



## Article

# Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs)

Tin Lai \*, Yukun Shi , Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou and Ziqi Wang

School of Computer Science, The University of Sydney, Camperdown, NSW 2006, Australia; yshi7396@uni.sydney.edu.au (Y.S.); zidu5377@uni.sydney.edu.au (Z.D.); jiwu9957@uni.sydney.edu.au (J.W.); kefu2224@uni.sydney.edu.au (K.F.); ydou2446@uni.sydney.edu.au (Y.D.); zwan9927@uni.sydney.edu.au (Z.W.)  
\* Correspondence: tin.lai@sydney.edu.au

**Abstract:** The demand for psychological counselling has grown significantly in recent years, particularly with the global outbreak of COVID-19, which heightened the need for timely and professional mental health support. Online psychological counselling emerged as the predominant mode of providing services in response to this demand. In this study, we propose the Psy-LLM framework, an AI-based assistive tool leveraging large language models (LLMs) for question answering in psychological consultation settings to ease the demand on mental health professions. Our framework combines pre-trained LLMs with real-world professional questions-and-answers (Q&A) from psychologists and extensively crawled psychological articles. The Psy-LLM framework serves as a front-end tool for healthcare professionals, allowing them to provide immediate responses and mindfulness activities to alleviate patient stress. Additionally, it functions as a screening tool to identify urgent cases requiring further assistance. We evaluated the framework using intrinsic metrics, such as perplexity, and extrinsic evaluation metrics, including human participant assessments of response helpfulness, fluency, relevance, and logic. The results demonstrate the effectiveness of the Psy-LLM framework in generating coherent and relevant answers to psychological questions. This article discusses the potential and limitations of using large language models to enhance mental health support through AI technologies.



**Citation:** Lai, T.; Shi, Y.; Du, Z.; Wu, J.; Fu, K.; Dou, Y.; Wang, Z. Supporting the Demand on Mental Health Services with AI-Based Conversational Large Language Models (LLMs). *Biomedinformatics* **2024**, *4*, 8–33. <https://doi.org/10.3390/biomedinformatics4010002>

Academic Editor: Alexandre G. De Brevin, Pan Zheng and Bin Wang

Received: 13 July 2023

Revised: 22 August 2023

Accepted: 7 December 2023

Published: 22 December 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** psychological counselling; artificial intelligence; large language models (LLMs); natural language processing (NLP); mental well-being; natural language generation (NLG)

## 1. Introduction

The field of AI utilising dialogue technology has witnessed significant growth, particularly in the domain of automatic chatbots and ticket support systems [1]. This application of dialogue technology has emerged as a cutting-edge and increasingly popular approach in the realm of AI-powered support systems. With changing global dynamics, the severity of the ongoing pandemic, and an upsurge in psychological challenges faced by the public, the mental well-being of young individuals, in particular, is a cause for concern. The pressures of urbanisation and the internet have led to various psychological issues [2], including depression, procrastination, anxiety, obsessive-compulsive disorder, and social phobia [3], which have become prevalent ailments of our time.

Psychological counselling involves the utilisation of psychological methods to provide assistance to individuals experiencing difficulties in psychological adaptation and seeking solutions. The demand for psychological counselling has witnessed a significant surge in recent years [4], while the availability of professional psychological consultants remains insufficient. The profession of psychological consulting imposes high standards and qualifications. For instance, registered psychologists within psychological associations require students to possess a Master's degree in psychology-related disciplines, undergo a minimum of 150 h of direct counselling, and receive face-to-face supervision by registered

supervisors for no less than 100 h [5]. Additionally, the burnout rate among mental health professionals further exacerbates this shortage [6].

In 2020, the global outbreak of COVID-19 exacerbated the need for timely and professional psychological counselling due to the tremendous stress it imposed on society [7]. Consequently, online psychological counselling through the internet has progressively become the dominant mode of delivering counselling services [8]. AI-based assistive psychological support not only addresses the severe supply–demand gap in the consulting industry, but also enhances the responsiveness of online psychological counselling services, thereby promoting the implementation of mental health strategies. Such an assistive tool serves to ease the shortage of mental health support when no human counsellors are available to help.

In light of these circumstances, our team is determined to develop an assistive mental health consulting framework to serve as a constant source of support. Creating an AI-powered framework can allow users to engage with it comfortably, given its non-human identity, thereby reducing feelings of shame among users [9]. In particular, with the absence of available psychological support, our framework serves as the second-best approach to providing timely support to patients. Amid the challenges posed by the pandemic, online psychological counselling has proven instrumental and has gradually become the predominant form of counselling. However, the growing disparity between supply and demand within our society's psychological consultation industry is a pressing concern. The application of AI technology to mental health and psychological counselling is an emerging and promising field. Conversation frameworks, chatbots, and virtual agents are computer programs that simulate human conversation [10]. They can engage in natural and effective interactions with individuals, providing them with emotional experiences through the incorporation of emotional and human-like characteristics. In practical terms, dialogue frameworks hold significant potential for supporting the demand in online consultations and addressing supply–demand imbalances.

In this study, we propose an AI-based **Psychological Support with Large Language Models (Psy-LLM)** framework designed for question answering, with the purpose of providing online consultation services to alleviate the demand for mental health professionals during pandemics and beyond. Psy-LLM is an online psychological consultation model pre-trained with large language models (LLMs) and further trained with questions-and-answers (Q&A) from professional psychologists and large-scale crawled psychological articles. The framework can provide professional answers to users' requests for psychological support. In particular, Psy-LLM can provide mental health advice, both through recommendations for health professionals and as standalone tools for patients when no human counsellors are available due to time constraints or staff shortages. Our model is built upon large-scale pre-training corpus models, specifically PanGu [11] and WenZhong [12]. The PanGu model, developed by Huawei's Pengcheng Laboratory, and the WenZhong model, developed by the Idea Research Institute, served as the basis for our work. For data acquisition, we collected a substantial number of Chinese psychological articles from public websites. Additionally, we obtained permission from the Artificial Intelligence Research Institute of Tsinghua University to utilise the PsyQA dataset, which comprises many question–answer pairs related to psychological counselling. Each answer in the dataset was reviewed and adjusted by professionals holding master's degrees or above in psychological counselling to ensure its quality. We fine-tuned the model in downstream tasks using the acquired dataset and PsyQA [13]. As part of the evaluation process, we established a dedicated website and deployed the fine-tuned model on a server, allowing users to provide timely ratings. Based on the scoring results, we iteratively refined and re-fine-tuned the model.

Our contribution includes proposing a framework for AI-based psychological consultation framework and an empirical study on its effectiveness. We successfully developed a mental health consulting model that effectively provides clear and professional responses to users' psychological inquiries. Empirically, we tested deploying the model on a server,

and the model responded to users within seconds. Our framework has the potential to offer a practical tool for professionals to efficiently screen and promptly respond to individuals in urgent need of mental support, thereby addressing and alleviating pressing demands within the healthcare industry.

## 2. Related Works

In recent years, there has been increasing interest in utilising AI for tackling difficult problems in traditional domains like adopting AI in the construction industry [14], localisation in robotic applications [15], assistance systems in the service sector [16], financial forecasting [17], improving workflow in the oil and gas industry [18], planning and scheduling [19], monitoring ocean contamination [20], remote sensing for search and rescue [21], and it has even been used in the life cycle of material discovery [22]. The health care industry has adopted AI-based machine-learning techniques for classifying medical images [23], guiding cancer diagnosis [24], as screening tools for diabetes [25], and ultimately to improve the clinical workflow in the practice of medicine [26].

One area of research focuses on using conversational agents, also known as chatbots, for mental health support. Chatbots have the potential to provide accessible and cost-effective assistance to individuals in need. For example, Martinengo et al. (2022) [27] qualitatively analysed user-conversational agents and found that these types of chatbots can offer anonymous, empathetic, and non-judgemental interactions that align with face-to-face psychotherapy. Chatbots can utilise NLP techniques to engage users in therapeutic conversations and provide personalised support. The results showed promising outcomes, indicating the potential effectiveness of chatbots in delivering mental health interventions [28]. Pre-trained language models have also gained attention in the field of mental health counselling. These models, such as GPT-3 [29], provide a foundation for generating human-like responses to user queries. Wang et al. (2023) [30] explored the application of LLMs in providing mental health counselling. They found that LLMs demonstrated a certain level of understanding and empathy, providing responses that were perceived as helpful by users. However, limitations in controlling the model's output and ensuring ethical guidelines were highlighted.

Furthermore, there is a growing body of research on using NLP techniques to analyse mental-health-related text data [31]. Researchers have applied machine-learning algorithms to detect mental health conditions [32], predict suicidal ideation [33], and identify linguistic markers associated with psychological well-being [34]. For instance, de Choudhury et al. (2013) [35] analysed social media data to predict depression among individuals. By extracting linguistic features and using machine-learning classifiers, they achieved promising results in identifying individuals at risk of depression. Additionally, several studies have investigated the integration of modern technologies into existing mental health interventions. For instance, Lui et al. (2017) [36] investigated the use of mobile applications to support the delivery of psychotherapy.

Shaikh and Mhetre (2022) [37] developed a friendly AI-based chatbot using deep learning and artificial intelligence techniques. The chatbot aimed to help individuals with insomnia by addressing harmful feelings and increasing interactions with users as they experienced sadness and anxiety. In another line of research, chatbots have been extensively studied in the domain of customer service. Many companies have adopted chatbots to assist customers in making purchases and understanding products. These chatbots provide prompt replies, enhancing customer satisfaction [38]. Furthermore, advancements in language models such as BERT and GPT have influenced the development of conversational chatbots. Researchers have leveraged BERT-based question-answering models to improve the accuracy and efficiency of chatbot responses [39]. The GPT models, including GPT-2 and GPT-3, have introduced innovations such as zero-shot and few-shot learning, significantly expanding their capabilities in generating human-like text [29]. However, limitations in generating coherent and contextual responses and the interpretability of the

models have been identified. The model incorporated a 48-layer transformer stack and achieved a parameter count of 1.5 billion, resulting in enhanced generalisation abilities [29].

In summary, previous work in AI and NLP for mental health support has demonstrated the potential of chatbots, pre-trained language models, and data analysis techniques. These approaches offer new avenues for delivering accessible and personalised mental health interventions. Nonetheless, further research is needed to address ethical, privacy, and reliability issues and to optimise integrating AI technologies into existing counselling practices.

