

FINAL REPORT

DEPARTMENT OF COMPUTING

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

A Multi-language Virtual Guide for Self-attachment Therapy

Author:
Ruoyu Hu

Supervisor:
Dr Anandha Gopalan
Prof Abbas Edalat

Second Marker:
Prof Lucia Specia

Submitted in partial fulfillment of the requirements for the MEng Degree in
Computing (Artificial Intelligence and Machine Learning) of Imperial College
London

Abstract

Recent advances in the use of transformers in facilitating empathetic dialogue systems have led to increasing applications of conversational agents in the field of psychotherapy interventions. However, much of existing work is only in English, and a lack of in-language, in-domain data has restricted the scalability of innovations in digital mental healthcare to other languages.

In this work, we investigate the feasibility of creating an effective virtual agent for the delivery of self-attachment therapy in simplified Chinese using existing in-domain English data with limited in-language data and minimal human translations. To this end, we propose a framework that could be replicated for different languages without the need for large-scale human translations whilst maintaining safety and reliability. We obtain a new state-of-the-art (SOTA) in monolingual Chinese emotion classification performance and provide quantitative evaluation of two generative models for use in our empathetic rewriting task.

We evaluate the effectiveness of our framework against the previous work in two non-clinical trials with a total $N = 15$ participants, where participants are required to interact with the platform for at least 30 minutes a day over the course of 5 days. We find that our platform is evaluated higher than previous works in this area in user engagement and perceived usefulness, and is scored comparatively in other metrics.

Lastly, we provide suggestions with the aim to guide future work by reflecting on the limitations of the framework identified in the non-clinical trials.

Acknowledgements

The completion of this project would not have been accomplished without the incredible support from my supervisors: Dr Anandha Gopalan and Professor Abbas Edalat, who gave me incredible creative freedom to explore. To them, I owe my sincerest gratitude. I would like to thank Dr Anandha Gopalan for his ceaseless enthusiasm and invaluable feedback throughout the entirety of the project, and whose encouragement allowed me to maintain focus even in trying times. I would like to thank Professor Abbas Edalat for his patience in helping me develop my understanding of self-attachment therapy, and his guidance in the considerations required for safe deployment of digital mental health interventions.

I would like to extend my gratitude to Lisa Alazraki, who created the previous SAT chatbot, for generously explaining the details of her work, sharing the challenges faced in her project, and for her time in reading my draft report; Neophytos Polydorou, for his assistance in the many forms required for the ethical approval application and participant screening, and for having the patience to read through my drafts to provide valuable feedback.

I would like also like to thank Professor Lucia Specia, for sharing her expertise in Natural Language Processing, and providing technical guidance early in the project that pointed me in the direction of possible methods.

I would like to thank the volunteers that took part in the trials of my project, whose time and effort was instrumental to the successful completion of this project, and whose feedback has, and will hopefully continue to, led to improvements to the delivery of self-attachment therapy intervention to a greater number of people.

Contents

1	Introduction	1
2	Background	3
2.1	Self-Attachment Therapy and Previous Work	3
2.2	Importance of Empathy in Digital Mental Health Intervention	3
2.3	Virtual Personas and User Engagement	4
2.4	Large Pretrained Language Models	4
2.4.1	BERT	5
2.4.2	RoBERTa	5
2.4.3	GPT-2	5
2.4.4	T5	5
2.5	Toxicity and Bias of Language Models	6
2.6	Cross-lingual Transfer	6
2.6.1	Contrastive Alignment	6
2.7	Datasets	7
2.7.1	EmpatheticPersonas	7
2.7.2	NLPCC	9
2.8	Sentence Fluency Evaluation	9
3	Methodology	10
3.1	Dataset Translation	10
3.2	Emotion Classification	12
3.3	Empathetic Response Generation	18
3.3.1	Empathy Classification	18
3.3.2	Empathetic Rewriting	19
3.4	Conversation Flow	26
3.5	Web Platform and User Interface	29
4	Evaluation	31
4.1	Non-clinical Trial Setup	31
4.1.1	Overview	31
4.1.2	Participant Screening	32
4.1.3	Evaluation Questionnaire	32
4.2	Participant Feedback	32
4.2.1	Model Revisions	33
4.2.2	Results	33

4.3	Limitations of the Study	39
5	Conclusions and Future Work	41
6	Ethical Considerations	43
	References	45
	Appendices	52
A	Participant Trial Feedback	52
A.1	First Trial	52
A.2	Second Trial	53
B	Questionnaire Responses	54
C	Example Conversations	59
D	Overview of SAT Protocols	63

Chapter 1

Introduction

With the onset of the COVID-19 pandemic, there has been significant negative impact on the mental health conditions of the population from a variety of environmental stimuli (38), with the effect most severe among the 18-34 sociodemographic group in the UK but visible in all age demographics (11). Cases of patients suffering mental health issues associated with a range of negative emotions such as defeat, entrapment and loneliness increased significantly from pre-pandemic levels (33).

As such provision of mental health support has become more imperative in addressing mental health problems arising as a result of public health emergencies. Yet, availability and cost remain major factors in the delivery of mental health assistance, it is therefore desirable for the incorporation of digital technologies and novel techniques to supplement existing methods.

Self-attachment technique (SAT) (13) is a recent development in psychotherapy that has shown promise in early pilot projects (14), but crucially is self-administered, allowing for significant scalability, tackling a significant shortfall of common delivery of psychotherapy. SAT focuses on attachment theory, specifically secure attachment. It consists of a series of protocols that can be delivered in an algorithmic format to assist patients with emotion regulation, allowing for enhancement of positive emotions and minimisation of the effect of negative emotions.

Prior work has been conducted by previous students that incorporate technologies into the delivery of SAT protocols, with the most recent producing a chatbot assistant (1). Conversational agents (25) have significant potential in application to psychotherapy (49, 42), as recent advancements in the field of neural natural language processing with large neural pretrained language models (12, 40) using transformer-based architecture (53) have reached new state-of-the-art on a variety of tasks that facilitate greater capability of human-computer interaction.

However, prior works are limited only to English, a situation emblematic of much of the recent progress in application of machine learning models to Natural Language Processing (23, 57). Monolingual NLP for certain languages can encounter the problem of resource-availability, where there is a lower volume of available task-specific data to train a model to the same level of performance as higher-resource languages such as English. Recent advancements in cross-lingual alignment methods (43, 20, 19) that leverage training conducted on a high-resource language in a low-resource setting (3) have reduced the gap between monolingual performances.

In this thesis, we present a computational framework for the delivery of self-attachment technique (SAT) in a simplified Chinese setting, leveraging existing data along with open-domain, language specific data. Our approach aims to investigate the data-requirement and the efficacy of deploying mental health interventions in a different language setting. We then deploy our chatbot as a web platform and conducted human evaluation with a cohort of $N = 15$ subjects from the non-clinical population with native level reading fluency in simplified Chinese over two trials. Lastly, we show that our approach is scored highly for ability to engage empathetically with the user, ability to identify user’s emotional state, perceived usefulness for the platform and quality of responses. We offer a comparison between our platform and the previous monolingual English implementation (1). We conclude with reflections based on feedback from the human evaluation, identifying areas of improvement and provide suggestions for targets of future work. Our contributions are summarised as follows:

- We introduce a pipeline for leveraging existing task-specific data from the previous work (1) to create in-language, fluent and empathetic responses with quantitative comparison to basic machine-translated data, as well as qualitative analysis of common issues encountered in machine-translated data (Section 3.1).
- We introduce a framework using GPT2 (40) to perform empathetic rewriting as a form of response generation and data augmentation, with quantitative comparison of GPT2 against mT5 (57). We justify our use of GPT2 in our final implementation as it outperformed mT5 on our task in metrics for generated sentence fluency (LM perplexity: 18.46 vs 20.20), empathy (percentage rated ‘Strongly Empathetic’: 93.07% vs 83.61%) and novelty (sentence pair-wise similarity: 0.641 vs 0.697). (Section 3.3.2).
- We produce a multilingual emotion classifier model using an `infxlm-large` model(8) that attains state-of-the-art (SOTA) monolingual emotion classification performance (+1.52% acc., +1.54% macro-f1) in Chinese, whilst achieving comparable monolingual English performance to the previous work (Section 3.2).
- We create a 120-utterance emotion classification test dataset of native level Chinese utterances, balanced across 4 emotion classes (Section 3.2). We train an empathy classifier using the same setup to yield comparable monolingual performance in empathy classification (+0.73% acc., +0.53% macro-f1) (Section 3.3.1).
- Qualitative analysis aimed at guiding future research based on trial feedback and limitations (Section 4.3), as well as suggestions for future work (Chapter 5).

Chapter 2

Background

2.1 Self-Attachment Therapy and Previous Work

Self-attachment therapy (SAT) is a psychotherapy framework consisting of a series of self-administered protocols with aims to reinforce neural patterns associated with secure attachment (13). The series of 20 SAT protocols focus on establishing a parent-child attachment between a patient’s current ‘self’ and their ‘childhood self’, through which it aims to enhance a patient’s positive emotions and helps them to develop and recognise methods to further control negative emotion. A recent pilot study investigating the efficacy of self-attachment techniques in treating resistant depression and anxiety (14) showed promising results.

SAT protocols can be administered in an algorithmic manner, creating the potential for the recommendation of protocols to be conducted by a smart assistant based on a user’s current condition through conversational interaction.

The work conducted in this paper will be building upon prior work conducted by a previous student (1, 2) which produced an English-capable empathetic guide for self-attachment therapy, as well as produced the EmpatheticPersonas dataset (Section 2.7.1). The previous chatbot was capable of emotion recognition at 94.96% accuracy, using this information within a rule-based conversation mechanism to generate responses corresponding to 5 virtual personas. Chatbot responses were generated from selecting existing sentences through maximising a weighted sum of sentence fluency, empathy, and novelty scores.

2.2 Importance of Empathy in Digital Mental Health Intervention

According to psychotherapy research, an important component in the efficacy of psychotherapy interventions is the capability of the therapist to engage in an empathetic manner to the patient (15). Similar to prior works (1, 46), we focus on Godfrey T. Barrett-Lennard’s second phase of empathetic dialogue with the aim of producing empathetic responses demonstrating compassion towards the user.

Prior works on empathetic dialogue systems (45, 42, 46) have highlighted the importance of empathy in digital mental health support. However, prior works mostly focus on English, and at the time of writing, there is little open-domain, language-specific and task-specific data for open empathetic response generation in Chinese. Furthermore, large generative language models have been shown to suffer from toxic degeneration (17, 64), making them unsuitable for open-ended generation in mental health applications without significant detoxification or sufficient in-domain fine-tuning. We opt for a more controlled approach of empathetic rewriting in the style of (47), leveraging existing, in-domain, monolingual training data to produce empathetic responses.

2.3 Virtual Personas and User Engagement

Prior works in empathetic dialogue systems have shown the perceived level of empathy expressed correlates with the relative personalities of the conversation participants (62). That is, a higher degree of user satisfaction is obtained when an individual converses with a persona that is suited to their preferences. The previous work (1) made use of this by creating 4 named human personas (Arman, Olivia, Robert, Gabrielle) across two sexes (Male, Female) and age groups (18-39, 40-69), and a fifth persona ‘Kai’, a less-human-like persona incorporating the available responses of all 4 aforementioned personas. Trials conducted in the previous work showed the addition of personas to provide a significant improvement in user engagement and perceived level of empathy, but the differences in personas proved inconclusive due to the small number of feedbacks received. We opt to include all 5 personas for our chatbot to allow a greater degree of choice for the user.

2.4 Large Pretrained Language Models

In the advent of transformers (53), large pretrained transformer-based language models have continued to yield increasing performance on a variety of NLP tasks. The increasing availability of state-of-the-art pretrained language models via platforms such as Huggingface (54) has drastically increased the accessibility of highly complex language models.

A **language model** (LM) is a statistical model capable of computing the probability of a word w_i conditioned on a context or sequence of words w_1, \dots, w_m , such that $P(w_i|w_1, \dots, w_m)$ can be used to generate a sequence by selecting the next word with the highest conditional probability. A **multilingual** language model pertains to the model’s ability to handle multiple languages simultaneously (3).

The use of multilingual language models opens the possibility of utilising a single model (i.e. encoder and emotion classifier module) across multiple languages, to increase the volume of applicable task-specific training data but also allow for deployment in multiple languages from a single model.

2.4.1 BERT

BERT (12), is a transformer-based encoder model trained using Masked-Language-Modelling (MLM), where a model learns to predict masked word(s) within a sentence. BERT achieved state-of-the-art in Natural Language Understanding tasks of Question Answering and Sentence Prediction. (48) outlines 3 methods for fine-tuning BERT which are also applicable to fine-tune other BERT-based models: (a) Further pretraining of the model using task-relevant data; (b) Fine-tuning the model using multiple target-relevant tasks; (3) Fine-tuning on the target task. We note that for the purpose of this project, it is unlikely we will have the processing power nor the volume of task-specific data that will allow us to perform (a).

2.4.2 RoBERTa

RoBERTa (30) builds upon BERT’s pretraining method, whilst maintaining the architecture, allowing it to be fine-tuned using the same principles as BERT. For our cross-lingual transfer (Section 2.6) objective, we evaluate **XLNet** (9) or **XLNet-R**, a multilingual RoBERTa model trained using translation language modelling (10), where the model predicts the translated utterance rather than masked tokens (MLM). **InfoXLNet** (8) further added contrastive pretraining (52) which has been shown to outperform XLNet-R on cross-lingual transfer benchmarks.

2.4.3 GPT-2

GPT-2 (40) is a decoder-only transformer-based language model, capable of continuously generating a sequence of tokens given a context. GPT-2 architecture differs from BERT in that it is unidirectional, that is, only words in one direction of the current token (i.e. left of) will be considered in its generation. It is capable of generating fluent and correct text, though it is not without problems that restrict its application in mental health assistance (Section 2.5). We evaluate a variant of **GPT-2-Chinese**, a monolingual GPT-2 model obtained from the UER framework (61) pretrained on a 14G subset of the 100G Chinese CLUECorpus2020 corpus (55).

2.4.4 T5

T5 (41) is a text-to-text transformer utilising an encoder-decoder architecture which was trained to explore the hypothesis that any task can be formulated as a text-to-text task, where given a text input context it produces a textual output. The architecture of T5 resembles the original encoder-decoder structure proposed by (53), and uses 12 each of encoder and decoder blocks. T5 achieved SOTA on a variety of NLP tasks, with the notable exception of machine translation tasks, which the authors hypothesised as due to its English-only training set. (57) released a massively multilingual version of T5, named **mT5** capable of handling 101 unique languages, whilst maintaining the same T5 architecture.

2.5 Toxicity and Bias of Language Models

Whilst there have been advancements into the application of transformer encoder models in a mental health context (24), the same cannot be said for decoder-based generative models such as GPT-2 (40). A major reason for this is the apparent capacity for such models to generate biased or toxic texts, due to inherent bias and toxicity in its web-crawled training corpus, that makes it unsafe for use in mental health contexts. It has been shown that such models may degenerate into toxicity (17). Recent works (64, 4, 32) have made strides into detoxification and debiasing of language models. However, the reduction in toxicity and bias may not yet be sufficient to ensure safe deployment for open generation in mental health settings. For this work, we investigate the efficacy of generative language models in the task of conditional rewriting.

2.6 Cross-lingual Transfer

Since (53), there has been significant gains in the application of large pretrained language models to a variety of NLP tasks. However, a majority of these models remain monolingual, and are trained on high resource languages, usually English (57). To facilitate multilingual applications, a past solution has been to train a language model for each language, which raises issues of space complexity, and a correlation of model performance to the size of training data available, leading to low resource languages models having poorer performance that can lead to worse user experience.

It can be therefore, advantageous to train multilingual language models capable of handling multiple languages simultaneously (3), and allowing for task-specific information learnt on a high resource language to transfer to a low resource language (10).

A method of facilitating cross-lingual transfer is to perform cross-lingual alignment of monolingual embedding spaces in a joint cross-lingual embedding space (43). This allows for utterances in different languages to learn similar representations, allowing for downstream task-specific parameters to be shared across languages. There are three main paradigms of alignment: 1. word alignment; 2. sentence alignment; and 3. document alignment.

For this thesis we will focus mainly on 2: cross-lingual sentence alignment, as there is a significant difference in language typology between English and Chinese that may not always guarantee the presence of equivalent tokens in parallel sentences making it unsuitable for 1, and the nature of our dataset being disconnected sentences makes it unsuitable for 3.

2.6.1 Contrastive Alignment

A popular form of cross-lingual sentence alignment is contrastive alignment (7, 52), which aims to maximise the similarity between positive examples of sentence representations whilst simultaneously minimising the similarity between negative exam-

ples. This maximises the margin between representations of ‘dissimilar’ sentences, improving sentence classification performance.

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(x_i, x_j)/\tau)}{\sum_k \exp(\text{sim}(x_i, x_k)/\tau)} \quad (2.1)$$

Some methods use the mean of token embeddings as sentence representations (22), whilst others utilise the embeddings of the [CLS] token (35), produced by BERT-based models for the purpose of classification, as sentence representation. Most methods utilise the vector dot product as similarity function.

$$\ell_{\text{XeroAlign}} = \text{MSELoss}(x_{\text{src}}, x_{\text{tar}}) \quad (2.2)$$

$$\ell_{\text{Total}} = \ell_{\text{Task}} + \ell_{\text{XeroAlign}} \quad (2.3)$$

XeroAlign (20, 19) furthers this concept whilst simplifying the procedure by performing alignment as an auxiliary task during fine-tuning by minimising the mean-squared-error of the high resource sentence [CLS] embedding and a machine translated sentence [CLS] embedding to facilitate zero-shot cross-lingual transfer that yielded state-of-the-art zero-shot sentence classification results on several multilingual task-oriented benchmarks (28, 56). Although prior work have shown this yields a comparable level of performance to contrastive alignment (19).

2.7 Datasets

2.7.1 EmpatheticPersonas

In-order to fine-tune any pretrained language models for use on our task, we utilise a machine-translated version of the EmpatheticPersonas dataset (1) which we will refer to as EmpatheticPersonasZH. The English version of this dataset was collected by a previous student, which we translate to Chinese using a publicly available machine translation tool. The dataset was then manually inspected by a native speaker with significant semantic errors being rewritten to be more faithful to the original text.

