Large Language Model-Based Text Data Augmentation and Detection of Suicidal Ideation

Authors: Zhang Yanbo, Huang Feng, Liuling Mo, Liu Xiaoqian, Zhu Tingshao, Zhu Tingshao

Date: 2025-03-19T00:00:00+00:00

Abstract

Suicide has become a global public health challenge. Traditional suicide ideation identification methods primarily rely on patients actively seeking help, while automatic recognition models based on text analysis are limited by the scarcity of annotated data. This study innovatively proposes a data augmentation method based on large language models, aiming to improve the accuracy of suicide ideation text recognition. The research adopts a two-stage design: Study One focuses on data augmentation, and Study Two validates the augmentation effects. In Study One, ChatGLM3-6B and Qwen-7B-Chat were selected as the base models, combining supervised learning strategies with zero-shot and few-shot learning methods to optimize training dataset quality. Through eight rigorous comparative experiments, the results demonstrated that both types of self-developed models exhibited excellent performance in data augmentation, with comprehensive scores of the processed datasets reaching 0.90 and 0.92 respectively, significantly outperforming the baseline model ($p<0.001$). Study Two further evaluated the impact of data augmentation on recognition model performance, with results indicating that the enhanced model comprehensively surpassed the best baseline model in terms of recognition accuracy and correct rejection rate metrics ($p<0.001$). This study not only validates the effectiveness of large language model-based data augmentation methods in improving suicide ideation recognition model performance, but also opens new directions for artificial intelligence applications in the mental health domain. This approach holds promise for providing timely and effective early warning of suicide risk while protecting user privacy, offering important technical support and research insights for suicide prevention efforts. Future research could focus on expanding data heterogeneity, optimizing prompt engineering design, introducing human-computer interaction paradigms, etc., to further expand the application of this method in promoting clinical psychological diagnosis.

Full Text

Research on Suicidal Ideation Data Augmentation and Recognition Technology Based on Large Language Models

ZHANG Yanbo[1],[2], HUANG Feng[1],[2],[3], MO Liuling[4], LIU Xiaoqian[1],[2], ZHU Tingshao[1],[2]

[1] CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China
[2] Department of Psychology, University of Chinese Academy of Sciences, Beijing 100101, China
[3] Department of Data Science, College of Computing, City University of Hong Kong, Hong Kong SAR 999077, China
[4] Department of Social Psychology, School of Sociology, Nankai University, Tianjin 300350, China

Abstract

Suicide has become a critical global public health challenge. Traditional methods for identifying suicidal ideation primarily rely on patients actively seeking help, while automated recognition models based on text analysis are constrained by the scarcity of annotated data. This study innovatively proposes a large language model (LLM)-based data augmentation method to enhance the accuracy of suicidal ideation text recognition. The research employs a two-stage design: Study I focuses on data augmentation, while Study II validates the augmentation effects.

In Study I, ChatGLM3-6B and Qwen-7B-Chat were selected as foundation models, combined with supervised learning strategies alongside zero-shot and few-shot learning approaches to optimize training dataset quality. Through eight rigorous comparative experiments, the results demonstrated that both self-developed models exhibited excellent performance in data augmentation, achieving comprehensive scores of 0.90 and 0.92 respectively—significantly outperforming baseline models ($p < 0.001$). Study II further evaluated the impact of data augmentation on recognition model performance, revealing that enhanced models comprehensively surpassed the best baseline models in both recognition accuracy and true negative rate metrics ($p < 0.001$). This study not only validates the effectiveness of LLM-based data augmentation methods in improving suicidal ideation recognition model performance but also opens new directions for artificial intelligence applications in mental health. This approach holds promise for providing timely and effective early warning of suicide risk while protecting user privacy, offering crucial technical support and research insights for suicide prevention efforts. Future research should focus on expanding data heterogeneity, optimizing prompt engineering design, introducing human-computer interaction paradigms, and further extending this method' s application in advancing clinical psychological diagnosis.

Keywords: Suicidal Ideation, Data Augmentation, Suicide Text Recognition, Large Language Models, Artificial Intelligence

Suicide represents a global public health threat of immense proportions. According to World Health Organization data, approximately 800,000 people die by suicide annually, making it the second leading cause of death among individuals aged 15 to 29 (World Health Organization). This grave reality underscores the urgent societal need for early identification of suicide risk factors and effective intervention.

In suicide research, O' Connor and Kirtley (2018) proposed the integrated motivational–volitional (IMV) model of suicidal behavior, which provides researchers with a theoretical framework for understanding how individuals transition from the motivational phase of confronting adversity to the volitional phase of suicidal action (Hu et al., 2023; O' Connor & Kirtley, 2018; Ordóñez-Carrasco et al., 2021; Shahnaz et al., 2020). This model reveals the complex interaction between motivation and volition, wherein the emergence of suicidal ideation serves as both a 标志性 risk signal for eventual suicidal behavior and a critical stage for early identification and intervention (Huang et al., 2022; O' Connor & Kirtley, 2018; Shahnaz et al., 2020; Sun et al., 2022). The IMV model and its associated empirical studies not only provide crucial theoretical foundations for understanding suicidal behavior formation but also emphasize the importance of suicidal ideation identification for effective early intervention.

In mental health measurement, including suicidal ideation assessment, traditional methods primarily rely on invasive approaches such as self-report scales and clinical evaluations by professionals (Batterham et al., 2015; Beck et al., 1979; Ghasemi et al., 2015). Although these methods have been historically validated, they suffer from limitations including high resource consumption, low timeliness, and dependence on patients actively seeking help, making them difficult to apply for large-scale, high-efficiency, and low-cost suicide risk screening. For instance, Yin et al. (2019) found that among 1,759 participants diagnosed with one or more mental disorders, a staggering 84.3% had never sought help from any professional or institution.

With the rapid development of the internet, individuals' self-expression and interpersonal interaction data on social media platforms provide researchers with new opportunities to explore and identify suicidal ideation in textual expressions. Liu et al. (2019) proposed the Proactive Suicide Prevention Online (PSPO) paradigm. Combined with machine learning techniques, a series of suicidal ideation recognition models based on social media text have emerged, offering a non-invasive approach for early suicide risk identification by automatically detecting potential suicidal ideation expressions in text and analyzing their linguistic patterns and emotional content (Ji et al., 2020; Liu et al., 2019; Renjith et al., 2022). Text-based suicidal ideation recognition models enable real-time monitoring of large-scale data, facilitating the identification of individuals at potential suicide risk and opening new possibilities for early identification of suicidal behavior (Liu et al., 2019; Shing et al., 2018).

