

Supervised Learning and Large Language Model Benchmarks on Mental Health Datasets: Cognitive Distortions and Suicidal Risks in Chinese Social Media

Hongzhi Qi^a, Qing Zhao^a, Jianqiang Li^a, Changwei Song^a, Wei Zhai^a, Dan Luo^b, Shuo Liu^b, Yi Jing Yu^b, Fan Wang^b, Huijing Zou^b, Bing Xiang Yang^b, and Guanghui Fu^{*c}

^aFaculty of Information Technology, Beijing University of Technology, Beijing, 100124, China

^bCenter for Wise Information Technology of Mental Health Nursing Research, School of Nursing, Wuhan University, Wuhan, China

^cSorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, Paris, France

ABSTRACT

In the realm of social media, users frequently convey personal sentiments, with some potentially indicating cognitive distortions or suicidal tendencies. Timely recognition of such signs is pivotal for effective interventions. In response, we introduce two novel annotated datasets from Chinese social media, focused on cognitive distortions and suicidal risk classification. We propose a comprehensive benchmark using both supervised learning and large language models, especially from the GPT series, to evaluate performance on these datasets. To assess the capabilities of the large language models, we employed three strategies: zero-shot, few-shot, and fine-tuning. Furthermore, we deeply explored and analyzed the performance of these large language models from a psychological perspective, shedding light on their strengths and limitations in identifying and understanding complex human emotions. Our evaluations underscore a performance difference between the two approaches, with the models often challenged by subtle category distinctions. While GPT-4 consistently delivered strong results, GPT-3.5 showed marked improvement in suicide risk classification after fine-tuning. This research is groundbreaking in its evaluation of large language models for Chinese social media tasks, accentuating the models' potential in psychological contexts. All datasets and code are made available at: <https://github.com/HongzhiQ/SupervisedVsLLM-EfficacyEval>.

Keywords: Large language model, Deep learning, Natural language processing, Mental health, Social media

1. INTRODUCTION

The omnipresent specter of mental illness, particularly depression, continues to impose significant challenges globally.¹ According to the World Health Organization (WHO), an estimated 3.8% of the global population experiences depression.¹ Specifically in China, the prevalence of depression is notably high, with estimates around 6.9%,² underscoring the escalating mental health concerns in the nation. Such severe depression can often precipitate suicidal behaviors.³ As digital avenues for communication flourish, social media platforms like Twitter and Sina Weibo have evolved into reflective mirrors, offering glimpses into the emotional landscapes of countless users.⁴ Within these platforms, a specific subset of topics recurrently surfaces, with users frequently conveying deep-seated negative emotions and, alarmingly, pronounced suicidal inclinations.^{5,6}

Artificial intelligence (AI), especially the branches of deep learning and natural language processing technique, is an avenue that holds promise in addressing this challenge.⁷ Over recent years, AI research has resulted in the formulation of several algorithms tailored for emotion recognition within textual data.⁸ However, these advancements are not without obstacles.⁹ Constructing a potent deep learning model often demands considerable time and financial resources. The intricacies of data labeling, predominantly the need to enlist domain experts and the model's variance in performance when shifted across different application areas, highlight pressing

Further author information: (Send correspondence to Guanghui Fu, guanghui.fu@inria.fr)

challenges.¹⁰ This highlights a compelling need for more agile and adaptable algorithmic solutions especially in medical domain.¹¹ It is in this context that the emergence and proliferation of large language models are particularly noteworthy.

Large language models, characterized by their expansive parameters and the depth of their training datasets, stand as the state-of-the-art in the framework of computational linguistics.¹² Their potential lies in their ability to comprehend and emulate human-like text nuances. Despite their promising potential, several studies have sought to validate their practical implications. For instance, Xu et al.¹³ examined four public datasets related to online social media sentiment detection. However, their study focused solely on English data, and the classification granularity was relatively broad. To date, there is a notable gap in research concerning the Chinese context, particularly in the area of fine-grained emotion recognition, which is often of greater significance. The lack of comprehensive evaluations and practical tests has inadvertently led to a cautious approach, especially in sectors demanding high reliability, like medicine and healthcare.¹⁴

Motivated by the need to better understand mental health sentiments on Chinese social media platforms, our research embarks on a rigorous evaluation of supervised learning and large language models. We offer the following contributions:

- We introduce and publicly release two new expertly-manual annotated social media datasets in the mental health domain, specifically focusing on cognitive distortion and suicide risk classification. These datasets not only serve as valuable resources for the community but also have profound real-world implications, potentially informing strategies for suicide prevention and interventions for cognitive distortions.
- We propose a comprehensive benchmark using both traditional supervised learning and large language models on these datasets. By employing a variety of strategies, including zero-shot, few-shot, and fine-tuning, we seek to determine the most effective methods for leveraging these models in the context of mental health tasks on Chinese social media.
- Lastly, our study pioneers the exploration of fine-tuning capabilities of GPT-3.5, leveraging real-world data. This endeavor seeks to determine the adaptability and specialized performance enhancements possible with the model currently unexplored in the literature.

2. RELATED WORK

The intertwining of artificial intelligence (AI) with different fields has spurred innovations and transformations at an unprecedented scale. An example of this is the fusion of natural language processing (NLP) tools, notably deep learning based model, with domains as critical as the mental health field.¹⁵ Additionally, as digital interactions burgeon, especially on social media, the urgency to understand and analyze human sentiments becomes paramount. In this section, we will introduce deep learning techniques for sentiment analysis utilizing text data (Section 2.1). Subsequently, we will discuss the evolution, potential, and current research on large language models in this domain (Section 2.2).

2.1 Text sentiment analysis

In the swiftly evolving digital era, social networking platforms have emerged as pivotal channels for expressing emotions globally. These platforms generate vast amounts of unstructured data every second. Accurately and promptly discerning the emotions embedded within this data presents a formidable challenge to computational algorithms.⁸ Fu et al.¹⁶ presented a distant supervision method designed to build systems that classify high and low suicide risk levels using Chinese social media data. This approach minimizes the need for human experts of varying expertise levels to perform annotations. By integrating this model with crucial psychological features extracted from user blogs, they attained an F1 score of 77.98%. Singh et al.¹⁷ employed a BERT-based model for sentiment analysis on tweets sourced globally and another dataset specifically from India, both focusing on the topic of COVID-19. They reported achieving an accuracy of 94%. Wan¹⁸ introduced a method for sentiment analysis of comments on Weibo platforms, leveraging deep neural networks. The data undergoes feature extraction through multilevel pooling and convolution layers. Comments are preprocessed and transformed into

text representations using the word2vec algorithm. Subsequently, key features are extracted from the feature matrix using a CNN. For the final classification and sentiment analysis, the softmax logistic regression method is employed. Zhang et al.¹⁹ explored the correlations among emotion labels, social interactions, and temporal patterns within an annotated Twitter dataset. They introduced a factor graph-based emotion recognition model that seamlessly integrates these correlations into a unified framework. This model adeptly identifies multiple emotions by applying a multi-label learning approach to Twitter datasets. Wang et al.²⁰ introduced a topic modeling technique, termed LDA, to examine the primary concerns expressed on Weibo during the COVID-19 pandemic. They assessed the emotional inclinations of these topics, determined their proportional distributions, and conducted user behavior analysis based on metrics such as likes, comments, and retweets. Furthermore, they explored shifts in user concerns and variations in engagement among residents from different regions of mainland China. Such insights guide public sentiment and actions during health emergencies, emphasizing the importance of vigilant social media monitoring.

