Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

A Multilingual Chatbot for Self-Attachment Therapy (Algorithmic Human Development)

Author:
Alicia Law

Supervisor:
Dr Anandha Gopalan
Dr Josiah Wang

Submitted in partial fulfillment of the requirements for the MSc degree in Computing (Software Engineering) of Imperial College London

September 2022

Abstract

The digitisation of mental healthcare services has seen an increase in conversational agents used to administer psychotherapeutic treatment. Such applications often leverage the use of deep learning transformer models to enhance user experience. However, due to the lack of available in-domain data in other languages, these chatbots are primarily English-based, thereby limiting access for non-English speaking communities.

To tackle this, we devise a new framework for developing a deep learning assisted chatbot for self attachment therapy (SAT) in Mandarin, which can be replicated to develop chatbots in other languages. Rather than sourcing for in-domain data in Mandarin which is costly and time consuming, we challenge ourselves to use machine translations of existing English data and instead explore different methodologies to attain quality outcomes using large pretrained language models.

In this work, we develop a multilingual classifier for emotion classification, as well as a text generation model capable of producing empathetic rewritings that are fluent and semantically accurate. For both tasks, we report comparable performance, and in some aspects even outperform, previous works during the evaluation.

As pretrained language models are computationally expensive and memory intensive to operate on edge devices, we leverage the use of Knowledge Distillation to achieve performant yet compressed models suitable for use in actual applications. We also made improvements to the user interface to enhance user engagement and experience.

We conduct a human trial with N = 27 bilingual (English-Mandarin) participants to formally assess the effectiveness of our chatbot. Overall, the trial findings report improved performance in terms of emotion classification and utterance quality over previous works, and scored comparatively in other aspects.

Finally, we note several areas for improvement and provide suggestions on how to address these, as well as discuss interesting opportunities for exploration in future works.

Contents

Chapter 1

Introduction

1.1 Motivation

According to the Global Burden Disease study, mental disorders have been ranked amongst the top ten leading causes of burden [1] worldwide since 1990 [1]. This trend is expected to worsen in the future, exacerbated by the Covid-19 pandemic which caused economic and social stresses, resulting in a 25% increase in anxiety and depression rates worldwide [2, 3].

Despite the concerning statistics, there persists a "mental health treatment gap", which describes the large disparity between the need for and availability of mental healthcare services [1]. This can be attributed to the following reasons: (i) stigma on mental health, (ii) unaffordable treatment and (iii) limited and unequal distribution of mental healthcare resources [4].

In hopes to alleviate the above issues, this paper presents a chatbot for Self Attachment Therapy (SAT), also known as *SATbot*. SAT is a new psychotherapeutic treatment that can be self-administered in a structured manner, allowing it to be easily delivered via digital technology [5, 6]. By leveraging the abundance of digital devices and global connectivity facilitated by the internet, such a chatbot can enable the vast and cheap distribution of mental healthcare that can be freely administered anywhere, even privately at home.

While numerous works have been done by the Algorithmic Human Development Team to develop "emphatic virtual agents and emotional recognition platforms for SAT" [7], these have been English focused. Such is the case for most digital mental healthcare applications, where studies have shown extremely limited online resources that serve non-English speaking communities [8, 9, 10]. These trends are concerning as they pose as barriers to a large population from accessing mental healthcare services.

As such, the present study aims to devise a framework for extending exiting En-

[1] Burden is defined according to a disease's prevalence and harm [1]

glish deep-learning assisted virtual conversational agents for psychotherapy to non-English languages. As a proof-of-concept, we focus on developing the *SATbot* in Mandarin. These methodologies can be replicated across languages and applications, and we hope that this contributes to achieving equitable access to mental healthcare for non-English speaking communities in the future.

1.2 Self-Attachment Therapy (SAT)

Self-Attachment Therapy (SAT) is a new psychotherapeutic treatment based on John Bowlby's Attachment Theory. It attributes affect dysregulation [2] disorders to sub-optimal emotional attachments formed between an individual and their primary caregivers during their early childhood. For instance, individuals who experienced secure attachments (i.e. had available and responsive caregivers) in their childhood tend to exhibit stronger self-esteem and self reliance, and hence healthier mental states as adults [5].

The SAT treatment is comprised of a series of self-administered protocols, which guides individuals to develop new secure attachments through imagining their current self caring and attending to their inner childhood self. This stimulates optimal neural growth, allowing individuals to better navigate and regulate their negative emotions, thereby tackling mental disorders stemming from insecure attachment [12].

1.3 Related Works

There are currently 20 SAT protocols. Given that the protocols follow a structured delivery pattern [6], the use of a smart chatbot to recognise emotions and provide protocol recommendations accordingly via conversational interaction can significantly value-add to the SAT process for patients. Such a chatbot can ensure that patients practice the correct protocols, as well as ensure users navigate through their treatment without being overwhelmed. As of present, there have been several *SATbot*s done by previous students.

Rule-based Chatbot The first *SATbot* was by Ghachem [13] who devised the chatbot's rule-based conversation flow. In this study, only 28% of trial participants agreed that the chatbot provided useful suggestions and was engaging, and only 29% of participants found the chatbot to be empathetic. At present, most psychotherapeutic chatbots are rule-based and suffer from similar issues, as they are bounded to a limited set of pre-defined inputs and responses, making them appear monotonous and robotic [14, 15]. However, such rigidity allows for more secure and responsible chatbots, which is especially crucial in this application for mental healthcare [16].

[2] Affect dysregulation is defined as the "impaired ability to regulate and/or tolerate negative emotional states" [11]

Neural Chatbot (English) To address the limitations of Ghachem's chatbot, Alazraki [6] introduced the idea of Empathetic Rewritings to create more varied and human-like utterances. This was done by crowd sourcing rewritings of base utterances followed by a split-and-merge data augmentation methodology to extend the dataset. Utterances were then selected based on a multi-objective function that maximised fluency, empathy and novelty, helping ensure that conversations with the chatbot be safe and engaging. Alazraki also adopted the use of neural emotion classifiers, which enhanced the accuracy of emotion predictions in comparison to Ghachem's [13] rule-based classifier. Overall, the project yielded promising results, with 69% of participants finding the chatbot to be engaging and 81% finding it to be empathetic.