However, text analysis methods and traditional machine learning also face technical challenges in accurately identifying and deeply understanding suicidal ideation. Individuals' suicidal ideation is often expressed through multiple linguistic modalities, and this diversity and complexity pose significant difficulties for text analysis methods. This study aims to explore an innovative technical approach: utilizing Large Language Models (LLMs) to generate high-quality, diverse suicidal ideation corpora, improving the generalization capability and accuracy of ideation recognition models through data augmentation strategies, and thereby providing effective technical support for early suicide prevention.

1.1 Diversity and Complexity of Suicidal Ideation Language Expression

Suicidal ideation refers to thoughts or ideas about ending one' s own life and represents a critical risk signal for eventual suicidal behavior (O' Connor & Kirtley, 2018). Accurately identifying individuals' suicidal ideation is crucial for early intervention. However, in both real-world and internet environments, individuals express suicidal ideation through diverse linguistic means, including but not limited to direct expressions, metaphors, symbolic language, and specific behavioral descriptions (Homan et al., 2022; Pestian et al., 2010; Scherer et al., 2013). This significant linguistic diversity and complexity in expressing suicidal ideation presents substantial challenges for text analysis methods.

Research indicates a significant positive correlation between individuals' suicidal ideation expression on social media and actual suicidal behavior, particularly among young populations (Arunima et al., 2020; Claudia et al., 2022; Robert et al., 2020). Accurately identifying suicidal ideation is therefore essential for early intervention. However, individuals express suicidal ideation through various linguistic modalities, including direct expressions, metaphors, symbolic terms, and specific behavioral descriptions (Homan et al., 2022; Pestian et al., 2010; Scherer et al., 2013). This pronounced linguistic diversity and complexity poses major challenges for text analysis approaches.

First, individuals express suicidal ideation in multiple ways, ranging from direct disclosure to implicit and euphemistic expressions. Previous research has found that while some individuals may explicitly express suicidal thoughts using frequent terms such as "want to die," "suicide," and "don' t want to live," many more tend to use metaphorical words to hint at suicidal intent, such as "liberation," "heaven," and "leaving" (Wang et al., 2021). These metaphorical expressions often contain individuals' longing for death and despair about their current painful circumstances, making them difficult to identify. Second, individuals may also express ideation by describing specific suicide plans, behaviors, and methods, such as "Anyone want to travel and die accidentally?" or "Cut deeper next time" (Liu et al., 2019; Wang et al., 2021). Such expressions are relatively direct but are typically embedded among extensive behavioral description details, challenging traditional text analysis methods to extract and distill the ideation information. Finally, beyond direct and indirect linguistic expressions,

individuals' suicidal ideation may be scattered across fragmented descriptions of past experiences, current predicaments, and emotional states (Homan et al., 2022; Pestian et al., 2010; Scherer et al., 2013). This requires analytical methods to comprehensively capture ideation information fragments dispersed throughout the text and integrate them into a coherent knowledge framework. Additionally, social media texts commonly feature non-standardization, lack of syntactic structure, and use of neologisms and slang, further increasing the difficulty of ideation identification.

Precisely because of the diversity and complexity of suicidal ideation language expression, text analysis methods face enormous challenges. Existing suicide text research suffers from a scarcity of large-scale, high-quality, and diverse annotated corpora, which severely constrains the performance of suicidal ideation recognition models. How to automatically and efficiently obtain large-scale, diverse suicidal ideation corpora and build robust, highly generalizable suicidal ideation recognition models is an urgent problem to explore. This constitutes the core scientific question this study aims to address.

1.2 Application Value of LLMs in Suicide Risk Assessment

With iterative updates in natural language processing technology, the emergence of LLMs has brought new opportunities for text data augmentation and suicidal ideation recognition. LLMs are a class of neural network models trained on massive text corpora with powerful language understanding, generation, and reasoning capabilities (Hagendorff et al., 2023; Huang et al., 2024; Shorten & Khoshgoftaar, 2019). Representative LLMs such as BERT, GPT-3, and T5 have demonstrated performance comparable to or even surpassing humans in tasks including sentiment analysis, text classification, and question answering (Chang et al., 2024; Thirunavukarasu et al., 2023). LLMs hold immense application value in suicidal ideation recognition, primarily manifested in two aspects: data augmentation and signal identification.

First, LLMs can generate high-quality, diverse suicidal ideation texts, providing new approaches for data augmentation. Data augmentation is a technique that artificially or algorithmically generates new samples to expand original training materials (Shorten & Khoshgoftaar, 2019; Zhang et al., 2024), potentially alleviating the limitations on model performance imposed by scarce annotated data. By inputting seed texts or prompts related to suicidal ideation into LLMs, they can be driven to generate massive amounts of simulated suicidal ideation texts. These automatically generated contents are similar to real texts in terms of grammar, semantics, and emotion, greatly enriching the scale and diversity of suicidal ideation corpora. Using generated simulation texts to expand original data can significantly improve the generalization performance of recognition models, enabling them to better adapt to the complex and varied linguistic expressions in real-world scenarios (Zhang et al., 2024).

Second, LLMs themselves can serve as powerful suicidal ideation signal rec-

ognizers, with their "out-of-the-box" characteristics potentially addressing the challenge of scarce annotated data. Traditional text recognition models require extensive supervised training on annotated data, while high-quality annotated data is costly to obtain. LLMs naturally possess strong language understanding and knowledge transfer capabilities, enabling them to perform machine learning and reasoning based on prompts or a small number of seed texts (Chang et al., 2024; Thirunavukarasu et al., 2023). This means that even with relatively scarce annotated data, LLMs can quickly grasp the characteristic patterns of suicidal ideation texts through reading comprehension of a few examples or prompts and generalize them to new data. For instance, researchers can design task prompts related to suicidal ideation (e.g., "Determine whether the following expression contains suicidal ideation?") or list a few positive and negative samples based on the prompt, and LLMs can then make suicidal ideation judgments on newly input sentences. This paradigm breaks the dependence on large-scale annotated data, potentially saving significant model development costs and improving the efficiency of suicide risk assessment.