Although deep learning algorithms typically demonstrate impressive results, they often require a significant volume of labeled data to perform optimally. The distant supervision approach highlighted in Fu et al.’s research¹⁶ aims to reduce the need for labeling, but it still requires the involvement of three different expert groups at various expertise levels to yield desired results. Nonetheless, when applying these models to new datasets or tasks, domain adaptation issues often arise. These trained models can see a decline in their efficacy, making deep learning algorithms both costly and inflexible. Given these hurdles, there’s a growing demand for efficient and user-centric methods to assist individuals in emotion detection on social media platforms. The recent advancements in large language models present a potential solution to this challenge, but their precise impact still warrants examination from multiple perspectives and specialists.

2.2 Large language model and its applications in medical domain

The advent of Large Language Models (LLMs), such as OpenAI’s ChatGPT,¹² has revolutionized the field of natural language processing.²¹ These LLMs demonstrate emergent abilities that significantly outperform those of their smaller, pre-trained models.²² Initially conceived for understanding and generating human-like text, LLMs have found diverse applications ranging from content generation,²³ medical report assistant,²⁴ coding assistance,²⁵ education,²⁶ and answering medical related questions.²⁷ The sheer scale of these models enables them to generate complex, contextually relevant content. LLMs have garnered significant attention in medical domain.¹⁴ For instance, Jiang et al.²⁸ developed a clinical LLM named NYUTron to assist physicians and health-care administrators in making time-sensitive decisions. This model can process on unstructured clinical notes from electronic health record. And it can achieve good performance with AUC score ranging from 78.7–94.9%. The model has been successfully deployed in a prospective trial, indicating its potential for real-world application in providing point-of-care guidance to physicians.

Concurrently, research in psychology-related domains has also been conducted by other researcher.²⁹ Qin et al.³⁰ devised an interpretable and interactive depression detection system employing large language models (LLMs). This innovative approach allows for the detection of mental health indicators through social media activity and encourages users to interact with the system using natural language. While this facilitates a more personalized understanding of an individual’s mental state, it also raises ethical concerns. The absence of human oversight could lead to biased outcomes, thereby posing potential risks to users. Additionally, if this system were to become a foundational diagnostic tool for future psychological counseling, issues related to user privacy could become a point of concern. Chen et al.³¹ developed a tool designed to improve the realism of psychiatrist-patient simulations using ChatGPT-based chatbots. Their approach involved using distinct prompts to enable large language models (LLMs) to emulate the roles of both a depressed patient and a psychiatrist. The study confirmed the feasibility of utilizing ChatGPT-driven chatbots in psychiatric contexts. However, the research also acknowledged limitations: individual patients and counselors have unique communication styles, and some patients may be reluctant to engage in conversation. These nuances present a challenge for achieving truly realistic simulations with ChatGPT. Addressing the simulation of diverse personalities in a meaningful way remains a key area for further investigation. Fu et al.³² developed a counseling support system designed to augment the capabilities of non-professional counselors. The system provides multiple features, including mental health analysis, evaluation of therapist responses, and suggested interventions. This application serves as a valuable use case for language models in the mental health sector. Ten professional psychologists assessed the system on five

critical dimensions, and the findings were favorable, with a 78% expert approval rate indicating that the system can deliver effective treatment strategies. Ayers et al.³³ developed a ChatGPT-based chatbot and compared its responses with those of physicians to patient inquiries on a social media forum. Notably, 78.6% of the evaluators preferred the chatbot’s responses, citing their speed and greater empathetic tone. However, a key limitation of this study lies in its exclusive focus on interactions within online forums. Such settings may not accurately reflect the nuances of real-world patient-physician dialogues, as physicians often tailor their responses based on pre-existing relationships and the context of a clinical setting. In summary, there is active research into the utilization of LLMs in the field of psychology, and these research demonstrate considerable potential. However, delineating the limitations of LLMs remains a crucial issue that warrants further investigation. Additional studies are needed to comprehensively evaluate the capabilities and boundaries of LLMs in psychological applications.

Xu et al.¹³ present a pioneering evaluation of multiple Large Language Models (LLMs) across various mental health prediction tasks using four publicly available online text datasets. Their insights offer guidance to practitioners on optimizing the use of LLMs for specific applications. While their research stands as a monumental verification of LLMs’ potential in the mental health domain, it is noteworthy that their datasets are exclusively in English and do not address multi-label classification tasks. Yang et al.³⁴ assessed ChatGPT’s capabilities in mental health analysis and emotional reasoning by evaluating its performance on 11 datasets across five tasks. The study also investigated the impact of different emotion-based prompting strategies. Experimental results indicate that while ChatGPT surpasses traditional neural network-based approaches, it still lags behind more advanced, task-specific methods. Nevertheless, ChatGPT demonstrates significant potential in the area of explainable mental health analysis. In conclusion, while the integration of LLMs in medicine presents compelling prospects, there’s an imperative to ensure privacy and uphold ethical standards. Responses generated may not always be flawless.³⁵ Particularly in mental health, relying solely on LLM-driven systems for diagnosis or support introduces numerous unpredictable variables. It’s crucial to recognize that LLMs warrant meticulous scrutiny and validation.³⁶ Evaluation should be considered an essential discipline to facilitate the more effective development of large language models (LLMs).³⁷

3. METHODS

We conducted experiments to classify suicide risk and cognitive distortions on Chinese social media data using supervised learning methods and large language models (LLMs). Within the framework of supervised learning, we explored two models BERT³⁸ and LSAN³⁹ as baseline, detailed in Section 3.1. For the large language models, we utilized zero-shot prompt, few-shot prompt, and fine-tuning methods. Subsequent sections provide a comprehensive introduction of these methods.

3.1 Baseline supervised learning model

We experimented with two representative models: LSAN³⁹ and BERT.³⁸ LSAN is adept at uncovering the relationships between labels, making it particularly suitable for our cognitive distortion recognition task. On the other hand, BERT represents a groundbreaking pre-trained model architecture that had achieved state-of-the-art (SOTA) on 11 distinct NLP tasks. We discuss each in detail below:

- **LSAN:** The LSAN model is engineered to utilize label semantics for identifying the relationships between labels and documents, thereby creating a label-specific document representation. The model also employs a self-attention mechanism to focus on this representation, which is derived from the document’s content. An adaptive fusion strategy integrates these components effectively, facilitating the generation of a comprehensive document representation suitable for multi-label text classification. The LSAN model has proven effective, particularly in predicting low-frequency labels.
- **BERT:** Bidirectional Encoder Representations from Transformers (BERT) has been a pivotal development in natural language processing (NLP). Unlike traditional NLP models that process text unidirectionally, BERT uses a bidirectional approach, facilitated by the Transformer architecture, to understand the full context of each word. It is pre-trained using a masked language model objective, where random words are replaced with a ‘[MASK]’ token and the model predicts the original word. This design has enabled BERT to set new performance standards in diverse NLP tasks, such as question-answering and sentiment analysis, especially when fine-tuned on specific task data.

3.2 Large language models

Given that our data is in Chinese, we explored the open-source models ChatGLM2-6B and GLM-130B,⁴⁰ both of which support Chinese language processing. The primary distinction between these two models lies in the number of parameters they possess. GPT-3.5⁴¹ stands as a flagship large-scale language model. We experimented with various prompt word constructions and sought to integrate prior knowledge from the psychological domain, along with the most recent public fine-tuning functionalities. GPT-4,⁴² being the latest iteration, was also included in our assessment. Detailed introduction on these models are provided in the subsequent sections.