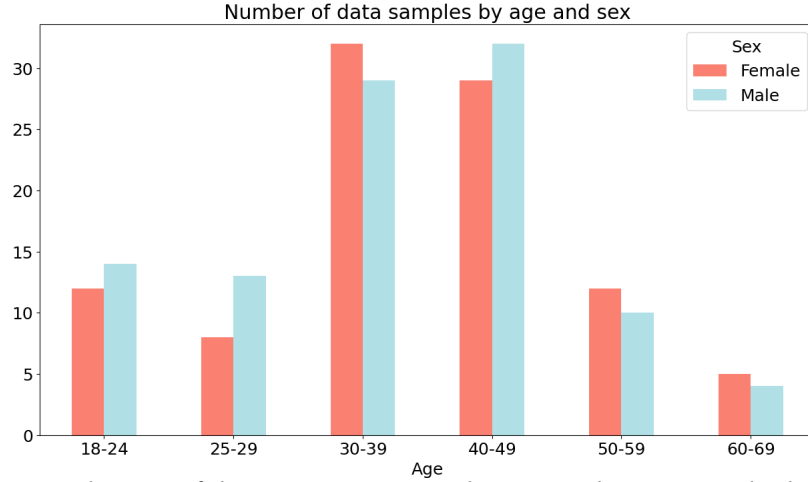


Figure 2.1: Distribution of dataset participants by age and sex. Note the higher concentration of participants in the 30-39 and 40-49 categories. Note: the same figure was used in the previous work (1), reconstructed for this paper as we utilise the same dataset.

The dataset consists of responses from 200 participants, with a 98 female and 102 male, the distribution of participant ages and sex is illustrated in Figure 2.1. There are 1181 utterances of emotional expressions, balanced across 4 emotions: 300 expressions of Sadness (悲伤), 297 expressions of Anger (愤怒), 284 expressions of Fear/Anxiety (焦虑) and 300 expressions of Joy (快乐), the distribution of participants by emotional utterances is illustrated in Figure 2.2. In addition, there are 2144 empathetic rewritings of 45 base utterances, with similar numbers of rewritings for each base utterance.

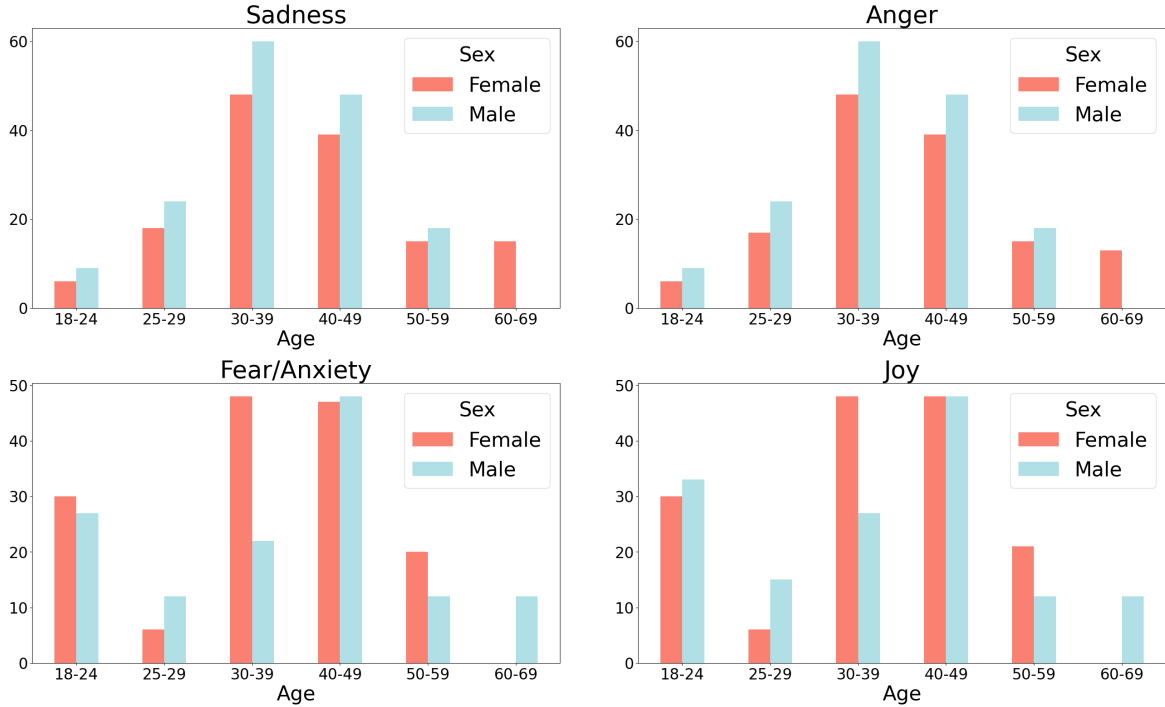


Figure 2.2: Distribution of participants for sample utterances exhibiting emotion by age and sex from the EmpatheticPersonas dataset.

2.7.2 NLPCC

The NLPCC dataset (63) consists of 40,132 utterances distributed in an unbalanced manner across 6 labels, the distribution is shown in Table 2.1. This dataset only shares 3 of its 6 labels with EmpatheticPersonasZH, combined with its unbalanced nature makes it unsuitable for direct use in fine-tuning in our emotion classification task. We can group the positive emotions (Like, Joy) for 11647 utterances and negative emotions (Sadness, Anger, Disgust) for 14493 utterances to form a coarse-grain fine-tuning task where the model is trained to classify positive, negative, or neutral emotion from a given utterance.

Label	Like (喜好)	Joy (高兴)	Sadness (悲伤)	Disgust (厌恶)	Anger (愤怒)	N/A (无情绪)
No. Expr.	6697	4950	5348	5978	3167	13993
Combined	11647		14493			13993

Table 2.1: Distribution of emotional expressions in the NLPCC Emotion Classification Dataset. Emotion label and number of corresponding utterances.

2.8 Sentence Fluency Evaluation

For the purpose of evaluating the fluency of utterances within our translated dataset, we investigate a variety of existing evaluation metrics utilised in machine translation tasks.

BLEU (36) and ROUGE (29) are some of the most commonly used metrics for evaluating the quality of machine translated text, by comparing the translated utterance against reference human translated utterance(s) and utilising its respective similarity as a metric for the quality of translation. This can be inhibitive slow due to the requirement for (sometimes multiple) human translation for each utterance.

It can therefore be desirable for automatic evaluation of machine translation systems, particularly reference-free methods of evaluation (16). Methods such as SLOR (26) utilises the negative-log-perplexity of a sentence given by a language model to evaluate its fluency, by calculating the joint probability of the words of a sentence $S = w_1, \dots, w_t$, with each word w_i conditioned on the words preceding it. SLOR (26) also utilises the unigram (single word) log-probability as a regularisation term, which some other methods do not use.

$$\text{SLOR}(S) = \frac{1}{|S|} \ln \prod_{i=1}^t p_M(w_i | w_{<i}) - \ln \prod_{i=1}^t p(w_i) \quad (2.4)$$

Another reference-free method is PRISM-SRC (50), which leverages multilingual LMs to compute, for a translated utterance S_{tar} of length t , its log-probability conditioned on the source utterance S_{src} . PRISM-SRC achieved SOTA in reference-free NMT evaluation with correlation to human evaluation.

$$\text{PRISM-SRC}(S_{\text{src}}, S_{\text{tar}}) = \frac{1}{|S_{\text{tar}}|} \sum_{i=1}^t \ln p_M(w_{\text{tar},i} | w_{\text{tar},<i}, w_{\text{src}}) \quad (2.5)$$

We can calculate this metric using a multilingual language model (i.e. mT5 (57)).

Chapter 3

Methodology

In this section, we provide some technical details regarding the steps taken to facilitate our framework in order to produce the emotion classification and empathetic response generations components of our chatbot in simplified Chinese with limited available native data.

3.1 Dataset Translation

The previous work (1, 2) sourced examples of emotional expressions and empathetic utterances via crowd-sourcing. The process can be time-consuming and expensive, unsuitable for scaling to different languages. We opt to leverage the existing English data for the purpose of this thesis. However, manual translation of the entire dataset would likely also require crowd-sourcing, and lead to the associated issues. To this end, we propose a translation pipeline that can be used for different languages without large-scale human translation.

We first translate the existing English dataset into simplified Chinese using publicly available machine translation tools¹, then manually identify and fix semantic errors in the machine translated corpus. Figure 3.1 shows an example of obvious semantic errors arising from literal translations in the machine translated text; such errors were remedied by rewriting part or all of the sentence to be more coherent with the original intent of the text.

EN, Original	ZH, Machine Translation	ZH, Edited
I've got a bad case of the blues	我有一个糟糕的 蓝调案例 。	我心情有点低落。 (I'm feeling low)
case (noun): situation blues (adj, inf): sad	case (noun): incident blues (adj): (colour)	

Figure 3.1: Example of translation errors resolved in version 1 of the translated dataset. Note the literal translations of the words ‘blue’ and ‘case’ in the machine translated text. Manual edits removed the specific tokens but rewrote the translation overall to be more coherent with the original intent.

¹Tools such as Google Translate, Baidu Translate etc. are publicly available and free to use.

We produce three versions of the translated dataset: **1.** The machine translated text without edits (raw), **2.** The translated dataset after a single round of edits (v1), **3.** The translated dataset after two rounds of edits, with colloquialisms injected (v2).

The ‘v1’ revision of the dataset was created by manually searching over the dataset, spotting any sentences that appeared strange or grammatically incorrect, cross-referencing with the source sentence and making small adjustments to remedy significant translation errors (Figure 3.1). We define ‘significant translation errors’ as errors that causes a significant change in the sentence’s intent, or renders it incomprehensible.

En, Original	ZH, Machine Translation
I’m sorry to hear you feel this way	我很抱歉听到您有这样的感觉
抱歉 (noun): Regret, Apology	
ZH, Edited	
听到您有这样的感受我很遗憾	

Figure 3.2: An example of translation errors resolved in version 2 of the translated dataset. Note the translation of ‘sorry’ → ‘抱歉’, which the noun-form equivalent to ‘be sorry’ or ‘I apologise[to hear that ...]’, a result of the EN phrase being translated literally into ZH. The manual edits to the utterance replaces the incorrect phrase with one more appropriate in native speech. Additionally, the translation for ‘feel’ → ‘感觉’(to feel [physically]) was replaced with ‘感受’(to feel [emotional]) as it was more appropriate. Note also that the final sentence order differs to the original EN sentence.

Revision ‘v2’ was created following feedback from the first human trial evaluation of the chatbot platform (Chapter 4), where trial participants identified issues with the response generated; such as having carried certain English figures-of-speech verbatim to Chinese (Figure 3.2) and lacking Chinese colloquialisms. Another pass of the dataset was conducted, focusing on mistranslation patterns identified in the trial, as well as making more significant structural changes to the responses to be more natural sounding.

Evaluation

We apply two reference-free fluency evaluation metrics to empathetic responses of three major revisions of the EmpatheticPersonasZH dataset to compare the relative fluency of the translated texts. Results are shown in Table 3.1 alongside the language model perplexity².

Revision	SLOR	σ_{SLOR}	PRISM-SRC	$\sigma_{PRISM-SRC}$	PPL	σ_{PPL}
raw	3.84	0.43	39.05	7.34	19.77	12.45
v1	3.87	0.36	34.71	7.13	18.08	8.14
v2	3.92	0.37	35.04	7.25	18.83	9.02

Table 3.1: Average sentence SLOR scores (higher is better), PRISM-SRC scores (lower is better) and Perplexity (PPL) scores (lower is better) for different revisions of the dataset.

²SLOR and PPL calculated using GPT2, PRISM-SRC using mT5-small

We observe from Table 3.1 that both revisions yield a better fluency score across all three metrics, suggesting that the manual edits did improved the fluency of the utterances over raw machine translated text, and improved translation quality with respect to the source sentence. We note that ‘v2’ yields a slightly higher improvement over the ‘raw’ version in SLOR than ‘v1’, whilst ‘v1’ scores higher on PRISM-SRC and Perplexity. We hypothesise this to be due to the edits made in ‘v2’ being of a finer-grain nature, using ‘rarer’ tokens that appears more in colloquialisms.

3.2 Emotion Classification

In order for the chatbot to formulate a conversation flow appropriate to the user, it is imperative for SAT chatbot to be able to accurately and reliably determine the user’s emotional state from their response to the question of ‘How are you feeling?’. For the purpose of this project, we consider four primary emotional states of ‘Sadness’, ‘Anger’, ‘Fear/Anxiety’ and ‘Happiness’.

To this end, we experiment with a variety of methodologies focusing on two SOTA multilingual language models: XLM-RoBERTa(9) and InfoXLM(8), and compare their performance with existing monolingual solutions. Additionally, we explore the cross-lingual zero-shot capability of our implemented solution.

Contrastive Alignment

We note that whilst the edits made to the translated EmpatheticPersonas dataset resolved the issue of blatant mistranslations, the translated utterances generally contain less variety than human utterances (Figure 3.3). This appears to be an inherent problem with the original English dataset, especially with regards to shorter emotional utterances. This could lead to issues with generalisation, and negatively impact the task performance.

我感觉不太好 (I don't feel great)	我感觉很好 (I feel great)	我感觉很生气 (I feel really angry)
<i>label: Sadness</i>	<i>label: Joy</i>	<i>label: Anger</i>

Figure 3.3: Example of emotional expressions in the dataset from different classes but share a significant portion of tokens. This may cause the model to learn similar sentence representations that can lead to issues with classification.

To this end, we perform contrastive alignment (7, 52, 43) between source (EN) and target (ZH) sentence embeddings as an auxiliary task (19) to fine-tuning for our task in the source language.

Given a batch of pairs of source language utterances x_{SRC} and a machine translated utterances x_{TAR} in the target language, we formulate a contrastive alignment objective based on (7)

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(x_{\text{SRC},i}, x_{\text{TAR},j}))}{\sum_k \mathbb{1}_{[k \neq i]} \exp(\text{sim}(x_{\text{SRC},i}, x_{\text{TAR},k}))} \quad (3.1)$$

where $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is an indicator function that masks out the similarities of utterances which share the same label. The intuition is to maximise the difference between utterances of different classes, whilst simultaneously minimising the difference between the CLS embeddings of the same utterances in different languages. We utilise this contrastive object as an auxiliary task to the classification task trained on the source language.

$$\ell_{\text{TOTAL}} = \ell_{\text{TASK}} + \ell_{\text{CTR}} \quad (3.2)$$

For our fine-tuning step using the NLPCC Emotion Classification dataset, we also use this contrastive objective as an auxiliary objective to minimise the similarity for each utterance against other utterances with different labels.

Data Augmentation

Sentence 1 Sentence 1
最近几天我的心情一直很低落我一点都不舒服一切事情都不顺
(I've been feeling down for the last couple of daysI'm not feeling well, nothing's going my way)

Figure 3.4: Example of data augmentation by concatenating two emotional expressions of sadness, that should maintain an expression of the same emotion. We omit separators to force the model to consider the concatenated sentences as a single sentence.

We perform an additional data augmentation step on the training set by concatenating sentences with identical emotion labels together (Figure 3.4). The assumption being that those utterances exhibiting the same emotion would still exhibit the same emotion when concatenated together but still introduce variety to the sentence structure. We perform this augmentation twice for each sentence in each class to increase the size of the EmpatheticPersonas training set from 944 to 2832 utterances. We also omit any separators between the two sentences to encourage the model to consider the entire sentence instead of only a single sub-phrase.

Baseline

We compare the performance of multilingual models (InfoXLM (8), XLM-R (9)) in our setup against the monolingual performance of the emotion classifier module used by a previous student (1). The baseline model used a RoBERTa-base model (30), fine-tuned on a subset of the Emotion dataset (44) and then on EmpatheticPersonas.

Experiment Setup

For our experiments, we train our selected pretrained language model in three configurations:

1. Fine-tune directly on the translated EmpatheticPersonas dataset. (*Single Fine-tuning*)

2. Fine-tune using EN EmpatheticPersonas dataset with contrastive alignment as an auxiliary objective, followed by a small amount of fine-tuning on translated data. (*Single Fine-tuning w. Contrastive Alignment*)
3. Fine-tune first on a subset of the NLPCC Emotion Classification dataset (63), followed by the steps in 2. (*Double Fine-tuning*)

We split all datasets into train (80%), validation (10%) and test (10%) sets, with hyperparameters being tuned on the validation set, and selecting the best performing combinations. Each model is then evaluated on the test dataset. Results are shown in Table 3.2.

We also created an additional test set of human utterances in Chinese consisting of 120 utterances distributed equally across the 4 emotion classes. The utterances conform more accurately to native speech and are generally harder than the translated utterances. Results of our model tested on this dataset is shown in Table 3.3.

Hyperparameters

We search for optimal hyperparameters for each task from a set using a process of beam search; by conducting experiments starting with a fixed set of hyperparameters with a single tuning target, we pivot between different hyperparameters by selecting the setup with the current highest performance. Doing so reduces the number of experiments required compared to grid search and was desirable due to limited compute available. In all tasks, we utilise an AdamW optimiser with a tuned epsilon value and a linear learning rate scheduler with a tuned warmup period.

For our emotion classification experiments using infoXLM-large, we tune from the following set of hyperparameters (**bold** denotes final chosen hyperparameters):

- **Learning rate:** 5×10^{-4} , 1×10^{-5} , **5×10^{-5}** , 1×10^{-6}
- **Adam epsilon:** 1×10^{-4} , 1×10^{-5} , 1×10^{-6} , 1×10^{-8}
- **Batch size:** 1, 2, 4, 8, 16, **32**
- **Warmup steps:** 0, 20, **50**, 100

Results

Model	Configuration	Accuracy	Macro-F1
roberta-base ³	Single Fine-tuning Baseline (1)	89.92	90.26
	Double Fine-tuning Baseline (1)	94.96	95.10
xlm-roberta-large	Single Fine-tuning	86.55	86.51
	Single Fine-tuning w. Contrastive Alignment	88.24	88.31
infxlm-large	Single Fine-tuning	88.24	88.68
	Single Fine-tuning w. Contrastive Alignment	89.92	89.88
	Single Fine-tuning+Alignment (Augmented)	91.67	91.67
	Double Fine-tuning	94.96	95.18
	Double Fine-tuning (Augmented)	96.48	96.64

Table 3.2: Emotion classification results of different models trained with different configurations using the EmpatheticPersonas datasets.