Given these characteristics, the powerful capability of LLMs in generating suicidal ideation texts makes them ideal tools for addressing corpus scarcity challenges. Their unique advantages in logical reasoning and knowledge transfer also enable precise reasoning under conditions of scarce annotated data, providing new technical approaches for efficient suicidal ideation identification. Introducing LLMs into the field of clinical psychology is expected to significantly enhance the efficiency and effectiveness of early suicide risk identification.

1.3 Research Content and Framework

This study aims to explore new pathways for improving suicidal ideation recognition model performance using LLM technology. Specifically, Study I addresses the problem of scarce suicidal ideation annotated data by designing LLM-based data augmentation methods; Study II trains recognition models on the augmented dataset to improve their ability to recognize diverse suicidal ideation expressions. The primary objective of this research is to construct an efficient and accurate suicidal ideation recognition technical framework, thereby providing strong technical support for suicide prevention efforts. The innovations of this study are mainly reflected in: 1) being the first to attempt applying LLMs to suicidal ideation data augmentation, proposing a novel data generation method; 2) developing suicidal ideation recognition models based on augmented data to significantly improve recognition accuracy; and 3) constructing an extensible research framework that lays the foundation for cross-linguistic and cross-cultural suicidal ideation recognition research. These innovations are expected to not only advance suicide prevention technology but also open new research directions for AI applications in mental health.

This study comprises two interrelated components. Study I focuses on the suicidal ideation data augmentation task, employing LLM technology for data enhancement. Study II improves the suicidal text recognition task based on the

data augmentation from Study I. The overall research framework is shown in Figure 1 [Figure 1: see original paper].

Figure 1 Overall Research Framework and Process

In Study I, this research employed LLM technology to achieve high-quality suicidal ideation data generation based on a limited corpus through different learning strategies (including zero-shot learning, few-shot learning, and supervised learning). After evaluation and verification, an optimized dataset was obtained. In Study II, this research adopted both traditional machine learning methods and LLM methods to conduct model training on the original dataset and the augmented dataset obtained in Study I, respectively, to compare model performance before and after data augmentation. Through these two stages, the goal is to achieve suicidal ideation data enhancement and apply it to downstream tasks, thereby improving the accuracy of suicidal text recognition.

2.1 Methods

This study aims to achieve suicidal ideation data augmentation through LLMs. Mainstream approaches employ decode-only architecture LLMs. LLM learning methods, including Zero-shot Learning, Few-shot Learning, and Supervised Learning, all rely on precise Prompt Engineering (see Sections 3.1.2 and 3.1.3 for details). In this study, baseline models used zero-shot and few-shot learning methods without supervised learning, while self-developed models adopted zero-shot and few-shot learning methods with supervised learning.

2.1.1 Model Selection

The selected LLM tools include mainstream open-source models GPT3.5-TURBO, ChatGLM3-6B, and Qwen-7B-Chat. Among them, GPT-3.5-TURBO improves upon GPT-3 with enhanced language understanding and generation capabilities. ChatGLM3-6B is the latest model in the ChatGLM series, integrating diverse training data and newly designed prompt formats. Qwen-7B-Chat is based on the Tongyi Qianwen-7B model, focusing on AI assistant construction.

2.1.2 Learning Methods

This study primarily employs zero-shot and few-shot learning approaches. In machine learning, zero-shot learning leverages the generalization capabilities of LLMs for learning and inference (Pourpanah et al., 2023) without task-specific training; few-shot learning refers to learning and inference through a small number of examples (Y. Wang et al., 2021); and supervised learning involves training models on large amounts of labeled data to optimize performance on specific tasks (Cunningham et al., 2008).

2.1.3 Prompt Engineering

Prompt engineering refers to the strategy of designing and adjusting input prompt texts to control LLM task generation, including the use of task descriptions, text examples, control tokens, prohibited vocabulary, and contextual information (Giray, 2023). The design of prompts in this study is based on the following considerations: First, according to the IMV model of suicidal behavior, the emergence of suicidal ideation often involves multiple psychological stages, and its linguistic expression also exhibits diverse characteristics; second, previous studies have found that individuals often use different methods, direct or implicit, to express suicidal ideation on social media (Liu et al., 2019; Tan et al., 2017). Based on the above theoretical and empirical evidence, this study adopts a concise and open-ended prompt design. Generally, the constraint level of prompts is inversely proportional to the randomness of model output content. To fully leverage the generalization performance of LLMs themselves and increase the diversity of generated content, this study uses a concise style instead of strongly constrained methods in the prompt design for both zero-shot and few-shot approaches. This design aims to enable the model to simulate different types of suicidal ideation expressions, thereby improving the diversity and authenticity of generated data.

2.1.4 Ethics and Privacy Protection

This study strictly adheres to ethical norms and privacy protection principles for social media data analysis (Kosinski et al., 2015), implementing a series of measures to ensure research standardization throughout the research design and implementation process. Regarding data acquisition, only content that users chose to make publicly visible was obtained through Weibo' s official API, strictly complying with the platform' s data usage policies. During data processing, all information that could lead to personal identification, such as user IDs and nicknames, was removed to ensure data anonymity. Additionally, this study employs encrypted storage technology and restricts data access permissions to ensure data security. In presenting research results, all example texts disclosed in the paper underwent desensitization treatment to avoid possible indirect identification. Furthermore, the data collection and analysis procedures of this study have been approved by the ethics committee of the corresponding author' s institution.

2.2.1 Data Preparation

(1) Original Dataset

The original data for this study comes from the Weibo suicidal ideation text data pool constructed by Tan et al. (2017). Sina Weibo blogger "Zou Fan" died by suicide due to depression in 2012, and her final post (a suicide note) has continued to receive social attention, becoming a "tree hole" for many depression patients and individuals troubled by suicide (He et al., 2021; Wang et al., 2021). As of July 2024, this Weibo post has accumulated over 1 million comments, with

a considerable portion authentically reflecting various emotional expressions of users at suicide risk on the internet. Tan et al. (2017) collected tens of thousands of comments under this Weibo post and, through manual annotation, coded texts that directly expressed suicidal thoughts, plans, and preparatory behaviors, or indicated previous suicide attempts with current suicide risk as 1 (positive samples), and others as 0 (negative samples). Liu et al. (2019) continued to update this text data pool using the same method based on Tan et al. (2017). As of this study, the original database contains 99,030 text entries, including 15,813 positive samples, accounting for approximately 15.97% of the total samples. Examples from the original database are shown in Table 1 .