### 3. Mental Health and Social Well-Being in Overly Populated Cities

The availability of mental health professionals has always been a major problem in overpopulated cities such as China. The World Health Organisation reported that the prevalence of depression in China exceeded 54 million people even before the onset of the COVID-19 pandemic [40]. The situation was exacerbated by the implementation of quarantine measures and social distancing, leading to a worsening condition [41]. Unfortunately, only a small fraction of the affected population receives adequate medical treatment, as there are only 2 psychiatrists per 100,000 people in China [42]. Consequently, there is a pressing need for a dynamic system that can assist patients effectively. Contemporary conversational chatbots have demonstrated their ability to emulate human-like conversations.

Hence, it is imperative to develop a user-friendly AI-based chatbot specifically designed to address anxiety and depression, with the aim of improving the user's emotional well-being by providing relevant and helpful responses. This project aims to construct a Chinese psychological dialogue model capable of comprehending the semantic meaning of a consultant's request and offering appropriate advice, particularly to address the shortage of mental health workers during demanding periods. The trained model will be integrated into a website, featuring a user interface (UI) that ensures ease of operation, thereby enhancing the efficiency of psychological counselling.

#### 3.1. Research Questions

Psychology is an intricate and advanced discipline gaining increasing significance as society progresses. However, because of its high barriers to entry, resources for psychological counselling have long been scarce. In numerous cases, individuals face challenges accessing adequate mental health support [43]. Furthermore, the high cost of psychological counselling often prevents many individuals from prioritising their mental well-being. This issue is particularly prominent in China, a country with a large population where psychological problems have been historically overlooked. China needs a robust foundation for psychological counselling, addressing challenges such as a deficient knowledge base and limited data. Consequently, intelligent assistance in psychology must be improved in the Chinese context.

Traditional psychological counselling primarily focuses on privacy and employs a one-on-one question-and-answer approach, inherently leading to inefficiencies. However, in today's high-pressure society, where mental health issues are pervasive, relying solely on scarce psychologists is arduous. Additionally, influenced by traditional culture, individuals often hesitate to acknowledge and address their psychological problems due to feelings of shame and perceiving such discussions as signs of weakness [44]. This reluctance is especially prominent when conversing with real humans, let alone seeking assistance from unfamiliar psychologists. Furthermore, with the advancement of modern natural language processing (NLP) artificial intelligence models, there is a possibility to optimise the conventional and widely adopted question-and-answer model specifically for the field of psychology. When interacting with AI, people are more inclined to express their true thoughts and emotions without fear of prejudice and discrimination compared with interactions with real humans. However, the lack of verbal cues and continuous monitoring of patients' emotional progress is also a cause of concern in an online psychological consultation context, even when performed by human counsellors [45].

### 3.2. Research Scope

Through our project, we aim to provide a meaningful contribution to the field of mental well-being. By leveraging our AI model and proposing a framework for an internet-accessible consultation, we intend to enhance the accessibility of mental health support, making it more affordable and providing an avenue for psychological question-and-answer interactions, particularly to address the shortage of human counsellors for mental health support. To achieve this goal, we gather professional counselling question-and-answer data and psychologically relevant knowledge data to construct a robust question-and-answer model specific to this domain. The success of our project relies on the utilisation of high-quality question-and-answer models. We plan to employ established Chinese pre-trained models with exceptional human-computer interaction and communication skills, characterised by fluent language, logical reasoning, and semantic understanding. However, these existing pre-trained models need more specialised psychological expertise and emotional understanding for counselling purposes. To address this limitation, we intend to integrate two models, namely the WenZhong model and the PanGu model, and evaluate their performance to determine the more suitable choice as our final model.

Subsequently, it is imperative to ensure our model is accessible to a broader audience. Leveraging the internet provides the most effective means to accomplish this objective. Together with the guidance of professional mental health experts, our model can provide an additional venue for the general public to access mental health support by easing the stress and demand on mental health staff through an open and online platform.

## 4. Psy-LLM Framework

The Psy-LLM framework aims to be an assistive mental health tool to support the workflow of professional counsellors, particularly to support those who might be suffering from depression or anxiety.

### 4.1. Target Audience and Model Usage

The contributing factor of Psy-LLM—an AI-powered conversational model—is two-fold. (1) Firstly, Psy-LLM is trained with a corpus of mental health supportive Q&A from mental health professionals, which enables Psy-LLM to be used as an assistive tool in an online consultation context. When users want to seek support from an online chat, Psy-LLM can provide suggestive answers to human counsellors to ease the staff's workload. Such an approach eases the entry barrier for newly trained mental health staff to provide useful and supportive comments for those in need. (2) Furthermore, in the absence of human counsellors (e.g., during off-hours or high-demand periods), Psy-LLM can also be a web front-end for users to interact with the system in an online consultation manner. Providing timely support to help-seeking individuals is especially important among suicidal individuals [46]. Therefore, an AI-powered online consultation might be the next best venue to respond to the absence of human counsellors.

### 4.2. Large-Scale Pre-Trained LLMs

Our project involves leveraging two large-scale pre-training models, namely WenZhong and PanGu, to develop the question-answering language model. The utilisation of pre-training models offers several advantages, including the following: (1) Enhanced language representations: Pre-training on extensive unlabelled data enables the model to acquire more comprehensive language representations, which in turn can positively impact downstream tasks. (2) Improved initialisation parameters: Pre-training provides a superior initialisation point for the model, facilitating a better generalisation performance on the target task and expediting convergence during training. (3) Effective regularisation: Pre-training acts as an effective regularisation technique, mitigating the risk of overfitting when working with limited or small datasets. This is especially valuable as a randomly initialised deep model is susceptible to overfitting on such datasets. By harnessing the

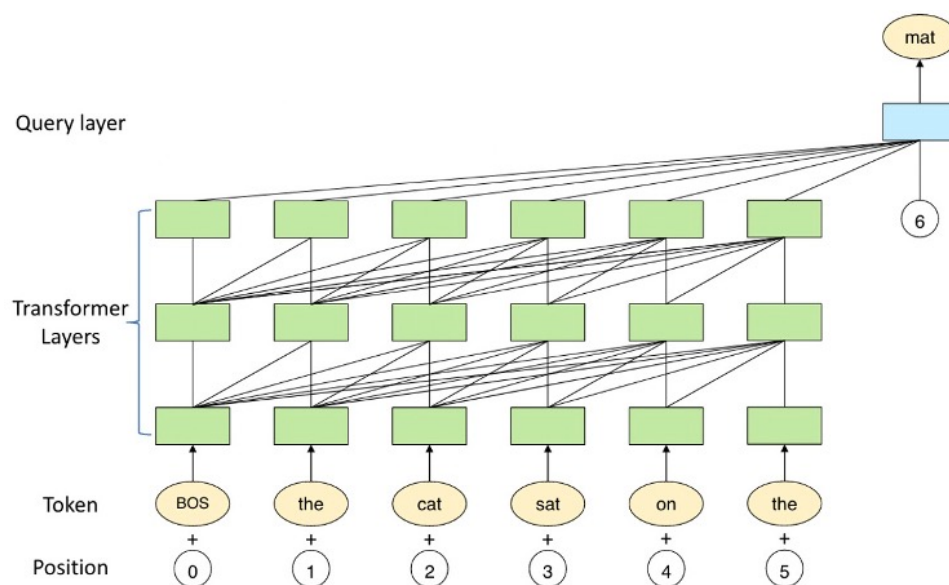


advantages of pre-training models, we aim to enhance the performance and robustness of our question-answering language model for psychological counselling.

#### 4.3. PanGu Model

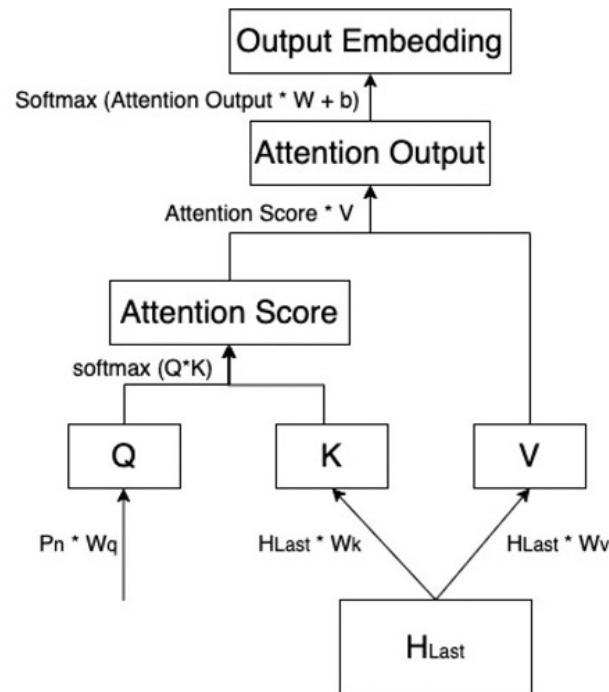
The PanGu model is the first Chinese large-scale pre-training autoregressive language model with up to 200 billion parameters [11]. In an autoregressive model, the process of generating sentences can be likened to a Markov chain, where the prediction of a token is dependent on the preceding tokens. The PanGu model, developed within the MindSpore framework, was trained using 2048 Ascend AI processors provided by Huawei and was trained on a high-quality corpus of 1.1 TB. It was officially released in April 2021 and has achieved the top rank in the Chinese Language Comprehension Benchmark (CLUE), a widely recognised benchmark for Chinese language comprehension [47].

The architecture of the PanGu model follows a similar structure to that of GPT-3, employing standard transformer layers (Figure 1). Each transformer layer comprises two sub-layers: multi-head attention (MHA) and a fully connected feed-forward network (FFN). MHA involves three primary steps: calculating the similarity between the query and key, applying a softmax function to obtain attention scores, and multiplying the attention scores with the value to obtain the attention output. The attention output then passes through a linear layer and undergoes softmax to generate the output embedding. The output embedding is combined with the FFN input through a residual module. The FFN consists of two linear layers with a GeLU activation function between each consecutive layer. MHA and FFN utilise the pre-layer normalisation scheme, facilitating faster and easier training of the transformer model.



**Figure 1.** The model layers and architecture of the PanGu model [11].

However, the last layer of the PanGu model deviates from the standard transformer layer structure. Instead, it incorporates a query layer designed to predict the next token, thereby enhancing the model's positional awareness and improving generation effectiveness. The query layer is a narrow yet powerful decoder that relies solely on position information. The structure of the query layer is illustrated in Figure 2. The primary distinction between the query layer and the transformer layer lies in the query input of self-attention. While the inputs of query, key, and value in other self-attention layers of the transformer remain standard, the query layer introduces a query embedding, which functions similarly to position embedding, as the query input for self-attention in the last layer.