Our Double Fine-tuning model (ZH) with data augmentation outperforms the previous SOTA (EN) setup in monolingual classification (+1.52% Acc, +1.54% Macro-F1). Additionally, the same model yields (90.76, 91.06) on English test data, outperforming the previous best-performing English ‘Single Fine-tuning’ model.

We note from Figure 3.2 that the single fine-tuning performance for both multilingual models underperformed the baseline single fine-tuning setup despite larger models (roberta-base: 123M, xlm-roberta-large and infxlm-large: ~355M). We hypothesise this to be due to the significantly larger number of languages supported by the multilingual languages models, a lower amount of model capacity is available per language (39). We do note that the (90.76, 91.06) attained on English data by the best performing setup was accomplished with only a single fine-tuning stage on English data, suggesting a degree of cross-lingual transfer from the additional Chinese data to English performance.

Ablations (Figure 3.2) show the biggest improvements in performance comes from the use of additional native data (+~5%) followed by data augmentation (+~1.6%) and contrastive alignment (+~1.5%) suggesting that the model performance is still reliant on exposure to in-language training data, but doing so does not appear to have a significant negative impact on performance in other languages. Such reliance on native data can be problematic in situations where such data is scarce, but data augmentation and contrastive alignment using existing data has shown promising gains in classification performance, and could potentially supplement such cases.

³Prior work is English only. We compare against the monolingual performance of the multilingual model.

Model	Configuration	Accuracy	Macro-F1
infoxlm-large	Single EN Fine-tuning Zero-shot	87.50	87.63
	Single Fine-tuning	88.11	88.23
	Single Fine-tuning w. Contrastive Alignment	90.67	90.73
	Single Fine-tuning+Alignment (Augmented)	92.50	92.57
	Double Fine-tuning	91.67	91.58
	Double Fine-tuning (Augmented)	93.33	93.33

Table 3.3: Emotion classification results of different models trained with different configurations tested on our native test set.

We note that the zero-shot (Table 3.3) performance is encouraging, being slightly below the ‘Single Fine-tuning’ performance that makes use of translated data.

Our ‘Double Fine-tuning with Data Augmentation’ configuration yielded performance on the native test set comparable to the previous monolingual SOTA (1), though we note that the native test set contains utterances that are in general more difficult than that in EmpatheticPersonas due to incorporating colloquialisms. We note from Table 3.3 that both data augmentation and fine-tuning using native utterances (63) yield improvements in performance on native test set, though the methods yield different amount of improvement compared to testing on EmpatheticPersonas. We also note that exposing the model to the additional NLPCC data does not appear to yield as significant an increase in performance compared to before, we hypothesise this to be due to the differences in the nature of emotional expressions between the two datasets in addition to the added complexity of the native test set. We suggest for the test set to be further expanded in the future to incorporate more utterances from different speakers with different tones, as more test examples may be necessary to cover possible user responses.

NLPCC Fine-tuning Configurations

The NLPCC Emotion Classification dataset shares three (Sadness, Anger, Joy) of emotion four classes with our EmpatheticPersonas dataset, in order to find an optimal way of utilising the available data, we experiment with different means using this dataset to perform in-language fine-tuning:

1. Binary classification using all emotional utterances from the NLPCC Emotion Classification dataset, with ‘Sadness’, ‘Disgust’ and ‘Anger’ labelled as ‘Negative’ emotions, and ‘Like’ and ‘Joy’ labeled as positive emotions. (26140 Utterances)
2. Multi-class classification using the shared labels ‘Sadness’, ‘Anger’ and ‘Joy’, with down-sampling to the smallest class to maintain balance. (9501 Utterances)
3. Multi-class classification similar to 2., but with an equal number of neutral utterances to provide negative example as class 5 to the three existing classes. (12668 Utterances)

Model	Configuration	Accuracy	Macro-F1
roberta-base	Double Fine-tuning Baseline (1)	94.96	95.10
xlm-roberta-large	Single Fine-tuning	86.55	86.51
	Single Fine-tuning w. Contrastive Alignment	88.24	88.31
infxlm-large	Single Fine-tuning	88.24	88.68
	Single Fine-tuning w. Contrastive Alignment	89.92	89.88
	Single Fine-tuning+Alignment (Augmented)	91.67	91.67
	Double Fine-tuning (binary)	95.02	94.80
	Double Fine-tuning (3-labels)	94.12	94.47
	Double Fine-tuning (4-labels)	94.96	95.18
	Double Fine-tuning (4-labels, Augmented)	96.48	96.64

Table 3.4: Emotion classification results of different models trained with different fine-tuning configurations.

Results are shown in Table 3.4. We note that all setups that involved fine-tuning on this larger dataset outperformed setups without, leading to our choice to use this step in our final model. Our early experiments also showed this larger corpus as most beneficial to our classifier model when it is used first in the fine-tuning pipeline prior to task-specific fine-tuning, but can negatively impact performance when used later. We hypothesise to be due to inherent differences in the nature of the utterances contained in this dataset from our task, as this dataset contained web-crawled data, and is much noisier in general than EmpatheticPersonas.

We also note that despite the significantly lower number of utterances, the double fine-tuning setup with 4 labels (12668 vs 26140 utterances) performs similarly to the binary classification setup. Additionally, the 4-label setup using an additional neutral class leads to better performance compared to the 3-label setup (Table 3.4) when the 4th emotion class ‘Fear/Anxiety’ is introduced despite ‘Neutral’ being designated a different label. We hypothesise this to be due to a notion of ‘reserving’ encoder model capacity for a 4th class, allowing more efficient introduction of the ‘Fear/Anxiety’ class as a label different to the shared 3.

3.3 Empathetic Response Generation

3.3.1 Empathy Classification

The ability to demonstrate empathy in conversation is crucial to success in mental health intervention, it is therefore paramount for our model to be able to identify empathy in utterances (46, 49), and to produce empathetic speech.

Base utterance	Was this caused by a specific event/s?
Rewriting exhibiting ‘Little to None’ empathy	Is your mood caused by anything specific?
Rewriting labelled ‘Weakly Empathetic’	Did something in particular happen that made you feel sad? Feel free to tell me openly.
Rewriting labelled ‘Strongly Empathetic’	I’m so sorry to hear that. Can you identify a specific event that caused you to feel that way?

Table 3.5: Example of utterance rewriting from the EmpatheticPersonas dataset of the same base utterance under the user’s ‘Sadness’ emotional context, exhibiting different levels of empathy. Note the higher degree of compassion shown for the user’s emotional state in the utterance labelled ‘Strongly Empathetic’, but also the subjectivity of the utterance labelled ‘Weakly Empathetic’; whilst the rewriting acknowledges the user’s emotion, it does not demonstrate significant compassion to the user’s emotion, hence the lower score.

We utilise the edited translations of a labelled subset of EmpatheticPersonas(1) with 1100 utterances displaying varying levels of empathy (331 *Little to None*, 388 *Weakly Empathetic*, 381 *Strongly Empathetic*). Table 3.5 shows examples of rewritings in the labelled subset of EmpatheticPersonas dataset demonstrating different levels of perceived empathy. We train an InfoXLM-large model using identical hyperparameters as Emotion Classification (Section 3.2) with results shown in Tables 3.6 and 3.7.

(EN Empathetic)

I am sorry to hear that you are feeling that way. I would like to ask you if this was caused by something specific, if you do not mind.

(ZH Empathetic)

我很遗憾听到您有这种感觉。如果您不介意，我想问您这是否是由特定原因引起的。

Figure 3.5: Example of the same empathetic utterance in English and in Simplified Chinese. Note the different order of phrases expressing the same sentiment within both sentences.

Figure 3.5 shows two examples of the same empathetic utterance expressed in EN and ZH. We see that similar phrases exist in both sentences to facilitate empathetic expression, however, we note that the sentence structure differs, as well as the choice of words to express the same sentiments. For example, the phrase “I’m sorry

to hear that[...]" is equivalent to the phrase “我很遗憾听到[...]”, but the word 遗憾 typically translated as ‘遗憾 (*adj*): regret, pity’. Whilst the phrases may be semantically equivalent, their tokens do not individually carry the same meaning. This could limit the amount of transferable information in empathetic utterances across languages.

Results

Model	Accuracy	Macro-F1
T5-base (1)	80.18	80.66
InfoXLM-large	80.91	81.19

Table 3.6: Empathy classification result in Chinese compared to previous work in English.

Level of Empathy	Precision	Recall	Macro-F1
Little to None	82.35	77.78	80.00
Weakly Empathetic	70.73	76.32	73.42
Strongly Empathetic	91.43	88.89	90.14

Table 3.7: Empathy classification results for utterances displaying different levels of Empathy. Note that the model is much better at recognising when a sentence exhibits strong levels of empathy.

We observe from Table 3.6 that our model yields marginally better monolingual performance than previous SOTA. We also note from Table 3.7 that the model is much better at identifying sentences displaying high levels of empathy than distinguishing between sentences displaying some or little empathy. It is for this reason we opt to use this model as a binary classifier for identifying if a sentence is strongly empathetic, to support our empathetic rewriting task.

We also experimented with data augmentation via backtranslation. However, we did not pursue this further as the level of empathy was inconsistent across backtranslations and the sentence fluency degraded significantly.

3.3.2 Empathetic Rewriting

In order to produce sentences which are fluent, novel and displayed empathy, prior work (1) utilised a weighted sum of fluency, novelty and empathy scores, of which fluency and empathy are static, and the weights manually tuned. At inference time, pre-written sentences are sampled, with the highest scoring sentence by this metric presented to the user. This method could pose problems for adaptation to different languages, namely, that the weights of the objective function would have to be manually tuned, and can be highly subjective with regards to the choice of most appropriate response. Additionally, we would need to repeat the procedure of sourcing native level responses, increasing the cost of scaling to more languages. Prior work also utilises a split-and-merge data augmentation mechanic to increase the amount of candidate sentences that can be generated; however, this method is reliant on the

sentence structure of English utterances (with sentiment-specific text at the middle or end, separated from empathetic expressions by commas) which is not applicable to different sentence structures in different languages (Figure 3.5).

We explore the potential for empathetic responses to be generated as empathetic rewritings of base utterances by leveraging empathetic expressions in existing data and the capability of state-of-the-art generative transformer language models to generate fluent sentences (47). Additionally, we also investigate the capacity for the model to produce empathetic utterances as a means of data augmentation.

Experiment Setup

We first apply the empathy classifier to the entire EmpatheticPersonasZH dataset, selecting rewritings that are classified as ‘Strongly Empathetic’, which yielded 780 utterances exhibiting high levels of empathy (377 Male, 403 Female) for use as training data.

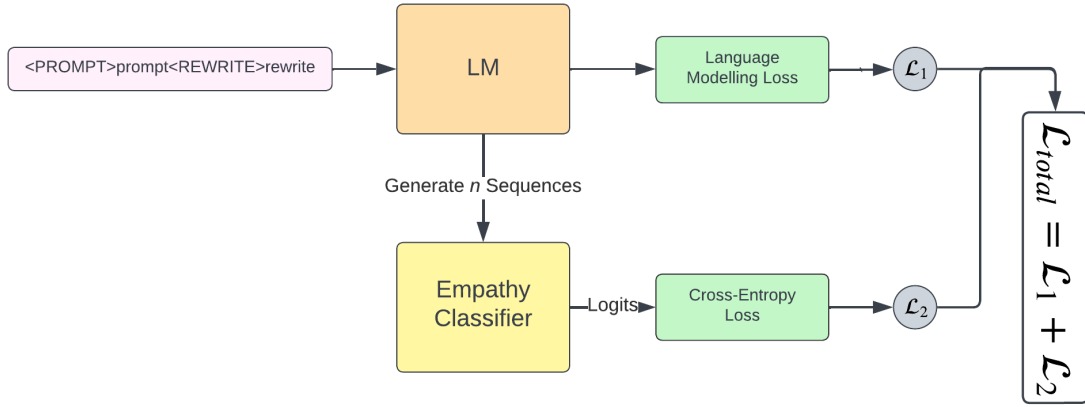


Figure 3.6: The training setup used to fine-tune a language model for empathetic rewriting of base utterances, inspired by (47). This setup aims to produce fluent sentences by minimising the language modelling loss and sentence empathy by utilising the empathy classifier as an auxiliary training objective.

We evaluate two state-of-the-art language models for the purpose of empathetic rewriting: GPT2-Chinese (40, 61), and the small version of mT5 (57). The specific models were chosen as the respective versions closest in size, though the mT5-small model is almost 3 times the size of the GPT2-Chinese model. Figure 3.6 shows the training setup, utilising language modelling as the primary objective to produce fluent utterances, and an auxiliary empathy objective to train the model to produce empathetic utterances in the style of the EmpatheticPersonasZH dataset.

Hyperparameters

Hyperparameters for the generator models were tuned in a similar manner to the classifier models (Section 3.2). We note from analysis of the high empathy utter-

ances in English and Chinese that the number of tokens present in the parallel sentences in both languages differed significantly, with an average ‘Strongly Empathetic’ sentence in ZH containing 42.4 tokens, and the equivalent in English containing 27.5. The difference in number of tokens led to the inclusion of maximum sequence length as a tunable hyperparameter.

For our empathetic generation task using GPT2-chinese-cluecorpussmall, we tune from the following set of hyperparameters (**bold** denotes final chosen hyperparameters):

- **Learning rate:** 5×10^{-4} , 1×10^{-5} , **5×10^{-5}** , 1×10^{-6}
- **Batch size:** 1, 2, **4**, 8
- **Warmup steps:** 0, 20, **50**, 100
- **Maximum sequence length:** 50, **100**, 120, 150, 200

Personas and Prompts

Prior works have highlighted the contributions of personas to user satisfaction (62, 1), wherein users are more accepting of empathy displayed by specific personas. We encode persona information within the prompt using the 4 human personas from (1) with differing sexes (S_s) and ages (S_a) in addition to the emotion (S_e) and base utterance (S_b).

$$\langle \text{PROMPT} \rangle S_s [\text{SEP}] S_a [\text{SEP}] S_e [\text{SEP}] S_b \langle \text{REWRITE} \rangle S_r$$

Figure 3.7: Example prompt setup, consisting of persona-specific sex and age information, with emotion and base utterance. For GPT2, the prompt and target are concatenated into a single sentence. `bos` and `eos` tokens are added by the tokenizer and not shown here.

The rewritten utterance (S_r) is added following a `<REWRITE>` token at training time as the learning target. At inference time, S_r is omitted and the model is expected to generate a rewriting conditioned on the given prompt.

Results

In order to provide quantitative evaluation on the performance of the models, we train both GPT2 and mT5 models for 450 epochs, and generate 20 rewritings for each of 45 base utterances for each persona, removing duplicate generations, then performed fluency and empathy evaluations on the generated utterances. Results are shown in Table 3.8 and Table 3.9.

Model	Personas	SLOR	σ_{SLOR}	Perplexity	σ_{PPL}
gpt2-chinese	Male (18-39)	3.92	0.34	18.23	9.03
	Female (18-39)	3.93	0.37	19.04	9.41
	Male (40-69)	3.96	0.34	18.16	8.37
	Female (40-69)	3.96	0.35	18.41	8.11
	Overall	3.94	0.35	18.46	8.80
mT5-small	Male (18-39)	3.96	0.44	19.98	11.13
	Female (18-39)	3.93	0.43	20.56	10.14
	Male (40-69)	3.93	0.44	20.62	10.70
	Female (40-69)	3.95	0.42	19.63	9.19
	Overall	3.94	0.43	20.20	10.33

Table 3.8: Fluency evaluation of generated rewritings by model, evaluated using SLOR (26) (higher is better) and the given language model perplexity (PPL) (lower is better) with standard deviations σ provided for each metric.

Table 3.8 shows fluency evaluation scores for persona-specific utterances generated using GPT2 and mT5. We use both the SLOR metric (26) and the language model perplexity to evaluate the comparative fluency of utterances from both models. Despite a significantly smaller model, the utterances generated using GPT2 outperforms mT5 on the perplexity metric, and yields an equal score on the SLOR metric, with lower standard deviation for both metrics.

Model	Personas	High Empathy Utt. (%)
mT5-small	Male (18-39)	82.82
	Female (18-39)	85.34
	Male (40-69)	81.07
	Female (40-69)	85.12
Overall: 83.61		
gpt2-chinese	Male (18-39)	91.19
	Female (18-39)	92.70
	Male (40-69)	92.70
	Female (40-69)	94.49
Overall: 93.07		

Table 3.9: Percentage of generated utterances from each model classified as ‘*Strongly Empathetic*’. We note that GPT2 generated rewritings exhibiting higher levels of empathy at a significantly higher rate than mT5. We also note that for both models, stronger empathy was exhibited in the rewritings of the female personas than the male personas across all age groups.

We also perform empathy evaluation on the generated utterances by computing the percentage of generated utterances that are rated to be ‘*Strongly Empathetic*’

by our empathy classifier. Table 3.9 shows that the GPT2 model generated sentences with high levels of empathy at a significantly higher rate than mT5 (93.07% vs 83.61%). We also calculate average pairwise sentence similarity scores using BERTScore (60) as a metric of variety within the generated sentences. Table 3.10 shows that utterances generated by the GPT2 model scores a lower similarity score than those of the mT5 model, thus demonstrating higher variety.

Model	Personas	Mean Sim.	σ_{sim}
mT5-small	Male (18-39)	0.695	0.047
	Female (18-39)	0.696	0.046
	Male (40-69)	0.699	0.045
	Female (40-69)	0.699	0.044
	Overall:	0.697	0.045
gpt2-chinese	Male (18-39)	0.643	0.060
	Female (18-39)	0.643	0.058
	Male (40-69)	0.633	0.066
	Female (40-69)	0.645	0.053
	Overall:	0.641	0.060

Table 3.10: Mean pairwise sentence similarity (lower is better) of rewritings generated by each model, calculated from BERTScore (60) using an `infolm-large` without fine-tuning. We note that the rewritings produced by GPT2 quantitatively contains more variant responses, which could contribute to conversation engagement and avoid responses becoming stale. This trend is visible upon manual inspection of the generated responses.