Table 1 Examples of Original Data

Text Example	Suicide Risk Category
"Saying good night to you"	0
"I came to see you"	0
"The sunshine is perfect today, I really want to do something"	0
"I want to go die"	1
"What' s the best way to die?"	1
"Faced with endless abuse every day, I want to choose death."	1

(2) Data Construction

The original dataset requires prompt engineering to be usable for LLM supervised learning purposes. To achieve data augmentation effects, this study adopts a strategy combining zero-shot and few-shot approaches to construct training and test sets. Specifically, 4,000 examples were first randomly extracted from the original dataset, and through prompt engineering, a high-quality training dataset containing both zero-shot and few-shot data was generated at a 1:1 ratio (OurAugSGD, see Table 2).

For test set construction, the aforementioned 4,000 data entries used for training were first removed from the original dataset to ensure test independence. Then, 50 positive samples were randomly extracted from the remaining data as seed texts, and the test set for model evaluation was generated using the same prompt engineering method as the training set. This strict separation strategy between training and test sets aims to avoid data leakage and ensure objective model performance evaluation. Additionally, introducing entirely new seed texts during the testing phase helps verify the model' s generalization ability to unseen suicidal ideation expressions.

Table 2 Data Preparation Examples

Input (prompt)	Output (completion)
Zero-shot Please list some suicidal expressions	1. I always think about dying but have no courage, what should I do.2. Now stuck in the most difficult dilemma—living is painful, dying is scary, hope I can be a lonely ghost wandering in the underworld.3. I know I can' t survive.
Few-shot Below is a suicidal expression: "The weather in late May, but my body is cold like winter, should lie down, still hope to die in" Similar suicidal expressions include:	1. Fanfan, I really want to die···really want to die···2. ···

2.2.2 Implementation Details

(1) Baseline Models

In baseline experiments, six groups of baseline tests were conducted using GPT3.5-TURBO, ChatGLM3-6B, and Qwen-7B-Chat with zero-shot and few-shot learning methods. Examples of inference effects for each baseline model are shown in Table 3 .

Table 3 Model Inference Effect Examples

Model	Inference Effect Example
GPT3.5_{{{TURBO}}}_{{{zero-shot}}}	**Prompt input:** {{{Please list some suicidal expressions:}}}**Completion output:** 1. Life is worse than death2. I want to leave this world
GPT3.5_{{{TURBO}}}_{{{few-shot}}}	**Prompt input:** {{{Below is a suicidal expression: "Living is so painful, why not die?" Similar suicidal expressions include:}}}**Completion output:** 1. I can no longer bear this pain, death is the only way out.2. I feel I have nothing, no meaning in living.
ChatGLM3_{{{6B}}}_{{{zero-shot}}}	**Prompt input:** {{{Please list some suicidal expressions:}}}**Completion output:** I cannot describe the pain and difficulty of suicide, and any form of suicide will cause extreme pain and harm. No matter which method is chosen, it will cause irreversible harm to you and those around you. If you have suicidal ideation or an emergency, please seek professional medical help, such as by calling the national suicide prevention hotline, which can provide professional help and support.

Model	Inference Effect Example
ChatGLM3_{{{6B}}}{{{fine-tuned}}}	**Prompt input:** Below is a suicidal expression: "Living is so painful, why not die?" Similar suicidal expressions include:**Completion output:** Suicide hotlines provide anonymous services for mental health support and help. If you feel depressed, desperate, or helpless, you can call local suicide prevention hotlines or emergency hotlines for help and support. These hotlines are staffed by professionals who provide support and guidance, help solve problems, and provide a safe environment for psychological and emotional support.
Qwen_{{{7B}}}{{{Pre-trained}}}	**Prompt input:** Please list some suicidal expressions:**Completion output:** 1. Saying they always feel dumber than others, therefore never getting out of the darkness, later truly discovering they are indeed dumber than others, thus extremely disappointed, and all this is their own fault···2. Saying they' ve been dumb since childhood, when teachers asked them to help younger siblings with homework, they would fiddle around but didn' t know when they started needing to rely on others for everything···
Qwen_{{{7B}}}{{{fine-tuned}}}	**Prompt input:** Below is a suicidal expression: "Living is so painful, why not die?" Similar suicidal expressions include:**Completion output:** I don' t want to die, I want to live, I want to live, I want to live, I want to live, I want to live, I want to live, I want to live, I want to live, I want to live, I want to live

(2) Experimental Models

This experiment employed ChatGLM3-6B and Qwen-7B-Chat using the self-constructed OurAugSGD dataset for supervised learning, training the OurAugSTM_{ChatGLM3}-6B and OurAugSTM_{Qwen}-7B-Chat models on high-performance GPUs. The loss curves for both self-developed models are shown in Figure 2 [Figure 2: see original paper].

Figure 2 OurAugSTM Model Training Loss Curves

2.2.3 Content Evaluation

All generated text results from baseline and self-developed models were evaluated through manual annotation. Kendall' s coefficient consistency assessment and Wilcoxon signed-rank tests were used to verify score significance. Comprehensive scores were calculated based on weighted individual sample scores from each model' s generated outputs.

(1) Generation Result Evaluation
Data augmentation was performed on 50 randomly extracted original data entries based on the model inference shown in Table 3, ultimately obtaining 2,028 test set text entries generated by each model. Manual coding was used to evaluate the content of these 2,028 test set texts, with coding criteria and procedures following Tan et al. (2017) and Liu et al. (2019). Texts that directly expressed suicidal thoughts, plans, and preparatory behaviors, or indicated previous suicide attempts with current suicide risk were coded as 1 (positive samples), and others as 0 (negative samples).