**Figure 2.** The query layer in the PanGu model.

The PanGu model is available in four distinct variations, each characterised by different parameter sizes (Table 1). These variations include PanGu 350 M, PanGu 2.6 B, PanGu 13 B, and PanGu 200 B (which is not open source). The parameter sizes differ across these models, reflecting their varying levels of complexity and capacity for language understanding and generation.

**Table 1.** The parametric size of the various settings in the PanGu model.

Model	Parameters	Layers	Hidden Size	Head	Seq Length
PanGu 350 M	350 M	24	1024	16	1024
PanGu 2.6 B	2.6 B	32	2560	40	1024
PanGu 13 B	13.1 B	40	5120	40	1024
PanGu 13 B	207.0 B	64	16,384	128	1024

#### 4.4. WenZhong Model

In addition to the PanGu model, we also incorporated the WenZhong model as one of the models used. The WenZhong model is a pre-trained model based on the GPT-2 architecture and trained on a large-scale Chinese corpus. Over the past few years, pre-trained models have become the foundation of cognitive intelligence, enabling advancements in natural language processing and computer-vision algorithms.

The scale of pre-trained models has been rapidly increasing, growing by a factor of 10 each year, starting from the initial BERT model with 100 million parameters, to the more recent GPT models with over 100 billion parameters. Given the nature of our task, which requires a generation model with expertise in different professional domains, we opted for the WenZhong model.

Because of the model size of LLMs like GPT, computing resources are the limiting factor hindering further progress in the field. Universities and research institutions often need more computing power to train and utilise large-scale pre-trained models. This limitation impedes the broader implementation of AI technologies. Hence, we adopted the WenZhong model, which is built upon a large pre-trained model trained on a Chinese corpus, so as to avoid training the model from scratch.

The WenZhong model series consists of one-way language models dominated by a decoder structure and a series of powerful generation models. The WenZhong-3.5 B model, with 3.5 billion parameters, employs 100 G data and 256 A100 GPUs for 28 h of training, exhibiting strong generation capabilities. Thus, the WenZhong model is highly powerful, featuring 30 decoder layers and billions of parameters. We also utilised the WenZhong-GPT2-110M version in this project, comprising 110 million parameters and 12 layers. It is important to note that the WenZhong model was pre-trained on the Wudao Corpus (300 G version).

#### 4.5. Collecting Large Scale Dataset

Two types of data sources were obtained for this project. The first dataset, PsyQA [13], consisting of question and answer pairs, focuses on Chinese psychological health support. The authors authorised us to use this dataset, which contains 22,000 questions and 56,000 well-structured, lengthy answers. The PsyQA dataset includes numerous high-quality questions and answers related to psychological support, and it had already undergone basic cleaning before we received it. We selected a test set of 5000 samples from this PsyQA dataset for our experiments.

##### 4.5.1. Data Crawling

The second dataset was obtained by crawling various Chinese social media platforms, such as Tianya, Zhihu, and Yixinli. These platforms allow users to post topics or questions about mental and emotional issues, while other users can offer support and assistance to help-seeking individuals. The Yixinli website specifically focuses on professional mental health support, but only provides approximately 10,000 samples. Other types of datasets collected from these platforms included articles and conversations, which we converted into a question-and-answer format. However, we excluded the articles from our fine-tuning training due to the model's input limitations and the fact that our predictions focused on mental health support answers. The articles were often lengthy, and many of them were in PDF format, requiring additional time for conversion into a usable text format. Consequently, we only obtained around 5000 article samples. In order to address the lack of emotional expression in the text of these articles, we incorporated text data from oral expressions. We crawled audio and video data from platforms like Qingting FM and Ximalaya, popular audio and video-sharing forums in China. However, converting audio and video data into text format was time-consuming, resulting in a limited amount of data in our dataset. We utilised the dataset obtained from websites for fine-tuning training. Ultimately, our entire dataset consisted of 400,000 samples, each separated by a blank line, i.e., “\n\n”.

Table 2 shows the time spent on data crawling from different websites. It is evident that most of the samples in this dataset were obtained from Tianya, resulting in a data size of approximately 2 GB. The datasets from Zhihu and Yixinli were 500 MB and 200 MB, respectively. Overall, we spent approximately 70 h on data collection. Although the data collected from the internet were abundant and authentic, the cleaning process could have been smoother due to inconsistencies in the online data.

**Table 2.** Dataset crawled from different platforms.

Platform	Data Size	Crawling Time
Tianya	2 GB	40 h+
Zhihu	500 Mb	20 h+
Yixinli	200 Mb	8 h+

To address the time-consuming nature of web crawling, we implemented a distributed crawl technology that utilised idle computers connected to the internet or other networks, effectively harnessing additional processing power [48]. Our approach involved obtaining sub-websites from the main website and saving them using custom crawling code. This



code primarily relied on Python libraries such as “requests”, “BeautifulSoup”, and “web-driver”. In addition, we employed dynamic web crawlers capable of collecting clickable elements, simulating user actions, comparing web page states, manipulating the DOM tree, and handling various user-invoked events [49]. Unlike static page structures that cannot handle dynamic local refresh and asynchronous loading [50], dynamic crawlers can extract data from behind search interfaces.

The process of the dynamic crawler involved leveraging web developer tools within the browser to obtain XHR (XMLHttpRequest) information, which included requests containing headers, previews, and responses. We acquired relevant files by systematically searching through these layers of data and capturing network packets. After obtaining the sub-websites using static and dynamic crawling methods, we distributed them across multiple idle computers. Each computer was assigned specific sub-websites, and we collected project-related data using a combination of static and dynamic crawling techniques. Ultimately, we utilised eight computers for the crawling process, which took approximately 70 h.

#### 4.5.2. Data Cleaning

In line with the PanGu paper [11], we adopted the original data cleaning method utilised in the PanGu model. Additionally, we incorporated some additional cleaning steps. The following are the cleaning steps we employed:

1. *Removal of duplicate samples*: We eliminated any duplicate samples in the dataset to ensure data uniqueness.
2. *Removal of samples containing advertised keywords*: We excluded samples that contained specific keywords associated with advertisements or promotional content.
3. *Deletion of data with less than 150 characters*: Samples with less than 150 characters were removed from the dataset, as they were deemed insufficient for effective model training.
4. *Removal of URLs*: Any URLs present in the samples were eliminated to maintain the focus on the text content.
5. *Removal of user names and post time*: User names, such as “@jack”, and post timestamps were removed from the samples, as they were considered irrelevant to the text content.
6. *Removal of repeated punctuation*: Instances of repeated punctuation marks, such as “!!!” or “.....”, were removed from the samples to ensure cleaner and more concise text.
7. *Conversion of traditional Chinese to simplified Chinese*: All traditional Chinese characters were converted to simplified Chinese characters to standardise the text.

Following the data cleaning process, the dataset could be directly inputted into the PanGu model. However, for training with the WenZhong model, the samples needed further processing. Specifically, all punctuation marks were removed, and the samples were tokenised to ensure a consistent length of 1000 tokens for compatibility with the WenZhong model.

#### 4.5.3. Data Analysis

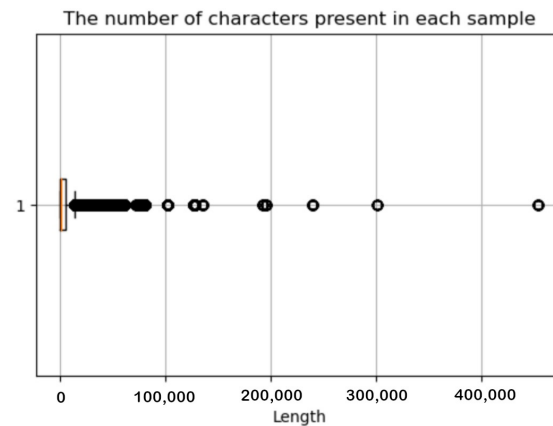
Data analysis plays a crucial role in understanding the fundamental characteristics of textual data. In the context of the Chinese language, the exploratory data analysis (EDA) methods may be less diverse than those used for English. In this study, we primarily employed two common methods: word frequency analysis and sentence length analysis, to gain insights into the dataset.

To analyse the distribution of characters in each sample, we referred to the character number data presented in Table 3. By visualising this information using a box chart, we examined the range of character counts across the samples. Some samples were empty after the data cleaning, which we then pruned from our dataset. Figure 3 displays the distribution of sample lengths, indicating that the majority of samples fell within the range of 10,000 characters.

Overall, these preliminary analyses allowed us to gain initial insights into the dataset and provided a foundation for further exploration and understanding of the textual data.

**Table 3.** Data distribution of the length of each sample.

Count	Mean	Std	Min	25%	50%	70%	Max
371,434	5343	11,335	0	653	1835	6039	454,611



**Figure 3.** The number of characters in each sample.

To examine the word frequency in our dataset, we conducted an analysis after removing the stop words. The word cloud visualisation in Figure 4 illustrates the most frequent words in the dataset. Notably, the prominent words observed included “anxiety”, “appearance”, “marriage”, “relationship”, “family”, and “stressful”. These words are highly relevant to the topic of mental health, indicating that our dataset is robust and aligns well with the focus of our task.

The presence of these mental-health-related terms further underscores the suitability of our data for addressing the objectives of our study. It suggests that our dataset encompasses significant content related to psychological aspects, allowing us to effectively explore and address relevant topics in the context of our research.



**Figure 4.** Word cloud of the frequent words within our dataset, containing words such as anxiety, appearances, looks, felings, scars, self-esteems, etc.

#### 4.6. Model Training

**Model Size:** We used PanGu 350 M to generate language considering the computational power. It contains 350 million parameters, 24 layers, 1024 hidden sizes, and 16

attention heads. Furthermore, we also trained the WenZhong-110 M model, which contains 12 layers and has 110 M parameters.

**Training Data:** We employed the 2.85 GB psychology corpus data crawled from psychology platforms like Yixinli and Tianya, to train the original PanGu 350 M model. After that, we used 56,000 question–answer pairs from PsyQA dataset to fine-tune the model.

**Training Platform:** We trained the PanGu model on the OpenI platform with a free V100 graphics card GPU because OpenI is the open source platform of the PanGu model, and it was convenient for us to deploy the required files, images, and GPU. The batch size was set to 8, and the training iteration was set to 100,000 because we found that 50,000 iterations was not enough for the model’s loss to converge. We trained the WenZhong model in Jupyter Notebook. To fine-tune this model, we tokenised the data, which transformed words into tokens. Further, we also isolated the max length of each sentence as 500.

#### 4.7. Dataset Evaluation

Determining the cleaning rules and data filtering thresholds are important aspects of the data cleaning process. We employed a data quality evaluation method that combined both manual and model-based evaluations to evaluate the dataset obtained from website crawling.