Since the GPT2 model has demonstrated stronger fluency (Table 3.8), expressed empathy (Table 3.9) and variety (Table 3.10) in its rewritings, we opt to use this model for empathetic response generation. To this end, we generate a total of 30 rewritings for each persona-base-prompt pair, yielding 2207 responses after removing duplicates (or significantly similar sentences), compared to the initial 780 training instances. Table 3.11 shows some example generated responses from each persona and emotional context.

We note from observation of the generated utterances (Table 3.11) that, consistent with our qualitative evaluation (Section 3.3.2), the generated sentences maintain a good degree of fluency, expressed empathy and sentence novelty across all personas and emotions. From the examples in Table 3.11, we note different manners of speech utilised by different personas, such as ‘哦，真对不起’ (Oh no, I’m so sorry), that is more commonly used by the Female (18-39) persona as an apologetic expression than by either of the (40-69) personas, who uses ‘我很抱歉’ (I’m sorry/I apologise) more frequently, which is a more formal tone of speech. This is likely a result of the same differences in the original English data that is maintained in the translation and edit processes.

Persona	Emotion	Base sentence	Rewriting
Male (40-69)	Anxiety	谢谢。现在我会问一些问题以了解您的情况。 (Thank you. Now I will ask some questions to understand your situation)	感谢您与我分享。我可以再问您一些问题，希望我能帮助您处理这种负面情绪吗？ (Thank you for sharing. I would like to ask you some more questions in the hopes of helping you deal with these negative emotions, is that alright?)
Female (40-69)	Sadness	这是由特别事件引起的吗？ (Was this caused by a specific event/s?)	谢谢您让我知道，我非常抱歉您有这种感觉。请问这种感觉是不是有什么特殊原因？ (Thank you for letting me know, I'm very sorry to hear you're feeling this way, please could I ask if there's a reason you're feeling like this?)
Female (40-69)	Joy	那很好！让我推荐一个您可以尝试的练习。 (That's good! Let me recommend a protocol you can attempt)	我很高兴听到这个。我认为这将是尝试可能让您感觉更好的练习的最佳时机！ (I'm really glad to hear this. I feel this could a great opportunity to try something that can make you feel even better!)
Male (18-39)	Anger	这是由最近还是以以前事情引起的？ (Was this caused by a recent or distant event (or events)?)	谢谢，我想知道您认为您生活中的什么事件引发了这种感觉。这是最近发生的事件还是更遥远的过去？ (Thank you, I would like to know if there's a particular event in your life that's caused these feelings. Is this a recent event or something more distant?)
Female (18-39)	All emotions	我很抱歉。请从下面的情绪中选择最能反映您感受的情绪： (I'm sorry. Please select from the emotions below the one that best reflects what you are feeling)	哦，真对不起。您能选择您感受到哪种情绪，或者最接近它的一种情绪吗？ (Oh no, I'm so sorry. Could you select which emotion you're feeling, or the one closest to it?)

Table 3.11: Example GPT2 generated responses for 4 human personas across a variety of emotions, with the original text and native level translation below. We note the different figures of speech used by different personas as well as the different levels of empathy expressed in the responses.

Rewriting of Low Empathy Utterances as Bases

We also investigated the potential of performing empathetic rewriting using the 1364 utterances that were classified as displaying empathy levels of either ‘Little to None’ or ‘Weakly Empathetic’. Their classified levels of empathy made them unsuitable for use as training data on our task for generation purposes; however, we experimented with using low-empathy utterances as bases for empathetic rewriting. The resulting utterances showed generally lower levels of empathy than the base set, with only 85.26% (1163/1364) utterances classified as ‘Strongly Empathetic’. Additionally, inspection of the generated utterances showed a number of instances where the model generated fluent and empathetic text that were unrelated to the given base prompt. We hypothesise this to be related to ineffective usage of the semantic information in the base prompt, with the model only utilising the base utterance for indicating the

type of desired rewriting.

Consequently, the current model may be ill-suited to open prompting, and future work should enforce explicit semantic equivalence between the base and rewriting towards facilitating effective empathetic rewriting of general text. Nevertheless, our current setup is still an effective means of significantly increasing the number of possible empathetic responses from the original 780 whilst maintaining a good level of empathetic expression (Table 3.9).

3.4 Conversation Flow

The conversation flow used by the chatbot is primarily based on the conversation flow used in the previous work (1), with some structural changes to facilitate more appropriate response generation. Example conversations are provided with translations in Appendix C.

Once the user has logged into the platform, they are prompted to select a persona to converse with, out of 4 human personas (Robert, Gabrielle, Arman and Olivia) and ‘Kai’, a more abstract persona whose response pool is composed of the other 4. The chatbot would then load the response pool associated with the select persona, allocating a subset to be produced as response candidates (Figure 3.8).

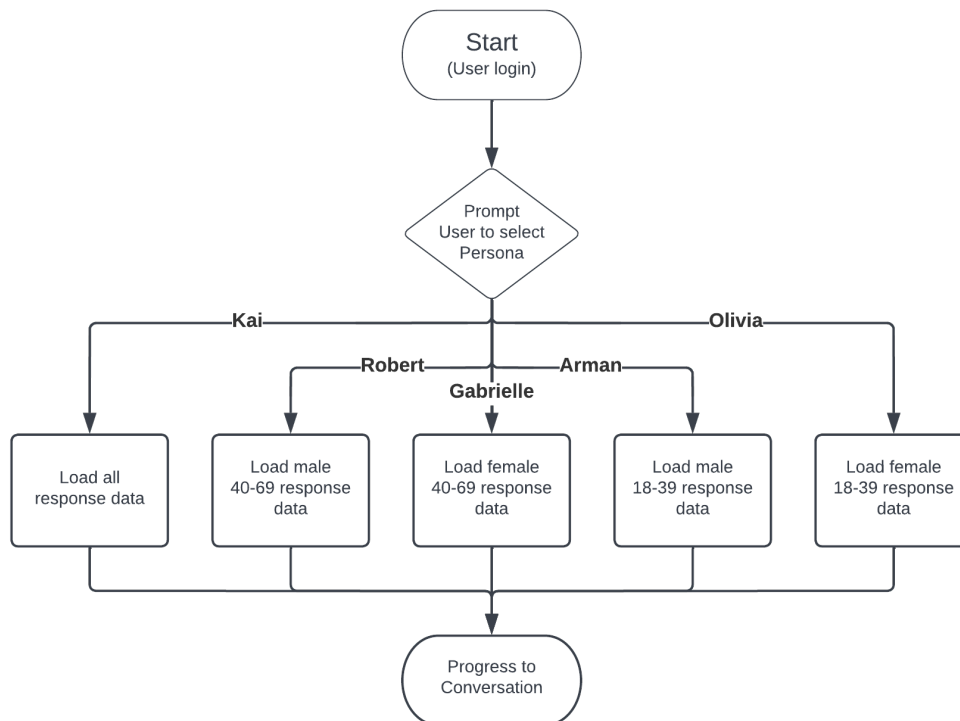


Figure 3.8: The model loads responses depending on the user’s choice of persona.

Once a persona has been selected, the chatbot progresses to the conversation phase, starting with emotion classification. As seen in Figure 3.9, the conversation flow mainly consists of two main branches dependent on user emotional state: positive emotional state (happiness/content) and negative emotional state (sadness/anger/fear/anxiety). Due to the differences in the two main conversation branches, user satisfaction is dependent in large part to the correctness of the emotion classification prediction. The chatbot will classify user emotion based on the user response to the question of ‘How are you feeling?’, the chatbot will then confirm the predicted emotion with the user before progressing onto a conversation path.

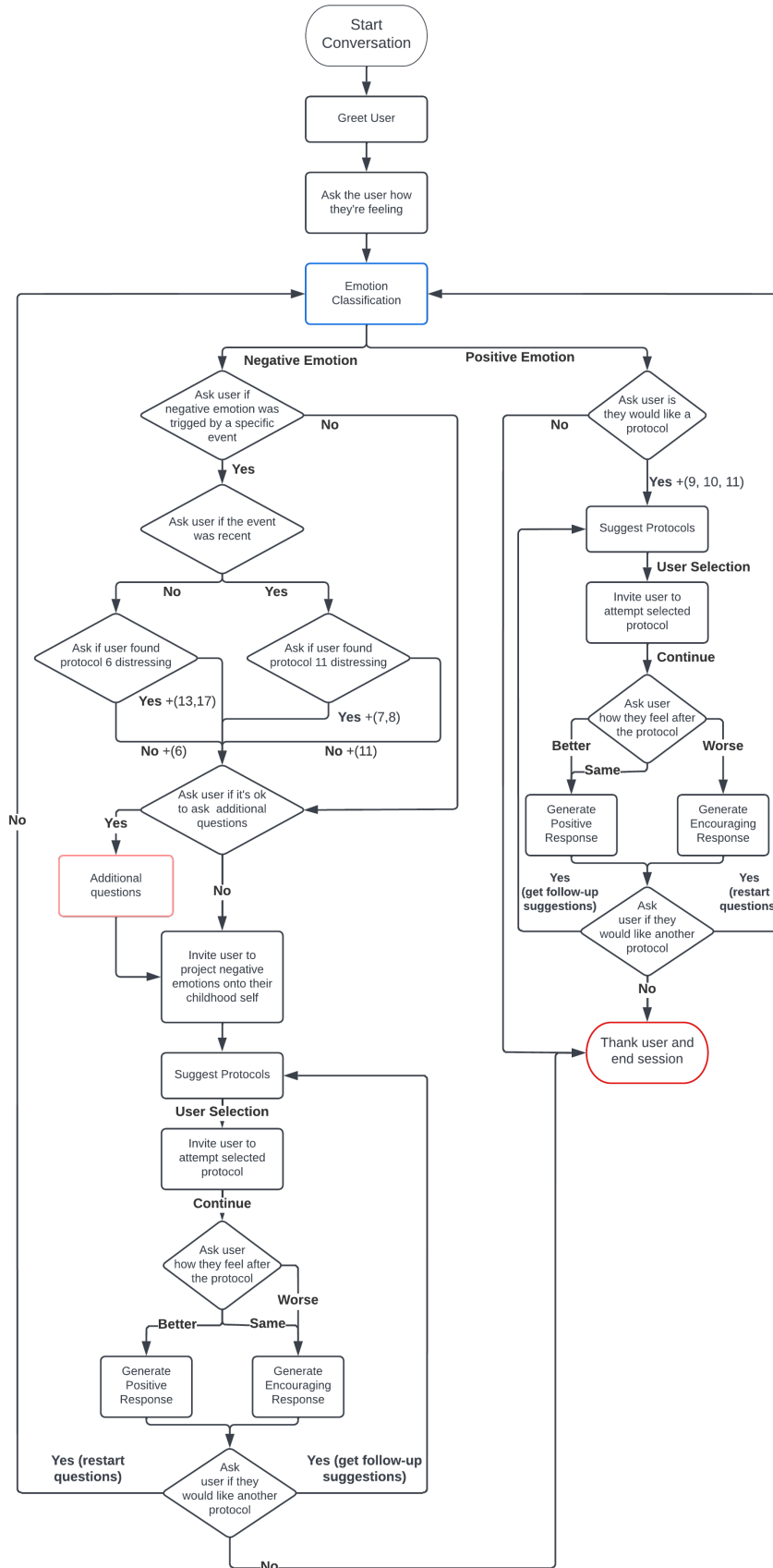


Figure 3.9: Conversation flow for the chatbot, where $+(n, \dots)$ denotes the addition of the following protocols into the suggestion pool.

In the negative emotion branch, the chatbot can, with the user's permission, ask a series of randomly selected questions to ascertain additional context for appropriate protocols. Questions are selected from a pool without replacement in random order until the user answers 'yes' to a question or all additional questions have been exhausted. Figure 3.10 illustrates this process, Table 3.12 shows the additional questions that can be presented to the user. All chatbot responses are generated conditioned on the respective persona, and thus can be articulated differently. This includes the additional questions, in which generated rewritings are displayed to the user.

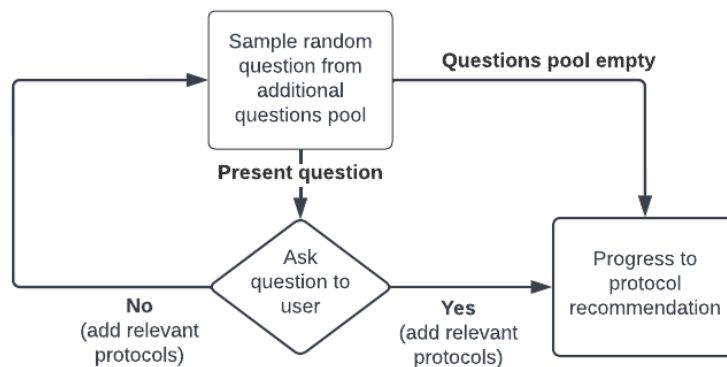


Figure 3.10: Random additional question retrieval strategy

Base question	Action on positive response	Action on negative response
Have you strongly felt or expressed any of the following emotions towards someone: envy, jealousy, greed, hatred, mistrust, malevolence, or renegefulness? Do you believe that you should be the saviour of someone else?	Add protocols 13, 14, progress to recommendation Add protocols 8, 15, 16, 19, progress to recommendation	Add protocol 13, sample next question Add protocol 13, sample next question
Do you see yourself as the victim, blaming someone else for how negative you feel?	Add protocols 8, 15, 16, 19, progress to recommendation	Add protocol 13, sample next question
Do you feel that you are trying to control someone?	Add protocols 8, 15, 16, 19, progress to recommendation	Add protocol 13, sample next question
Are you always blaming and accusing yourself for when something goes wrong?	Add protocols 8, 15, 16, 19, progress to recommendation	Add protocol 13, sample next question
Is it possible that in previous conversations you may not have always considered other viewpoints presented?	Add protocols 13, 19, progress to recommendation	Add protocol 13, sample next question
Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?	Add protocols 13, 19, progress to recommendation	Add protocol 13, sample next question

Table 3.12: Additional questions from which the chatbot samples to obtain additional context of negative emotions to aid in protocol recommendation. All questions undergo empathetic rewriting before being presented to the user.

3.5 Web Platform and User Interface

The chatbot is accessible to users through a web platform. We base the design of the web interface largely on the previous work (1) to allow for visual coherence across different language implementations of the chatbot.

The web platform consists of a ReactJS frontend with a Flask backend, accessible using web browsers, the exact architecture is shown in Figure 3.11. Despite visual similarity with the previous work (1), the frontend and backend components were rewritten from the ground up to accommodate a different response generation scheme, with the aim of improving the ease with which new languages and conversation logic can be added in the future.

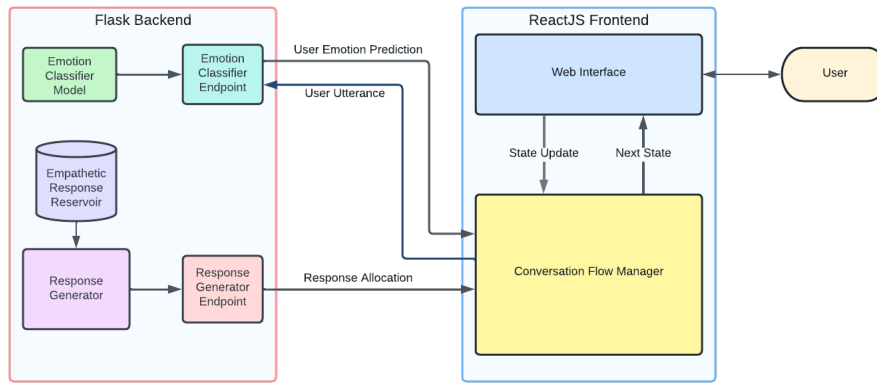


Figure 3.11: Web platform architecture.

We leverage the react-chatbot-kit library (34) for the chatbot component of our platform, with the addition of some custom widgets to facilitate persona-based conversation. Figures 3.12 and 3.13 shows the visual appearance of the web interface as shown to users during evaluation trials.

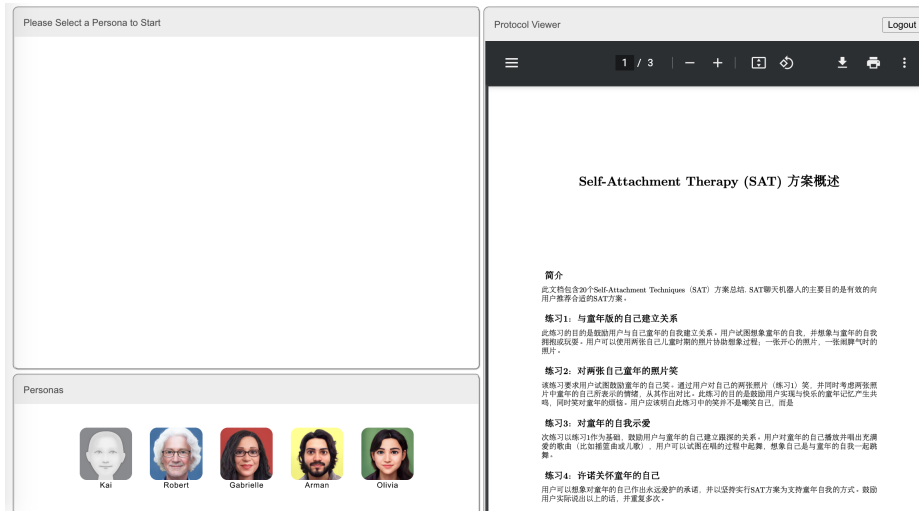


Figure 3.12: Web interface prior to persona selection.

A viewing panel is added alongside the chatbot interface to allow the user to simultaneously view information regarding the protocols during conversation. A persona selector is also present to allow for the users to select a persona. Based on feedback from trial participants of the previous work (1), the persona selector will highlight the selected persona, in addition to showing the persona's avatar beside the chatbot text bubbles, to give the selected persona additional prominence, as some previous trial participants reflected being unsure which persona was selected during conversation. Figure 3.13 shows this in action.



Figure 3.13: Web interface after persona selection, engaged in conversation.

In addition, a login system was implemented to control access to the web platform, with trial participants being allocated unique keys prior to the start of the trial. The login interface is shown in Figure 3.14.