Neural Chatbot (Mandarin) As an extension to Alazraki's project, Hu [17] developed a Mandarin version of the *SATbot*. While mostly similar in terms of implementation, the key difference with Hu's chatbot (other than being in Mandarin) lie in the chatbot's empathetic rewriting component. Rather than crowd-sourcing native empathetic rewritings, Hu leveraged the use of a generative language model (GPT-2 Chinese) to generate rewritings.

While Hu's methodology yielded generally fluent rewritings, trial participants noted presence of English-style figures of speech and a lack of colloquialism in the utterances, which was the result of using an English translated dataset during training. Another issue noted in Hu's implementation was with regards to high latency in the emotional classifier, which was the result of using a BERT-large sized language model. Finally, Hu's chatbot application was purely in Mandarin and did not incorporate the English language. We look to address these specific points in our implementation.

1.4 Legal, Social and Ethical Considerations

This project presents a series of legal, social and ethical concerns that need to be addressed with care. We present the areas of concern and how we take these into consideration throughout the study.

Patient Safety

Firstly, the *SATbot* is a mental health application targeted at patients suffering from mental health conditions. Hence, patient safety must be treated with utmost priority. The *SATbot* ensures this by including the following in its implementation:

1. *Safe & Non-Toxic Chatbot Conversations (see Chapter 6)*
   Empathetic rewritings using generative language models are trained using a controlled dataset that has been vetted for safety. Utterances also go through an empathy check via an empathy classifier and is only permitted after having attained a high empathy score. Finally, as an additional precaution, all utterances are manually vetted for toxic content before being displayed.

2. *Terminating Therapy*
   SAT Protocols involve users interacting with their childhood self, which can inadvertently trigger strong emotions within patients suffering from childhood trauma. Should patients be uncomfortable with the suggested protocol at any point, they are given the option to decline treatment. The application will also take note of the protocol and omit it in the remainder of the session.

It should also be stressed that *SATbot*, while a mental health application, is not equipped to treat patients suffering from serious mental health conditions such as depression. During human trials, participants are provided with a participant information sheet which declares the above. It also outlines details of the study and the potential risks entailed. Only when consent is received, are participants given credentials required to interact with the chatbot.

Data Protection

While patients do not need to provide personal information to interact with the chatbot, the *contents* that patients discuss with the chatbot in itself are considered personal data under UK's Data Protection Act (DPA) [18] and General Data Protection Regulations (GDPR) [19]. To ensure adherence to data protection laws, *SATbot* does not store user interactions beyond the treatment sessions. We also do not store metadata from user's devices (eg. geolocations, IP/MAC addresses, IMEI codes etc.).

Data compliance is also ensured during the human trial. The human trial conducted in this paper has been approved by the Imperial College Research Ethics Committee. Prior to the trial, participants are informed on how their personal information is handled, and will require to provide consent before participating. Moreover, participant responses are anonymous, and all responses collected are used strictly for the purposes of the current study.

Data Sourcing

Many datasets are involved in the training of the machine learning models used by the *SATbot*. On this end, we ensure that all datasets used have been ethically sourced and comply with data protection laws. For instance, the `EmpatheticPersonas` dataset (see Chapter 3) is approved by Imperial College's Research Ethics Committee.

1.5 Aims and Objectives

To summarise, the present study aims to devise a framework to develop virtual conversational agents for non-English languages. As a proof-of-concept, we show this through developing a **strongly bilingual** (English (EN) and Mandarin (ZH)) chatbot for SAT that is built upon previous works [6, 13, 17]. The chatbot will similarly be safe (non-toxic), reliable and engaging. Ultimately, the project aims to extend mental health access to non-English speaking users and help combat the global mental

health crisis.

The detailed objectives and key distinguishing contributions of the present study are:

- To develop a multilingual emotion classifier, with emphasis on Mandarin (Simplified), that achieves comparable performance to previous works (Chapter 5),
- To improve emotion classification inference speed via Knowledge Distillation to combat high latency issues faced in previous chatbot versions (Chapter 5),
- To explore alternative training methodologies for response generation, namely transformer reinforcement learning via Proximal Policy Optimisation (PPO) to train an empathetic, fluent and accurate generation model for empathetic rewritings (Chapter 6),
- To improve the *SATbot* interface - making it more interactive, engaging and user-friendly (Chapter 7),
- To fully integrate the Chinese version of the chatbot with previous [6] English works to obtain a fully bilingual application (Chapter 7),
- To formally evaluate the chatbot performance via a non-clinical trial, and identify future design improvements (Chapter 8 and 9).

Chapter 2

SATbot Overview

In this chapter, we will provide a brief overview of the *SATbot*. This will also help introduce the key components of the project which will be discussed in greater detail in the subsequent chapters.

The motivation for developing the *SATbot* was to help patients navigate through the wide range of SAT Protocols. Ghachem [13] devised the chatbot's conversational flow structure, illustrated in Figure 2.1, which aims to help the chatbot build a clear understanding of the patient situation in order to suggest relevant protocols. In this project, we will adhere to the same conversational flow. There are 2 distinct dialogues that can occur and these are dictated by the user's emotional state, namely positive or negative.

This introduces the first key element of the project - **emotion classification** (see Chapter 5). In order to present the appropriate conversation, the chatbot needs to correctly identify the user's emotions. The *SATbot* does this by presenting users with a prompt asking how they are feeling. It then needs to process the user's response and accurately determine their emotion. Following Alazraki's [6] implementation, we develop a neural classifier that can classify four emotion categories: sadness, anger, happiness and fear. As such, positive emotional state is defined by happiness, and the remaining are considered as negative emotional states.