- **ChatGLM2-6B:** ChatGLM2-6B is an open-source bilingual language model with 6.2 billion parameters, optimized for Chinese question-answering and dialogue. It employs similar technology to ChatGPT and is trained on roughly 1TB of Chinese and English text data. The model can be fine-tuned through various techniques like supervised learning and human feedback. It also features an efficient tuning method based on P-Tuning v2, requiring at least 7GB of GPU memory for customization. Due to quantization techniques, it can run on consumer-grade graphics cards with only 6GB of memory.
- **GLM-130B:** GLM-130B is a bilingual pre-trained language model optimized for both English and Chinese, boasting a substantial 130 billion parameters. This model aims to provide an open-source alternative of a scale comparable to GPT-3, while shedding light on the complexities of training such large-scale models. Impressively, GLM-130B surpasses GPT-3 175B on multiple English benchmarks and outperforms ERNIE TITAN 3.0 260B,⁴³ the largest existing Chinese language model, on relevant benchmarks. A distinctive feature of GLM-130B is its capability for INT4 quantization without substantial performance degradation, thus facilitating efficient inference on widely available GPUs.
- **GPT-3.5:** GPT-3.5 is a cutting-edge language model developed by OpenAI, designed to offer enhanced conversational capabilities. Building on the foundation of its predecessor, GPT-3, this iteration introduces improvements in both performance and cost-efficiency. OpenAI’s commitment to refining and advancing the capabilities of their models is evident in GPT-3.5, which provides users with a more coherent, context-aware, and responsive conversational experience. As part of OpenAI’s mission to ensure that artificial general intelligence benefits all of humanity, GPT-3.5 is a testament to the organization’s dedication to innovation and excellence in the realm of natural language processing.
- **GPT 4:** GPT-4 is a groundbreaking multimodal model capable of processing both image and text inputs to generate text-based outputs. Marking a significant advancement over its predecessors, GPT-4 exhibits human-level performance across a range of professional and academic benchmarks, including a top 10% score on a simulated bar exam. Built upon the Transformer architecture, the model is initially trained to predict subsequent tokens in a given sequence and later undergoes a post-training alignment process to improve its factuality and behavior. A critical component of the project involved the development of scalable infrastructure and optimization techniques that function consistently across various sizes, allowing the team to extrapolate GPT-4’s performance metrics based on smaller models. Despite its notable capabilities, GPT-4 does inherit certain limitations from earlier versions, such as occasional content “hallucinations” and a constrained context window.

The large language model is widely recognized as being pre-trained on vast amounts of text data. However, the manner in which prompt are inputted is crucial, as it directly influences the LLM’s comprehension and output for a given task. In light of this, we have formulated the following prompts.

LLM Zero-shot Prompting We initiate our exploration with prompt design tailored for tasks within a zero-shot paradigm. This process encompasses various strategies, including direct task requests (acting as the basic), role-definition, scene-definition, and hybrid approaches. For illustrative purposes, the cognitive distortion classification task serves as the focal point. The design is elaborated as follows:

1. **Basic:** A direct task directive devoid of specific contextual emphasis.

- (a) English translation: “Please conduct a multi-classification task to ascertain if it encompasses any of the specified 12 cognitive distortions ([list of cognitive distortions]).”
 - (b) Formulaic representation: $M(T, 12CD)$, where M stands for multi-classification, T symbolizes the task, and $12CD$ represents the 12 cognitive distortions.
2. **Role-definition Prompting:** The prompt delineates the role of the respondent (in this case, a psychologist) and emphasizes reliance on psychological insights.
- (a) English translation: “Assuming the role of a psychologist and leveraging psychological insights, please conduct a multi-classification task to discern if it integrates any of the 12 cognitive distortions ([list of cognitive distortions]).”
 - (b) Formulaic representation: $R(M(T, 12CD))$, where R embodies the role-definition of being a psychologist.
3. **Scene-definition Prompting:** The context of a social media setting is introduced, highlighting user identifiers to preclude ambiguity.
- (a) English translation: “Considering the provided user ID and the associated posts on social media, please based on the post content, engage in a multi-classification task to determine the presence of any of the 12 cognitive distortions ([list of cognitive distortions]).”
 - (b) Formulaic representation: $S(M(T, 12CD))$, with S denoting the scene, which in this scenario, pertains to the user’s ID and corresponding social media posts.
4. **Hybrid Prompting:** A synthesis of both role and scene definitions, offering an integrative instruction.
- (a) English translation: “With the given user ID and their respective social media posts, and adopting the role of a psychologist fortified with psychological expertise, please execute a multi-classification task to verify the inclusion of any of the 12 cognitive distortions ([list of cognitive distortions]).”
 - (b) Formulaic representation: $S + R(M(T, 12CD))$, intertwining the scene context (S) with the role-definition (R).

LLM Few-shot Prompting In this segment, few-shot prompting is construed as the provision of prior knowledge or a batch of n training instances to LLMs, thereby enabling them to internalize this information and adeptly execute the stipulated task. This methodology unfolds as:

- 1. **Background Knowledge:** The model is furnished with psychological definitions supplemented by emblematic cases, followed by one of the four prompting strategies devised from zero-shot prompting. Prompts that integrate background knowledge and employ the hybrid strategy from zero-shot prompting are detailed as follows:
 - (a) English translation: “Given the definitions of cognitive distortions denoted by D and the prototypical cases represented by C , and in light of the supplied user ID and associated social media posts, you are assumed to be a psychological expert well-versed in the aforementioned definitions and cases. Drawing from this backdrop, please conduct a multi-classification task to evaluate the correlation with any of the 12 cognitive distortions ([list of cognitive distortions]).”
 - (b) Formulaic representation: $D + C + S + R(M(T, 12CD))$, where D encapsulates the background definition, and C signifies the prototypical instances from academic literature, S represents scene-definition and R stands for role-definition.
- 2. **Training with n Samples per Category:** In this approach, n training instances are randomly selected for each category to train the LLM, followed by one of the four prompting strategies designed from zero-shot prompting. These instances are represented as $train_n$ in the following tables. Prompts that incorporate the training instances employ the hybrid strategy from zero-shot prompting are detailed as follows:

- (a) English translation: “You are provided with learning samples denoted by T . In light of the supplied user ID and associated social media posts, and assuming your role as a psychologist with the relevant expertise. Drawing from this backdrop, please conduct a multi-classification task to evaluate the correlation with any of the 12 cognitive distortions ([list of cognitive distortions]).”
- (b) Formulaic representation: $T + S + R(M(T, 12CD))$, integrating T as the training set with the scene-definition (S) and role-definition (R).

3. **Background knowledge and training with n samples per category:** This approach investigates whether enhancing sample diversity in few-shot prompting augments the LLM’s comprehension of psychological health tasks. It incorporates psychological definitions, symbolic examples, and provides n training instances per category for LLM training. A command is subsequently issued using a previously described few-shot prompting strategy. The following example integrates background knowledge and training instances, and poses a query using the hybrid strategy from zero-shot prompting:

- (a) English translation: “Given the definitions of cognitive distortions represented by D and the prototypical cases denoted by C , you are also provided with learning samples represented by T . Assuming your expertise as a psychological expert familiar with the aforementioned definitions and cases, and in consideration of the supplied user ID and associated social media posts, please conduct a multi-classification task to evaluate the correlation with any of the 12 cognitive distortions ([list of cognitive distortions]).”
- (b) Formulaic representation: $D + C + T + S + R(M(T, 12CD))$, where D encapsulates the background definition, C signifies the prototypical instances from academic literature, T is integrated as the training set, S denotes the scene-definition, and R represents the role-definition.