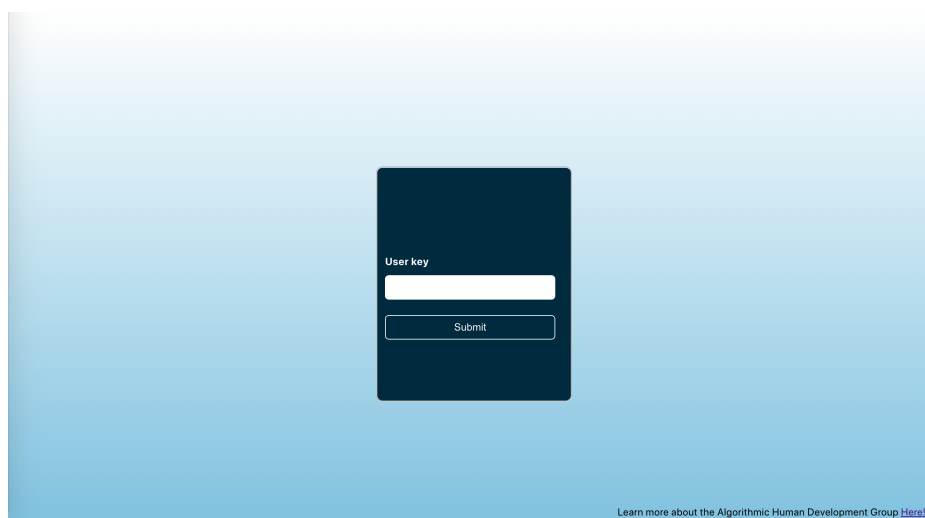


Figure 3.14: Web platform login interface.

Chapter 4

Evaluation

In this section we provide details of the non-clinical trials used to evaluate our chatbot, as well as qualitative analysis of the user responses. At the end of this section, we summarise the limitations of our trial, and provide suggestions for improvements in future work.

4.1 Non-clinical Trial Setup

4.1.1 Overview

We formally evaluate our chatbot through a human trial. The trial aims to recruit participants from the non-clinical population with reading fluency in simplified Chinese, and knowledge of the SAT protocols. Participants are required to interact with the chatbot for a minimum of 30 minutes each day over the course of 5 days. Participants are requested to note down any mistranslations or errant sentences encountered during the trial to provide as feedback to aid in improving the chatbot responses. At the end of the trial period, participants are sent a feedback questionnaire with Likert scale questions evaluating the performance of the chatbot, and open questions for broader feedback. Participant input during interaction with the platform is not recorded and the responses to the feedback questionnaires are anonymised.

Prior to the start of the trial, the chatbot platform is deployed as a web application accessible via web browsers. Each participant is given a unique key with which they can access the web platform.

We conduct a total of two trials, of equal lengths, with the aim of using participant feedback from the first trial to devise improvements to the chatbot, the effectiveness of which would be evaluated in the second trial. Due to the small size of the population with both native level fluency in Chinese and knowledge of SAT protocols, our recruitment procedure prioritises language fluency; with participants being issued information documents regarding self-attachment with the initial correspondence to maximise the time they have to familiarise themselves with SAT prior to the start of the trial. As of the time of writing, 17 prospective participants have consented to participate in the study, of which 15 were deemed eligible after the screening pro-

cess and issued access to the platform. Our 15 trial participants across the two trials consisted of 10 female participants and 5 male participants aged in between 18-60. Due to the anonymised nature of the trial responses, the specific demographic of each trial is unknown.

4.1.2 Participant Screening

Due to certain SAT protocols requiring users to recall childhood memories, which can lead to users reliving painful childhood traumas, we screen prospective trial participants for severe depression or anxiety conditions to ensure participant suitability to undertake our intervention. To this end, we use the Chinese version of the Beck's Depression Inventory (CBDI) and Beck's Anxiety Inventory (CBAI) which have been used for screening in both domestic and international Chinese populations (6, 59, 58). We omit any participants satisfying the exclusion criteria from participating in the trial.

4.1.3 Evaluation Questionnaire

The feedback questionnaire allows for the trial participants to give feedback regarding their experience in interacting with the chatbot. The focus of the survey is to gauge the participant's observations regarding several key points (i) whether the chatbot was able to accurately determine the participant's emotion during interactions; (ii) how well did the chatbot display empathy during interactions; (iii) the quality of the chatbot responses; (iv) was the conversation with the chatbot engaging; (v) how was the chatbot-delivered SAT experience as a whole.

The survey response provides an idea of the capability of the chatbot in providing empathetic conversation, the efficacy of using machine translated data as a basis for response generation, and whether such a platform benefited participants in a psychotherapy context. This will also allow for direct comparison between this chatbot and previous works conducted by previous students. Participant responses to the feedback questionnaires are anonymised, and completed through a combination of open questions and Likert scale questions; wherein participants select answers from multiple-choice questions best corresponding to their opinions. Specific participant responses are provided in Appendix A where applicable. Qualitative analysis of the questionnaire responses is provided in Section 4.2, along with examples of participant textual responses from the trials.

4.2 Participant Feedback

We analyse the participant responses to the evaluation questionnaire collected at end of both trials, and compare them, where applicable, to results obtained in the previous evaluation trial (1). The exact distribution of participant responses in both trials can be found in Appendix A.

4.2.1 Model Revisions

Feedback was collected from trial participants at the conclusion of the first trial, which led to some changes made prior to commencing the second trial focused primarily on improving the quality of generated responses. Participant feedback highlighted certain patterns of semantic errors; mostly a result of English figures-of-speech translated directly to Chinese whilst maintaining syntactic correctness. Certain participants also reflected a lack of Chinese colloquialisms, making the responses sound less natural (Figure 4.1).

有不少“我很抱歉听到这个”，应该“听到这个我很遗憾”更符合
(There's quite a few [literal translation of 'I'm sorry'], should be more [I'm sorry' in an empathetic sense])

根据场合的不同，或者可以设计一些更趋于口语的词句？这样不会感觉特别拘谨
(The chatbot maybe could have benefited from some slang/colloquialisms based on the conversation context. This will help the conversation to feel less formal.)

Figure 4.1: Examples of participant textual feedback identifying translation errors or unnatural sounding text.

A second pass of the EmpatheticPersonasZH dataset was conducted (Section 3.1) with the aim of addressing the issues identified in the first trial. Erroneous patterns were identified and replaced with human translation, as well as in-language colloquialisms. The updated revision of the response data is referred to as ‘v2’ here, and was used to retrain the response generator model to produce an updated response set.

4.2.2 Results

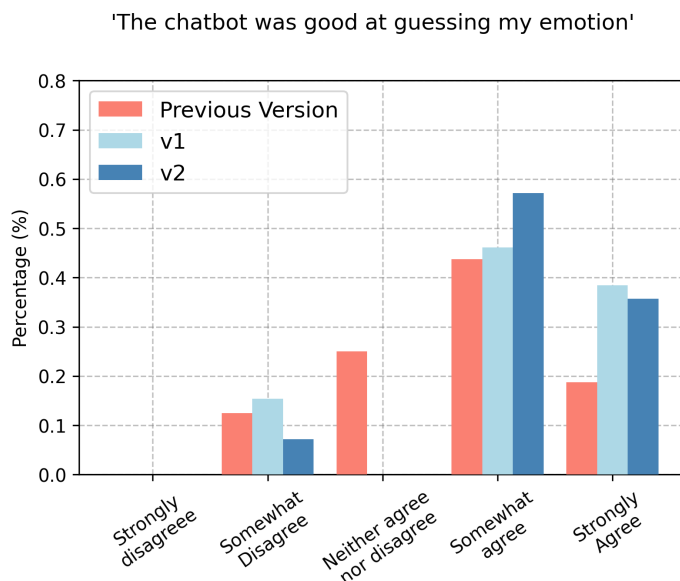


Figure 4.2: Emotion classification evaluation of the two revisions presented in the two trials compared to the previous version. We note the improvement in perceived emotion recognition performance, but also the portion of participants who disagreed.

Figure 4.2 shows participant responses when asked to evaluate the effectiveness of the platform at ‘guessing their emotion’. 84% of participants agreed that the chatbot was good at guessing their emotions in the first trial, and 92% agreed in the second trial. This is a significant increase over the previous work at 62%, which appears to stem primarily from the reduction of participants reporting ‘Neither agree nor disagree’.

However, the portion of participants disagreeing to some extent on the above notion (15% for v1, 7% for v2) remains similar to the previous work (13%). Open text participant feedback from the participants (Figure 4.3) suggested this sentiment was due to finding emotion classification being limited to four emotions to be too narrow to be an appropriate reflection of their emotional state, which can also commonly be composed of a multitude of different emotions. This was also part of feedback received in the previous trial, and is currently part of a project worked on by other students, which is why it was not tackled in this project. Nevertheless, we do provide suggestions to tackle this problem in Chapter 5.

情绪分类比较准确，但是类型不太多，所以不太符合人的真实情绪，总体而言挺有意思的
(Emotion classification is quite accurate for the types of emotions, but there are not many types, so the final classification doesn't entirely match real human emotions. Overall quite interesting)

应该有多种情绪才更真实，而且人的情绪是复杂的，并不是可以确切分类的
(There should be more emotions to be more realistic; also, human emotions are complex and not completely discretised.)

只把感情归类到四种有点简单。比如心烦/烦躁，AI 识别为愤怒，其实归类不太准确，心烦更偏向于是焦虑
(Limiting emotions to only one of four categories, such as frustration being classified as anger, is somewhat of an oversimplification, and isn't entirely accurate, as frustration can lean more towards anxiety depending on the context.)

Figure 4.3: Examples of participant textual feedback regarding present issues with emotion classification.

Trial participants were asked to evaluate the ability of the chatbot to demonstrate empathy in conversation by selecting their level of agreement with the phrase ‘When interacting with the personas, I found that they displayed empathy in their responses throughout the conversation’.

We observe from Figure 4.4 that while 85% of trial participants agreed with the statement, 14% selected ‘Neither agree nor disagree’. This distribution is consistent with results from the previous work of 81% and 18% respectively, although our version has a higher portion of participants who selected ‘Strongly Agree’ (29% vs 16%). This implies that under the currently undertaken procedures, the level of empathy expressed in the source data can be maintained across translations from English to Chinese with limited application of rewriting focusing on remedying syntactic and semantic errors.

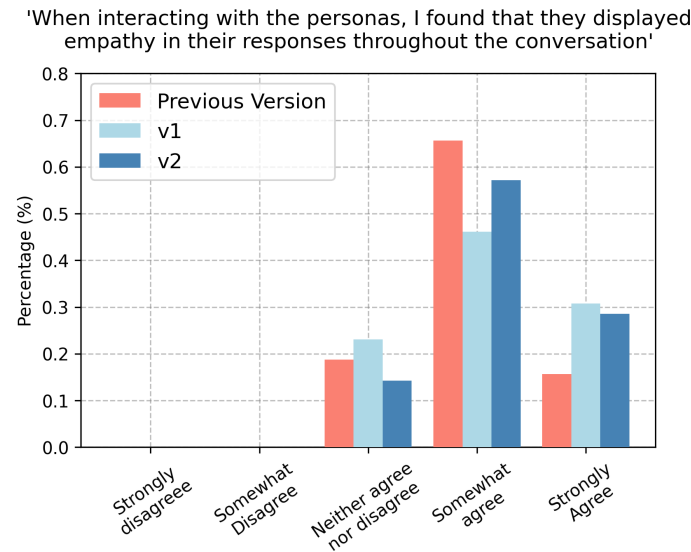


Figure 4.4: Participant perception of empathy displayed by the chatbot during interaction. We note a similar distribution compared to the previous version and the lack of disagreement over the chatbot displaying empathy.

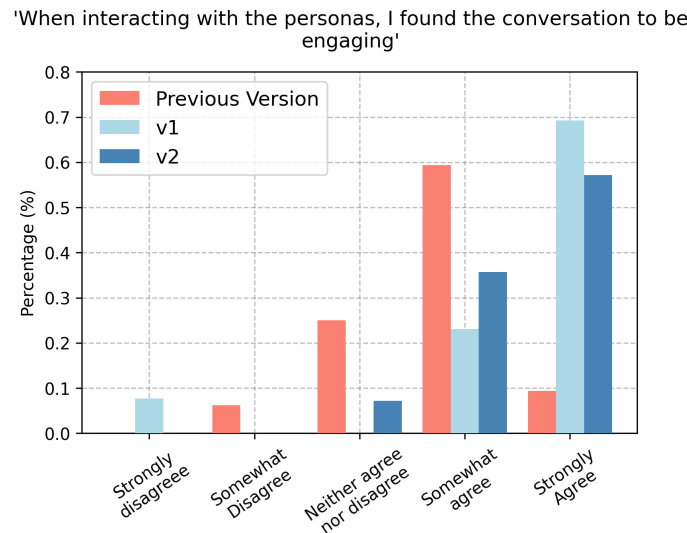


Figure 4.5: Participant engagement levels from interaction with the chatbot. Comparison with the previous work shows significant increase in user engagement, though we note an isolated instance from the first trial wherein a participant reflects significantly poor engagement.

User engagement was evaluated as part of the participant feedback with results shown in Figure 4.5. 92% of participants across both of our trials reported varying degrees of agreement with the phrase: 'I found the conversation to be engaging', significantly higher than the 68% shown in the previous work. The remaining 7% of participants selected 'Neither agree nor disagree' in the second trial, whilst the participant(s) in the first trial selected 'Strongly Disagree'.

Due to the anonymised nature of the study, we are unable to confirm whether trial participants from the first trial completed the feedback questionnaire for the second trial. However, the feedback provided with the questionnaire from the first trial

suggested the reason for the poor engagement was due to a lack of open dialogue beyond the initial emotion recognition (Figure 4.6), where the participant felt too constricted in having to select conversation options. This sentiment was also raised in the feedback to prior work, and again, is currently subject to work conducted by another student concurrent to this project, which is why it was not tackled in this project. We offer suggestions for future work to tackle this in Chapter 5.

只有第一句话可以自由输入，其余的沟通都是选项，不是很自然
(Open conversation is only permitted at the start of the conversation, all other actions are choice-based, not very intuitive.)

希望能多聊一些 (Wished I could converse with the chatbot more)

比如问题：有什么特别的事情引起了你的感受吗？我一般反应就是回答某某事情，而不是选择是/否。结果回答的就不能被识别。最后只能选择是/否
(For questions such as ‘was this caused by a specific event’, I would normally answer with a description of the event(s), rather than discretising to a yes/no answer. But open conversation is not used at these states, so only a binary choice is available.)

Figure 4.6: Examples of participant textual feedback regarding the lack of open conversation capabilities that hindered user engagement.

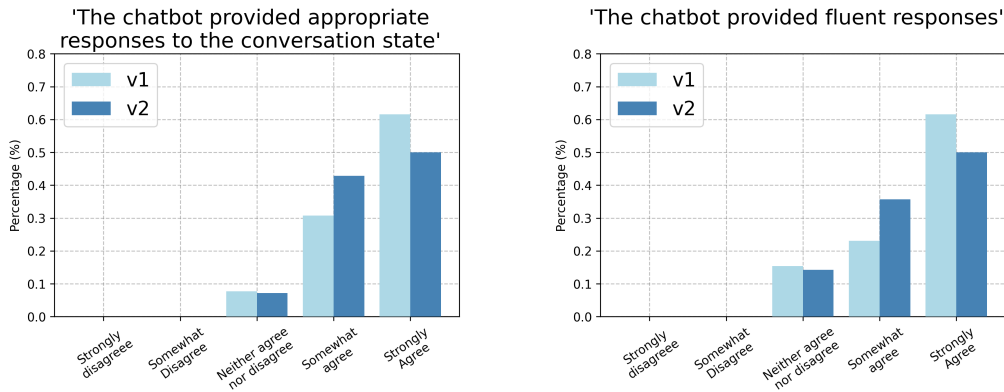


Figure 4.7: Participant evaluation of the quality of chatbot responses with regards to response fluency and how appropriate the responses were to the conversation state.

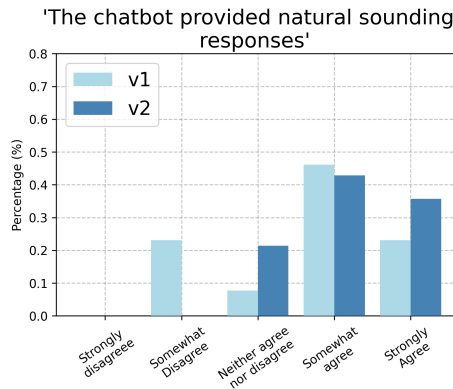


Figure 4.8: Participant evaluation of the quality of chatbot responses with regards to the naturalness of the responses. We note of mixed perception in trial 1 of the naturalness of chatbot responses and an improvement in reception in trial 2.

Participants were asked to evaluate the quality of the chatbot responses as part of the study by giving feedback on whether the chatbot provided responses ‘appropriate’ their prior conversation, whether the responses were ‘fluent’ and ‘natural sounding’, results are shown in Figures 4.7 and 4.8 for both trials. Participants response for both revisions of the chatbot reflected similarly with regards to the appropriateness and fluency, with 92% of participants in the first trial agreeing that the provided responses were ‘appropriate’, and 85% agreeing to ‘fluency’; results for the second trial were 92% and 86% respectively.

Participant reception for ‘natural sounding’ responses was far more mixed, with only 23% of participants selecting ‘Strongly Agree’ compared to 23% who selected ‘Somewhat Disagree’. From participants’ textual feedback (Appendix B), this appears to stem from two primary issues:

Issue 1.: Semantic errors which were not corrected in the initial edit to the response dataset due to sentences being grammatically correct.

Issue 2.: A lack of in-language colloquialisms which leads to responses feeling overly formal or robotic.

In issue 1, common semantic errors include direct translations of English figures-of-speech into Chinese, which, whilst grammatically correct, does not entirely sound natural in Chinese (Figure 4.1). More examples of such types of mistranslations can be seen in Appendix B. This issue was also reported by another student undertaking this project in another language, suggesting this to be an inherent issue with using machine translated response data.

In response to the trial feedback from the first trial, further edits were made to the EmpatheticPersonasZH dataset (Section 3.1). For issue 1, the semantic patterns identified in the first trial was targeted, and rewritten in line with more natural sounding speech. Where applicable, phrases reflecting colloquialisms in English were replaced with Chinese counterparts. In certain instances, this would lead to significant change in sentence structure/order in order to produce a more natural sounding sentence (issue 2.). The resulting ‘v2’ revision of the dataset was then used to train a new generative model and produce a different response pool for use in the second trial. The results from both trials are evaluated against one another in Figures 4.7 and 4.8.