For the model-based evaluation, we utilised the PanGu 350 M model and calculated the perplexity metric after each data cleaning stage. A lower perplexity value indicated a more effective cleaning process and higher dataset quality. In addition to the model-based evaluation, we sought input from experts in psychology. We invited two members from our University’s School of Psychology, Faculty of Science, to perform a random sample check on the dataset after it had undergone the cleaning process. While this method did not cover the entire corpus comprehensively, it provided valuable insights and played a role in data cleaning and quality evaluation.

The evaluation process involved the following steps: First, we provided the experts with a sample of the cleaned dataset and asked them to assess its quality based on their expertise and domain knowledge. They evaluated the dataset for accuracy, relevance, and coherence, providing feedback and suggestions for further improvements.

Next, we conducted a comparative analysis between the model-based evaluation and the expert evaluation. We examined the perplexity scores obtained from the PanGu 350 M model and compared them with the feedback provided by the experts. This allowed us to identify any discrepancies or areas of improvement in the dataset.

Overall, the combination of model-based evaluation and expert assessment comprehensively evaluated the dataset quality. It allowed us to identify and address any issues or shortcomings in the data cleaning process, ensuring that the final dataset used for training and evaluation was high quality and suitable for our research purposes.

#### 4.8. Model Training Setting

**Models:** For our training, we utilised the PanGu 350 M model, considering the available computational resources. This model consists of 350 million parameters, 24 layers, a hidden size 1024, and 16 attention heads. Additionally, we targeted the WenZhong-110 M model, which contains 12 layers and 110 million parameters.

**Training Data:** We collected a psychology corpus dataset totalling 2.85 GB, which was crawled from psychology platforms such as Yixinli and Tianya. This dataset was used for training the original PanGu 350 M model. Subsequently, we fine-tuned the model using 56,000 question–answer pairs from the PsyQA dataset.

**Training Platform:** The PanGu model was trained on the OpenI platform, utilising a free 1 V100 graphics card GPU. OpenI is an open-source platform specifically designed for the PanGu model, allowing us to easily deploy the necessary files, images, and GPU resources. For training with the V100 graphics card (32 GB memory), the minimum recommended configuration is one card, while the recommended configuration is two. The graphics card requirements can be adjusted based on the memory size (for example, a 16 GB memory

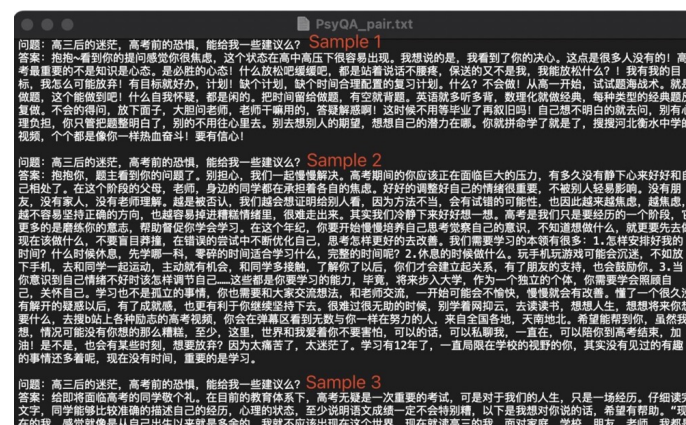
card would require twice as many cards as the V100). Increasing the number of graphics cards can help improve the training speed if the dataset is large. We set the batch size to 8 and performed training for 100,000 iterations, as we observed that 50,000 iterations were insufficient for the model's loss to converge. For the WenZhong model, we used Jupyter Notebook to run the pre-trained model and fine-tuned it on a system with 64 GB memory and an RTX3060 graphics card. The version details of the hardware and software components are listed in Table 4.

**Table 4.** Hardware and software versions.

Hardware and Software	Version
Operating System	Windows 10
Numpy	1.18.5
Pandas	1.3.4
Torch	1.11.0
Tokenizers	0.13.1
tqdm	4.64.1
Jupyter	1.0.0
Transformers	4.23.1

### Training Process

According to the guide for training the PanGu model with GPU, the first step is environment configuration. We prepared Pytorch, PanGu image, one V100 graphics card, and some PanGu model files like vocabulary. The second step is data preprocessing. We placed the training corpus into a text file, where each sample is a paragraph separated by two new lines. Then, we converted it into a binary file because that is the required input format of the training PanGu model. The third step is model training. We uploaded the PanGu model and the bin file to the OpenI platform and set some parameters like iteration to train it. Our training procedure for the PanGu model consisted of two steps. Firstly, we trained the original PanGu 350 M model with all the crawling data for 100,000 iterations. The model started to converge at about 60,000. This model learned psychology-domain knowledge based on pre-trained data. Secondly, we fine-tuned it with the PsyQA dataset to improve the model's capability to provide useful answers to users about mental health support. Figure 5 contains some Q&A samples of the training corpus.



**Figure 5.** Preparing the training corpus. The sample question is same for the 3 samples. Question: Experiencing confusion after completing the third year of high school and feeling anxious before the college entrance examination—do you have any guidance or advice to offer? The answers refers to concepts such as: Recognising your anxiety stemming from high school pressure, it's crucial to foster a mindset of determination for the college entrance exam, emphasising the importance of a goal-oriented approach, structured planning, persistent practice, seeking guidance, and believing in your potential.

We used the early stop method to choose appropriate iterations. Stopping the training of the network before the validation loss increased effectively prevented the model from overfitting. For example, when the model had more than 9000 training iterations, the validation loss of the model started to rise, which means the phenomenon of overfitting occurred. A similar approach was also used for the WenZhong model.

#### 4.9. Model Evaluation

In this section, we assess the performance and effectiveness of our proposed language model for online psychological consultation. We employed a combination of intrinsic and human evaluation metrics to evaluate the model's capabilities comprehensively. We began by utilising perplexity, ROUGE-L, and Distinct-n metrics to measure the model's language generation quality, similarity to the reference text, and diversity. Additionally, we recognised the limitations of these metrics and emphasised the importance of human evaluation in providing subjective assessments of the model's outputs, considering factors such as coherence, relevance, and overall quality. Through this comprehensive evaluation approach, we aimed to gain a comprehensive understanding of our model's strengths, weaknesses, and suitability for its intended purpose in the context of online psychological consultation.

##### Metric-Based Evaluation

Perplexity is a widely used intrinsic evaluation metric that measures how well a language model predicts a given sample. Mathematically, perplexity is defined as the reciprocal of the average probability assigned to each token in the dataset by the language model [51]. In simpler terms, a lower perplexity value indicates a better language model performance. As perplexity is based on the average log-likelihood of the dataset, it can be computed quickly and is statistically robust, as it is not easily affected by outliers.

The formula for calculating perplexity is given by

$$PP(W) = \mathbb{P}(w_1 w_2 \dots w_N)^{-1/N} \quad (1)$$

$$= \sqrt[N]{\prod_{i=1}^N \frac{1}{\mathbb{P}(w_i)}} \quad (2)$$

where  $PP$  is the perplexity,  $\mathbb{P}$  is the probability of the  $i$ th word, and  $N$  is the length of a sentence. It is important to note that perplexity tends to decrease as the dataset size increases, indicating a better performance.

However, it is crucial to understand that low perplexity does not necessarily equate to high accuracy. Perplexity is primarily used as a preliminary measure and should not be solely relied upon for evaluating model accuracy. Additionally, comparing the performance of models on different datasets with varying word distributions can be challenging [51]. Therefore, while perplexity provides valuable insights into model performance, it should be complemented with other evaluation metrics and considerations when assessing model accuracy.

ROUGE-L (longest common subsequence) is an evaluation metric that measures the number of overlapping units between the predicted text generated by a language model and the actual reference text [52]. ROUGE-L measures how closely the generated text matches the desired output by quantifying the similarity between the predicted and reference texts.

Distinct-1 and Distinct-2 are evaluation metrics that assess the diversity of the generated text. Distinct-1 calculates the number of distinct unigrams (individual words) divided by the total number of generated words. In contrast, Distinct-2 calculates the number of distinct bigrams (pairs of adjacent words) divided by the total number of generated bigrams [53]. These metrics reflect the degree of diversity in the generated text by quantifying the presence of unique unigrams and bigrams.



The formulas for calculating Distinct-n are as follows:

$$\text{Distinct-n} := \text{Distinct}(n) = \frac{\text{Count}(\text{unique}, n\text{-gram})}{\text{Count}(\text{word})} \quad (3)$$

Here,  $\text{Count}(\text{unique}, n\text{-gram})$  represents the number of  $n$ -grams that are not repeated in a reply and  $\text{Count}(\text{word})$  indicates the total number of  $n$ -gram words in the reply. A higher value of  $\text{Distinct}(n)$  indicates a greater diversity in the distinct generations.

These evaluation metrics, including perplexity, ROUGE-L, Distinct-1, and Distinct-2, provide insights into the quality, similarity, and diversity of the generated text by the language model. They serve as valuable tools for assessing the performance and effectiveness of the model in generating accurate and diverse outputs.

While perplexity and Distinct-n provide insights into the language model's performance in language generation, they do not necessarily indicate high accuracy. Therefore, in order to evaluate models more convincingly, human evaluation is still necessary. Human evaluators can provide subjective assessments of the generated text, considering factors such as coherence, relevance, and overall quality, which are important aspects that cannot be fully captured by automated evaluation metrics alone.

### Human Evaluation

For human evaluation, we developed an online marking system to assess the performance of our language model in the context of online psychological consultation. This evaluation system aims to streamline the process and ensure effective assessment by focusing on four key metrics: *Helpfulness*, *Fluency*, *Relevance*, and *Logic*. Each metric is scored on a scale of 1 to 5, allowing evaluators to provide a quantitative assessment of each aspect. The four metrics are defined as follows:

1. *Helpfulness*: This metric evaluates whether the generated response is helpful for patients seeking psychological support.
2. *Fluency*: Fluency refers to the degree of coherence and naturalness exhibited in the generated response.
3. *Relevance*: Relevance assesses the extent to which the response's content directly relates to the posed question.
4. *Logic*: Logic examines the logical consistency and coherence of the meaning conveyed in the generated response.

To conduct the human evaluation, we invited six students from the psychological faculty to assess a set of 200 question-answer pairs generated by our model. We employed two evaluation methods to understand the model's performance comprehensively.

In the first method, evaluators compared responses generated by the PanGu model and the WenZhong model in response to the same question. They assigned scores to these answers based on the predetermined metrics, allowing for a direct comparison between the two models. The second method involved incorporating the actual answers alongside the predicted responses as a whole, allowing evaluators to assess the differences and similarities between the generated and actual responses.