When the conversation begins, users are asked a series of questions in a rule-based manner. This serves to build a selection of protocols tailored to the user's current predicament. These protocols are then recommended to users at the end of the conversation. Based on the current conversational flow, there are 45 questions/ statements, also referred to as base utterances, that can be presented to users at each stage of the chatbot conversation flow (see components highlighted in purple and yellow in Figure 2.1). To increase engagement, these base utterances will undergo **empathetic rewritings** (see Chapter 6). This is the second component of the project, which aims to transform low empathy base utterances to contain higher levels of empathy. The task will also help generate variation in conversation and make the *SATbot* appear more human-like.

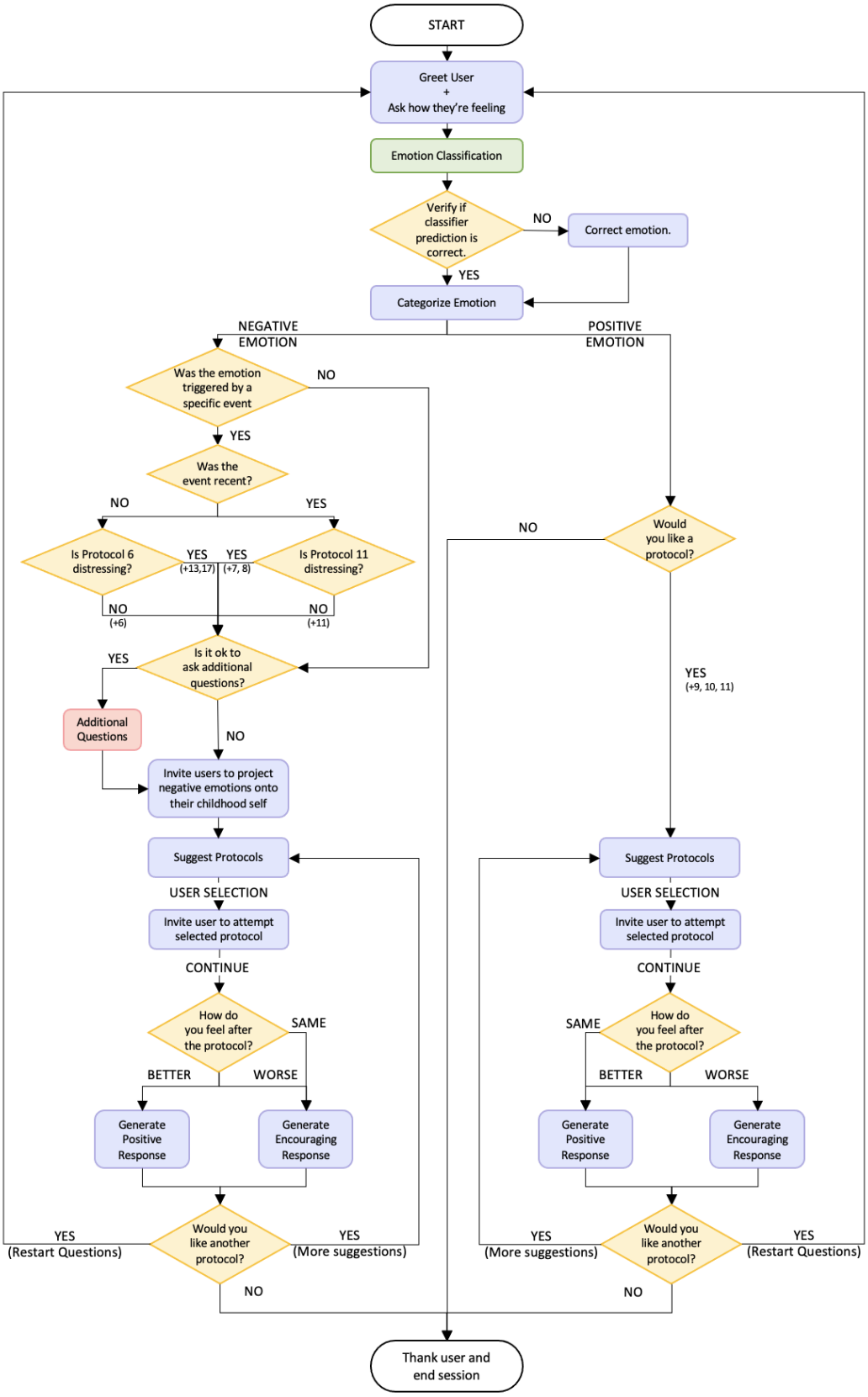


Figure 2.1: Chatbot conversation flow, whereby +(n, . . .) represents the protocols added into the suggestion pool. For the section on additional questions (in pink), please refer to Figure 2.2. Figure is redrawn based on figures from [6] and [17].

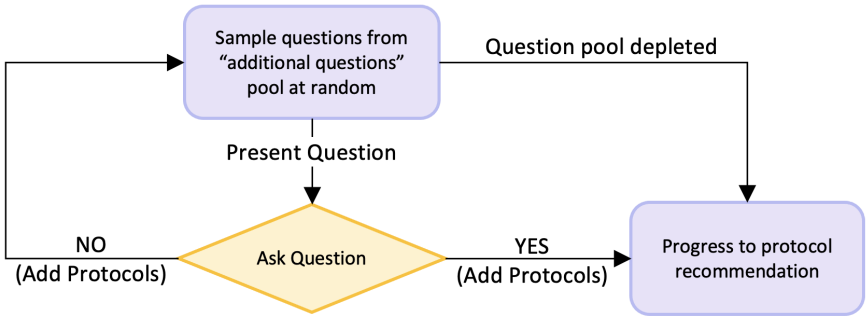


Figure 2.2: Additional questions conversation flow. Questions are randomly selected from a pool of "additional questions" without replacement until users respond "yes" or until the question pool is exhausted. At each stage, protocols are added into the suggestion pool accordingly. For the full list of questions and corresponding protocol additions, see Table 2.1.