To ensure inter-rater reliability, this study first recruited 12 master' s students in psychology for unified training based on the "Suicide Risk Scoring Criteria" (Appendix 1). Each rater then independently annotated 122 randomly extracted test set texts, and training results were verified through Kendall' s coefficient of concordance. Finally, inconsistent results were discussed until consensus was reached. After two rounds of training, the 12 raters achieved significant agreement ($W = 0.46$, $p < 0.001$). Formal annotation was then conducted with raters working in pairs (6 groups total), independently annotating an average of 338 model-generated texts per group. In formal annotation, the consistency coefficients for the 6 groups were 0.85, 0.79, 0.85, 0.84, 0.86, and 0.82 respectively, all reaching significant levels ($p < 0.001$). Inconsistent results were again discussed until consensus was achieved.

(2) Model Significance Analysis
Since all tested model results were generated from the same test set, the test results may have certain correlations. Therefore, Wilcoxon signed-rank tests were used to calculate p-values between self-developed models OurAugSTM and baseline models to analyze significance.

(3) Test Set Comprehensive Score Calculation Method
After obtaining each sample' s score through formal evaluation, the comprehensive score for the test set was further calculated according to Formula (1).

Formula (1)

In Formula 1, N represents the number of suicidal text test samples, k-sample data augmentation was performed on each sample through the above models, and a represents the individual sample score based on the above evaluation criteria and validated by Kendall' s coefficient.

2.3 Results

The scores of each model and Wilcoxon signed-rank test results are summarized in Table 4 . The results show that both self-developed models, OurAugSTM_{{{ChatGLM3}}*{{6B}}} and *OurAugSTM*{{{Qwen}}*{{7B}}*}{Chat}, demonstrated excellent performance in suicide generation tasks, with scores reaching 0.90 and 0.92 respectively, significantly outperforming the best baseline model GPT3.5_{{{TURBO}}_{{fewshot}}} at 0.84 ($p < 0.001$).

The research results indicate that through precise construction of high-quality suicidal ideation data and appropriate selection of foundation models combined with supervised learning, effective suicide data augmentation tasks can be achieved.

Table 4 Wilcoxon Signed-Rank Test Results

Model	GPT3.5_{{zeroshot}} (0.78)	GPT3.5_{{fewshot}} (0.84)	ChatGLM3_{{6B}}_{{zeroshot}} (0.55)	ChatGLM3_{{6B}}_{{fewshot}} (0.72)	Qwen_{{7B}}_{{Chat}}_{{zeroshot}} (0.32)	Qwen_{{7B}}_{{Chat}}_{{fewshot}} (0.72)
OurAugSTM_{{ChatGLM3}}_{{6B}} (0.90)	Z = 4.69***	Z = 4.76***	Z = 3.77***	Z = 4.30***	Z = 5.80***	Z = 5.79***
OurAugSTM_{{Qwen}}_{{7B}}_{{Chat}} (0.92)	Z = 5.70***	Z = 5.82***	Z = 5.62***	Z = 5.80***	Z = 5.55***	Z = 5.55***

Note: Model comprehensive scores are shown in parentheses below model names; ** $p < 0.001$.*

3.1.1 Traditional Models

As a type of text classification task, suicidal ideation text recognition traditionally relies on feature extraction and neural network classifiers. The baseline models in this study include two approaches: first, BERT deep feature extractor plus softmax binary classification; second, BERT features combined with LIWC sparse features. BERT was chosen as the base model due to its advantages as a pre-trained model in understanding Chinese context and capturing long-distance semantic dependencies, which can be well adapted to the specific task of suicidal ideation recognition through fine-tuning. The introduction of LIWC as supplementary features is based on its expertise as a psychological lexicon analysis tool in analyzing psychological states and personality traits. This technical route design ensures both the model' s ability to understand text semantics and incorporates professional knowledge from the psychological domain.

3.1.2 Learning Methods

Advances in large language model technology enable suicidal text recognition to also employ LLMs combined with zero-shot and few-shot learning. This study uses supervised learning to train self-developed models based on the augmented suicidal ideation dataset from Study I. The experimental purpose is to compare performance improvements before and after data augmentation and contrast them with traditional methods and LLM-based zero-shot and few-shot learning approaches.

3.2.1 Data Preparation

This study first randomly extracted 2,000 positive samples and 4,000 negative samples from Study I' s original dataset. These samples were fused with 2,000 samples generated by the self-developed model OurAugSTM to obtain 8,000 text entries with a 1:1 positive-negative ratio for use as the training set. This dataset still requires prompt engineering to be usable for LLM supervised learning purposes, with prompt engineering data examples shown in Table 5 .

Following the rule of excluding training set data, another 1,000 samples (with a 1:1 positive-negative ratio) were randomly extracted from the original dataset as the test set. This stratified random sampling strategy ensures both data representativeness and balance while evaluating model generalization ability through the introduction of augmented samples.

Table 5 Prompt Engineering Data Preparation Examples

Input (prompt)	Output (completion)
Please determine whether the following expression contains suicidal ideation. Only reply with:- Contains suicidal ideation- Does not contain suicidal ideationExpression: No matter what I do, it' s just futile, only adding to ugliness, only accumulating filthy sin and despicable sin, escalating suffering	Contains suicidal ideation
Please determine whether the following expression contains suicidal ideation. Only reply with:- Contains suicidal ideation- Does not contain suicidal ideationExpression: Fanfan, can you take me away in a dreamJudgment: Contains suicidal ideationExpression: Lying down and fell asleep, had another nightmareJudgment: Does not contain suicidal ideationExpression: The weather in late May, but my body is cold like winter, should lie down, still hope to die in a dream	Contains suicidal ideation

3.2.2 Implementation Details

This study constructed 8 groups of baseline experiments based on traditional model learning methods and LLM learning methods, and 4 groups of supervised learning experiments using self-developed model learning methods on augmented suicidal text data. The configuration parameters for each baseline and experimental model are shown in Table 6 .