By employing these human evaluation methods, we aimed to gain valuable insights into the performance of our language model, particularly in terms of the disparities between the predicted and actual responses. This comprehensive evaluation approach will provide a deeper understanding of the model's capabilities and guide further improvements in its performance for online psychological consultation.

## 5. Experimental Results

In this section, we present the findings and outcomes of the evaluation and experimentation conducted to assess the performance and effectiveness of our proposed language model for online psychological consultation. This section provides a comprehensive analysis of the model's performance based on intrinsic and human evaluation metrics. We

discuss the results obtained from metrics such as perplexity, ROUGE-L, and Distinct-n, which shed light on language generation quality, similarity to reference text, and diversity of the generated responses. Additionally, we present the outcomes of the human evaluation, which includes scores given by evaluators based on metrics such as *Helpfulness*, *Fluency*, *Relevance*, and *Logic*. Through these rigorous evaluations, we aim to provide an in-depth understanding of the strengths and weaknesses of our language model and its suitability for the task of online psychological consultation.

### 5.1. Result of Intrinsic Evaluation

The results of the intrinsic evaluation comparing the performance of the PanGu model and the WenZhong model are presented in Table 5. The metrics used for evaluation included perplexity, ROUGE-L, Distinct-1, and Distinct-2.

As shown in Table 5, the PanGu model outperformed the WenZhong model across all metrics. The PanGu model achieved a lower perplexity value of 34.56 compared with 38.40 for the WenZhong model, indicating that the PanGu model better predicted the sample probabilities in the dataset.

Furthermore, the ROUGE-L score, which measures the similarity between the generated responses and the reference text, was higher for the PanGu model (28.18) than the WenZhong model (23.56). This suggests that the PanGu model generated responses more aligned with the expected answers.

In terms of diversity in generated responses, the PanGu model also exhibited higher Distinct-1 (4.57) and Distinct-2 (12.74) scores compared with the WenZhong model (3.55 and 9.67, respectively). This indicates that the PanGu model produced a wider range of distinct unigrams and bigrams in its responses, leading to more diverse language generation.

The intrinsic evaluation results consistently demonstrate that the PanGu model performed better than the WenZhong model across all metrics. These findings highlight the superior language generation capabilities of the PanGu model in the context of online psychological consultation.

**Table 5.** The result of the intrinsic evaluation between the two models.

Model	Perplexity	Rouge-L	Distinct1	Distinct2
WenZhong	38.40	23.56	3.55	9.67
PanGu	34.56	28.18	4.57	12.74

### 5.2. Case Study: Examining Psy-LLM's Responses on Coping with Depression

In the following, we qualitatively evaluate the outputs of both models by asking our Psy-LLM a common question for individuals suffering from depression.

#### 5.2.1. PanGu Model

To demonstrate the capabilities of the PanGu model, we provide an example of question answering using the model on coping with depression (Figure 6), which we have translated into English. The question posed to the model is,

*“Question: How to deal with depression?”*

and the model's response is as follows:

*“Answer: Hello, first let me give you a small hug. I hope it brings you some warmth. Typically, individuals with depression are often misunderstood and even scolded by their families, which can be really disheartening. I understand that you are feeling very sad and disappointed with your family situation. I want you to know that you are a remarkable person, and you are not at fault. Remember, change starts with you, and you are incredibly strong and sensible”.*

This response showcases the model's fluency and ability to employ different strategies. Analysing the response in Figure 6, we can observe that the model utilised various

strategies. The initial part, offering a virtual hug, demonstrates a reassurance strategy [13]. It provides emotional support and encouragement to individuals seeking consultation, helping them relax. The second part, addressing family dynamics, employs a restatement strategy by clearly reiterating the question to provide a more realistic understanding of the individual's mental health condition. The final part utilises a guidance strategy, advising the individual on steps to improve their situation. Overall, the model's response is coherent and consistent with normal human logic, making it highly valuable.

The training of the PanGu model consisted of two steps. Initially, the model was trained with a vast amount of crawled data, enabling it to learn extensive domain knowledge in psychology. Subsequently, the model underwent fine-tuning using the downstream PsyQA dataset. During this stage, the model primarily focussed on learning the pattern of "Question:" and "Answer:" to generate responses in the same format as the downstream tasks. When we input a question in the format "Question: . . ." to the model, the model output a response in the format "Answer: . . .". This demonstrates that the model successfully learned the desired pattern and could generate appropriate responses.

```

successfully loaded /dataset/model_all_psy/model_all_psy/iter_0100000/mp_rank_00/model_optim_rng.pt
Building prefix dict from the default dictionary ...
Loading model from cache /tmp/jieba.cache
Loading model cost 0.600 seconds.
Prefix dict has been built successfully.
Input is: 问题: 抑郁症怎么办?
Output is:
答案: 楼主你好, 首先给你一个小小的拥抱。希望可以给你带来一些温暖。抑郁症的患者基本都是得不到家人的理解, 自己也没有得到家人的理解, 甚至还被他们骂, 这真是会让人感到难过又无助。我知道你现在很难过, 你对你的家庭状况十分失望, 也很心疼你。但是我想告诉你, 你很棒, 你没有错, 你的所有改变都是你自己开始的, 你是一个很坚强的人。你也是一个很懂事的人,

```

**Figure 6.** A demonstration of the PanGu model's Q&A capability. Translation: Question: How to deal with depression? Answer: Hello, to begin with, allow me to offer you a gentle embrace, hoping it provides some comfort. Individuals experiencing depression often face a lack of understanding from their families, which is disheartening when they feel misunderstood or criticised. I understand you're currently feeling profoundly sad, and it's causing you deep pain. Nevertheless, I want you to know that you're amazing, you're not at fault, and the journey towards any changes you desire begins with your own inner strength; you're resilient and perceptive.

### 5.2.2. WenZhong Model

In WenZhong model, after fine-tuning the model using psychology data, the model could output very suitable answer. We also utilised the WenZhong model to answer the same question (Figure 7), and the generated response was as follows:

```

抑郁症怎么办?? 答: 我们的经验告诉我们, 有很多时候只要靠自己的意愿去做事, 那么就会有许多的不适应。比如说你在学习上, 每天的学习时间, 学习地点, 学习态度, 学习能力等等, 这些都是要慎重对待的, 尤其是这些都要尽可能的避免, 比如说, 那么如果自己不想去做一些消极的事情, 就不要管它了, 尽可能的去做一些抑郁症的预防措施, 以免会导致自己患上这种疾病。生活中我们也会看到一些患有

```

**Figure 7.** A demonstration of the WenZhong model's Q&A capability. Translation: Question: How to deal with depression? Answer: Our experiences often tell us that many times, if someone simply acts on their own emotions and feelings, there will be many unexpected consequences. For instance, in your studies, the time spent studying daily, the study environment, your attitude towards learning, learning capabilities, and so on—these aspects require careful consideration. It's especially crucial to try to avoid certain things as much as possible. For example, if you don't wish to engage in negative activities, then it's better to refrain from them. Instead, focus on taking preventive measures against depression to prevent oneself from developing such an illness. In life, we also encounter individuals who have been deceived.

*"Our experience tells us that there are many times when you just have to do things by your own will, then there will be a lot of discomfort. For example, if you are studying, the time of day, the place of study, the attitude of study, the ability to study, etc., these must be treated carefully, especially these should be avoided as much as possible, for example, then if you do not want to do something negative, leave it alone and do as much as possible to do some preventive measures for depression, so that it will not lead to the development of this disease".*

This example demonstrates that the WenZhong model could provide suitable answers in the field of psychology. However, one challenge we encountered is that some generated outputs did not directly address the question. This issue may be attributed to the limited fine-tuning of data specific to psychology. In order to further improve the performance of the WenZhong model in psychology-related tasks, a larger and more diverse dataset from the field of psychology could be incorporated during the fine-tuning process.

### 5.3. Human Evaluation

To conduct an empirical evaluation of Psy-LLM's effectiveness, we enlisted the participation of six students from the psychological faculty to assess a set of 200 question-answer pairs generated by our language model. In order to obtain a comprehensive understanding of the model's performance, we employed two evaluation methods for the participants to provide ratings on the responses. We created a web front-end for users to access our Psy-LLM platform, and their technical details are discussed in Section 6.

The first method directly compared responses generated by both the PanGu and the WenZhong models in response to the same question. Evaluators assigned scores to these answers based on predetermined metrics, enabling a clear and direct comparison between the performances of the two models. In the second method, we presented evaluators with a combined set of predicted and actual responses. This allowed them to evaluate and assess the differences and similarities between the generated responses and the ground truth answers.

By utilising these human evaluation methods, we aimed to gain valuable insights into the performance of our language model, particularly in terms of the disparities between predicted and actual responses. This comprehensive evaluation approach will provide a deeper understanding of the model's capabilities and guide further improvements in its performance for online psychological consultation.

The human evaluation results, using two different methods, are presented in Tables 6 and 7. These evaluate human-perceived metrics of *Helpfulness*, *Fluency*, *Relevance*, and *Logic*. Table 6 shows the results of the first human evaluation method, where evaluators provided scores for each metric. Consistent with the findings from the intrinsic evaluation, the PanGu model outperformed the WenZhong model in terms of *Helpfulness* (3.87 vs. 3.56), *Fluency* (4.36 vs. 4.14), *Relevance* (4.09 vs. 3.87), and *Logic* (3.83 vs. 3.63). These results indicate that human evaluators generally considered the PanGu model's generated responses more helpful, fluent, relevant, and logical than the WenZhong model.

However, a notable observation is made when comparing the scores obtained in Table 6 with the scores from Table 7. Table 7 presents the scores for the predicted answers of both models, as well as the actual answers. Interestingly, the scores for the actual answers were significantly higher than those for the predicted answers of both models across all metrics. This discrepancy suggests that the evaluators, who had the opportunity to compare the actual answers with the predicted answers, marked the predicted answers relatively lower. This finding highlights the importance of incorporating human evaluation when assessing the performance of language models and the need for further improvement in generating more accurate and satisfactory responses.

In summary, the human evaluation results align with the intrinsic evaluation findings, indicating that the PanGu model performs better than the WenZhong model. However, it is important to note that the scores for the actual answers are considerably higher than those for the predicted answers, implying room for improvement in the generated responses of the language models.