Additional Questions	**Protocols added if users responded:**	
	Yes	**No**
Have you strongly felt or expressed any of the following emotions towards someone: envy, jealousy, greed, hatred, mistrust, malevolence, or revengefulness?	13, 14	13
Do you believe that you should be the saviour of someone else?	8, 15, 16, 19	13
Do you see yourself as the victim, blaming someone else for how negative you feel?	8, 15, 16, 19	13
Do you feel that you are trying to control someone?	8, 15, 16, 19	13
Are you always blaming and accusing yourself for when something goes wrong?	8, 15, 16, 19	13
Is it possible that in previous conversations you may not have always considered other viewpoints presented?	13, 19	13
Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?	13, 19	13

Table 2.1: List of additional questions that can be asked and the protocols added based on user response (yes/no).

Chapter 3

Dataset

3.1 EmpatheticPersonas (EP)

The EmpatheticPersonas dataset [6] is a crowd-sourced dataset collected by Alazraki, intended for *SATbot* development. This dataset is comprised of 2 main components. Firstly, it contains 1181 written expressions of **emotions**, intended to train an emotion classifier. The data is relatively evenly distributed across 4 emotion classes: 284 for Fear/Anxiety, 297 for Anger, 300 for Sadness and 300 for Joy/Contentment.

Secondly, it contains 2144 **empathetic rewritings** of the 45 base utterances. 1100 of the empathetic rewritings have been annotated for empathy on a 0 to 2 scale to represent non-empathetic, slightly empathetic and highly empathetic utterances in ascending order. This is intended for training an empathy classifier.

Taking Alazraki's EmpatheticPersonas dataset [6], Hu **translated** the dataset fully from English into Simplified Mandarin using a machine translation tool [17]. However, as translations lack colloquialisms and contain inaccuracies, Hu [17] also crowd-sourced a small EmpatheticPersonas emotions dataset in **native Mandarin**. This is intended for model evaluation only, to assess classifier performance when faced with native input. It is comprised of 120 datasets, evenly distributed across the 4 emotions (i.e. 30 each).

Finally, in this paper, we also create a **code-switching** EmpatheticPersonas emotions dataset, where both English and Mandarin are used in the same sentence/(s). Original sentences from the EmpatheticPersonas test set were modified, with certain English words replaced by Mandarin characters, and sentence structures modified (where applicable) to resemble those used by code-switchers. This dataset was then manually inspected by three bilingual speakers and rewritten where necessary. This dataset is comprised of 100 utterances, evenly distributed across the 4 emotions as well (i.e. 20 each), and is only intended for model evaluation.

Given the many variations of the EmpatheticPersonas dataset, the remainder of this report will follow the following naming conventions listed in Table 3.1. To

summarise:

Naming	Language	Size and Use Case
`EmpatheticPersonasEN` [6]	English	1181 emotions 2144 empathy
`EmpatheticPersonasZH` [17]	Mandarin (translated)	1181 emotions 2144 empathy
`EmpatheticPersonasZH` (Native) [17]	Mandarin (native)	120 emotions (for testing only)
`EmpatheticPersonasCS`	English-Mandarin mixed	100 emotions (for testing only)

Table 3.1: `EmpatheticPersonas` Dataset Summary

3.2 Emotional Chatting Machine (ECM) Dataset

The Emotional Chatting Machine (ECM) dataset [20] is a collection of emotional datasets in Mandarin sourced from the NLPCC 2013 [1] and NLPCC 2014 [2] emotion classification task. The dataset is comprised of 40132 utterances spread across 6 emotions: Anger, Disgust, Happiness, Like, Sad and Others, with detailed distributions shown in Table 3.2.

Given the limited size of the `EmpatheticPersonas` dataset, this sizeable corpus will be used in addition to coarsely finetune the emotional classifier in Chapter 5. However, since only 3 of these emotions (anger, happiness and sad) coincide with those used in the emotion classifier, we will include 'others' as a filler for the fourth emotion during training to maintain a four label configuration.

	Anger	Happiness	Sad	Disgust	Like	Others
Quantity	3167	4950	5348	5978	6697	13997

Table 3.2: ECM Data distribution across emotions.

[1] http://tcci.ccf.org.cn/conference/2013
[2] http://tcci.ccf.org.cn/conference/2014

Chapter 4

Large Pretrained Language Models (PLMs)

Language Models (LMs) are probabilistic models that compute the conditional probability of a word (w_i) given the history/ sequence of words ($w_1, w_2, ..., w_{i-1}$). The word with the highest conditional probability $P(w_i|w_1, w_2, ..., w_{i-1})$ will be assigned next in the sequence [21].

The introduction of Transformers by Vaswani et al. [22] marked a new era for language models, replacing recurrent neural networks (RNN) as the predominant architecture of its time [23]. LM capabilities were then furthered through applying Transfer Learning [24].

By first applying unsupervised **pretraining**, pretrained models can then be fine-tuned on smaller task-specific labelled data to learn task-specific patterns. The result is a finetuning process that is relatively quick and inexpensive, yet still able to yield state-of-the-art performance for various downstream tasks [25, 26, 27, 28]. This is attributed to the pretraining process, which has demonstrated to be critical in improving the language model's understanding of the general language system [29].

At present, the use of **p**retrained **l**anguage **m**odels (PLMs) with transformer-based architecture are now the industry standard for NLP tasks [23]. As such, we will focus on implementations using PLMs only in the *SATbot*. More specifically, given the current project's focus on the Mandarin language (and more), we will focus only on **monolingual Mandarin and multilingual** PLMs.

Chapter 2 briefly introduced the 2 key components to the *SATbot*:

i. **Emotion Classification**
   This is a *multi-class sequence classification* task, whereby the model takes in a sequence (i.e. sentence(s)) as input and outputs a corresponding class label [30].

ii. **Empathetic Rewritings**
   This is a *conditional text generation* and *style-transfer* task.

In the following sections, we will describe the models evaluated in the development of the *SATbot* at each of these phases and provide a brief background for better understanding.

4.1 PLMs for Classification

4.1.1 Foundational PLMs

Before introducing the PLMs used in the *SATbot* for classification, it is important to introduce the foundational models they were built upon - BERT and RoBERTa.