We note a significant reduction in negative responses in the second trial, where no participants disagreed with the statement ‘The chatbot provided natural sounding responses’ compared to the previous 23%. We do, however, note that 21% of participants selected ‘Neither agree nor disagree’, suggesting that, whilst the changes made to the dataset improved user reception, it may still not be equivalent to entirely human-level text. The portion of participants agreeing with the sentiment increased in the second trial to 79% compared to the previous 69%.

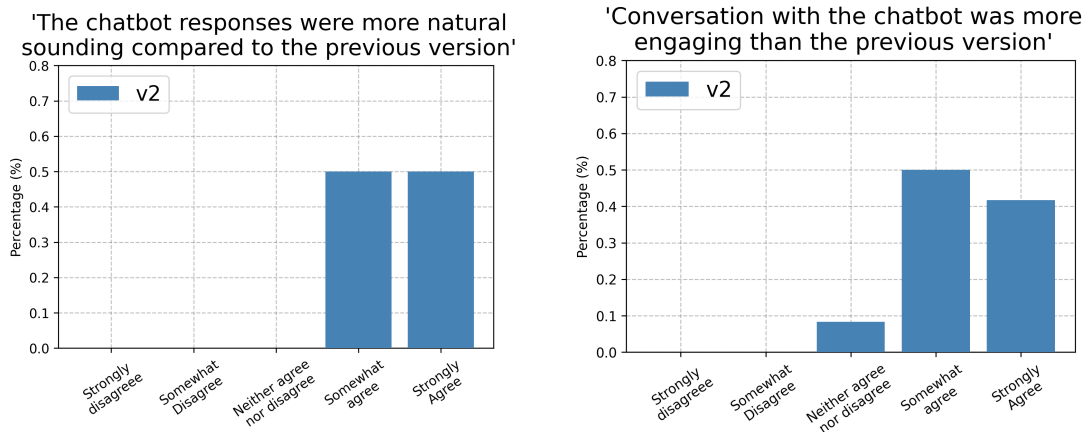


Figure 4.9: Feedback from participants of both trials regarding the relative quality of conversation of the second revision of the model against the first.

Figure 4.9 shows responses from the second trial by participants who participated in both trials. Results show overwhelmingly that the edits improved how ‘natural sounding’ the response were and improved perceived user engagement. Curiously, despite participants perceiving a higher degree of engagement, the user engagement figures shown in Figure 4.5 shows weaker expression of engagement, as 57% of participants of the second trial selected ‘Strongly agree’ compared to 69% from the first trial. Overall, 92% of trial participants responded in agreement in both trials. The difference in level of perceived engagement across both trials could be due to differences in specific participants across the two trials, as we note trial two had two new participants and one participant who took part in the first trial but not the second.

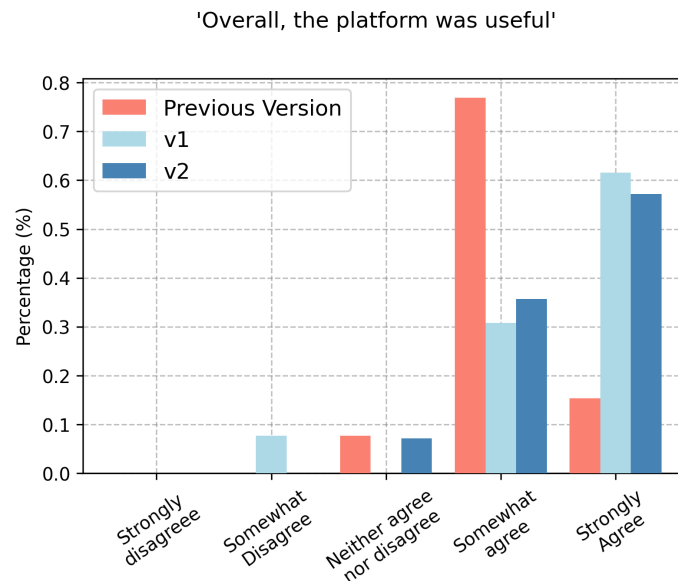


Figure 4.10: Participant feedback on the usefulness of the current implementations of the chatbot compared to the previous work. Note the positive perception to the usefulness of the platform.

Overall usefulness of the platform was evaluated by trial participants according to their agreement with the sentence ‘Overall, the platform was useful’. 92% of participants responded in agreement to the sentiment with varying strength in both trials, compared to 92% of the previous work. 62% of trial participants of our second trial selected ‘Strongly Agree’ compared to 15% from the previous work.

4.3 Limitations of the Study

Due to the nature of the project, study participants from the non-clinical population required reading fluency in simplified Chinese and knowledge to carry out SAT protocols. However, this is a very small population, and, as a result of prioritising reading fluency in simplified Chinese, a significant portion of the trial participants had little to no prior experience with SAT protocols. We attempt to remedy this issue by compiling a simplified Chinese document guide to SAT protocols and presenting it, alongside original English guides and tutorials as a comprehensive SAT information package to participants up to three weeks prior to the start of the first trial. We encouraged participants to familiarise themselves with the 20 SAT protocols by attempting them prior to the start of the trial. Despite this, it is likely that the general participant familiarity with SAT protocols during our trial was lower than that of participants in the previous study (1). Future studies should allow participants a longer period of time, with more significant guidance in carrying out SAT protocols prior to the start of the trial to ensure that all participants have a good understanding of practising SAT protocols.

In total, 17 volunteers completed the consent form to participate in our study, of which, 15 overall participated in at least one of the trials. Of the 15 total participants, 5 were male, 10 were female. Future studies should ideally be conducted on a larger sample size balanced to reflect population demographics. Though we note, due to the aforementioned issues regarding requiring participants to have both in-language reading fluency and prior familiarity with SAT, this may be infeasible in the immediate future without an additional means of teaching participants to practise SAT protocols.

We conducted two trials of five days each, with trial participants encouraged to interact with the platform for at least 30 minutes on each of the five days. Whilst we note that participant feedback on engagement and unchanged components such as emotion classification seemed consistent across both trials, participant reception may be influenced by the perceived novelty of the platform. Future studies should encourage additional interactions over a longer period to allow participants to build a more informed opinion of the platform’s qualities. Additionally, we requested participants to note down their opinions across interactions to report in the feedback questionnaire at the end of each trial period. However, from observation of questionnaire responses, it appears certain participants did not abide by this, potentially due to the length of the trial period. It may be beneficial in future studies to supplement the final feedback questionnaire with incremental feedback after each interactive session i.e. a rating out of 5 stars for fluency or engagement in said conversation, as a portion of participants may write their feedback off their impressions off of only

the last few interactions instead of across the entire trial period.

We note that whilst a similar portion of participants agreed to having found conversation with the chatbot engaging in both trials, a smaller portion of participants selected ‘Strongly Agree’ in the second trial compared to the first (Figure 4.5). Due to the anonymised nature of feedback from both trials, we are unable to directly compare individual participant sentiment between the trials. Being able to view participant engagement across different versions to associate different feedback with changes in engagement, could prove informative on the changes necessary to improve user engagement over a longer period. This could be achieved by assigning each trial participant a unique key at the conclusion of the first trial, then requiring the participant to input the key in subsequent trials. Doing so could preserve participant anonymity whilst giving an insight into participant sentiment over time.

Chapter 5

Conclusions and Future Work

Our framework builds upon previous work (1) to provide a foundation for future multilingual development to deploy computational methods for mental health support to more languages, with a more scalable pipeline and less reliance on in-domain, in-language data. The human trials of our platform showed promising results with respect to the perceived empathy, usefulness, user engagement, quality of responses and ability to correctly identify user emotion. Nevertheless, limitations of the trial, along with feedback received have revealed multiple areas where future work could be conducted to make an impact.

Following on from limitations of the study identified in Section 4.3, future studies can be conducted with a large cohort of participants with in-language reading fluency and prior experience in SAT, or given ample time to gain familiarity with practising SAT protocols. Future studies should also ideally be conducted over a longer period of time to allow participants to form a more informed impression of the platform, and to identify any issues arising from longer-term interaction with the system. To alleviate to a certain degree the requirement for participants with prior knowledge of SAT and its protocols, it is worth exploring a tool for teaching participants about SAT that is integrated into the platform. Future work in this area could leverage the existing documentation for self-attachment techniques and large pretrained language models as knowledge bases (37) to produce a QA component. This could be beneficial as a portion of trial participants have highlighted a lack of interaction with the platform during the period where users are expected to carry out the protocols.

Despite participant perception of the quality of rewriting being higher in the second trial following further edits, we note that the edits made in both versions of the translated dataset were conducted principally by a single individual, who characterises only one of the personas available in our platform. Whilst effort was made to replicate the tone of speech present in the original data when performing the manual edits, it is nevertheless possible that the tone of the colloquialisms used in the second version lacks diversity with respect to the other persons. It is advisable that future work should attempt to perform the rewriting task with more people, of different ages and sexes to ensure that the diversity of personas is preserved across translations.

Two principal issues identified by trial participants are the limited set of emotions

recognised by the emotion classifier, and the lack of capability for open conversation. Both issues were also identified in the previous work, and have subsequently resulted in projects pursued by other students concurrently to this one. However, both of the solutions currently remain English-only, and future work could investigate the potential for such systems to perform in a multilingual setting.

Our emotion classifier model outperforms previous SOTA (+1.52% Acc., +1.54% F1) on monolingual performance, we recommend future work to investigate the performance of the same model on more languages, and the feasibility of deploying emotion classification for different languages using the same model. An extension to the project could also be conducted to expand the range of emotions recognised by the classifier, and to investigate the efficacy of formulating the emotion classification task as a multi-label classification task, as human emotions can be complex and are typically not mutually exclusive. Additionally, the current emotion classifier model is quite large (2.1G), which resulted in a small number of trial participant complaints of latency. Future work could also investigate the potential of distilling the large emotion classifier model (21, 51) and the corresponding impact on multilingual and monolingual performance.

Finally, future work can be conducted into increasing the depth of conversation through introducing the capability for open conversation with the chatbot. We have previously discussed the toxicity of large, pre-trained language models (Section 2.5), however, recent work on debiasing large pre-trained LMs (32) have shown promise in reducing language model bias whilst maintaining comparable downstream task performance. Future work could leverage such debiasing techniques to safely generate more natural sounding conversation. Recent works also include methods such as: prompting (5), style transfer (27), rationale-based generation (31) and strategic empathetic generation (49) that can potentially be used to achieve more natural sounding, empathetic conversation. It is paramount, however, that future work ensures response generation remains safe for the users due to the sensitive nature of computational methods for mental health support, and that appropriate considerations are made in this regard.

Chapter 6

Ethical Considerations

This project focuses on technologies applied in mental health settings, and as such, presents a variety of legal, social, and ethical issues.

The human trial of the chatbot collects participant feedback for the purpose of evaluation, during the process of which certain participant data may be collected that qualifies as personal data under the UK Data Protection Act (18) and the UK General Data Protection Regulation (GDPR). Measures were taken at every stage to ensure that the collection of any participant personal data was kept to a minimum, and that any personal data collected should be held for no longer than is necessary for its processing. The collection and processing process of any collected personal data were conducted in adherence to the processes outlined in the aforementioned regulations. Participants were also made aware prior to the start of the trial of any data that will be collected. We collect participant personal data only when strictly necessary, such as collecting participant consent. We do not store user inputs to the platform, which may contain sensitive information, and under no circumstances do we, during interaction, store persona information or insights such as geolocation data, IP or MAC addresses, IMEI codes or any other type of metadata from the users' devices.

A recent pilot study has shown promising results for self-attachment therapy in the treatment of chronic anxiety and depression in Iranian women(14), however, it is currently untested for other serious mental health conditions or on a different demographic. Trial participants are screened for severe anxiety/depression and a history of severe mental health conditions prior to involvement in the trial to ensure that the platform will be safe for them to use. Certain protocols also require users to recall past events, which may be unsuitable to users who suffered from childhood trauma. Care was taken to screen trial volunteers to ensure safe participation in the human trial, and participants are given the option to avoid being recommended certain protocols that may worsen their condition.

Considerations were also taken to ensure that the data used in the creation of the chatbot is sourced ethically and in accordance with data protection laws, this applies to the EmpatheticPersonas dataset used by the previous student which was approved by the Research Ethics Committee of Imperial College London; and any other publicly available dataset used as supplementary training data.

Ethical approval

The non-clinical trial for the evaluation of this SAT chatbot received ethical approval from Imperial College's Research Ethics Committee. In accordance to the current regulations, we limit information collection to the minimum required for the successful completion of the study, including anonymising trial feedback. Participants are made aware, prior to giving consent to participation, of which information will be collected, how it will be handled, and withhold the right to withdraw from the trial at any time.

References

- [1] Lisa Alazraki (2021). A deep-learning assisted empathetic guide for self-attachment therapy. In *MSc Thesis, Department of Computing, Imperial College London*.
- [2] Lisa Alazraki, Ali Ghachem, Neophytos Polydorou, Foaad Khosmood, and Abbas Edalat (2021). An Empathetic AI Coach for Self-Attachment Therapy. In *Proceedings - 2021 IEEE 3rd International Conference on Cognitive Machine Intelligence, CogMI 2021*, 78–87.
- [3] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama (2020). On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4623–4637.
- [4] Soumya Barikeri, Anne Lauscher, Ivan Vulic, and Goran Glavaš (2021). RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 1941–1955.
- [5] Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren (2022). Fine-Grained Controllable Text Generation Using Non-Residual Prompting. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:6837–6857.
- [6] Lu Chen, Lin Wang, Xiao Hui Qiu, Xiu Xian Yang, Zheng Xue Qiao, Yan Jie Yang, and Yuan Liang (2013). Depression among Chinese University Students: Prevalence and Socio-Demographic Correlates. In *PloS one*, 8:e58379.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). A Simple Framework for Contrastive Learning of Visual Representations. In *37th International Conference on Machine Learning, ICML 2020*, 1575–1585.
- [8] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou (2020). InfoXLM:

- An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3576–3588.
- [9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2020). Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.
- [10] Alexis Conneau and Guillaume Lample (2019). Cross-lingual Language Model Pretraining. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 7059–7069.
- [11] Michael Daly, Angelina R Sutin, and Eric Robinson (2020). Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK Household Longitudinal Study. In *Psychological Medicine*, 1–10.
- [12] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.
- [13] Abbas Edalat (2015). Introduction to Self-attachment and its Neural Basis. In *International Joint Convergence on Neural Networks (IJCNN)*, 1–8.
- [14] Abbas Edalat, Massoumeh Farsinezhad, Makhsoos Bokharaei, and Fateme Judy (2022). A Pilot Study to Evaluate the Efficacy of Self-Attachment to Treat Chronic Anxiety and/or Depression in Iranian Women. In *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 6376*, 19:6376.
- [15] Robert Elliott, Arthur C. Bohart, Jeanne C. Watson, and David Murphy (2018). Therapist Empathy and Client Outcome: An Updated Meta-Analysis. In *Psychotherapy*, 55:399–410.
- [16] Kawin Ethayarajh and Dorsa Sadigh (2020). BLEU Neighbors: A Reference-less Approach to Automatic Evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, 40–50.
- [17] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–3369.
- [18] UK Government (2018). Data Protection Act 2018, Queen’s Printer of Acts of Parliament, *Date Accessed: 21/01/2022*.
URL <https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted>

- [19] Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci (2022). CrossAligner & Co: Zero-Shot Transfer Methods for Task-Oriented Cross-lingual Natural Language Understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, 4048–4061.
- [20] Milan Gritta and Ignacio Iacobacci (2021). XeroAlign: Zero-Shot Cross-lingual Transformer Alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 371–381.
- [21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean (2015). Distilling the Knowledge in a Neural Network. In *arXiv preprint arXiv:1503.02531*.
- [22] Junjie Hu, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig (2021). Explicit Alignment Objectives for Multilingual Bidirectional Encoders. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3633–3643.
- [23] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (2020). XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization. In *37th International Conference on Machine Learning, ICML 2020*, PartF168147-6:4361–4371.
- [24] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria (2021). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *arXiv preprint arXiv:2110.15621*.
- [25] Tianbo Ji, Yvette Graham, Gareth J. F. Jones, Chenyang Lyu, and Qun Liu (2022). Achieving Reliable Human Assessment of Open-Domain Dialogue Systems. In *arXiv preprint arXiv:2203.05899*.
- [26] Katharina Kann, Sascha Rothe, and Katja Filippova (2018). Sentence-Level Fluency Evaluation: References Help, But Can Be Spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, 313–323.
- [27] Kalpesh Krishna, Deepak Nathani, Xavier Garcia, Bidisha Samanta, and Partha Talukdar (2022). Few-shot Controllable Style Transfer for Low-Resource Multilingual Settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:7439–7468.
- [28] Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad (2021). MTOP: A Comprehensive Multilingual Task-Oriented Semantic Parsing Benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2950–2962.
- [29] Chin-Yew Lin (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, 74–81.

- [30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Paul G Allen (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv preprint arXiv:1907.11692*.
- [31] Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang (2022). A Rationale-Centric Framework for Human-in-the-loop Machine Learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1:6986–6996.
- [32] Nicholas Meade, Elinor Poole-Dayana, and Siva Reddy (2021). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In *arXiv preprint arXiv:2110.08527*.
- [33] Rory C. O'Connor, Karen Wetherall, Seonaid Cleare, Heather McClelland, Ambrose J. Melson, Claire L. Niedzwiedz, Ronan E. O'Carroll, Daryl B. O'Connor, Steve Platt, Elizabeth Scowcroft, Billy Watson, Tiago Zortea, Eamonn Ferguson, and Kathryn A. Robb (2021). Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 Mental Health & Wellbeing study. In *The British Journal of Psychiatry*, 218:326–333.
- [34] Fredrik Oseberg (2021). React-chatbot-kit, *Technical Documentation*, Date Accessed: 05/05/2022.
URL <https://fredrikoseberg.github.io/react-chatbot-kit-docs/>
- [35] Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah, Saloni Potdar, and Mo Yu (2021). Multilingual BERT Post-Pretraining Alignment. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 210–219.
- [36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.
- [37] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel (2019). Language Models as Knowledge Bases? In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2463–2473.
- [38] Betty Pfefferbaum and Carol S. North (2020). Mental Health and the Covid-19 Pandemic. In *New England Journal of Medicine*, 383:510–512.
- [39] Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder (2020). MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7654–7673.