LLM Fine-tuning Fine-tuning represents a potent paradigm provided by OpenAI, enabling users to optimize the performance of pre-trained models, such as GPT-3.5. While GPT-3.5 is inherently trained on an expansive text corpus, the fine-tuning process sharpens its proficiency for specialized tasks by exposing it to additional task-specific instances. Following the fine-tuning, our evaluation paradigm retained the role, scene and hybrid definitions from the zero-shot prompting for consistency and comparative assessment:

1. **Role-definition Prompting:** Post fine-tuning with relevant training samples, we employed the prompt delineated in the role-definition section (refer to Section 3.2).
2. **Scene-definition Prompting:** Analogously, after the fine-tuning process, we reverted to the prompt illustrated in the scene-definition segment of the zero-shot prompting.
3. **Hybrid Prompting:** Similarly, after the fine-tuning process, we adopted the prompt presented in the hybrid strategy segment of the zero-shot prompting.

4. EXPERIMENTS AND RESULTS

4.1 Datasets and Evaluation Metrics

We undertook two psychology-related classification tasks: suicide risk and cognitive distortion. The suicide risk task primarily differentiates between high and low suicide risks, while the cognitive distortion task focuses on classifications defined by Burns.⁴⁴ We sourced our data by crawling comments from the “Zoufan” blog within the Weibo social platform. Subsequently, a team of qualified psychologists were enlisted to annotate the data. Given that this data is publicly accessible, there are no concerns related to privacy breaches.

For the suicide detection data, there were 648 records with low suicide risk and 601 records with high suicide risk. The dataset for cognitive distortion consists of a total of 910 entries. The classification labels employed for this data are as follows: all-or-nothing thinking, over-generalization, mental filter, disqualifying the positive, mind reading, the fortune teller error, magnification, emotional reasoning, should statements, labeling and mislabeling, blaming oneself and blaming others. For both sets of data, the training set and test set are divided according

Table 1. Summary of Experimental Datasets

Task	N_{train}	N_{test}	L	\bar{L}	\bar{W}
cognitive distortion	728	182	12	1.27	53
suicide detection	999	250	1	1	47.79

to the ratio of 4:1. The statistics of these two datasets are listed in Table 1, where N_{train} and N_{test} denote the number of training and test samples, respectively. L is the total number of classes, \bar{L} is the average number of labels per sample, and \bar{W} is the average number of words per sample. We utilize three evaluation metrics to measure the performance of different algorithms for our two tasks: precision, recall, and F_1 score. Precision is the ratio of correctly predicted positive observations to the total predicted positives and recall (or sensitivity) represents the ratio of correctly predicted positive observations to all the actual positives. These two metrics provide a comprehensive view of the algorithm’s performance in terms of its positive predictions. The F_1 score offers a more holistic view of the model’s performance, especially when the distribution of the true positive and true negative rates is uneven.

4.2 Experiment design

Our experimental methodology is both hierarchical and greedy. Using cognitive distortions as an example to show our points, our evaluations spanned several dimensions:

- **Prompt Design Perspective:** Initially, we assessed four prompting strategies within the zero-shot learning framework. Subsequently, based on their performance metrics, the top two strategies were selected for further evaluation in the few-shot learning setting across various LLMs.
- **LLM Performance Perspective:** Across all zero-shot prompts, ChatGLM2-6B’s performance was found to be lacking, resulting in our decision to omit it from subsequent few-shot prompting experiments. For GPT-3.5, its token limitation prevented us from entering five samples for each category during few-shot prompting. Consequently, we reserved the *train₅* approach exclusively for GPT-4.
- **Fine-tuning Perspective:** A discernible performance gap exists between GPT-3.5 and GPT-4. However, OpenAI’s recent introduction of fine-tuning capabilities for GPT-3.5 and reports from official channels suggest that, under specific conditions, GPT-3.5 might outperform GPT-4 post fine-tuning. Consequently, our attention was centered on the fine-tuning of GPT-3.5. Regrettably, the current iteration of GPT-4 lacks fine-tuning functionalities, curtailing our capacity to assess its potential in this dimension.

The detailed experimental setup is as follows:

- **LSAN:** We used word2vec to train 300-dimensional embeddings for both document and randomly-initialized label texts. The attention mechanism helped us compute word contributions to labels and create label-specific document representations. Dot products between these document and label vectors refined these relationships further. These two types of document representations were then fused using weighted combinations. For predictions, we employed a fully connected layer, followed by RELU and a sigmoid function. Losses were calculated using a cross-entropy function during training.
- **BERT:** We employ BERT to extract 768-dimensional vectors from Chinese sentences. To mitigate overfitting, a dropout function is applied to these sentence vectors. Subsequently, a fully connected layer is introduced to independently classify suicide risk and cognitive distortions. The sigmoid function serves as the activation function for the output layer. Both the BERT layer and the fully connected layer are trained simultaneously.
- **LLM-zero shot:** Both GPT-3.5 and GPT-4 are closed-source and available through API provided by OpenAI. We picked the gpt-3.5-turbo, one of the most capable and cost-effective models in the GPT-3.5 family, and the gpt-4, more capable than any GPT-3.5 model, able to do more complex tasks, and optimized for chat. As for the GLM models, we employed the smaller, open-source variant, ChatGLM2-6B, suitable

for deployment on consumer-grade hardware. Given the extensive parameter count of GLM-130B, it posed deployment challenges due to its elevated operational costs. Furthermore, its API lacked the capability to handle cognitive distortion multi-label classification task, leading us to conduct tests via its official website. Acknowledging the inherent variability in LLM’s outputs, our experimental design involved averaging the outcomes over five runs. For GPT-3.5, GPT-4, and ChatGLM2-6B, we adjusted the temperature to values of 0.1, 0.3, 0.5, 0.7, and 0.9, conducting experiments at each setting. Given the absence of a temperature setting for GLM-130B on its platform, we simply executed five repeated runs and computed the mean performance. For zero-shot evaluations, we initiated performance validation on the basic strategy across the LLMs, subsequently examining the efficacy of role-definition, scene-definition, and hybrid strategies, aiming to discern the influence of domain-specific information on LLM’s performance.

- **LLM-few shot:** We conducted an assessment using the top two performing prompt strategies from the zero-shot tests, determined by their F1-scores. The impact on performance was assessed when augmenting these strategies with background, $train_n$, and their combination (background + $train_n$). Specifically, background strategy denotes the incorporation of prior knowledge, $train_n$ represents the addition of training samples, where n is the number of positive samples chosen for each category. background + $train_n$ suggests simultaneous enrichment with prior knowledge and training samples. Given the varying token input constraints among different models, the sample size selected for each model differed. In addition, we also experimented with the integration of basic, role, scene, and hybrid strategies in the zero-shot prompting scenario.
- **LLM-fine-tuning:** We fine-tuned the GPT-3.5 Turbo model for predicting suicide risk and cognitive distortions using the API interface provided by OpenAI. We utilized three types of prompts: role-based, scene-based, and hybrid strategies.

5. RESULTS

In our study, we focused on two specific tasks: suicide classification and multi-label classification of cognitive distortions. And the results can be seen in Table 2 and Table 3 respectively. Our analysis examined these two tasks in Section 5.1 and Section 5.2 respectively from three distinct aspects: training strategy, the construction of prompt, and a comparative evaluation across various LLMs. Ultimately, we assessed and compared the model’s performance on these two psychological tasks to draw conclusions in Section 5.3. Considering the intricate nature and distinctiveness of the cognitive distortion task, LLMs demonstrate suboptimal performance. We have included a human evaluation stage conducted by psychology experts regarding the predictions of the large models in Section 5.4.