**Table 6.** Average human ratings of Psy-LLM responses, with only the two AI-powered versions.

Metrics	WenZhong	PanGU
Helpfulness	3.56	3.87
Fluency	4.14	4.36
Relevance	3.87	4.09
Logic	3.63	3.83

**Table 7.** Average human ratings of Psy-LLM responses, alongside ground truths from the datasets.

Rating Metrics	WenZhong	PanGU	Ground Truth
Helpfulness	3.45	3.54	4.52
Fluency	3.95	4.12	4.83
Relevance	3.77	3.96	4.72
Logic	3.61	3.75	4.56

## 6. Web Interface for Accessible Online Consultation

One of the primary objectives was to explore the provision of online AI-powered consultation and Q&A services in psychology. We adopted a distributed architecture, separating the model's front-end, back-end, and computing servers into modular components. Each module was developed with distinct responsibilities, allowing for easier upgrades and interchangeability of combinations. Communication between the modules was achieved through API interactions, enabling them to function independently without relying on the internal functionality of other modules.

Furthermore, we placed a strong emphasis on security during the design process. We implemented measures to encrypt and protect our modular systems at a product level. The common API interface was productised and encrypted, ensuring secure communication between the components. Additionally, we implemented HTTPS web-system architecture, enhancing security by encrypting each cloud server with TLS (SSL). By adopting a distributed and modular approach and prioritising security, we aimed to address the challenges of hosting a large-scale online consultation service model. These design choices allowed for flexibility, scalability, and enhanced security in our system architecture, contributing to our project's overall success and reliability.

### Web Technologies

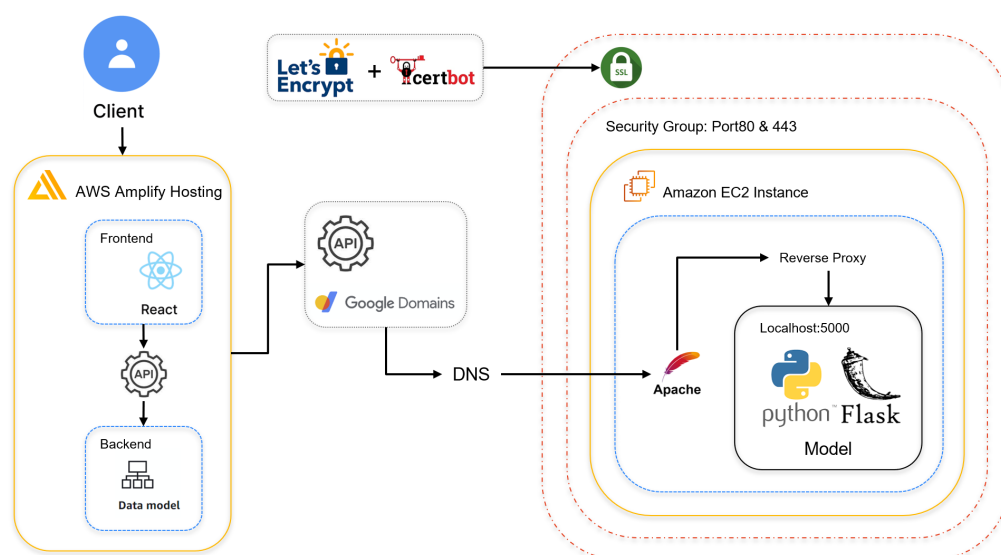
We utilised the following services and technologies for our website development:

- *ReactJS*: ReactJS was our front-end framework, because of its extensive library support. ReactJS offers a wide range of reusable components and follows a modular, component-based architecture, making designing and enhancing the front-end easier. ReactJS is responsive and provides excellent cross-platform support.
- *AWS Amplify*: AWS Amplify is a rapid front-end deployment service provided by Amazon. It enable us to quickly deploy the front-end of our website and seamlessly communicate with other system components. Amplify provides fully managed CI/CD (continuous integration/continuous deployment) and hosting, ensuring fast, secure, and reliable encryption services.
- *Google Domain*: We utilised Google Domain services for secure encapsulation of our EC2 host DNS.
- *Amazon EC2*: EC2 provides virtual server instances with highly available underlying designs. It offers reliable, scalable, and flexible access in terms of cost and performance. EC2 provides powerful computing resources and pre-configured environments, making it an excellent choice for running large models. Its robust network performance and high-performance computing clusters allow for high throughput and low-latency online processing. We used simple Flask-based scripts to handle concurrent requests.



- *Python and Flask*: We used Python as our scripting language to run the models and build APIs. Flask, a web framework written in Python, was used for creating API endpoints and handling request–response interactions.
- *Apache*: We used Apache, an open-source web server software, for configuring port forwarding, reverse proxies, and listening.
- *Let's Encrypt and Certbot*: We employed Let's Encrypt and Certbot for TLS (transport layer security) encryption, ensuring secure communication between the website and users.

The diagram in Figure 8 illustrates the architecture of our independently developed web system for the cloud-based site. The website's user interface is accessible through the front-end, deployed on the AWS Amplify service. Built on the ReactJS framework, the front-end communicates with the back-end database through an internal API, enabling storage of the user evaluation data for model effectiveness optimisation. The database is hosted within Amplify Hosting. The standalone website interacts with the model runtime server via a public API. To ensure privacy and protect the host address, we registered a public domain name through Google Domains and link it to the host server's DNS.



**Figure 8.** Online web front-end architecture.

The pre-trained model is deployed on an Amazon EC2 instance host configured as an AWS Linux virtual server. The model uses Python code and Flask scripts, allowing for local server calls. Apache is used for HTTP reverse proxy communication, forwarding the external model input data to the local server where the model is waiting and generating results. To provide secure HTTPS encryption for the web products deployed on AWS Amplify, we employed TSL encryption for the EC2 instance DNS addresses. This was achieved using Let's Encrypt and Certbot as cost-effective alternatives to commercial SSL certificates.

The website's front-end was designed with simplicity, featuring an input box for users to enter Chinese questions, as depicted in Figure 9. Upon submission, the system communicates with the back-end model through the API. It awaits the completion of model processing (Figure 10a) before returning the results to the output box, as depicted in Figure 10b. Users can rate the results using the built-in rating system (Available at <https://www.wjx.cn/vm/OJMsMXn.aspx> (accessed on 12 July 2023)), and there is a link to an additional evaluation site at the bottom of the page.

# A Chinese Psychological Consultation Dialogue System Based on Large Corpus Pre-training Model

Please enter your question in Chinese and end with a question mark

Submit

output

Poor+

Please take a few moments to comment on our answers!

**Figure 9.** Website initial status.

# A Chinese Psychological Consultation Dialogue System Based on Large Corpus Pre-training Model

每次和对象呆在一起就没什么话说？

Submit

Loading... This will usually take one minute

output

Poor+

Please take a few moments to comment on our answers!

每次和对象呆在一起就没什么话说？

Submit

每次和对象呆在一起就没什么话说？可是，在这样一个充满情感的社会，哪有不同性格的男人和女人联系在一起？只是，如果看一个人，能够从他身上，知道他的生活习惯，他的家庭环境，和他的生活习惯，就能够真正的知道他的内心世界，而这个人只说他自己，他也不知道自己，那么，有一个女人，她有一个可爱的妹妹，喜欢和他在一起，她可以不向她的生活，她也可以在每次下班回家的时间，对他哭，哭得很温柔，而她的妹妹，可以很平静，很温柔，给予她很多，给予她爱。在这样一个和谐的社会，如果

Excellent+

Please take a few moments to comment on our answers!

**Figure 10.** Web front-end for the online psychological consultation. (a) Website loading status. (b) Website result status. The user is using the interface to answer the question of: What should be done when there's a lack of conversation every time spent with the partner?

## 7. Discussion

The discussion section provides a comprehensive analysis of the project outcome, product perspective, website perspective, model perspective, and evaluation perspective.

### 7.1. Project Outcome

We have successfully developed and implemented an effective chatbot for mental health counselling. Through training and fine-tuning large-scale Chinese pre-training models on mental health datasets, the chatbot has acquired valuable knowledge of psychology, enhancing its ability to provide counselling services. The deployment of the chatbot on a website interface has created a convenient and accessible platform for users seeking mental health support. Although the chatbot is currently in its prototype stage, our project demonstrates the feasibility of building an AI-based counselling system. It is a valuable reference for future research and development in this area.

From a model perspective, our evaluation results demonstrate the superiority of the PanGu model over the WenZhong model, as expected because of its larger size and advanced architecture. The PanGu model's design contributes to its outperformance, particularly its incremental learning ability and enhanced natural language understanding capabilities. However, both models fall short of achieving human-level performance, which can be attributed to the quality of the training dataset and the inherent limitations of autoregressive language models. Enhancing the dataset quality and exploring alternative

language model architectures hold promise for addressing these limitations and further improving performance.

### *7.2. Evaluation Perspective*

The human evaluation results indicate that both the PanGu model and the WenZhong model have yet to reach human-level performance. Despite training the models on our dataset crawled from websites, the predicted answers strongly focussed on psychological content but lacked logical coherence. One potential reason for this is the quality of our dataset, which may need to be higher to provide comprehensive and reliable training examples. Although we conducted human evaluation during the data cleaning stage, the sheer volume of data made it challenging to cover every instance. To address this, we recommend thoroughly evaluating the website data before crawling to ensure a higher-quality dataset.

Another factor impacting human evaluation is the limited computing conditions during model training. Our model requires a specific training environment and numerous parameters, making it time consuming to adjust and fine-tune it effectively. We could not optimise the parameters and achieve optimal testing results with our current resources. Consequently, the model's performance may have been hindered by these limitations. Furthermore, the autoregressive nature of both the PanGu model and the WenZhong model poses challenges in comprehending contextual information. As autoregressive language models, their training processes are unidirectional, focused on modelling the joint probability from left to right. The next predicted word is solely based on the preceding word, limiting their ability to capture information from broader contexts. This lack of a contextual background reference makes it difficult for language models to handle reading comprehension tasks in the same way as humans.

In summary, the evaluation results shed light on the areas where improvements can be achieved. Enhancing the dataset quality through pre-evaluation and addressing the limitations of our computing conditions are crucial steps toward advancing the model's performance. Additionally, exploring alternative language model architectures that can effectively capture contextual information may contribute to bridging the gap between model-generated responses and human-level performance.

### *7.3. Product and Practicality Perspective*

The performance of the online consultation service indicates its significant potential for streamlining mental health support with minimal resources. The user experience has been a priority in product design, and cloud infrastructure deployment ensures easy access via mobile devices. As part of our future improvements, we plan to incorporate an automatic emotion recognition system into the website, enabling the identification of users in distress and facilitating timely intervention. The design and development of our product hold substantial societal value in the mental health support field, providing a promising avenue for further exploration and refinement.