Bidirectional Encoder Representation from Transformers (BERT)

As its name suggests, BERT [31] enables textual representations that consider both left and right contexts. This is vastly different from prior language models, which use unidirectional representations and hence can only encode a restricted amount of information. Bidirectional representations are learnt through unsupervised pretraining on two tasks:

1. *Masked Language Modelling (MLM)*
   A percentage of words are randomly masked during training and the model is tasked to predict the masked words, thereby training the model to learn context.

2. *Binarised Next Sentence Prediction (NSP)*
   Given 2 sentences A and B, the model is tasked to predict if Sentence B follows Sentence A, allowing the model to learn discourse coherence.

While new and better PLMs have been developed over the years, BERT often forms the basis to these models and is also adopted as the baseline performance in many studies [32].

Robustly Optimised BERT Approach (RoBERTa)

While adopting the same architecture as BERT, RoBERTa [28] uses an improved pretraining approach, namely:

1. Longer pretraining with larger batches and a dataset 10 times that of BERT's,

2. Dynamic masking instead of static masking during MLM,

3. No NSP, and

4. Longer input sequences.

This change gave RoBERTa significant performance improvement over not only BERT, but also previously published models including XLNet, in many NLP benchmarks including GLUE, RACE and SQuAD [28, 29].

4.1.2 Monolingual Chinese PLMs

BERT models have been replicated over multiple languages, these include Chinese, the language of interest in this paper. One of the first Chinese models developed was **Chinese BERT** by Google [33], which is a BERT-base model pretrained on Wikipedia corpora in Simplified and Traditional Chinese.

As of current, the best performing bert-base size model is **hfl/chinese-bert-wwm-ext** [1] [35], which is built upon Chinese Bert. Its key difference is that it:

1. uses an extended dataset during training (ext), and
2. adopts whole word masking (wwm).

4.1.3 Multilingual PLMs

Multilingual Models are single LMs pretrained on data in multiple languages, with the aim that the model will learn a generalised representation that is universally applicable to any language. More specifically, sentences and words with similar contexts should be mapped to the same shared vector space across all languages (as illustrated in Figure 4.1) [36].

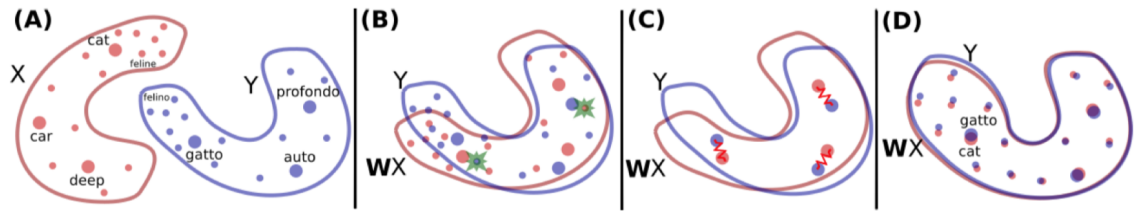


Figure 4.1: Gradual alignment of word embeddings in one language to another [37].

XLM-Roberta (XLM-R)

XLM-R [38] is a RoBERTa model pretrained on monolingual corpora in 100 languages using Multilingual Masked Language Modelling (MMLM). This is the most standard multilingual pretraining method, which extends the MLM training objective (see Section 4.1.1) to multiple languages.

Despite not using any explicit cross-lingual training objectives (eg. parallel corpora), XLM-R is able to generalise surprisingly well across languages, learning multilingual representations that are language-agnostic [39, 40]. Studies [36, 39, 40, 41] attribute this to the "bridge-effect", which is facilitated by the use of a common shared vocabulary, position embedding and special tokens across languages. Common word

[1] This is based on the CLUE benchmark for Chinese NLU tasks as of 26 May 2022 [34].

pieces that occur between languages (eg. numbers and punctuation) act as "anchor-points", encouraging co-occurring word pieces to be mapped within the same space, gradually converging the embedding space for different languages.

InfoXLM

InfoXLM [41] builds upon XLM-R, utilising the same model architecture and hyperparameters. The key difference lies in the pretraining objectives, where InfoXLM utilises *explicit* cross-lingual training objectives, deliberately training representations of similar contexts into the same multilingual vector space. These objectives are:

1. **Translation Language Modelling (TLM)**
   TLM is similar to MLM, differing only in the use of parallel corpora. This enables the model to predict masked words using context in *either* languages, thereby implicitly learning to align representations between languages [41, 42].

2. **Cross-Lingual Contrastive Learning (XLCO)**
   XLCO aims to minimise a contrastive loss function, which helps to encourage similar representations between positive examples (i.e. a sentence and its translation) and vice versa for negative examples (i.e. a sentence with a random translated sentence), thereby aligning cross-lingual representations [41, 43].

4.2 PLMs for Text Generation

4.2.1 Generative Pretrained Transformer (GPT-2)

GPT-2 [44] is a decoder-only language model that uses unidirectional representation. This means that GPT-2 can only consider words to the left of the current predicted token during text generation (i.e. a traditional language model). This is trained using Causal Language Modelling (CLM) with masked self-attention, whereby remaining words beyond the current word position are masked.

While GPT-2 has strong text generation capabilities, it is also known to display concerning toxic behaviour, such as sexist or racist remarks, as a result of its large web-crawled training dataset [45]. Given that the *SATbot* is intended for use in mental healthcare, active measures must be taken to ensure no toxic utterances are generated when using GPT-2.

In this paper, we will utilise **uer/gpt2-chinese-cluecorpussmall**, a monolingual GPT-2 model trained using the UER framework [46] on the Chinese CLUECorpus2020 [47] for Chinese text generation.

Chapter 5

Emotion Classification

Emotion Classification is the first key component of the *SATbot*. In order for the *SATbot* to suggest relevant and helpful protocols to its users, it must first identify the user's current emotional state. This is done via an emotion classifier.