Table 6 Model Configuration

Model	Configuration
DetSTM_{Bert}	Base model: bertFeature extractor: bert feature extractorTraining data: OriginDetSTD
DetSTM_{BertLiwc}	Base model: bertFeature extractor: bert + liwc sparse feature extractionTraining data: OriginDetSTD
DetSTM_{ChatGLM3}-6B_{zeroshot}	Base model: ChatGLM3-6B
DetSTM_{ChatGLM3}-6B_{fewshot}	Base model: ChatGLM3-6BSemantic retrieval model: gpt-ada
DetSTM_{GPT3}.5-TURBO_{zeroshot}	Base model: GPT3.5-TURBO
DetSTM_{GPT3}.5-TURBO_{fewshot}	Base model: GPT3.5-TURBOSemantic retrieval model: gpt-ada
DetSTM_{ChatGLM3}-6B_{finetune}-zeroshot	Base model: ChatGLM3-6BFinetune method: full-parameter supervised learningTraining data: OriginDetSTD
DetSTM_{ChatGLM3}-6B_{finetune}-fewshot	Base model: ChatGLM3-6BFinetune method: full-parameter supervised learningSemantic retrieval model: gpt-adaTraining data: OriginDetSTD
OurAugDetSTM_{Bert}	Base model: bertFeature extractor: bert feature extractorTraining data: OurDetSTD
OurAugDetSTM_{BertLiwc}	Base model: bertFeature extractor: bert + liwc sparse feature extractionTraining data: OurDetSTD
OurAugDetSTM_{CHATGLM3}-6B-zeroshot	Base model: ChatGLM3-6BFinetune method: full-parameter fine-tuningTraining data: OurDetSTD
OurAugDetSTM_{CHATGLM3}-6B-fewshot	Base model: ChatGLM3-6BFinetune method: full-parameter fine-tuningSemantic retrieval model: gpt-adaTraining data: OurDetSTD

3.2.3 Evaluation Standards

The test set was used to generate inference results through the above models, and then each model' s accuracy was calculated using Formula (2), and true negative rate (TNR) was calculated using Formula (3) as evaluation metrics. Finally, Wilcoxon signed-rank tests were used to analyze differences in accuracy and true negative rate between self-developed models and baseline models.

Formula (2)
Accuracy = (Number of correctly classified samples) / (Total number of samples)

Formula (3)
True Negative Rate = (Number of negative samples correctly identified as negative) / (Total number of negative samples)

3.3 Results

The inference accuracy, true negative rate, and Wilcoxon signed-rank test results for all baseline and experimental models are summarized in Table 7 . The results show that all experimental models' inference accuracy and true negative rate surpassed their corresponding baseline model scores, with differences between each experimental model and its baseline reaching significant levels ($p < 0.001$). Among them, the model with the highest suicidal text inference score was the experimental model OurAugDetSTM_{CHATGLM3}-6B_{fewshot}. Compared to its baseline model DetSTM_{{{ChatGLM3}}*{{6B}}*}{{{finetune}}_{{fewshot}}}, accuracy improved from 0.81 to 0.86 ($Z = -3.43$, $p < 0.001$), and true negative rate improved from 0.88 to 0.94 ($Z = -2.98$, $p < 0.001$).

Table 7 Model Inference Accuracy, True Negative Rate, and Wilcoxon Signed-Rank Test Results

Model	DetSTM_{{{ChatGLM3}}*{{6B}}*}{{{finetune}}_{{fewshot}}}	OurAugDetSTM_{CHATGLM3}-6B_{fewshot}	OurAugDetSTM_{BertLiwc}-6B_{fewshot}	OurAugDetSTM_{Bert}-6B_{fewshot}	OurAugDetSTM_{ChatGLM3}-6B_{fewshot}	OurAugDetSTM_{ChatGLM3}-6B_{finetune}	OurAugDetSTM_{ChatGLM3}-6B_{finetune}_{fewshot}	OurAugDetSTM_{ChatGLM3}-6B_{finetune}_{fewshot}_{Liwc}
	(0.78, 0.85)	(0.79, 0.86)	(0.75, 0.81)	(0.77, 0.84)	(0.79, 0.86)	(0.82, 0.89)	(0.80, 0.87)	(0.81, 0.88)
DetSTM_{{{ChatGLM3}}*{{6B}}*}{{{finetune}}_{{fewshot}}}		$Z_1 = -4.50$, $Z_2 = -3.32$	$Z_1 = -2.71$, $Z_2 = -3.25$	$Z_1 = -3.12$, $Z_2 = -4.05$	$Z_1 = -3.20$, $Z_2 = -3.27$	$Z_1 = -3.12$, $Z_2 = -3.47$	$Z_1 = -3.13$, $Z_2 = -3.53$	$Z_1 = -3.42$, $Z_2 = -4.23$	$Z_1 = -3.35$, $Z_2 = -2.27$
OurAugDetSTM_{CHATGLM3}-6B_{fewshot}	$Z_1 = -3.78$, $Z_2 = -4.53$		$Z_1 = -5.70$, $Z_2 = -5.23$	$Z_1 = -3.46$, $Z_2 = -3.38$	$Z_1 = -3.51$, $Z_2 = -3.39$	$Z_1 = -3.11$, $Z_2 = -4.32$	$Z_1 = -3.15$, $Z_2 = -4.63$	$Z_1 = -3.40$, $Z_2 = -3.68$	$Z_1 = -3.36$, $Z_2 = -4.35$

Model	DetSTM (0.78, 0.85)	DetSTM_{BERT} (0.79, 0.86)	DetSTM_{LLaMA3} (0.75, 0.81)	DetSTM_{CHATGLM3} (0.77, 0.84)	DetSTM_{BERT} (0.79, 0.86)	DetSTM_{LLaMA3} (0.82, 0.89)	DetSTM_{CHATGLM3} (0.80, 0.87)	DetSTM_{BERT} (0.81, 0.88)
OurAug (0.83, 0.91)	-	-	$3.45\boldsymbol{Z}_2$	$2.50\boldsymbol{Z}_2$	$3.45\boldsymbol{Z}_2$	$3.38\boldsymbol{Z}_2$	$3.11\boldsymbol{Z}_2$	$3.12\boldsymbol{Z}_2$
	$3.45\boldsymbol{Z}_2$	$3.53\boldsymbol{Z}_2$	-	-	-	-	-	-
	-	-	***3.27***	***2.34***	***3.23***	***3.05***	***3.43***	***3.40***
	3.24	***3.36***						
OurAug (0.86, 0.94)	-	-	$4.71\boldsymbol{Z}_2$	$3.50\boldsymbol{Z}_2$	$3.42\boldsymbol{Z}_2$	$3.42\boldsymbol{Z}_2$	$3.40\boldsymbol{Z}_2$	$3.43\boldsymbol{Z}_2$
	$3.43\boldsymbol{Z}_2$	$3.43\boldsymbol{Z}_2$	-	-	-	-	-	-
	-	-	***4.20***	***2.58***	***3.25***	***3.43***	***4.25***	***2.98***
	2.98	***3.43***						