-
- [40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). Language Models are Unsupervised Multitask Learners. In *OpenAI Blog*, 1:9.
- [41] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research*, 21:1–67.
- [42] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y. Lan Boureau (2019). Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1:5370–5381.
- [43] Sebastian Ruder, Ivan Vulić, and Anders Søgaard (2019). A Survey Of Cross-lingual Word Embedding Models. In *Journal of Artificial Intelligence Research*, 65:569–631.
- [44] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen (2018). CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3687–3697.
- [45] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, Tim Althoff, and Paul G Allen (2021). Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *arXiv preprint arXiv:2101.07714*.
- [46] Ashish Sharma, Adam S Miner, David C Atkins, Tim Althoff, and Paul G Allen (2020). A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276.
- [47] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas (2020). Counseling-Style Reflection Generation Using Generative Pretrained Transformers with Augmented Context. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*.
- [48] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang (2019). How to Fine-Tune BERT for Text Classification? In *China National Conference on Chinese Computational Linguistics*, 194–206.
- [49] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang (2021). PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1489–1503.
- [50] Brian Thompson and Matt Post (2020). Automatic Machine Translation Evaluation in Many Languages via Zero-Shot Paraphrasing. In *Proceedings of the 2020*
-

- Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 90–121.
- [51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou, and Facebook Ai (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357.
 - [52] Aaron van den Oord, Yazhe Li, and Oriol Vinyals (2018). Representation Learning with Contrastive Predictive Coding. In *arXiv preprint arXiv:180.03748*.
 - [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017-December:5999–6009.
 - [54] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush (2019). HuggingFace’s Transformers: State-of-the-art Natural Language Processing. In *arXiv preprint arXiv:1910.03771*.
 - [55] Liang Xu, Xuanwei Zhang, and Qianqian Dong (2020). CLUECorpus2020: A Large-scale Chinese Corpus for Pre-training Language Model. In *arXiv preprint arXiv:2003.01355*.
 - [56] Weijia Xu, Batool Haider, and Saab Mansour (2020). End-to-End Slot Alignment and Recognition for Cross-Lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 5052–5063.
 - [57] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 483–498.
 - [58] A. Yeung, S. Howarth, R. Chan, S. Sonawalla, A. Nierenberg, and M. Fava (2002). Use of The Chinese Version of the Beck Depression Inventory For Screening Depression in Primary Care. In *The Journal of Nervous and Mental Disease*, 190:94–99.
 - [59] A. Yeung, N. Neault, S. Sonawalla, S. Howarth, M. Fava, and A. A. Nierenberg (2002). Screening for major depression in Asian-Americans: a comparison of the Beck and the Chinese Depression Inventory. In *Acta Psychiatrica Scandinavica*, 105:252–257.
 - [60] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi (2019). BERTScore: Evaluating Text Generation with BERT. In *arXiv preprint arXiv:1904.09675*.

-
- [61] Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du (2019). UER: An Open-Source Toolkit for Pre-training Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 241–246.
- [62] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao (2020). Towards Persona-Based Empathetic Conversational Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6556–6566.
- [63] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu (2017). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 730–738.
- [64] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi (2021). Challenges in Automated Debiasing for Toxic Language Detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 3143–3155.

Appendix A

Participant Trial Feedback

A.1 First Trial

Statement	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
The chatbot was good at guessing my emotion (聊天机器人可以准确的识别我的情绪)	0/13	2/13	0/13	6/13	5/13
When interacting with the personas, I found that they displayed empathy in their responses throughout the conversation (聊天机器人在对话中表现出同理心)	0/13	0/13	3/13	6/13	4/13
When interacting with the personas, I found the conversation to be engaging (与聊天机器人对话有意思)	1/13	0/13	0/13	3/13	9/13
The chatbot provided appropriate responses to the conversation state (聊天机器人的语言符合先前对话)	0/13	0/13	1/13	4/13	8/13
The chatbot provided fluent responses (聊天机器人的对话语法正确)	0/13	0/13	2/13	3/13	8/13
The chatbot provided natural sounding responses (聊天机器人言语自然)	0/13	3/13	1/13	6/13	3/13
Overall, the platform was useful (聊天机器人总体而言是有用的)	0/13	1/13	0/13	4/13	8/13

Table A.1: Participant Likert scale feedback based on level of agreement with the given statements.

A.2 Second Trial

‘Have you participated in a previous trial associated with this project?
请问您是否曾经参加过此项目相关测试?’

Yes/是	No/否
12/14	2/14

Table A.2: Proportion of participants of the second trial who took part in the first trial.

Statement	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
The chatbot was good at guessing my emotion (聊天机器人可以准确的识别我的情绪)	0/14	1/14	0/14	8/14	5/14
When interacting with the personas, I found that they displayed empathy in their responses throughout the conversation (聊天机器人在对话中表现出同理心)	0/14	0/14	2/14	8/14	4/14
When interacting with the personas, I found the conversation to be engaging (与聊天机器人对话有意思)	0/14	0/14	1/14	5/14	8/14
The chatbot provided appropriate responses to the conversation state (聊天机器人的语言符合先前对话)	0/14	0/14	1/13	6/14	7/14
The chatbot provided fluent responses (聊天机器人的对话语法正确)	0/14	0/14	2/14	5/14	7/14
The chatbot provided natural sounding responses (聊天机器人言语自然)	0/14	0/14	3/14	6/14	5/14
Overall, the platform was useful (聊天机器人总体而言是有用的)	0/14	0/14	1/14	5/14	8/14

Table A.3: Participant Likert scale feedback based on level of agreement with the given statements.

Appendix B

Questionnaire Responses

Below we include participant open text responses from the trial feedback. Participants were given the option to complete the feedback form in either English or Chinese. '['] have been used to paraphrase certain parts of the original sentences to maintain the original sentiment in its translation, as most paraphrased sentences refer to mistakes arising from literal translations of English into Chinese, including direct translations below would not convey the original sentiment.

Please describe the overall emotional impact of your experience

请总体描述您和聊天机器人的沟通感受

- 只有第一句话可以自由输入，其余的沟通都是选项，不是很自然
(Open conversation is only permitted at the start of the conversation, all other actions are choice-based, not very intuitive.)
- 挺有意思的，某个话题能聊更深点回体验更好
(Quite fun and interesting, though certain discussions lack depth.)
- 可以通顺的对话，但是当然还是能够感觉到和真人对话的不同。对话过程没有压力，也觉得很新鲜，很有意思
(Capable of fluent and engaging conversation, but still feels different to conversations with real people. The experience is quite relaxing and feels very novel, quite fun.)
- 挺有意思及有意义的项目。沟通舒畅，却偶表达奇怪，有不像正常汉语会用的话语。总体对话感受不错，少数对话中遇到无法理解的句子
(It's an interesting and meaningful topic, the conversations were intuitive, with occasionally strange expressions, that don't sound like what you would normally use in Chinese conversation, overall quite good, there were very few occasions where it was completely incomprehensible)
- 情绪分类比较准确，但是类型不太多，所以不太符合人的真实情绪，总体而言挺有意思的
(Emotion classification is quite accurate for the types of emotions, but there are

not many types, so the final classification doesn't entirely match real human emotions. Overall quite interesting)

- 整体感受不错 (Overall experience was good.)
- 比较顺畅 (Quite intuitive.)
- 整体感受不错 (Overall experience was quite good)
- 稍微简单了点 (Conversation was somewhat basic.)
- 比较顺畅 (Quite intuitive.)
- 沟通顺畅 (Intuitive conversation.)
- 有意思 (Interesting.)
- The conversation is generally smooth, but it may take a long time for the chatbot to respond to some words
- I believe interacting with the chatbot has had an overall positive impact on my emotions.
- the chatbot was very useful in introducing me to SAT Protocols. the chatbot guided me through it, making it more easily understandable.
- I think overall it has had a positive impact on my mental state.
- Natural and relaxing.
- Great experience.

Do you have any further suggestions for how the platform can be improved?

对聊天机器人的意见和建议

- 如果聊天机器人能够用言语引导受试者进行各种选项，那将会更加自然
(If the chatbot could also guide the user through carrying out the protocols, that would make the whole conversation flow much better.)
- 推荐练习到提供反馈之间没有任何指导，需用户自己实行练习，感觉在这个阶段不太需要聊天机器人，所以后面又需要和它沟通感觉跳跃有点大。可以指导练习会对用户体验有所提升
(There are no instructions on how to carry out the protocols between recommendation and asking for feedback, you needed to carry out the protocols yourself. It feels like there's a big disconnect with the chatbot at this stage, and a big jump when you have to go back to the conversation afterwards; being able to direct the user through the protocols would improve user experience.)
- 希望能多聊一些 (Wished I could converse with the chatbot more)

- 根据场合的不同，或者可以设计一些更趋于口语的词句？这样不会感觉特别拘谨
(The chatbot maybe could have benefited from some slang/colloquialisms based on the conversation context. This will help the conversation to feel less formal.)
- 应该有多种情绪才更真实，而且人的情绪是复杂的，并不是可以确切分类的
(There should be more emotions to be more realistic; also, human emotions are complex and not completely discretised.)
- 聊天内容可以更有深度一些
(The conversation would have more depth or more opportunities for open conversation.)
- 只把感情归类到四种有点简单。比如心烦/烦躁，AI 识别为愤怒，其实归类不太准确，心烦更偏向于是焦虑
(Limiting emotions to only one of four categories, such as frustration being classified as anger, is somewhat of an oversimplification, and isn't entirely accurate, as frustration can lean more towards anxiety depending on the context.)
- 比如问题：有什么特别的事情引起了你的感受吗？我一般反应就是回答某某事情，而不是选择是/否。结果回答的就不能被识别。最后只能选择是/否
(For questions such as 'was this caused by a specific event', I would normally answer with a description of the event(s), rather than discretising to a yes/no answer. But open conversation is not used at these states, so only a binary choice is available.)
- 感情只归为四类有点简单，人类感情非常复杂，只归四类有点简单
(Emotions being limited to only 4 categories is a bit limited, people's emotions can be quite complex and restricting to 4 is a little simplistic.)
- 练习内容和方法可以更具具体 (The instructions for the exercises can be more specific)
- More guidance and interactions will make the chatbot more human-like after a protocol has been selected.
- the chatbot utterances can be improved to be more natural sounding.
- It would be better for the chatbot to respond faster and provide more specific recommendation
- Maybe can allow user to do the same exercise again?

Please list below any incorrect or unnatural sounding text you encountered

请指出任何在对话过程中遇到的不自然，或错误表达，以助改进

- 1. 下面这句话有点怪，比较难理解：有时，当我们通过某些练习并发现，正如我们所做的那样，这些练习可以揭示负面情绪和体验。可以考虑改为：根据我们的经验，有些联系可以帮助揭示负面情绪和体验。2. 你有没有尝试过练习11但发现它太令人痛苦了？-这句话也有点奇怪，比如没有练习过11（否）和练习过11发现令人痛苦（是），那练习过11但是发现不痛苦应该是选是还是否？可以考虑分成两次问答，一次问是否练习过11，一次问联系是否感觉痛苦。3. 提问最近发生的事情还是以前发生的事情，答案选择是最近和遥远。遥远这个词有点怪，不如改成以前。4. 最近尝试练习6是否让您回忆起这个事件，从而使您记忆犹新，并让您感到焦虑？-这个问题和反馈2一样，如果没参加过练习6怎么回答。

(1. The following line is a little strange and difficult to understand: [SENTENCE1], which can instead be expressed using: [SENTENCE2]. 2. 'Have you tried protocol 11 but found it to be too painful', is phrased strangely. For example, if you have tried exercise 11 but found it to be fine, should you select yes or no? Perhaps this could be split into two questions. 3. 'Was this caused by a recent or [distant] event?', the word [distant] is a little strange, could instead use [long ago]. 4. 'Have you recently tried protocol 6 and found it led to painful memories?', this has the same issue as 2.)

- 感谢您的开放is not a very natural expression in mandarin. One of the questions the chatbot asked me was 有时很容易只看到争论的一方面。您是否觉得自己总是以开放的心态进行对话？, which was kind of confusing. It would be nice if the question can be made a bit more clear.
- 有些例如‘我假设你是...’或者‘我很欣赏你是...’
(There are some sentences such as [I assume you're...] or [I appreciate that...])
- “我很欣赏这可能是新的”：语料库从英文转译成中文时有不准确的翻译
(‘[I appreciate this is new]’: sentences sometimes contain artifacts from the English to Chinese translation.)
- 有不少“我很抱歉听到这个”，应该“听到这个我很遗憾”更符合
(There's quite a few [literal translation of 'I'm sorry'], should be more ['I'm sorry' in an empathetic sense])
- 部分中文表达有点奇怪。比如：我拿错了棍子的一端
(Certain Chinese expressions are quite strange, such as: '[Holding the wrong end of the stick]')
- 当聊天机器人没有明白受访者的问题时，会让对话陷入死循环
(When the chatbot doesn't understand what you've said, it's not clear how to progress with the conversation.)
- 您的个人福祉目前是否受到特别严重的影响？It is natural to say well-being in English but not in Chinese
(Has your [literal translation of 'well-being'] been significantly impacted? It is natural to say well-being in English but not in Chinese)

- 1. 还是在(更远的过去)发生的? 很久以前would've been better. 2. 我有一些我认为(可能在这里工作)的建议。I think it meant to say "I have some suggestions that may work". It should've been 我有些建议可以对你有所帮助。
(1. 'did this happen [far away]?' [a long time ago] would've been better. 2. 'I have some suggestions [I think might work] here.' I think it mean to say "I have some suggestions that may work". It should've been [I have some suggestions which may help you])
- 不好意思, 不太记得了 (Sorry, can't remember.)

Please provide any additional comments you have

其他

- 1.练习过程中有提问: 是否和特定事件相关? 是最近还是发生的还是以前发生的? 然后后面就直接推荐练习, 看不出回答前面两个问题和后面推荐的练习是否有关联。比如如果是很久以前发生的事情推荐的练习和最近发生的事情推荐的练习是否不一样? 如果不一样, 可以在提示中反应一些依据, 比如我们推荐练习11, 这个练习会改善因为最近事件引起的焦虑情绪。2.目前没有对练习的引导, 就只提到请练习, 然后就问结果, 有点太简单。练习过程应该有所引导。
(1. The questions: 'Was this caused by specific events?', 'Was this event recent or distant?' immediately precede protocol recommendation, it isn't entirely clear what the link between the protocols and the aforementioned questions are. Maybe the chatbot could justify why it's recommending each protocol, for example 'I'm recommending protocol 11 because it can alleviate anxious emotions caused by recent events. 2. At the moment there are no step-by-step instructions for practising protocols, only recommendation, which is a little simplistic, there should be a guide in practising protocols.)
- Imho, the current version of the chatbot is definitely a step forward compared to the previous version. After sharing how I am feeling today, it asks me more questions, and I think it gives the impression that the chatbot is trying to make deeper connection with the users, which is a great improvement. However, I felt that the lack of interactions after a protocol has been chosen is still a minor problem.
- 挺有意思的, 如果能够全程自然对话就好了, 期待以后发展
(Quite interesting, would be much better if we could converse entirely using open text, excited to see where this project goes next.)
- 速度稍慢 (Can be somewhat slow)

Appendix C

Example Conversations

This section contains two example conversations between a human user and the SAT chatbot. EN translations are provided below the original text.

Conversation 1. Persona: Gabrielle, emotional context: happiness/content

Gabrielle:	您好，我是Gabrielle! (Hello, I'm Gabrielle) 您今天感觉如何？有什么样的感受？ (How are you feeling today? [What have you experienced]?)
User:	我今天感觉不错，没有什么特别的事情 (I'm feeling good today, nothing out of the ordinary)
Gabrielle:	谢谢您告诉我，我想确保我理解正确。我相信您有快乐/满足的感觉，您能确认一下吗？ (Thank you for telling me, I would just like to make sure I'm correct. I believe you have a happy/content feeling, can you confirm if that's right?)
User:	是，这是正确的。 (Yes, that is correct)
Gabrielle:	很好，让我们找到一种方法来保持这种感觉！我可以推荐一个练习供您尝试吗？ (Great, let us find a means to maintain this feeling! Can I recommend a protocol for you to try?)
User:	好的，我想尝试一些练习。 (Yes, I would like to try a protocol)
Gabrielle:	我有一些我认为可能对您有所帮助的建议。请继续看一看，您可以选择您觉得最合适的练习。可以查看右边的文档，其中包含各练习的简介。 (I have some protocols that I think might benefit you. Please take a look, you can select the one you would most like to try. You can also take a look at the document on the right, it contains information on all of the protocols)
User:	9. 对大自然与环境产生欣赏与爱 (9. Protocol for Attachment and Love of Nature)
Gabrielle:	请通读练习信息，当您觉得自己理解并准备开始时，请尝试通过它。完成后，请按继续。 (Please read the protocol information thoroughly, when you feel you're ready to attempt the protocol, please try it. Afterwards, please press continue)

User:	继续 (Continue)
Gabrielle:	感谢您的确认，我希望您不介意与我分享您在接受此练习后的感受。您觉得您感觉比以前更好还是更糟？ (Thank you for confirming, I hope you wouldn't mind me asking, could you share your feeling after you've attempted this protocol? Do you feel better or worse compared to before?)
User:	我感觉比之前更好 (I feel better than before)
Gabrielle:	我真的很高兴这个练习让您感觉好多了。您认为如果您尝试另一种练习会有帮助吗？如果是这样，您现在想尝试另一个吗？ (I'm really glad that you feel better after this protocol. Do you feel another protocol would benefit you? If so, would you like to try another one now?)
User:	否（结束对话） (No (End conversation))
Gabrielle:	我真的很感谢您的努力，感谢您的参与。让我们下次再谈。 (I really appreciate your efforts, thank you for participating. See you later.)
<i>End of Conversation</i>	