There are 2 key performance criteria to consider when developing the emotion classifier. Firstly, and more intuitively, the **accuracy/F1-score** of the classifier. The more closely we can predict a user's emotional state, the better user satisfaction will be as it provides the impression of an actual human-like chatbot capable of understanding complex emotions. This is addressed in Sections 5.1 and 5.2, where different models and finetuning methodologies are explored to improve model accuracy/F1 performance.

The second criteria is the classifier **inference speed**. If the *SATbot* takes too long to respond with an emotion prediction, such high latency will negatively impact user experience. This was noted as a limitation in Hu's paper [17], which we seek to address in Section 5.3 via Knowledge Distillation.

5.1 To Monolingual or Multilingual?

Chapter 4 introduced various PLMs, both monolingual and multilingual, for classification tasks in non-English languages. The most significant benefit of multilingual models is the ability for **Zero-Shot Cross-Lingual Transfer**, whereby models can be finetuned in one language but then subsequently utilised for tasks in another language [39].

This is extremely useful, especially for low-resource languages where insufficient labelled data is available, as the model can leverage on the additional pretraining data available in different languages to train performant models [48]. While transfer is most successful for structurally similar languages, it is possible even for languages in different scripts [39]. This finding is extremely beneficial for the present research, as Mandarin shares the same word order but not the same characters with many languages.

However, it is often argued that monolingual models tend to outperform multilingual models for a given language [36, 49, 50]. This is because multilingual models suffer from the **curse of multinguality**, which is the trade-off between the number of languages in a model with performance [38]. In our specific case for Mandarin, which is a high resource language, the effect of capacity dilution on multilingual models may potentially be more distinct in comparison [36, 50].

Whether monolingual or multilingual models are a better solution remains as a highly debated topic in the NLP space [36, 48, 49, 50]. To assess which model will yield the best performance for our *SATbot* Emotion Classifier, in this section, we explore both monolingual and multilingual models, more specifically [1]:

- Monolingual: *chinese-bert-wwm-ext*, *bert-base-chinese*
- Multilingual: *microsoft/infoxlm-base*, x*lm-roberta-base*

5.1.1 Investigation Methodology

To assess which model provides the best results, all models were finetuned on the `EmpatheticPersonas` dataset and evaluated on its emotion classification performance. For **monolingual** models, finetuning is straightforward, using only the `EmpatheticPersonasZH` (80%) train dataset. As it is a monolingual model, this trained model can only be used for emotion classification in Mandarin.

For **multilingual** models, we adopt the *translate-train-all* training methodology, noted to be the best training configuration for multilingual models [33, 38, 41, 51]. In this training methodology, English training data is machine-translated to the different languages of interest. These translated datasets are then concatenated and used to train a *single* multilingual model to be used for these languages. In this case, the training dataset will be a concatenation of the `EmpatheticPersonasZH` (80%) and `EmpatheticPersonasEN` (80%) datasets. As a result, this single model can be used for emotion classification in both English and Mandarin.

As each model will have different optimal hyperparameters, all models were tested across a range of epochs [0, 10] and learning rates [1e-05, 1e-04]. For each experiment, the models were ran 3 times and the average recorded. The highest average F1 score recorded on the held-out `EmpatheticPersonasZH` (10%) validation set is then taken as the model's performance.

5.1.2 Model Exploration Findings

Table 5.1 records the highest F1-scores for both monolingual and multilingual models across all learning rates and epochs.

[1]The following model names are the names as stated on HuggingFace. For more information on the models, please read Chapter 4.

Performance Summary	Model	F1-score (%)
Monolingual	bert-base-chinese	92.43
	chinese-wwm-ext	93.24
Multilingual	XLM-R	**94.06**
	InfoXLM	93.21

Table 5.1: Best recorded F1-scores (on validaton set) during model investigation.

Monolingual vs. Multilingual

Despite concerns of capacity dilution in multilingual models, multilingual models have shown to have comparable performance, and even outperformed monolingual models, in our investigation. This is likely due to the **additional information gain**, as multilingual models in the translate-train-all setting could utilise both English and Mandarin datasets during finetuning (i.e. two-fold increase in data availability).

Overall, our findings iterate those of the XLM-R paper [38], that it is possible to have a single multilingual model without having to trade-off per-language performance, at least in the case of Mandarin. This is extremely beneficial for the current project, which aims to spread SAT to non-English speaking communities. By adopting a single multilingual model over numerous single-language models, we can better scale and maintain our application, in the present study as well as for future works [36]. As such, the paper will move forward with multilingual models.

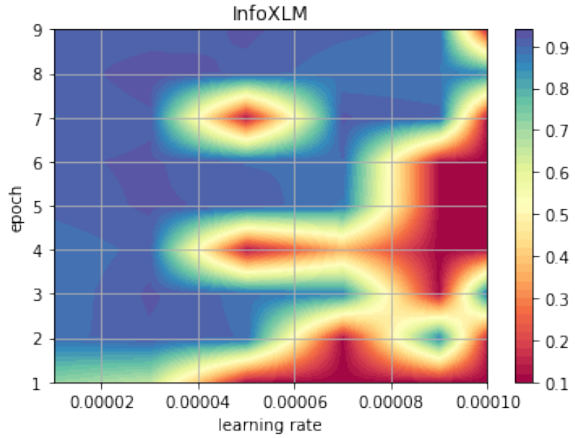
InfoXLM vs. XLM-R

As shown in Table 5.1, XLM-R outperforms InfoXLM. However, the key winning feature of XLM-R lies in its "stability". When performing multiple runs across differing learning rates and epochs, it was observed that **InfoXLM** displayed highly varied performance (ranging from 10% to 93%) with no strong patterns of a stable learning rate or epoch. This is illustrated in Figure 5.1(a) and 5.2(a) where the spread of red (low F1-scores of 10%) is staggered across the learning rates and epochs. This trend persists in different runs of InfoXLM, with the F1 pattern varying each time (see Figure 5.1(c)), making it difficult to establish a "safe" epoch and learning rate.