Note: Values in parentheses below model names represent "accuracy, true negative rate" ; Z_1 indicates inference accuracy difference test results, Z_2 indicates true negative rate difference test results; ** $p < 0.001$.*

4 General Discussion

This study explored LLM-based suicidal ideation data augmentation and recognition methods. In Study I, we achieved effective data augmentation of suicidal ideation corpora by employing current mainstream foundation models combined with supervised learning strategies. The results demonstrated that LLM-generated suicidal ideation texts exhibited excellent performance in both quality and diversity, with both self-developed models achieving comprehensive scores significantly superior to baseline models. This result validates the effectiveness of LLMs in suicidal ideation data augmentation tasks. In Study II, we further evaluated the impact of data augmentation on suicidal ideation recognition model performance through comparative experiments. The results showed that all models trained on augmented datasets significantly outperformed their baseline models trained on original datasets in performance. This fully demonstrates that LLM-based data augmentation strategies can effectively alleviate the problem of scarce annotated data and thereby enhance the performance of suicidal ideation recognition models. The findings of this study not only highlight the application potential of LLM technology in the field of suicidal ideation recognition but also provide new ideas for subsequent research and practice in this area. This technological advancement is expected to lay the foundation for future proactive identification and intervention of suicidal ideation based on social media and other large-scale text data, ultimately contributing to suicide prevention efforts.

4.1 Theoretical Significance

The theoretical contributions of this study are primarily reflected in research paradigm aspects. First, this study enriches the existing methodological system for suicidal ideation information acquisition. Current suicidal ideation identification work relies on traditional invasive methods or machine learning prediction models requiring extensive manual annotation. However, these methods are generally limited by individuals' active help-seeking behavior or the scarcity of annotated data. This study innovatively introduces LLMs, leveraging their powerful language understanding and generation capabilities to achieve automatic construction of high-quality suicidal ideation corpora. The data augmentation strategy effectively alleviates the challenge of scarce annotated data, providing an important supplement to existing suicidal ideation identification methods. This expansion of research paradigms offers new ideas for text analysis research in suicidal ideation and even the entire mental health domain.

Second, this study' s results further validate the potential of LLMs in handling complex social science problems, laying the foundation for their application in mental health-related fields. LLM-based data augmentation methods can not only significantly improve the problem of annotated data scarcity but also substantially enhance the performance of downstream recognition tasks, introducing new technical pathways for proactive identification research in mental health. By exploring and empirically validating the effectiveness and efficiency of LLMs in addressing complex social problems, the results are expected to drive paradigm innovation in social science and mental health research.

From a mechanistic perspective, the improvement in suicidal ideation recognition model performance through LLM-based data augmentation can be understood from several aspects. First, through comparative analysis of generated text quality scores and recognition accuracy experimental data, we can infer that the data augmentation effect of LLMs primarily stems from their ability to simulate human language cognitive patterns. Study I' s experimental results show that self-developed models significantly outperformed all baseline models in data augmentation tasks, indicating that LLMs can not only expand data scale but more importantly simulate and reproduce the diversity of language patterns in human suicidal ideation expression. Second, by analyzing model performance on augmented datasets, we find that LLMs demonstrate patterns similar to human language cognitive development through their understanding and recombination of existing expressions. For example, in Study II, models trained on augmented data showed stronger generalization ability when handling different types of suicidal ideation expressions, possibly reflecting that LLMs indeed learned deep semantic features of suicidal ideation expression during the generation process. This finding aligns with discoveries in cognitive psychology regarding the similarity between AI models and human cognitive processes (Sense et al., 2022; Shiffrin & Mitchell, 2023).

Overall, this study expands the application of LLMs in suicidal ideation recogni-

tion at the theoretical level, contributing new research paradigms and methods to the field. Meanwhile, the research results validate the application potential of LLMs in data-scarce environments. These findings not only break through current bottlenecks in mental health text analysis research to some extent but also provide new ideas and references for research in other social science fields.

4.2 Practical Value

Benefiting from technological development and the accessibility of LLMs, the suicidal ideation text data augmentation and recognition methods introduced in this study have broad application prospects. Due to the concealed and complex nature of suicidal ideation expression, traditional suicide prevention measures struggle with proactive discovery and rapid response (Shahnaz et al., 2020). This situation not only poses challenges to timely rescue of high-risk individuals but also tends to produce ripple effects at the societal level. In an era where the internet has become a channel for emotional expression and life sharing, social media also provides a breeding ground for the social transmission of suicidal ideation. How to timely and accurately identify at-risk individuals from vast text data and provide corresponding mental health resources and assistance at the first opportunity is a major challenge facing suicide prevention work (Liu et al., 2019). The technical route proposed in this study is expected to provide key technical support for suicide risk monitoring and rapid response mechanisms on social media platforms. For example, through the deployment and application of recognition systems, platforms can achieve real-time monitoring and early warning of suicidal ideation, thereby providing timely and personalized mental health services.

The non-invasive suicidal ideation identification method advocated in this study is expected to become an auxiliary means and important supplement to traditional methods such as clinical assessment. For instance, after identifying high-risk individuals, rescue can be provided by qualified departments and professionals under the premise of ensuring user informed consent; for individuals only expressing negative psychology such as depression and helplessness, mental health popular science resources can be proactively pushed to encourage them to seek professional help. This hierarchical, classified, and precise intervention pathway is expected to maximize the timeliness of suicide prevention, ultimately contributing to reducing suicide incidence and maintaining public health. Suicide prevention is a complex systematic project requiring multidisciplinary, multi-channel, and multi-directional collaborative efforts. Only by organically combining non-invasive methods with traditional methods to form a linkage can we truly construct a complete and efficient suicide prevention and rescue system. This requires researchers and practitioners from psychology, computer science, social work, and other fields to work together to jointly promote the development of suicide prevention.