Regarding the website, we have designed and implemented a modern, cloud-based network architecture that boasts lightweight, scalable, and highly secure features. This architecture allows for low-cost, large-scale model computing sites, enabling widespread accessibility to AI-based Q&A services. Our approach serves as a reference for small organisations and enterprises with limited resources, showcasing the possibilities of deploying AI capabilities effectively.

## **8. Limitations and Future Works**

While we have presented promising results with our Psy-LLM model for usage in assisting mental health workers, our study is exploratory in nature, and, hence, numerous limitations exist that we would like to raise, as follows.

### 8.1. Model Capability and Usage in Real-World

While there are numerous benefits in deploying an AI-powered large language model for supporting the demand in the mental health sector, several ethical and practical issues need to be considered. Firstly, as a language-based model, the model's output is based purely on the input text. However, studies have shown that nonverbal communication is one of the key factors in counselling outcomes [54]. In fact, a well-trained counsellor can often pick up subtle cues, even when there is a lack of response from the patient. Standalone LLM models like Psy-LLM cannot address such an issue (unless techniques like facial emotion detection from the computer-vision community are integrated as a unified system [55]). Furthermore, rapport building with clients is often a crucial step in clinical psychology. However, an AI-based model would face severe difficulties in building trusted client relationships. As a result, it is critical to realise that such an AI-powered system cannot replace real-world counselling setups. A practical approach would be to pair the model output under the supervision of a trained counsellor as a good psychoeducational tool. The model output can be used as an initial guideline or suggestion for assisting human counsellors in providing useful and trusted consultations with patients.

### 8.2. Data Collections

Several strategies can be implemented in future work to overcome the limitations in data collection. Firstly, to address the issue of anti-crawler rules on different websites, developing a more robust and adaptable crawler that can handle different anti-crawler mechanisms would be beneficial. The access limitation could involve implementing dynamic IP rotation or utilising proxies to avoid IP blocking. Machine-learning techniques, such as automatic rule extraction or rule adaptation, could also help automate handling anti-crawler mechanisms.

Incorporating more advanced data-cleaning techniques can also improve the quality of the crawled data. More advanced data-cleaning procedures may involve NLP methods, such as entity recognition, part-of-speech tagging, and named entity recognition to identify and filter out irrelevant or noisy data. Additionally, leveraging machine-learning algorithms, such as anomaly detection or outlier detection, can aid in identifying and removing low-quality or erroneous data points. In terms of dataset standardisation, establishing a unified standard for data generation in the online domain would greatly facilitate the cleaning process. This could involve collaborating with website administrators or data providers to develop guidelines or formats for data representation. Furthermore, using human annotators or experts in the domain to manually review and clean a subset of the dataset can provide valuable insights and ensure a higher-quality dataset.

However, it is important to acknowledge that achieving a completely clean dataset is challenging, particularly when dealing with large-scale datasets. As such, future work should strike a balance between the manual review and automated cleaning techniques, while also considering the cost and scalability of the data cleaning process.

### 8.3. Model Improvement

Increasing the scale of model training by utilising larger models or ensembles of models can enhance the performance and capabilities of the chatbot. Larger models can capture more nuanced patterns and relationships in the data, leading to more accurate and coherent responses. Exploring different model architectures beyond autoregressive language models may provide valuable insights. Bidirectional models (e.g., Transformer-XL) or models that incorporate external knowledge sources (e.g., knowledge graphs) can improve the chatbot's contextual understanding and generate more informative responses. Moreover, integrating feedback mechanisms into the training process can help iteratively improve the chatbot's performance. This could involve collecting user feedback on the generated responses and incorporating it into the model training through reinforcement learning or active learning.

Several disadvantages were also identified in the LLM architecture. Firstly, the maximum likelihood training approach of the WenZhong model is susceptible to exposure bias, which occurs when samples are drawn from the target language distribution. This bias can lead to errors for which researchers have yet to find effective solutions. Additionally, training the WenZhong model multiple times can significantly decrease its quality. Furthermore, the WenZhong model follows an autoregressive architecture, which models joint probability from left to right. This unidirectional training process limits its ability to capture information from all contexts, particularly hindering its performance in tasks requiring reading comprehension that rely on contextual background references. Similar to the WenZhong model, the PanGu model also exhibits autoregressive characteristics. Although it inherits the ability to estimate the joint probability of language models, it suffers from the same limitations of unidirectional modelling. It lacks bidirectional context information and may produce duplicate results requiring resolve deduplication.

We also have reservations about the Jieba tokeniser used in the PanGu model. Its performance and tokenisation ability need to handle complex Chinese tokenisation accurately. Furthermore, as neural networks and pre-trained models advance, Chinese NLP tasks increasingly demonstrate that tokenisation is only sometimes necessary. Large models can effectively learn character-to-character relationships without word segmentation. For instance, Google is considering discarding tokenisation and using bytes directly. Adopting a more flexible tokeniser could ensure the model is more suitable for various industrial applications, despite sacrificing some performance.

#### 8.4. User Experience and User Interface

Enhancing the chatbot's user experience and user interface can significantly impact its adoption and effectiveness. Future work should focus on improving the simplicity, intuitiveness, and accessibility of the website interface. This includes optimising response times, refining the layout and design, and incorporating user-friendly features such as autocomplete suggestions or natural language understanding capabilities.

Furthermore, personalised recommendations and suggestions to users based on their preferences and previous interactions can enhance the user experience. Techniques like collaborative filtering or user profiling can enable the chatbot to better understand and cater to individual user needs. Usability testing and user feedback collection should be conducted regularly to gather insights on user preferences, pain points, and suggestions for improvement. Iterative design and development based on user-centred principles can ensure that the chatbot meets user expectations and effectively addresses their mental health support needs.

#### 8.5. Ethical Considerations and User Privacy

As with any AI-based system, ethical considerations and user privacy are paramount. Future work should address these concerns by implementing robust privacy protection mechanisms and ensuring transparency in data usage. This includes obtaining explicit user consent for data collection and usage, anonymising sensitive user information, and implementing strict data access controls. Developing mechanisms to handle potentially sensitive or harmful user queries is crucial. The chatbot should have appropriate safeguards and guidelines to avoid providing inaccurate or harmful advice. Integrating a reporting system where users can report problematic responses or seek human intervention can help mitigate potential risks. Furthermore, monitoring and auditing the chatbot's performance and behaviour can help identify and rectify biases or discriminatory patterns. Regular evaluations by domain experts and user feedback analysis can improve the chatbot's reliability, fairness, and inclusivity.

While this project has made significant progress in developing an AI-based chatbot for mental health support, there are various limitations and areas for improvement. Overcoming challenges related to data quality, model performance, ethical considerations, and user experience will contribute to the overall effectiveness and reliability of the chatbot. By ad-



addressing these limitations and exploring future research directions, we can continue to advance the field of AI-powered mental health support systems and provide valuable assistance to individuals in need.

## 9. Conclusions

In conclusion, our project on Psy-LLM, an exploratory study on using large language models as an assistive mental health tool, has been successfully completed and implemented. While there are areas identified for improvement based on specific evaluation indicators, we are confident that with improved equipment conditions, we can enhance the performance of this platform. The experimental results obtained from this project hold significant potential to contribute to the fields of supportive natural language generation and psychology, driving advancements at the intersection of these domains. The deployment of such a system offers a practical approach to promoting the overall mental well-being of our society by providing timely responses and support to those in need.

**Author Contributions:** Conceptualization, T.L.; Methodology, T.L., Z.D., J.W., K.F., Y.D. and Z.W.; Software, Y.S. and Z.W.; Investigation, Y.S. and J.W.; Data curation, Y.S., Z.D., K.F. and Y.D.; Writing—original draft, Y.S., Z.D., J.W., K.F., Y.D. and Z.W.; Writing—review & editing, T.L.; Supervision, T.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** Data available in a publicly accessible repository at [13].

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

- Handoyo, E.; Arfan, M.; Soetrisno, Y.A.A.; Somantri, M.; Sofwan, A.; Sinuraya, E.W. Ticketing chatbot service using serverless NLP technology. In Proceedings of the 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE), Semarang, Indonesia, 27–28 September 2018; pp. 325–330.
- Trivedi, J.K.; Sareen, H.; Dhyani, M. Rapid urbanization-Its impact on mental health: A South Asian perspective. *Indian J. Psychiatry* **2008**, *50*, 161. [[CrossRef](#)] [[PubMed](#)]
- Tian, F.; Li, H.; Tian, S.; Yang, J.; Shao, J.; Tian, C. Psychological symptoms of ordinary Chinese citizens based on SCL-90 during the level I emergency response to COVID-19. *Psychiatry Res.* **2020**, *288*, 112992. [[CrossRef](#)]
- Chen, T.; Lucock, M. The mental health of university students during the COVID-19 pandemic: An online survey in the UK. *PLoS ONE* **2022**, *17*, e0262562. [[CrossRef](#)] [[PubMed](#)]
- Gay, J.L.; Swank, J.M. School counselor educators' experiences navigating practicum and internship during COVID-19. *Couns. Educ. Superv.* **2021**, *60*, 291–305. [[CrossRef](#)]
- Joshi, G.; Sharma, G. Burnout: A risk factor amongst mental health professionals during COVID-19. *Asian J. Psychiatry* **2020**, *54*, 102300. [[CrossRef](#)]
- Kontoangelos, K.; Economou, M.; Papageorgiou, C. Mental health effects of COVID-19 pandemic: A review of clinical and psychological traits. *Psychiatry Investig.* **2020**, *17*, 491. [[CrossRef](#)] [[PubMed](#)]
- Yurayat, P.; Tuklang, S. University Student Counselees' Attitudes and Experiences towards Online Counseling during the COVID-19 Pandemic: A Mixed Methods Study. *J. High. Educ. Theory Pract.* **2023**, *23*, 222–238.
- Prochaska, J.J.; Vogel, E.A.; Chieng, A.; Kendra, M.; Baiocchi, M.; Pajarito, S.; Robinson, A. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): Development and Usability Study. *J. Med. Internet Res.* **2021**, *23*, e24850. [[CrossRef](#)]
- Deryugina, O. Chatterbots. *Sci. Tech. Inf. Process.* **2010**, *37*, 143–147. [[CrossRef](#)]
- Zeng, W.; Ren, X.; Su, T.; Wang, H.; Liao, Y.; Wang, Z.; Jiang, X.; Yang, Z.; Wang, K.; Zhang, X.; et al. PanGu- $\alpha$ : Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation. *arXiv* **2021**, arXiv:2104.12369.
- Wang, J.; Zhang, Y.; Zhang, L.; Yang, P.; Gao, X.; Wu, Z.; Dong, X.; He, J.; Zhuo, J.; Yang, Q.; et al. Fengshenbang 1.0: Being the Foundation of Chinese Cognitive Intelligence. *arXiv* **2022**, arXiv:2209.02970.
- Sun, H.; Lin, Z.; Zheng, C.; Liu, S.; Huang, M. PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online Event, 1–6 August 2021; Association for Computational Linguistics: Toronto, ON, Canada, 2021; pp. 1489–1503. [[CrossRef](#)]