5.1 Suicide Risk

Training strategies In our training strategy comparison, we observed varying degrees of effectiveness across different models. The pre-trained BERT model exhibited a performance enhancement over the LSAN model, registering a 2.23% increase in F1-score. In contrast, fine-tuning GPT-3.5 led to a substantial performance gain, achieving an F1-score of 78.45%. This represented a notable 11.5% improvement in F1-score when compared to its base model (fine-tuning hybrid vs. zero-shot hybrid), bringing its performance closer to that of supervised learning models.

Design of prompts Our investigation into prompt design for large language models revealed nuanced outcomes across different strategies and models. In the context of zero-shot prompts, we found that while the hybrid strategy yielded satisfactory results, the performance differences among various types of prompts were not statistically significant. Upon enhancing the basic strategy with three additional strategies (role-define, scene-define, and hybrid), the performance differences in comparison to the basic strategy are illustrated in Table 4. For few-shot prompts, adding more data did not consistently improve performance; this was evident in the ChatGLM2-6B model where additional data sometimes reduced effectiveness. Conversely, GPT-4’s performance remained stable irrespective of the data size. Notably, the background+ $train_n$ +hybrid strategy emerged as the most effective across multiple models.

Table 2. Result for suicide binary classification task

Model category	Model name	Type	Sub-type	Train data	Test data	Precision	Recall	F1-score	
Supervised learning	LSAN BERT	train from scratch fine-tuning	-	999		74.59%	87.50%	80.53%	
			-	999		88.42%	77.78%	82.76%	
LLM	ChatGLM2-6B	zero-shot	basic	0	250	69.07%	37.10%	48.07%	
			role-define	0		65.77%	35.81%	46.15%	
			scene-define	0		64.52%	45.16%	53.01%	
			hybrid	0		65.68%	46.13%	53.74%	
		few-shot	background+scene-define	0		58.56%	47.26%	51.45%	
			background+hybrid	0		60.37%	70.64%	64.41%	
			train ₁₂ +scene-define	24		67.19%	59.52%	63.04%	
			train ₁₂ +hybrid	24		64.29%	49.20%	55.56%	
			background+train ₁₂ +scene-define	24		49.74%	56.61%	52.70%	
			background+train ₁₂ +hybrid	24		58.91%	73.23%	64.78%	
			train ₃₀ +scene-define	60		57.71%	26.77%	36.14%	
			train ₃₀ +hybrid	60		50.60%	24.84%	32.60%	
			background+train ₃₀ +scene-define	60		62.97%	52.90%	57.02%	
			background+train ₃₀ +hybrid	60		60.14%	47.10%	51.88%	
		GLM-130B	zero-shot	basic		0	54.58%	95.81%	69.52%
				role-define		0	55.51%	94.84%	70.02%
				scene-define		0	55.05%	93.87%	69.40%
				hybrid		0	57.37%	97.42%	72.20%
			few-shot	background+role-define		0	56.55%	90.32%	69.55%
				background+hybrid		0	56.91%	92.42%	70.43%
				train ₁₂ +role-define		24	53.18%	83.23%	64.89%
				train ₁₂ +hybrid		24	55.30%	88.39%	68.02%
		GPT-3.5	zero-shot	background+train ₁₂ +role-define		24	57.84%	83.38%	68.30%
				background+train ₁₂ +hybrid		24	60.88%	90.00%	72.61%
	basic			0		52.00%	88.23%	65.42%	
	role-define			0		53.31%	96.13%	68.59%	
	scene-define			0		52.16%	89.03%	65.76%	
	hybrid			0		52.55%	92.26%	66.95%	
	few-shot			background+role-define		0	54.90%	88.55%	67.76%
				background+hybrid		0	55.27%	89.03%	68.19%
				train ₁₂ +role-define		24	56.34%	83.39%	67.22%
				train ₁₂ +hybrid		24	57.19%	84.68%	68.27%
			background+train ₁₂ +role-define	24		59.37%	81.61%	68.71%	
			background+train ₁₂ +hybrid	24		58.26%	82.90%	68.41%	
	fine-tuning		role-define	999		84.76%	71.77%	77.73%	
			scene-define	999		84.11%	72.58%	77.92%	
			hybrid	999		84.26%	73.39%	78.45%	
	GPT-4		zero-shot	basic		0	57.43%	95.48%	71.72%
				role-define		0	57.29%	97.26%	72.10%
				scene-define		0	58.81%	97.58%	73.39%
				hybrid		0	57.47%	97.42%	72.30%
			few-shot	background+scene-define		0	64.91%	73.55%	68.86%
				background+hybrid		0	63.24%	84.03%	72.05%
				train ₁₂ +scene-define		24	60.70%	94.35%	73.87%
				train ₁₂ +hybrid		24	59.77%	84.19%	69.87%
				background+train ₁₂ +scene-define		24	65.44%	81.77%	72.63%
				background+train ₁₂ +hybrid		24	65.65%	78.87%	71.60%
train ₃₀ +scene-define				60	61.11%	92.42%	73.56%		
train ₃₀ +hybrid				60	60.79%	89.03%	72.22%		
background+train ₃₀ +scene-define				60	63.86%	83.06%	72.16%		
background+train ₃₀ +hybrid				60	70.16%	82.58%	75.81%		

We also studied the impact of extra training data in few-shot scenarios and observed that using role-define and train_n+role-define prompts often led to diminished performance. The role of background knowledge was model-dependent; in smaller models like ChatGLM2-6B, incorporating background knowledge led to a performance increase from 53.74% to 64.41%. However, this could not be universally verified due to token limitations. Finally, our comparison between few-shot and zero-shot prompts showed that few-shot prompts did not significantly outperform their zero-shot counterparts.

Comparison of LLMs In our comparative analysis of large language models, we observed several trends that highlight the complexities of model performance. Generally, GPT-4 outperformed GPT-3.5, and GLM-130B excelled over ChatGLM2-6B, suggesting the benefits of larger model architectures and more extensive training data. Yet, this trend was interrupted when GPT-3.5 underwent fine-tuning, outperforming GPT-4 by a differential of 2.64%. Additionally, GLM-130B demonstrated a performance comparable to GPT-4 and superior to GPT-3.5 for the specific task under study. These findings indicate that while larger models typically offer advantages, fine-tuning and task-specific capabilities can alter the performance landscape significantly.

5.2 Cognitive Distortion

Training strategies Our investigation into training strategies for large language models revealed nuanced performance outcomes. Initially, the pre-trained BERT model demonstrated a 2.83% performance advantage over LSAN trained from scratch. However, this difference was not statistically significant, implying that the

Table 3. Result for cognitive distortion multi-label classification task.

