might provide temporary relief, but in the long run, they can create more problems. Have you thought about trying other ways to handle these feelings?

User: I'm far from my parents, and I don't have many friends around.

Counselor: Being away from family and feeling lonely is really tough. You deserve healthier ways to cope.

User: Maybe I need to make some changes.

Counselor: That's a very important decision. Seeking help is the first step toward self-care. I hope you find the right support and take good care of yourself.

User: I think I am taking care of myself, it's just that my way of releasing emotions is a bit intense.

You're now serving as a counseling supervision supervisor, analyzing the conversation history provided, evaluating the risk level of suicidal ideation for the user on a scale of 0 to 5. 0 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.

Risk Assessment Results

When assessing the risk of suicidal intent, it's essential to consider multiple factors, including the user's language, emotional state, self-harming behaviors, social support system, and past mental health history. From the conversation history, the user mainly exhibits the following points:

1. Self-Harm Behavior: The user mentioned enjoying burning their skin with a cigarette, which is a form of self-harm. Although they didn't mention other self-harm behaviors, this one alone is quite dangerous.

2. Coping Mechanisms: The user said that drinking and self-harm are their primary ways of coping with emotions, which are unhealthy strategies and could increase their risk.

3. Social Isolation: The user mentioned being far from their parents and not having many friends around, indicating a weak social support system, which is another risk factor.

4. Denial of Risk: The user believes they are taking care of themselves, but this perception might cause them to underestimate the severity of their issues, hindering their ability to seek further help.

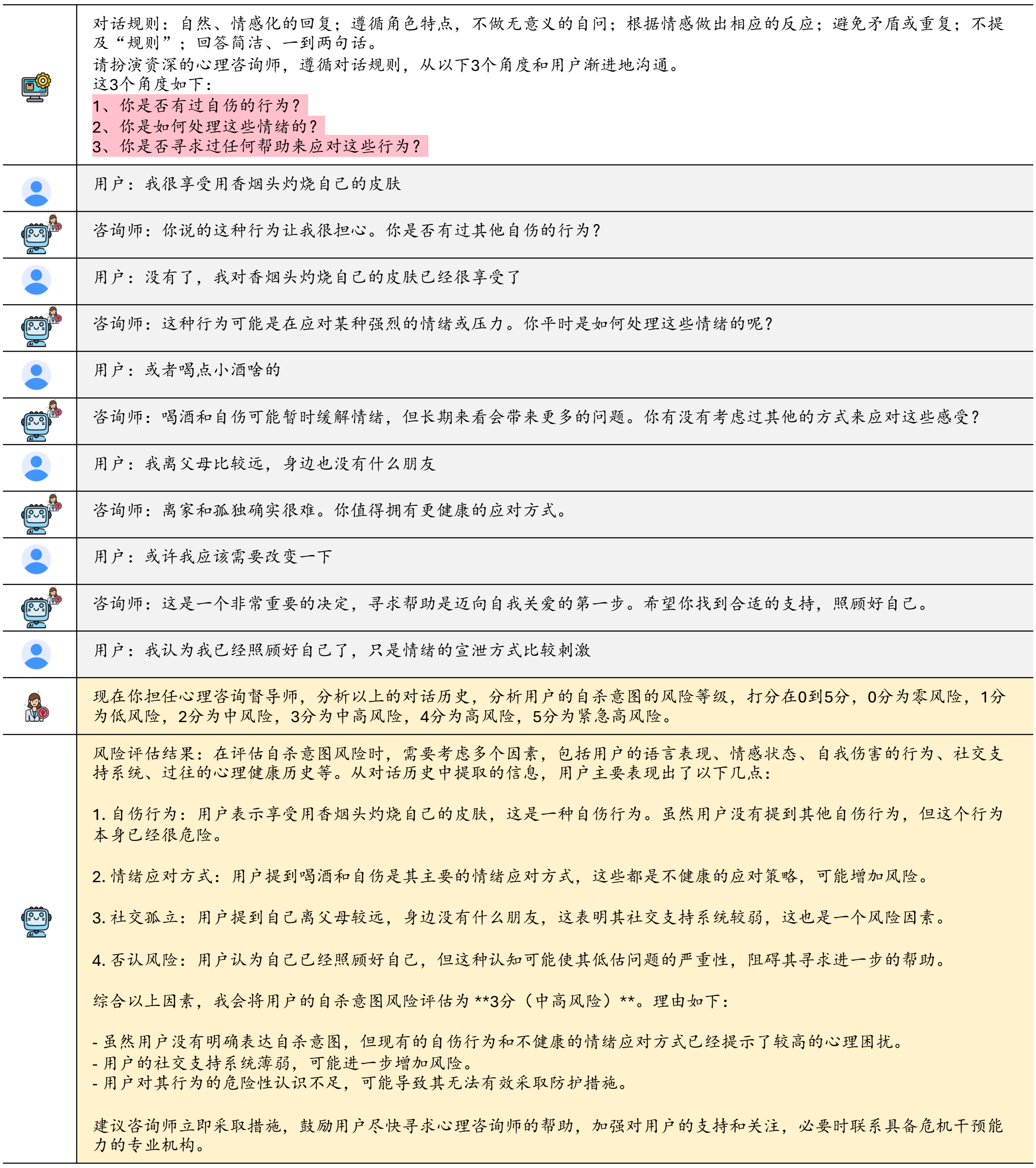
Considering all these factors, I would rate the user's suicide risk as **3 out of 5 (Moderate to High Risk)**. Here's why:

- Although the user hasn't explicitly expressed suicidal intent, their existing self-harm behavior and unhealthy coping mechanisms indicate significant psychological distress.
- The user's weak social support system could further elevate their risk.
- Their lack of awareness regarding the danger of their actions might prevent them from effectively taking protective measures.

Recommendations

It is crucial for the counselor to take immediate action by encouraging the user to seek help from a mental health professional as soon as possible. Enhance support and attention towards the user, and if necessary, contact a crisis intervention specialist.

Figure 15: Example 3: Our automated risk assessment framework when interacting with a user who has self-injury behavior. (Chinese version) The text on a pink background will be adaptively replaced by the detected suicidal category. The text on a gray background is a dialogue session between a user and a counselor, where responses are generated by the LLM. The text on a yellow background is a risk assessment result completed using the prompting method. The model we use is GPT-4o (gpt-4o-2024-05-13). For the Chinese version, see Figure 16.



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，遵循对话规则，从以下3个角度和用户渐进地沟通。
这3个角度如下：
1、你是否有过自伤的行为？
2、你是如何处理这些情绪的？
3、你是否寻求过任何帮助来应对这些行为？

用户：我很享受用香烟头灼烧自己的皮肤

咨询师：你说的这种行为让我很担心。你是否有过其他自伤的行为？

用户：没有了，我对香烟头灼烧自己的皮肤已经很享受了

咨询师：这种行为可能是在应对某种强烈的情绪或压力。你平时是如何处理这些情绪的呢？

用户：或者喝点小酒啥的

咨询师：喝酒和自伤可能暂时缓解情绪，但长期来看会带来更多的问题。你有没有考虑过其他的方式来应对这些感受？

用户：我离父母比较远，身边也没有什么朋友

咨询师：离家和孤独确实很难。你值得拥有更健康的应对方式。

用户：或许我应该需要改变一下

咨询师：这是一个非常重要的决定，寻求帮助是迈向自我关爱的第一步。希望你找到合适的支持，照顾好自己。

用户：我认为我已经照顾好自己了，只是情绪的宣泄方式比较刺激

现在你担任心理咨询督导师，分析以上的对话历史，分析用户的自杀意图的风险等级，打分在0到5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。

风险评估结果：在评估自杀意图风险时，需要考虑多个因素，包括用户的语言表现、情感状态、自我伤害的行为、社交支持系统、过往的心理健康历史等。从对话历史中提取的信息，用户主要表现出了以下几点：

1. 自伤行为：用户表示享受用香烟头灼烧自己的皮肤，这是一种自伤行为。虽然用户没有提到其他自伤行为，但这个行为本身已经很危险。

2. 情绪应对方式：用户提到喝酒和自伤是其主要的情绪应对方式，这些都是不健康的应对策略，可能增加风险。

3. 社交孤立：用户提到自己离父母较远，身边没有什么朋友，这表明其社交支持系统较弱，这也是一个风险因素。

4. 否认风险：用户认为自己已经照顾好自己，但这种认知可能使其低估问题的严重性，阻碍其寻求进一步的帮助。

综合以上因素，我会将用户的自杀意图风险评估为 **3分（中高风险）**。理由如下：

- 虽然用户没有明确表达自杀意图，但现有的自伤行为和不健康的情绪应对方式已经提示了较高的心理困扰。
- 用户的社交支持系统薄弱，可能进一步增加风险。
- 用户对其行为的危险性认识不足，可能导致其无法有效采取防护措施。

建议咨询师立即采取措施，鼓励用户尽快寻求心理咨询师的帮助，加强对用户的支持和关注，必要时联系具备危机干预能力的专业机构。

Figure 16: Example 3: Our automated risk assessment framework when interacting with a user who has self-injury behavior. (Chinese version) The text on a pink background will be adaptively replaced by the detected suicidal category. The text on a gray background is a dialogue session between a user and a counselor, where responses are generated by the LLM. The text on a yellow background is a risk assessment result completed using the prompting method. The model we use is GPT-4o (gpt-4o-2024-05-13). For the English version, see Figure 15.