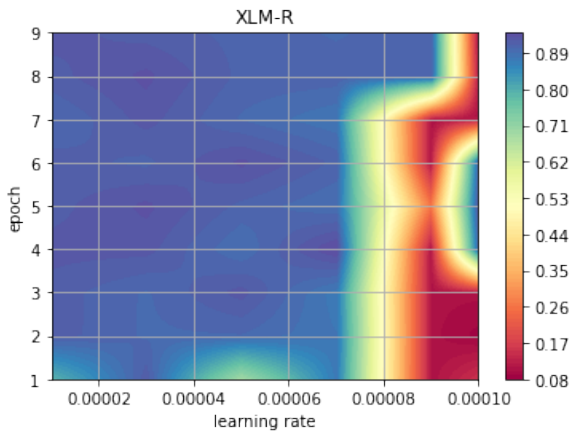
XLM-R, however, has noted to consistently yield F1-scores above 90% in learning rates less than 8e-05 across multiple runs of XLM-R (as indicated by the consistent blue plots in Figure 5.1(b) and 5.1(d)). As such, the paper will move forward with XLM-R, as it provides confidence of a strong performing model.

5.2 Double vs. Single Finetuning

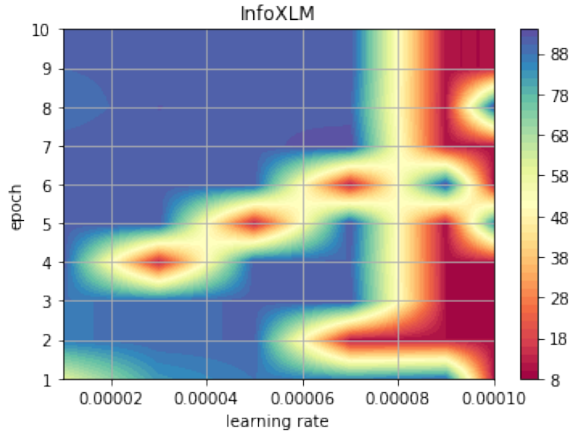
The model exploration concluded the use of a **multilingual XLM-R** model for classification tasks. In hopes to further enhance the model's performance, we investi-



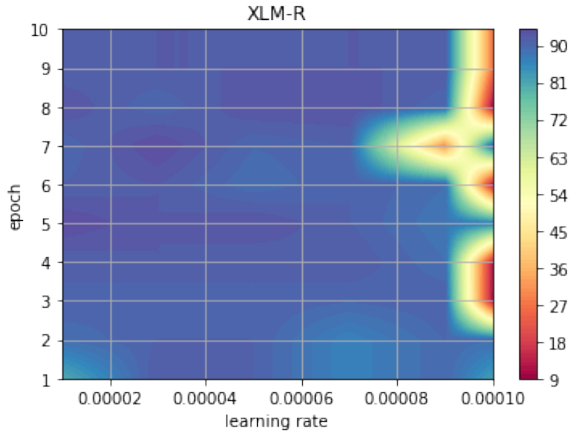
((a)) InfoXLM F1-scores.



((b)) XLM-R F1-scores.

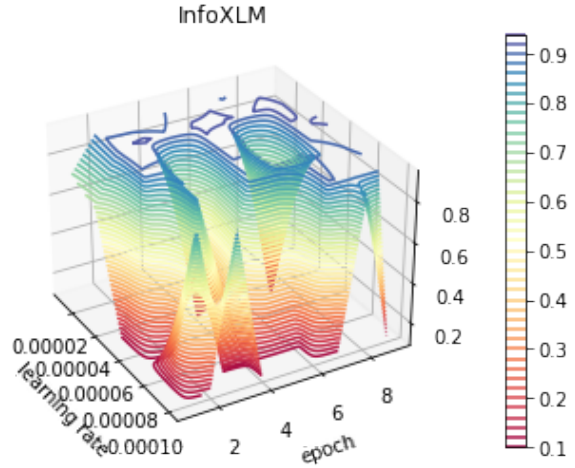


((c)) InfoXLM F1-scores in another run.

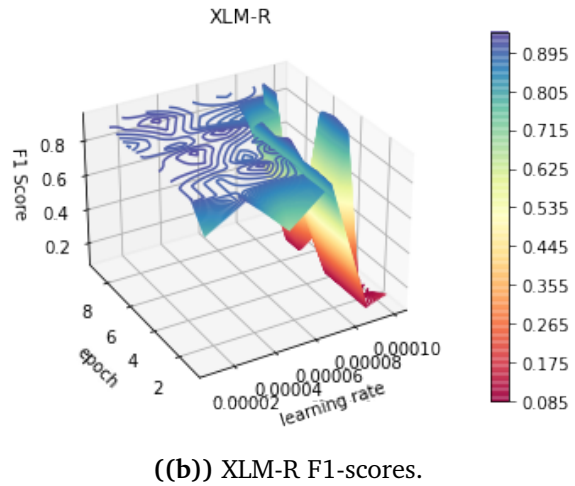


((d)) XLM-R F1-scores in another run.

Figure 5.1: 2D contour plot of F1-scores across different epochs and learning rates. Note that values have been interpolated for illustration purposes.



((a)) InfoXLM F1-scores.



((b)) XLM-R F1-scores.

Figure 5.2: 3D contour plot of F1-scores corresponding to Figure 5.1(a) and 5.1(b) respectively. Note that values have been interpolated for illustration purposes.

gate the use of double finetuning given the success of this training methodology in Alazraki's paper [6]. The training pipeline is as follows:

1. **First Finetuning - Emotional Chatting Machine (ECM) dataset**
   A XLM-R-base model is trained on 80% of the ECM dataset (see Section 3.2) and validated against 20% of the dataset to pick the best first finetuned model.

2. **Second Finetuning -** `EmpatheticPersonas (EP)` **dataset**
   The best ECM-tuned model is trained using the translate-train-all setting (i.e. 80% EN and 80% ZH concatenated). Information on the EP dataset is available in Section 3.1.