Model category	Model name	Type	Sub-type	Train data	Test data	Precision	Recall	F1-score
LLM	Supervised learning	LSAN BERT	train from scratch	728	182	76.79%	77.95%	76.08%
			fine-tuning	728		79.85%	80.49%	78.91%
	GLM-130B	zero-shot	-	0		9.56%	57.39%	16.39%
			basic	0		9.00%	65.31%	15.78%
			role-define	0		8.75%	60.43%	15.19%
			scene-define	0		9.93%	67.83%	17.31%
		few-shot	train ₁ +basic	12		7.59%	39.31%	12.69%
			train ₁ +hybrid	12		8.18%	44.09%	13.73%
		zero-shot	basic	0		10.25%	4.87%	6.49%
			role-define	0		12.58%	4.35%	6.18%
			scene-define	0		9.2%	3.91%	5.34%
			hybrid	0		8.61%	5.39%	6.38%
		few-shot	background+hybrid	0		2.39%	7.91%	11.63%
			background+basic	0		24.21%	10.09%	14.06%
			train ₂ +hybrid	24		13.32%	10.09%	11.46%
			train ₂ +basic	24		12.51%	10.00%	11.10%
		fine-tuning	role-define	728		10.80%	8.26%	9.36%
			scene-define	728		13.71%	10.43%	11.85%
			hybird	728		11.73%	10.00%	10.80%
	GPT-4	zero-shot	basic	0		16.46%	46.09%	24.18%
			role-define	0		16.69%	42.09%	23.86%
			scene-define	0		18.12%	43.22%	25.43%
			hybrid	0		16.26%	38.87%	22.84%
		few-shot	background+basic	0		21.96%	31.04%	25.54%
			background+scene-define	0		22.89%	34.18%	26.59%
			train ₂ +basic	24		31.25%	34.17%	32.47%
			train ₂ +scene-define	24		27.00%	27.74%	27.06%
			background+train ₂ +basic	24		24.35%	32.00%	27.63%
			background+train ₂ +scene-define	24		24.39%	28.09%	25.94%
			train ₅ +basic	60		25.62%	35.65%	29.57%
			train ₅ +scene-define	60		29.46%	34.61%	31.57%

observed discrepancy may not be meaningful. On the other hand, fine-tuning GPT-3.5 surprisingly led to a decrease in performance rather than the anticipated improvement. This underscores the complexity of model training and the need for careful consideration when implementing fine-tuning strategies.

Table 4. Performance Differences in Zero-Shot Enhancement Strategies Compared to Basic Strategy

Model	Suicide	Cognitive Distortion
Δ _ChatGLM2-6B_role-define	↓ -1.92%	—
Δ _ChatGLM2-6B_scene-define	↑ +4.94%	—
Δ _ChatGLM2-6B_hybrid	↑ +5.67%	—
Δ _GLM-130B_role-define	↑ +0.5%	↓ -0.61%
Δ _GLM-130B_scene-define	↓ -0.12%	↓ -1.2%
Δ _GLM-130B_hybrid	↑ +2.68%	↑ +0.92%
Δ _GPT-3.5_role-define	↑ +3.17%	↓ -0.31%
Δ _GPT-3.5_scene-define	↑ +0.34%	↓ -1.15%
Δ _GPT-3.5_hybrid	↑ +1.53%	↓ -0.11%
Δ _GPT-4_role-define	↑ +0.38%	↓ -0.32%
Δ _GPT-4_scene-define	↑ +1.67%	↑ +1.25%
Δ _GPT-4_hybrid	↑ +0.58%	↓ -1.34%

Design of prompts In the design of prompts for large language models, our study examined the performance of both zero-shot and few-shot prompts. For zero-shot prompts, we found that a meticulous design focusing on scene and role settings is crucial; otherwise, a basic task-oriented prompt is generally more effective. The changes in performance metrics for various strategies are shown in Table 4. In the realm of few-shot prompts, we observed that prompts providing specific data points outperformed those that simply offered background knowledge. Interestingly, increasing the training data in these prompts did not lead to better performance. A comparative analysis revealed that although few-shot prompts outperformed zero-shot prompts, they still fell

short of fully meeting the task requirements, as evidenced by GPT-4’s F1-score of approximately 30%.

Comparison of LLMs Consistently, larger models like GPT-4 outperformed their smaller counterparts such as GPT-3.5. When it came to the complex tasks in our study, The performance of ChatGLM2-6B was insufficient for handling complex tasks, while GLM-130B fared better but was still outdone by GPT-4. Given that our dataset consists of comments from social networks, the text is generally concise. As a result, token length did not substantially affect the performance of the models in our tasks. Rather, the selection of representative data for prompt construction emerged as a more crucial factor than merely increasing the number of tokens.

5.3 Cross-Task Comparison

As task complexity increased from binary to multi-label classification, large language models did not sustain their performance. In contrast, supervised learning models maintained a relatively stable F1-score close to 80% across both types of tasks. This highlights the limitations of large language models in replacing supervised learning for specialized tasks. While fine-tuning may benefit simpler tasks, it does not adequately address the challenges posed by complex tasks, calling for further investigation into fine-tuning mechanisms for large language models.

5.4 Expert evaluation and feedback

Owing to the subpar classification results of cognitive distortions by LLMs, we engaged in a manual analysis of these classification outcomes with the expertise of psychological scholars, focusing primarily on the most efficacious strategy in GPT-4, the train_2 +basic strategy. Based on the analysis, it was observed that, given sufficient textual information, GPT-4 can aptly identify cognitive distortion categories. The brevity and directness inherent in social media texts often deprive them of ample contextual information. However, GPT-4 can introduce relevant conjunctions and modality markers to infer the context (as delineated in Example 1 of Figure 1). Yet, in certain specific scenarios, the model does demonstrate errors:

- Most prominently regarding the categorization of "the fortune teller error" instances arise where patients articulate negative anticipations and feelings of desolation about their future, provide retrospectives of their past experiences, or convey apprehensions about potential challenges in forthcoming life events. Such articulations primarily embody the patients’ reflections and should not be deemed conclusive. Yet, GPT-4 has mistakenly classified these under the "the fortune teller error" category (refer to Example 2 of Figure 1).
- Additionally, challenges arose in the categorization of "should statements". Such statements predominantly manifest in patients’ regrets regarding past events. However, GPT-4 erroneously categorized patients’ expectations about the future as "should statements" as well (see Example 3 of Figure 1).
- In specific contexts, GPT-4 mistakenly classified patients’ negative self-assessments as "blaming oneself". However, such classifications lacked the reasoning that ascribes the responsibility for external events to oneself, leading to misjudgments. The appropriate labels for these instances might be "disqualifying the positive" or "mental filter" (refer to Example 4 of Figure 1).
- The model occasionally exhibits ambiguity among the categories of "over-generalization", "all-or-nothing thinking", and "magnification". Instances inherently aligned with Category A are often misclassified into Category B or Category C (refer to Example 5 of Figure 1).

Overall, due to the distinct characteristics of social media data, the task of discerning cognitive distortions within such data is inherently challenging. Even specialists within the domain of psychology inevitably introduce certain subjectivity when categorizing and discerning cognitive distortions in social media texts. In certain contexts, the LLMs can be more detailed, occasionally eliminating biases that may arise during human annotation (as illustrated in Example 6 of Figure 1).