Conversation 2. Persona: Kai, emotional context: Sadness

Kai:	您好，我是Kai！ (Hello, I'm Kai) 您今天感觉如何？有什么样的感受？ (How are you feeling today, [What have you experienced])
User:	不太好，跟朋友闹翻了 (Not great, had an argument with a friend)
Kai:	谢谢，我只是想确保我理解正确。我想您感觉悲伤/失落，这听起来对吗？ (Thank you, I would like just to ensure I'm correct on this. I think you're feeling sadness, does this sound correct?)
User:	是，这是正确的 (Yes, that is correct)
Kai:	我很遗憾听到您有所感受。是不是有什么特别的事情导致了这种感觉？ (I'm sorry to hear that you feel this way, was there a specific event that caused this feeling?)
User:	是有 (There is)
Kai:	我明白。您是说这件事发生在最近，还是您还记得更遥远的过去？ (I understand. Would you say this event occurred recently, or in a more distant past in your memory?)
User:	最近发生的 (Occurred recently)
Kai:	您是否注意到由于最近发生的事件尝试练习11（如果您尝试过）导致您有无法忍受的情绪？ (Have you noticed, if you have attempted protocol 11 as a result of this recent event, that it led to uncontrollable emotions?)
User:	没有 (No)

APPENDIX C. EXAMPLE CONVERSATIONS

Kai:	我理解您的情绪，我们将了解如何克服它。我想再问您一些问题，以便更好地理解您为什么会有这种感觉，这样可以吗？ (I understand your feelings, we can try to learn how to overcome them. I would like to ask you some more questions to better understand why you have these emotions, is that alright?)
User:	好的（详细问题） (Okay (Specific questions))
Kai:	感谢您的信任，我理解您所有的感受。您相信您可以或者应该成为别人的救世主吗 (Thank you for your trust, I understand. Do you feel you could or should be the saviour of someone else?)
User:	不是 (No)
Kai:	非常感谢您的回答。我现在想问您，您是否认为自己是这种情况的受害者，而您的痛苦感觉应该归咎于其他人。 (Thank you for your response. Now I would like to ask you: do you feel you are the victim in these kinds of situations, that your feelings are caused by others?)
User:	不是 (No)
Kai:	感谢您的耐心等待，我真的很感激。您能告诉我，您是否觉得在您生命中的这一刻正在经历个人危机，在工作或家庭中遇到严重的困难？ (Thank you for your patience. I really appreciate it. Could you tell me, whether you believe you're experiencing a personal crisis at the moment, experiencing difficulties at home or at work?)
User:	是的 (Yes)
Kai:	谢谢你，我现在会考虑哪种练习最适合你。请在这时候尝试将你当前的情绪投射到童年时期的自己身上。当您做到这点时，请按按“继续” (Thank you, I will now consider which protocols are suitable for you. Please take this opportunity to project your feeling onto your childhood self. Once you've achieved this, please press 'continue')
User:	继续 (Continue)
Kai:	我们可以尝试多种练习。您想探索以下哪些内容？可以查看右边的文档，其中包含各练习的简介 (We can try many different protocols, which one would you like to try? You can view additional information about the protocols in the document to the right)
User:	8. 放松脸部及身体肌肉 (8. Loosening Facial and Body Muscles)
Kai:	感谢您的选择，您现在应该可以看到练习了，请看看您是如何处理它的。完成后，您可以按继续。 (Thank you for your selection, you should be able to get started on the protocol, please consider how you would attempt the protocol. Once you've completed it, you can press continue)
User:	继续 (Continue)
Kai:	谢谢，我知道这可能要求很高。该练习是否让您感觉比以前更好或更糟？ (I understand this can be difficult, did the protocol make you feel better or worse than before?)
User:	不变 (No change)
Kai:	我很遗憾听到您有所感受。我们应该尝试另一个吗？ (I'm sorry you feel that way, should we try another protocol?)
User:	是（其他练习） (Yes (Other protocols))

Kai: 我有一些您可以尝试的推荐练习。您现在有什么想尝试的吗？如果是，请选择它。可以查看右边的文档，其中包含各练习的简介
(I have some protocols I can recommend you. Do you have any you would like to try? If so, please select it. You can view additional information on the protocols in the document to the right.)

User: 13.通过更换个人观点而处理负面情绪
(13. Changing Our Perspective for Getting Over Negative Emotions)

Kai: 我了解此练习可能具有挑战性，但请您尝试完成它。完成后，请按继续，然后我会问您几个后续问题。
(I understand this protocol could be challenging, but please attempt to finish it. Once you've completed it, please press continue, and I will ask a few more questions.)

User: 继续
(Continue)

Kai: 谢谢，我想知道您现在感觉如何。您认为该练习有任何正面或负面影响吗？
(Thank you, I would like to know how you're feeling now. Do you feel the protocol had any positive or negative effect?)

User: 我感觉比之前更好
(I feel better than before)

Kai: 这真的很好，做得很好。如果您愿意，您可以选择另一个练习尝试。
(That's great, well done. If you would like, you can try another protocol.)

User: 否（结束对话）
(No (End conversation))

Kai: 我要真诚地感谢您参与其中。我希望很快能见到您！
(I would like to wholeheartedly thank you for your participation. I hope to see you again soon!)

End of Conversation

Appendix D

Overview of SAT Protocols

Below is a brief overview of each of the 20 self-attachment techniques protocols that the SAT chatbot could recommend to the user. The English excerpts are lifted from the previous work (1) for the purpose of clarity, and we use a Simplified Chinese version produced for our platform.

练习1：与童年版的自己建立关系

1: Connecting with the Child

此练习的目的是鼓励用户与自己童年的自我建立关系。用户试图想象童年的自我，并想象与童年的自我拥抱或玩耍。用户可以使用两张自己儿童时期的照片协助想象过程：一张开心的照片，一张闹脾气时的照片。

This protocol aims to encourage the patient to connect with their own childhood self. The patient tries to visualise the childhood self and imagine that they are cuddling them or playing with them. The connection can be aided by two childhood photographs of the patient, where these are available: one where they were happy and smiling and one where they were sad and frowning.

练习2：对两张自己童年的照片笑

2: Laughing at our Two Childhood Pictures

该练习要求用户试图鼓励童年的自己笑。通过用户对自己的两张照片（练习1）笑，并同时考虑两张照片中童年的自己所表示的情绪，从其作出对比。此练习的目的是鼓励用户实现与快乐的童年记忆产生共鸣，同时笑对童年的烦恼。用户应该明白练习中的笑并不是嘲笑自己。

The objective of this exercise is to teach the patient's childhood self to laugh. To achieve this, the patient laughs at their own childhood pictures and at the contrast between the opposite emotions that they express. It should be noted that this laughter is meant to make light of life and its events, never to ridicule the patient or their childhood self.

练习3：对童年的自我示爱

3: Falling in Love with the Child

次练习以练习1作为基础，鼓励用户加强先前与童年的自己所建立的关系。用户对童年的自己播放并唱出充满爱的歌曲（比如摇篮曲或儿歌），用户可以试图在唱的过程中起舞，想象自己是与童年的自我一起跳舞。

Protocol 3 is aimed at establishing a deeper connection between the patient and their childhood self. The patient sings happy love songs out loud dedicating them to the childhood self. The patient is encouraged to increasingly use their whole body, dancing to the music as if they were dancing with their childhood self.

练习4：许诺关怀童年的自己

4: Vow to Adopt the Child as Your Own Child

用户可以想象对童年的自己作出永远爱护的承诺，并以坚持实行SAT方案为支持童年自我的方式。鼓励用户实际说出以上的话，并重复多次。

In this exercise, the patient makes a life-long pledge to adopt their childhood self as their own child. They vow to support and care for the child by continuing to practise the self-attachment protocols. When making this pledge, the patient should be speaking out loud.

练习5：与童年的自己保持爱护的关系

5: Maintaining a Loving Relationship with the Child

该练习目的为培养并保持用户与童年的自己建立的关系。用户可以选择一段向童年的自己示爱的短句，并向自己的两张童年照片念。在实行此练习的过程中鼓励用户同时播放歌曲，随其起舞（练习3）。

To nourish and maintain the relationship that they have established with the childhood self, patients select a short, loving phrase of their choice to repeat out loud. While focussing on their two childhood pictures, patients utter this phrase and recite love songs to the child. Again, patients are encouraged to use their whole body while doing so, moving and dancing to the melody of the song.

练习6：处理痛苦的童年回忆

6: An exercise to Process the Painful Childhood Events

该练习目的为帮助用户处理任何痛苦的童年回忆。此练习需要用户的想象童年的自己经历此事件，并经可能得关注细节，用户可以使用自己童年的照片而协助以上过程。然后用户试图想象成年的自己作为家长的角色去安慰并拥抱童年的自己。

鼓励用户说出任何童年人物与成年人物之间的对话，并且在以上过程中对自己面部按摩。

Protocol 6 aims to help patients process any difficult events that they may have endured during their childhood. In order to do so, a patient must try to visualise their childhood self experiencing the event in as much detail as possible, aided by their unhappy childhood photograph. They then must imagine their current, adult self comforting and embracing the distressed child, as a good parent would do. Any verbal interaction with the childhood self should preferably be spoken out loud, and the patient is encouraged to massage their own face while they picture themselves soothing and consoling the child.

练习7：对生活产生热情

7: Protocols for Creating Zest for Life

此练习要求用户在镜子前作出各种活动（唱歌，跳舞，朗诵）。鼓励用户在以上过程中想象镜子中的反射为童年的自我。次练习最好在一日中重复多次，并可以在不同情况下实行。

In this exercise, the patient looks at themselves in a mirror while dancing, reciting happy poems and singing love songs, imagining that what they see in the mirror is a reflection of their own childhood self. The exercise can also be repeated in different circumstances to aid its integration into daily life. For example, it can be carried out while doing housework.

练习8：放松脸部及身体肌肉

8: Loosening Facial and Body Muscles

此练习鼓励用户放松面部及身体肌肉（通过唱歌，跳舞），并同时想象与童年的自己玩耍或拥抱。

This protocol is designed to encourage patients to relax and loosen the muscles in their face and body, while imagining that they are interacting in a loving way with their childhood self.

练习9：对大自然与环境产生欣赏

9: Protocols for Attachment and Love of Nature

鼓励用户与大自然保持互助的关系，常访周边绿色环境，并欣赏大自然的美妙。此练习需要用户到一个户外环境，将注意力集中到一个对象（比如，一棵大树），观察此对象的细节（树枝，树叶）。鼓励用户长期在多个地点实施此练习。

Patients are encouraged to maintain a beneficial relationship with nature and the outdoors, by visiting green areas located near where they are. Once they are in such a location, they should focus their attention on their surroundings, feeling admiration and wonder for the beauty of the natural world (for example by admiring a beautiful tree). This exercise should preferably be carried out at different locations over time.

练习10：欣赏自己的成就，无论大小

10: Laughing at, and with One's Self

用户在此练习中选出自己的一个成就，无论大小，并对作出此成就微笑，逐渐从微笑发展到大笑。此练习的目的是通过对自己成就笑培养用户笑对任何情况的能力，以提高用户面对负面情绪的能力。在此重复此练习中的笑与练习2一样是来自内心的，自然的笑，并不是嘲笑。

Protocol 10 celebrates an accomplishment – no matter how big or small – that the patient feels they have achieved. The patient is encouraged to smile about this accomplishment, and to gradually turn that smile into an authentic, wholehearted laughter. Over time, this laughter should last increasingly longer. By being taught to laugh at their own accomplishments, patients gradually learn to laugh at any life event. As already noted in the description of Protocol 2, laughter in self-attachment technique is compassionate, and it is never meant to ridicule the patient.

练习11：通过自我安慰而面对负面情绪

11: Processing Current Negative Emotions

在此练习中，用户将任何负面情绪投射到童年的自我身上，想象童年的自我处在负面情绪状态下。用户接下来试图安慰并鼓励童年的自我，将任何安慰或鼓励的话语说出来。在负面情绪解除后，用户可以想象童年的自我感到满足的模样，可以使用自己童年的照片以助想象。

In this protocol, patients project any negative feeling that they may be experiencing onto their childhood self, picturing the child in an unhappy state. They can then soothe the child's emotions by consoling and reassuring them, speaking calming words out loud. Once the patient has managed to overcome the unhappy feeling, they can imagine the child being content again, by looking at the happy childhood photograph or picturing it in their head.

练习12：以笑减压

12: Continuous Laughter

此练习目的为帮助用户处理并减少心里压力。用户需在安静环境下实行次练习，用

户试图使用所有面部肌肉形成杜兴微笑 ‘eh, eh, eh, eh’; ‘ah, ah, ah, ah’; ‘oh, oh, oh, oh’; ‘ye, ye, ye, ye’ 又鼓励用户在实行此练习时尝试想象一些幽默/滑稽的事件以助于大笑。用户也可以对此练习的滑稽而笑。

This exercise is aimed at managing and reducing stress. It requires the patient to find a quiet place and smile using all the muscles in their face (i.e. form a Duchenne smile). They can then vocalise out loud any or all of the following phrases, as if laughing: ‘eh, eh, eh, eh’; ‘ah, ah, ah, ah’; ‘oh, oh, oh, oh’; ‘uh, uh, uh, uh’; ‘ye, ye, ye, ye’. The exercise also encourages patients to think of something funny to aid the laughter, be it an event that has happened or even just the silliness of the exercise itself.

练习13：通过更换个人观点而处理幅面情绪

13: Changing Our Perspective for Getting Over Negative Emotions

此练习需要用户观看一个鲁宾花瓶图案，试图将不同看图的视角投射到引起当前负面情绪的事情上，从而用过不同的视角观看问题。鼓励用户将黑瓶子联想成当前的负面情绪，而图中的两个白脸联想成先前以建立关系的成年自我与童年自我，看图时的视觉的转换联想为对事件的负面视角转为乐观视角。

Protocol 13 is to be performed with the visual aid of a picture of the Gestalt vase. Looking at the picture, the patient tries to associate their switching perception of the image in it with changes in their attitude toward their problems. The black vase in the picture represents the patient’s negative emotional patterns, whereas the two white faces – a symbol of the adult and childhood self now connected – are associated with victory in overcoming those patterns. Patients are encouraged to laugh victoriously when they manage to switch their perception of the image from the black vase to the white faces.

练习14：认识童年自我的不良行为

14: Protocols for Socializing the Child

在此练习中，鼓励用户从成年自我的角色将自己的行为,表现投射到童年自我的身上，从而认识任何童年自己的不良行为，并试图成熟得教导童年的自我如何克服不良行为。此练习的目的为通过这种方式试图帮助用户识别自我的不良行为并克服。

In this exercise the patient, acting as their adult self, is invited to recognise any anti-social tendencies or behaviours of the inner child, and attempt to contain them and discourage them in an empathetic and compassionate manner. The negative energy fueling these tendencies is thus redirected toward more constructive and creative activities.

练习15：认识并控制过度自恋或自责行为

15: Recognising and Controlling Narcissism and the Internal Persecutor

该练习需要用户熟悉心理创伤三点：迫害情结，受害者情结，救星情结，并以此作为基础而分析自己的与其相关的经历，特别是认识这些经历所导致的过度自恋或自责行为。通过分析自己的经历去识别并预防过度自恋或自责行为。

Protocol 15 is a self-reflection exercise aimed at recognising and analysing one's own feelings and behaviours relating to persecution, victimhood and rescue, as well as their negative effects. In light of this analysis, the patient re-evaluates past experiences informed by these tendencies, and learns to identify them and avoid them.

练习16：优化自我内心状态

16: Creating an Optimal Inner Model

用户在此练习中回忆自己童年时期所受的教养并分析自己与父母的关系，以及以上对用户今日的性格行为的影响。然后鼓励用户想象将现在的负面行为/性格投射到童年自我上，并努力教育童年自我如何改善自己的行为。

This protocol focuses on identifying and recognising any emotions and tendencies that may have been developed during one's upbringing. Here, the adult self thinks about the behaviours that the childhood self formed during the early years, as a result of interactions with a parent or main carer. The adult self is then tasked with compassionately teaching the child to change and improve these behaviours.

练习17：解决个人危机

17: Solving Personal Crises

此协议针对正在生活中经历个人危机的用户，例如与朋友/家人闹翻等。用户首先将现在的情绪投射到童年的自我，然后在练习15的基础上询问童年的自我此事件是否跟练习15的三点有关。完成后用户反省这次事件是否是个自我提升的机会。鼓励用户在实行以上步骤时尝试练习12。最终用户从成年自我的角度分析自己的经历，与曾经的经历做对比，并分析自己是否有可以提升的地方。

Protocol 17 builds on the previous ones to help the patient resolve a crisis that they may be experiencing in their private life, such as a fall-out with friends or loved ones. First, the patient asks the childhood self whether this crisis originates from the negative behaviours addressed in Protocol 15, and whether it can be an opportunity for growth. Once again, the patient is invited to laugh at their problems, by practising Protocol 12 at the same time as they go through this exercise. As a next step, the adult self evaluates the situation from their own perspective, comparing it with past experiences and trying to learn from it.

练习18：认出无害矛盾性并从中取乐

18: Laughing at the Harmless Contradiction of Deep-Rooted Beliefs/Laughing at Trauma

在此练习中用户阅读一段尼采的“权力意志”的话，并边读边笑。鼓励用户在读的过程中回忆曾经的负面经历，和这些经历对用户今日的自我的影响。此练习的目的是训练用户去笑对过去的负面事件，从而可以逐渐适应面对新的负面事件以及所产生的负面情绪。

In this exercise, the patient begins by reading out loud a quote from Nietzsche's *The Will to Power*, and is instructed to laugh while doing so. As they read through the quote, the patient is also invited to remember past negative experiences, and how these have made them into the person that they are today. After learning to laugh at distant trauma, patients can then gradually apply this exercise to recent and ongoing difficulties.

练习19：通过体验不同思想框架而提高想象力

19: Changing Ideological Frameworks for Creativity

练习19鼓励用户强迫自己考虑不同的思想模式，或在特别问题上挑战自己已有的想法，例如政治问题，社会问题（种族，性别问题，等）。鼓励用户在练习过程中形成两个角色，交换时辩论问题，并试图了解多个视角。

Protocol 19 encourages the patient to challenge their beliefs and convictions and learn to examine situations from multiple perspectives. The list of convictions that can be challenged include political leanings and personal ideas on cultural and social matters, such as issues pertaining race and sexuality. In each case, the patient debates these issues with themselves taking up the role of both proponent and opponent.

练习20：利用名言实施自我鼓励

20: Affirmations

此练习要求用户采集一系列名言，从其选出部分用户感觉深有同感的，通过朗读从中寻找鼓励，自信。

Patients are invited to draw a list of quotes, famous or otherwise, that they find powerful and which resonate with them. These quotes are to be read out loud whenever the patient needs to motivate themselves and find strength in the journey toward reaching their goals.