14. Regona, M.; Yigitcanlar, T.; Xia, B.; Li, R.Y.M. Opportunities and adoption challenges of AI in the construction industry: A PRISMA review. *J. Open Innov. Technol. Mark. Complex.* **2022**, *8*, 45. [\[CrossRef\]](#)
15. Lai, T. A Review on Visual-SLAM: Advancements from Geometric Modelling to Learning-Based Semantic Scene Understanding using Multi-Modal Sensor Fusion. *Sensors* **2022**, *22*, 7265. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Link, M.; Dukino, C.; Ganz, W.; Hamann, K.; Schnalzer, K. The Use of AI-Based Assistance Systems in the Service Sector: Opportunities, challenges and applications. In *Advances in Human Factors and Systems Interaction, Proceedings of the AHFE 2020 Virtual Conference on Human Factors and Systems Interaction, San Diego, CA, USA, 16–20 July 2020*; Springer: Cham, Switzerland, 2020; pp. 10–16.
17. Wang, X.; Zhang, H.; Zhang, Y.; Wang, M.; Song, J.; Lai, T.; Khushi, M. Learning Non-Stationary Time-Series with Dynamic Pattern Extractions. *IEEE Trans. Artif. Intell.* **2022**, *3*, 778–787. [\[CrossRef\]](#)
18. Koroteev, D.; Tekic, Z. Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future. *Energy AI* **2021**, *3*, 100041. [\[CrossRef\]](#)
19. Lai, T. Discover Life Skills for Planning as Bandits via Observing and Learning How the World Works. In *Proceedings of the IEEE/RISJ International Conference on Intelligent Robots and Systems (IROS)*, Kyoto, Japan, 23–27 October 2022; pp. 11360–11365. [\[CrossRef\]](#)
20. Xu, X.; Lai, T.; Jahan, S.; Farid, F.; Bello, A. A Machine Learning Predictive Model to Detect Water Quality and Pollution. *Future Internet* **2022**, *14*, 324. [\[CrossRef\]](#)
21. Lai, T. Real-Time Aerial Detection and Reasoning on Embedded-UAVs in Rural Environments. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–7. <https://doi.org/10.1109/TGRS.2023.3266360>.
22. Li, J.; Lim, K.; Yang, H.; Ren, Z.; Raghavan, S.; Chen, P.Y.; Buonassisi, T.; Wang, X. AI applications through the whole life cycle of material discovery. *Matter* **2020**, *3*, 393–432. [\[CrossRef\]](#)
23. Castiglioni, I.; Rundo, L.; Codari, M.; Di Leo, G.; Salvatore, C.; Interlenghi, M.; Gallivanone, F.; Cozzi, A.; D’Amico, N.C.; Sardanelli, F. AI applications to medical images: From machine learning to deep learning. *Phys. Med.* **2021**, *83*, 9–24. [\[CrossRef\]](#)
24. Chugh, G.; Kumar, S.; Singh, N. Survey on machine learning and deep learning applications in breast cancer diagnosis. *Cogn. Comput.* **2021**, *13*, 1451–1470. [\[CrossRef\]](#)
25. Hu, H.; Lai, T.; Farid, F. Feasibility Study of Constructing a Screening Tool for Adolescent Diabetes Detection Applying Machine Learning Methods. *Sensors* **2022**, *22*, 6155. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Brattain, L.J.; Telfer, B.A.; Dhyani, M.; Grajo, J.R.; Samir, A.E. Machine learning for medical ultrasound: Status, methods, and future opportunities. *Abdom. Radiol.* **2018**, *43*, 786–799. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Martinengo, L.; Lum, E.; Car, J. Evaluation of chatbot-delivered interventions for self-management of depression: Content analysis. *J. Affect. Disord.* **2022**, *319*, 598–607. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Denecke, K.; Abd-Alrazaq, A.; Househ, M. Artificial intelligence for chatbots in mental health: Opportunities and challenges. In *Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges*; Springer: Cham, Switzerland, 2021; pp. 115–128.
29. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
30. Wang, J.; Shi, E.; Yu, S.; Wu, Z.; Ma, C.; Dai, H.; Yang, Q.; Kang, Y.; Wu, J.; Hu, H.; et al. Prompt engineering for healthcare: Methodologies and applications. *arXiv* **2023**, arXiv:2304.14670.
31. Gonzalez-Hernandez, G.; Sarker, A.; O’Connor, K.; Savova, G. Capturing the patient’s perspective: A review of advances in natural language processing of health-related text. *Yearb. Med. Inform.* **2017**, *26*, 214–227. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Abd Rahman, R.; Omar, K.; Noah, S.A.M.; Danuri, M.S.N.M.; Al-Garadi, M.A. Application of machine learning methods in mental health detection: A systematic review. *IEEE Access* **2020**, *8*, 183952–183964. [\[CrossRef\]](#)
33. Ji, S.; Pan, S.; Li, X.; Cambria, E.; Long, G.; Huang, Z. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Trans. Comput. Soc. Syst.* **2020**, *8*, 214–226. [\[CrossRef\]](#)
34. Akstinaite, V.; Garrard, P.; Sadler-Smith, E. Identifying linguistic markers of CEO hubris: A machine learning approach. *Br. J. Manag.* **2022**, *33*, 1163–1178. [\[CrossRef\]](#)
35. De Choudhury, M.; Gamon, M.; Counts, S.; Horvitz, E. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, Cambridge, MA, USA, 8–11 July 2013; Volume 7, pp. 128–137.
36. Lui, J.H.; Marcus, D.K.; Barry, C.T. Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. *Prof. Psychol. Res. Pract.* **2017**, *48*, 199. [\[CrossRef\]](#)
37. Shaikh, T.A.H.; Mhetre, M. Autonomous AI Chat Bot Therapy for Patient with Insomnia. In *Proceedings of the 2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, Mumbai, India, 7–9 April 2022; pp. 1–5.
38. Tamrakar, M.K.; Badholia, A. Scientific Study of Technological Chatbot Adoption in Customer Service. In *Proceedings of the 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 17–19 August 2022; pp. 1117–1123. [\[CrossRef\]](#)
39. Kanodia, N.; Ahmed, K.; Miao, Y. Question Answering Model Based Conversational Chatbot using BERT Model and Google Dialogflow. In *Proceedings of the 2021 31st International Telecommunication Networks and Applications Conference (ITNAC)*, Sydney, Australia, 24–26 November 2021; pp. 19–22. [\[CrossRef\]](#)

40. World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*; Technical Report; World Health Organization: Geneva, Switzerland, 2017.
41. Gou, Y.; Wu, N.; Xia, J.; Liu, Y.; Yang, H.; Wang, H.; Yan, T.; Luo, D. Province-and individual-level influential factors of depression: Multilevel cross-provinces comparison in China. *Front. Public Health* **2022**, *10*, 893280. [[CrossRef](#)] [[PubMed](#)]
42. Xiang, Y.T.; Ng, C.H.; Yu, X.; Wang, G. Rethinking progress and challenges of mental health care in China. *World Psychiatry* **2018**, *17*, 231. [[CrossRef](#)] [[PubMed](#)]
43. Cataldo, F.; Chang, S.; Mendoza, A.; Buchanan, G. A Perspective on Client-Psychologist Relationships in Videoconferencing Psychotherapy: Literature Review. *JMIR Ment. Health* **2021**, *8*, e19004. [[CrossRef](#)] [[PubMed](#)]
44. Sandhu, D.S. An examination of the psychological needs of the international students: Implications for counselling and psychotherapy. *Int. J. Adv. Couns.* **1994**, *17*, 229–239. [[CrossRef](#)]
45. Novella, J.K.; Ng, K.M.; Samuolis, J. A comparison of online and in-person counseling outcomes using solution-focused brief therapy for college students with anxiety. *J. Am. Coll. Health* **2022**, *70*, 1161–1168. [[CrossRef](#)] [[PubMed](#)]
46. Hom, M.A.; Stanley, I.H.; Joiner, T.E., Jr. Evaluating factors and interventions that influence help-seeking and mental health service utilization among suicidal individuals: A review of the literature. *Clin. Psychol. Rev.* **2015**, *40*, 28–39. [[CrossRef](#)] [[PubMed](#)]
47. Xu, L.; Hu, H.; Zhang, X.; Li, L.; Cao, C.; Li, Y.; Xu, Y.; Sun, K.; Yu, D.; Yu, C.; et al. CLUE: A Chinese language understanding evaluation benchmark. *arXiv* **2020**, arXiv:2004.05986.
48. Thelwall, M.A. A web crawler design for data mining. *J. Inf. Sci.* **2001**, *27*, 319–325. [[CrossRef](#)]
49. Li, Y.; Han, P.; Liu, C.; Fang, B. Automatically Crawling Dynamic Web Applications via Proxy-Based JavaScript Injection and Runtime Analysis. In Proceedings of the 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, China, 18–21 June 2018; pp. 242–249.
50. Zhang, Y.; Wang, D.; Feng, S.; Zhang, Y.; Leng, F. An Approach for Crawling Dynamic WebPages Based on Script Language Analysis. In Proceedings of the 2012 Ninth Web Information Systems and Applications Conference, Haikou, China, 16–18 November 2012; pp. 35–38.
51. Chen, S.F.; Beeferman, D.; Rosenfeld, R. Evaluation Metrics For Language Models. 1998. Available online: <https://www.cs.cmu.edu/~roni/papers/eval-metrics-bntuw-9802.pdf> (accessed on 12 July 2023).
52. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 21–26 July 2004.
53. Li, J.; Galley, M.; Brockett, C.; Gao, J.; Dolan, W.B. A Diversity-Promoting Objective Function for Neural Conversation Models. In Proceedings of the North American Chapter of the Association for Computational Linguistics, San Diego, CA, USA, 12–17 June 2016.
54. Hill, C.E.; Siegelman, L.; Gronsky, B.R.; Sturniolo, F.; Fretz, B.R. Nonverbal communication and counseling outcome. *J. Couns. Psychol.* **1981**, *28*, 203. [[CrossRef](#)]
55. Jaiswal, A.; Raju, A.K.; Deb, S. Facial emotion detection using deep learning. In Proceedings of the 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 5–7 June 2020; pp. 1–5.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.