Example 1	<p>Original text: 走饭，我好累，好难过。不管怎么努力争取，都无法得到</p> <p>Translation: Zoufan, I'm so tired and sad. No matter how hard I try, I just can't get it.</p> <p>Ground truth: the fortune teller error</p> <p>GPT-4: the fortune teller error</p>
Example 2	<p>Original text: 饭，今天是5月了，我好像很烦、很迷茫。我现在又觉得世界不过如此我不想走这一遭了。我马上高考了，想到父母心真的真的会很痛。拜托、拜托，真的太难坚持了，我想睡了。</p> <p>Translation: Fan, it's already May. Feeling so overwhelmed and lost. The world seems so meh right now. I can't bear the thought of going through this, especially with college entrance exams coming up. Thinking of my parents just hurts so much. Please, it's really hard to keep going. I just wanna sleep.</p> <p>Ground truth: mental filter</p> <p>GPT-4: mental filter, disqualifying the positive, the fortune teller error, blaming oneself</p>
Example 3	<p>Original text: 我好累，真的好累。考试永远过不了，我不能拿我前途去赌，我好害怕。</p> <p>Translation: Feeling so drained. Exams never seem to end and I can't gamble with my future. Seriously scared.</p> <p>Ground truth: over-generalization</p> <p>GPT-4: the fortune teller error, should statements</p>
Example 4	<p>Original text: 我总是在逃避，我好没用。</p> <p>Translation: Always running away from things. I feel so useless.</p> <p>Ground truth: labeling and mislabeling</p> <p>GPT-4: all-or-nothing thinking, blaming oneself</p>
Example 5	<p>Original text: 有一天我在学校里不小心把头给洗了，本来隔天洗一次的，当时很晚了头没吹干就断电了。加上最近的压力，我崩溃了。我实在太蠢了，而我最亲的朋友居然还在我发疯的时候说：不就是头没吹干吗？干嘛这样？我终于知道人的真面目是什么了，对他人的痛苦不假思索地冷嘲热讽，连我都是这样。</p> <p>Translation: One day at school, I accidentally washed my hair when I usually wash it every other day. It was late, and before my hair could dry, the power went out. With all the recent stress, I broke down thinking how silly I was. And my closest friend? They actually said, "It's just wet hair, why overreact?" That's when I saw their true colors. Making light of someone's pain without a second thought. Even for me.</p> <p>Ground truth: magnification, labeling and mislabeling</p> <p>GPT-4: all-or-nothing thinking, mind reading, magnification, emotional reasoning, labeling and mislabeling, blaming others</p>
Example 6	<p>Original text: 感觉自己要被撕裂了，让我难受的东西像一汪水填在心里，没有一个出口。我不想活着了，但是我知道我不可能死，因为我死了爱我的人会痛苦。我恨他们的不理解，但我舍不得伤害他们，我希望自己得绝症，这样就可以给他们接受我要死去的时间。</p> <p>Translation: I felt like I was going to be torn apart, and the uncomfortable things filled my heart like a pool of water, without an outlet. I don't want to live anymore, but I know I can't die because the people who love me will suffer when I die. I hate them for not understanding, but I am reluctant to hurt them. I hope that I will be terminally ill so that I can give them time to accept that I am going to die.</p> <p>Ground truth: mental filter, magnification</p> <p>GPT-4: mental filter, magnification, emotional reasoning, blaming oneself, blaming others</p>

Figure 1. Typical examples of true labels versus GPT-4 predicted labels in cognitive distortion.

6. DISCUSSION

Our study systematically evaluated the effectiveness of large language models (LLMs) across two mental health related tasks on Chinese social media: suicide risk classification and cognitive distortion multi-label classification. Our results also reveal the nuanced role of prompt design. While the 'hybrid' prompt performed well in zero-shot settings, the benefits of increasing data in few-shot prompts were not universally beneficial. For more straightforward tasks, adding background knowledge appeared to help smaller models (ChatGLM2-6B), but its utility diminished in more complex models or tasks. This calls for a more customized approach to prompt engineering tailored to the specific task and the size of the model being used. If high-quality data is unavailable or prompt design proves challenging, allowing a LLM to directly handle the task may still yield acceptable performance. Larger language models like GPT-4 and GLM-130B generally outperform smaller variants such as GPT-3.5 and ChatGLM2-6B. However, it's important to note that these large models are not always competent at handling complex tasks and should not be seen as replacements for supervised learning algorithms. For simpler tasks, such as the suicide risk classification task examined in our study, the performance of LLMs is satisfactory. Interestingly, after fine-tuning, GPT-3.5 even outperforms GPT-4, achieving results that are nearly on par with those obtained through supervised learning methods. While there is often a preference for large input limits in large language models (LLMs), it's crucial to tailor these settings to the specific task at hand. For tasks involving shorter text, such as our study on sentiment analysis of social network data, the long input capability of an LLM may not be a primary concern. Our experiments indicate that extending the input data to construct few-shot prompts does not necessarily lead to improved performance. Therefore, it is important to carefully consider the nature of the task when configuring the input parameters of an LLM.

Our study does have some limitations. For instance, due to token constraints, we were unable to conduct certain tests—particularly those involving smaller models supplemented with background knowledge—across all tasks. Looking ahead, we plan to conduct more comprehensive studies that encompass a wider variety of tasks and models. This will allow us to draw more definitive conclusions regarding the comparative effectiveness of large language models and supervised learning algorithms. Additionally, the fine-tuning mechanisms of LLMs warrant further exploration, particularly for more efficient handling of complex tasks. The development of advanced prompt engineering techniques could also help optimize the performance of LLMs across various tasks.

7. CONCLUSION

In this study, we evaluated the performance of multiple large language models (LLMs) in two psychology-related tasks and compared their efficacy with that of supervised learning algorithms. Although LLMs show promise in various natural language processing applications, they are not yet a comprehensive substitute for supervised learning, particularly in domain-specific tasks. Fine-tuning LLMs can enhance performance on simpler tasks but is less effective for more complex challenges. The success of different training strategies and prompt designs is highly contingent on both the task and the size of the model, underscoring the necessity for task-specific customization. In summary, our research suggests that while LLMs offer considerable potential, significant work remains to make them universally effective across a broad array of complex tasks.

8. DATASET AND CODE AVAILABILITY

The experimental texts for our study are sourced from comments on a Sina Weibo post by the user "Zoufan," which can be viewed at: https://www.weibo.com/xiaofan116?is_all=1. An expert-annotated dataset for the study of cognitive distortions and suicide risk, along with the prompt of the large language model and the supervised learning model code, are now available at: <https://github.com/HongzhiQ/SupervisedVsLLM-EfficacyEval>.

Here are the models mentioned earlier along with their corresponding source code and online demo links:

- **ChatGLM2-6B:**

- Source code: <https://github.com/thudm/chatglm2-6b>
- Unofficial demo: <https://huggingface.co/spaces/mikeee/chatglm2-6b-4bit>

- **GLM-130B:**

- Source code: <https://github.com/THUDM/GLM-130B>
- Official online demo: <https://chatglm.cn/detail>

- **GPT series:**

- Web application: <https://chat.openai.com/>
- GPT-3.5 Fine-tuning details: <https://platform.openai.com/docs/guides/fine-tuning>

9. ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (grant numbers: 72174152, 72304212 and 82071546), Fundamental Research Funds for the Central Universities (grant numbers: 2042022kf1218; 2042022kf1037), the Young Top-notch Talent Cultivation Program of Hubei Province. Guanghui Fu is supported by a Chinese Government Scholarship provided by the China Scholarship Council (CSC).

10. AUTHOS CONTRIBUTIONS

Hongzhi Qi were responsible for the experiment design and programming. Qing Zhao and Jianqiang Li collaborated in the proposal of the AI-related aspects of the project, with Zhao focusing on data analysis and interpretation and Li serving as the leader of the computer science aspect of the project. Both also reviewed the manuscript. Dan Luo and Huijing Zou contributed to the manuscript writing, carried out experimental verification, and collected data. Changwei Song and Wei Zhai were responsible for code development and served as auxiliary programmers. Guanghui Fu proposed the central idea of the study and was a major contributor in writing the manuscript. Shuo Liu, Yi Jing Yu and Fan Wang took the lead in result evaluation and contributed psychological perspectives to the idea proposal. Bing Xiang Yang proposed the psychological aspects of the idea, performed experimental verification, and led the project from the psychology angle. All authors read and approved the final manuscript.

11. COMPETING INTERESTS

All authors declare no financial or non-financial competing interests.