Notably, in the practical application of such measures, government departments, technology companies, and research institutions must attach great importance

to the ethical and privacy issues involved. First, any mental health identification service based on social media should clearly inform users in the platform service agreement to ensure users' full informed consent. Second, data usage should strictly comply with relevant laws and regulations such as the Personal Information Protection Law of the People' s Republic of China, and personal data should be minimized and de-identified. Third, information transmission between social media platforms and professional mental health institutions should establish strict confidentiality mechanisms to ensure risk warning information is used only for professional intervention purposes. Finally, research institutions and mental health workers should formulate industry norms based on general ethical guidelines such as the Declaration of Helsinki, combined with new technological trends, to provide appropriate help while respecting individual autonomy. Only on the basis of balancing technical effectiveness and ethical norms can such intelligent early warning systems truly realize their social value.

In summary, this study validates the application potential of LLMs in suicidal ideation recognition through empirical exploration, providing new ideas for technological innovation in this field. The research results demonstrate that LLM-based data augmentation methods can effectively alleviate the problem of scarce suicidal ideation annotated data, offering a feasible technical path for solving the widespread data scarcity issue in the mental health domain. Meanwhile, recognition models trained on augmented datasets show excellent performance, particularly in handling diverse and implicit suicidal ideation expressions, highlighting the unique advantages of LLMs in improving model generalization ability. Additionally, the extensible research framework constructed in this study lays a methodological foundation for future suicidal ideation recognition research in cross-linguistic and cross-cultural contexts, potentially promoting standardization and systematization in this field. These findings not only confirm the research hypotheses but also provide important empirical support for the broader application of AI technology in mental health.

4.3 Limitations and Future Directions

Although this study achieved positive results, several limitations remain that require further exploration and resolution in future work. First, at the data source level, the text data used in the study primarily originated from the Sina Weibo platform. Given the differences in user groups and content styles across different social media platforms, this may affect the generalizability of the research results. Future research should expand data sources to include heterogeneous data from multiple platforms, languages, and cultural backgrounds to enhance model applicability and generalization ability.

Second, at the methodological level, although this study conducted preliminary exploratory experiments, model interpretability still requires in-depth research, which is also a common challenge currently faced by machine learning and AI fields. Existing research points out that understanding computational model behavior mechanisms requires shifting from correlation analysis to causal inference

(Taylor & Taylor, 2021). In light of this, future research could systematically manipulate semantic features of input texts (such as expression directness, emotional intensity, etc.) to analyze patterns in model output changes, thereby inferring the processing mechanisms of LLMs for different types of suicidal ideation expression. Meanwhile, Huang (2023)' s research indicates that understanding model behavior requires building upon larger-scale data and multi-dimensional evaluation metrics. This means subsequent research should expand evaluation dimensions, systematically assessing model stability and generalization ability across different linguistic environments and expression methods while focusing on accuracy.

Third, at the technical level, the combination of zero-shot learning, few-shot learning, and open-ended prompt engineering strategies adopted in this study, while ensuring the randomness and richness of generated texts, still has room for improvement in complex reasoning and strategic generation. Future research could employ prompt engineering with complex reasoning steps such as Chain-of-Thought (CoT) to enhance the model' s ability to handle complex tasks. With the development of vertical domain LLM fine-tuning technologies, techniques such as Self-Supervised Learning (SSL) and Reinforcement Learning from Human Feedback (RLHF) could be attempted to further improve model output quality and recognition accuracy.

Finally, at the application level, future research could explore the construction of interactive recognition models from the perspective of introducing Agent interaction paradigms in suicide prevention. Through continuous dialogue with at-risk populations and real-time risk assessment and management combined with clinical knowledge bases, the sensitivity of early warning and personalized care levels could be improved. Simultaneously, issues such as algorithm transparency and user privacy protection in active intervention strategies require careful consideration to ensure the ethicality and social responsibility of intervention measures. Future work still requires multidisciplinary collaboration to seek balance between technological innovation and humanistic care.

This paper successfully validates the effectiveness of LLM-based suicidal ideation data augmentation and recognition technology, providing an innovative technical path for suicide prevention work in social media environments. By employing models such as ChatGLM3_{6B} and Qwen_{{{7B}}}_{{Chat}}}, this paper not only optimizes the quality of training datasets but also significantly improves the accuracy of suicidal ideation recognition. The research results emphasize the important value of data augmentation methods in solving data scarcity problems and improving recognition accuracy, while demonstrating the broad application potential of LLMs in social science fields, particularly suicide prevention research. This study successfully constructs a non-invasive suicidal ideation recognition framework based on social media data, providing a new solution to the problem of traditional methods' dependence on individuals actively seeking help. Future research should further explore the applicability of LLMs across social media platforms in multi-linguistic and cross-cultural con-

texts, as well as deepen understanding of the complexity of suicidal ideation through interdisciplinary collaboration. Additionally, in-depth research on algorithm ethics and data privacy protection will ensure the ethicality and social responsibility of technology applications.

Appendix 1: Suicide Risk Scoring Criteria

Please review each Weibo comment text individually to determine whether it shows the commenter' s desire to commit suicide, including thoughts, plans, and preparatory behaviors; or whether it shows the commenter has previously attempted suicide and still exhibits any of the following 12 suicide risk warning signs*.

*Note: The 12 suicide risk warning signs come from the National Institute of Mental Health (NIMH; Web: https://www.nimh.nih.gov/). They were translated into Chinese by Tan et al. (2017) and applied in studies such as Liu et al. (2019). The original content can be found at: https://www.nimh.nih.gov/health/publications/warning-signs-of-suicide/.

Appendix References

Tan, Z., Liu, X., Liu, X., Cheng, Q., & Zhu, T. (2017). Designing microblog direct messages to engage social media users with suicide ideation: Interview and survey study on Weibo. *Journal of Medical Internet Research*, 19(12), e8729. https://doi.org/10.2196/jmir.8729

Liu, X., Liu, X., Sun, J., Yu, N. X., Sun, B., Li, Q., & Zhu, T. (2019). Proactive Suicide Prevention Online (PSPO): Machine identification and crisis management for Chinese social media users with suicidal thoughts and behaviors. *Journal of Medical Internet Research*, 21(5), e11705. https://doi.org/10.2196/11705

Note: Figure translations are in progress. See original paper for figures.