REFERENCES

- [1] World health organization, “Depressive disorder (depression),” (2023).
- [2] Huang, Y., Wang, Y., Wang, H., Liu, Z., Yu, X., Yan, J., Yu, Y., Kou, C., Xu, X., Lu, J., et al., “Prevalence of mental disorders in China: a cross-sectional epidemiological study,” *The Lancet Psychiatry* **6**(3), 211–224 (2019).
- [3] World health organization, “Suicide,” (2023).
- [4] Keles, B., McCrae, N., and Grealish, A., “A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents,” *International journal of adolescence and youth* **25**(1), 79–93 (2020).
- [5] Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., and Herrman, H., “Social media and suicide prevention: a systematic review,” *Early intervention in psychiatry* **10**(2), 103–121 (2016).
- [6] Luxton, D. D., June, J. D., and Fairall, J. M., “Social media and suicide: a public health perspective,” *American journal of public health* **102**(S2), S195–S200 (2012).
- [7] Coppersmith, G., Leary, R., Crutchley, P., and Fine, A., “Natural language processing of social media as screening for suicide risk,” *Biomedical informatics insights* **10**, 1178222618792860 (2018).
- [8] Nandwani, P. and Verma, R., “A review on sentiment analysis and emotion detection from text,” *Social Network Analysis and Mining* **11**(1), 81 (2021).
- [9] Acheampong, F. A., Wenyu, C., and Nunoo-Mensah, H., “Text-based emotion detection: Advances, challenges, and opportunities,” *Engineering Reports* **2**(7), e12189 (2020).

- [10] Saunders, D., “Domain adaptation and multi-domain adaptation for neural machine translation: A survey,” *Journal of Artificial Intelligence Research* **75**, 351–424 (2022).
- [11] Laparra, E., Bethard, S., and Miller, T. A., “Rethinking domain adaptation for machine learning over clinical language,” *JAMIA open* **3**(2), 146–150 (2020).
- [12] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., “A survey of large language models,” *arXiv preprint arXiv:2303.18223* (2023).
- [13] Xu, X., Yao, B., Dong, Y., Yu, H., Hendler, J., Dey, A. K., and Wang, D., “Leveraging large language models for mental health prediction via online text data,” *arXiv preprint arXiv:2307.14385* (2023).
- [14] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W., “Large language models in medicine,” *Nature Medicine*, 1–11 (2023).
- [15] Le Glaz, A., Haralambous, Y., Kim-Dufoir, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouguet, S., et al., “Machine learning and natural language processing in mental health: systematic review,” *Journal of Medical Internet Research* **23**(5), e15708 (2021).
- [16] Fu, G., Song, C., Li, J., Ma, Y., Chen, P., Wang, R., Yang, B. X., and Huang, Z., “Distant supervision for mental health management in social media: suicide risk classification system development study,” *Journal of medical internet research* **23**(8), e26119 (2021).
- [17] Singh, M., Jakhar, A. K., and Pandey, S., “Sentiment analysis on the impact of coronavirus in social life using the bert model,” *Social Network Analysis and Mining* **11**(1), 33 (2021).
- [18] Wan, F., “Sentiment analysis of weibo comments based on deep neural network,” in *[2019 international conference on communications, information system and computer engineering (CISCE)]*, 626–630, IEEE (2019).
- [19] Zhang, X., Li, W., Ying, H., Li, F., Tang, S., and Lu, S., “Emotion detection in online social networks: a multilabel learning approach,” *IEEE Internet of Things Journal* **7**(9), 8133–8143 (2020).
- [20] Wang, J., Zhou, Y., Zhang, W., Evans, R., and Zhu, C., “Concerns expressed by chinese social media users during the COVID-19 pandemic: content analysis of sina weibo microblogging data,” *Journal of medical Internet research* **22**(11), e22152 (2020).
- [21] Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R., “Challenges and applications of large language models,” *arXiv preprint arXiv:2307.10169* (2023).
- [22] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al., “Emergent abilities of large language models,” *arXiv preprint arXiv:2206.07682* (2022).
- [23] Liebreinz, M., Schleifer, R., Buadze, A., Bhugra, D., and Smith, A., “Generating scholarly content with ChatGPT: ethical challenges for medical publishing,” *The Lancet Digital Health* **5**(3), e105–e106 (2023).
- [24] Jeblick, K., Schachtner, B., Dexl, J., Mittermeier, A., Stüber, A. T., Topalis, J., Weber, T., Wesp, P., Sabel, B., Rieke, J., et al., “ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports,” *arXiv preprint arXiv:2212.14882* (2022).
- [25] Surameery, N. M. S. and Shakor, M. Y., “Use ChatGPT to solve programming bugs,” *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290* **3**(01), 17–22 (2023).
- [26] Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günemann, S., Hüllermeier, E., et al., “Chatgpt for good? on opportunities and challenges of large language models for education,” *Learning and individual differences* **103**, 102274 (2023).
- [27] Yeo, Y. H., Samaan, J. S., Ng, W. H., Ting, P.-S., Trivedi, H., Vipani, A., Ayoub, W., Yang, J. D., Liran, O., Spiegel, B., et al., “Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma,” *medRxiv*, 2023–02 (2023).
- [28] Jiang, L. Y., Liu, X. C., Nejatian, N. P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H. A., Laufer, I., Punjabi, P., et al., “Health system-scale language models are all-purpose prediction engines,” *Nature*, 1–6 (2023).
- [29] Farhat, F., “ChatGPT as a complementary mental health resource: a boon or a bane,” *Annals of Biomedical Engineering*, 1–4 (2023).
- [30] Qin, W., Chen, Z., Wang, L., Lan, Y., Ren, W., and Hong, R., “Read, diagnose and chat: Towards explainable and interactive LLMs-augmented depression detection in social media,” *arXiv preprint arXiv:2305.05138* (2023).

- [31] Chen, S., Wu, M., Zhu, K. Q., Lan, K., Zhang, Z., and Cui, L., “LLM-empowered chatbots for psychiatrist and patient simulation: Application and evaluation,” *arXiv preprint arXiv:2305.13614* (2023).
- [32] Fu, G., Zhao, Q., Li, J., Luo, D., Song, C., Zhai, W., Liu, S., Wang, F., Wang, Y., Cheng, L., et al., “Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals,” *arXiv preprint arXiv:2308.15192* (2023).
- [33] Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., et al., “Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum,” *JAMA internal medicine* (2023).
- [34] Yang, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., and Ananiadou, S., “Towards interpretable mental health analysis with ChatGPT,” (2023).
- [35] Vaishya, R., Misra, A., and Vaish, A., “ChatGPT: Is this version good for healthcare and research?,” *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* **17**(4), 102744 (2023).
- [36] Editorial, “Will ChatGPT transform healthcare?,” *Nature Medicine* , 505–506 (2023).
- [37] Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., et al., “A survey on evaluation of large language models,” *arXiv preprint arXiv:2307.03109* (2023).
- [38] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., “BERT: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805* (2018).
- [39] Xiao, L., Huang, X., Chen, B., and Jing, L., “Label-specific document representation for multi-label text classification,” in *[Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)]*, 466–475 (2019).
- [40] Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Xu, Y., Zheng, W., Xia, X., et al., “GLM-130b: An open bilingual pre-trained model,” *arXiv preprint arXiv:2210.02414* (2022).
- [41] OpenAI, “Introducing chatgpt,” (2023).
- [42] OpenAI, “Gpt-4 technical report,” (2023).
- [43] Wang, S., Sun, Y., Xiang, Y., Wu, Z., Ding, S., Gong, W., Feng, S., Shang, J., Zhao, Y., Pang, C., et al., “Ernie 3.0 titan: Exploring larger-scale knowledge enhanced pre-training for language understanding and generation,” *arXiv preprint arXiv:2112.12731* (2021).
- [44] Burns, D. D., [*Feeling good*], Signet Book (1981).