To serve as a baseline, a XLM-R model with single `EmpatheticPersonas` dataset finetuning was also conducted. Both models were hyperparameter tuned (refer to Appendix A) and the best models were selected based on performance on the validation set (10% of `EmpatheticPersonasZH`). These models have been tested against various `EmpatheticPersonas` test sets and the results are tabulated in Table 5.2. For more details about the different test sets, please see Section 3.1.

`EmpatheticPersonas` **Dataset**	**Finetuning**	**Accuracy** %	**F1-Score** %
ZH (translated)	double	93.86	93.86
	single	92.98	93.02
ZH (native)	double	90.00	90.09
	single	84.17	83.86
EN	double	91.23	91.40
	single	89.47	89.53

Table 5.2: Model results against different `EmpatheticPersonas` test sets with single and double finetuning.

As shown in Table 5.2, while gains were made across all datasets, the significant gains in double finetuning lie in the performance improvement over the **native** `EmpatheticPersonasZH` test set (+5.83% accuracy, +6.23% F1). This highlights the success of the first coarse grained finetuning using the native Mandarin ECM dataset, which appears to have taught the model general linguistic patterns of the Mandarin language. As a result, the model is able to yield relatively strong performance on the native `EmpatheticPersonas` test set despite being trained on a translated one.

Interestingly, performance over the English dataset (`EmpatheticPersonasEN`) also improved (+1.76% accuracy, +1.87% F1), once again illustrating the cross-lingual benefits of using a multilingual model, whereby training with data in one language (Mandarin) can benefit performance in another language (English).

Given that the current performance of the double-finetuned model is sufficiently strong (above 90% for all test sets), rather than attempting to further this performance as per Hu's paper [17], the present study diverges and explores Knowledge Distillation.

5.3 Knowledge Distillation

While large PLMs have allowed for state-of-the-art performance in various NLP tasks, their size makes them computationally expensive and memory intensive to train and operate. Adopting such models becomes impractical as they are not able to run efficiently at inference on edge devices [52].

To address the above limitations with large PLMs, several studies have investigated the use of Knowledge Distillation on BERT models as a compression technique with large success. These include DistilBERT [52], TinyBERT [53], and MobileBERT [54]. As such, guided by these papers, the present study will explore Knowledge Distillation on our double-finetuned emotion classifier. The following subsections provide a background into Knowledge Distillation, our devised distillation methodology and finally the distilled model's performance.

5.3.1 The Theory

What is Knowledge Distillation?

In Knowledge Distillation, a smaller compact "student" model is trained to learn the **generalisations** of a larger complex "teacher" model, thereby achieving a compressed but equally performant version of the larger model [55]. This differs from transfer learning, which aims to transfer the weights of a pre-trained model to another of the exact same architecture.

What are Generalised Representations?

A well-trained classification model will output a high probability estimate on the gold label, while setting near-zero probabilities on other labels. However, some of these near-zero probabilities are larger than others, indicating the model's knowledge regarding similarities between labels and hence its ability to generalise. This is also referred to as "Dark Knowledge", which is what the teacher model seeks to impart to the student model [52, 55].

Training Loss

To perform Knowledge Distillation, we adopt the Triple Loss training method used in DistilBERT by Sanh et al. [52]. As the name suggests, the training loss is comprised of 3 components: (i) classic supervised training loss, (ii) distillation training loss

and (iii) cosine embedding loss, the latter two being additional losses to aid in the "teaching".

1. **Classic Supervised Training Loss,** L_{ce}
   This is the cross entropy loss between the student model's predicted distribution (s_i) with the target training labels (q_i) which is in the form of a one-hot vector.
   $$L_{ce} = \sum_i q_i * log(s_i) \tag{5.1}$$

2. **Distillation Loss,** L_{dist}
   Defined by Hinton et al. [55], distillation training loss is the cross-entropy loss between the student model's *softened* predicted distribution (s_i) and the teacher's *softened* predicted distribution (t_i).
   $$L_{dist} = \sum_i t_i * log(s_i) \tag{5.2}$$

   These softened predictions are also known as the softmax-temperature probability distribution, given by:
   $$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{5.3}$$

   where:

   - T, temperature
     Controls the smoothness of the output distribution. This will also be an additional hyperparameter when tuning the distilled model.
   - z_i, the probability of class i

3. **Cosine Embedding Loss,** L_{cos}
   While most Knowledge Distillation methods use only losses 1 and 2, the cosine embedding loss is specific to Triple Loss. It aims to align the student's and teacher's hidden vector representations and is noted by Sanh et al. [52] to improve performance. The loss is as follows:
   $$L_{cos} = 1 - cos(T(x), S(x)) \tag{5.4}$$

The final training loss is taken as the average of the 3 losses:

$$L_{total} = \frac{L_{ce} + L_{dist} + L_{cos}}{3} \tag{5.5}$$

5.3.2 Multi-Stage Distillation Framework

Jiao et al. introduced a two-stage learning framework during the development of TinyBERT [53], whereby distillation occurs in each stage of the PLM training framework. More specifically, the two stages are referred to as:

- **Task-Agnostic (or Generic) Distillation**
  This occurs during the *pretraining* stage, during which a student model is distilled from a *pretrained* teacher model. This first stage is noted to be critical in improving the smaller student model's generalisation capability.
- **Task-Specific Distillation**
  This occurs during the *finetuning* stage. The pretrained student model from the first stage now learns from a *finetuned* teacher model. This stage enables the student to learn the task-specific knowledge of its teacher model.

Given the success of TinyBERT, the present paper will adopt a similar distillation framework.

Task-Agnostic Distillation

For the first stage, rather than distilling our own **task-agnostic** student model, the present paper will use **L6xH384 mMiniLMv2** [2] [56], a task-agnostic model distilled from XLM-R-large using self-attention relation.

Task-Specific Distillation

For the **task-specific** distillation stage, as our best performing model is one with double finetuning (see Section 5.2), the present study, inspired by TinyBERT's multi-stage distillation framework, will investigate various task-specific distillation pipelines.

Figure 5.3 gives an overview of the five distillation pipelines considered. Model 1 and 4 are models finetuned purely without Knowledge Distillation, and we use these as baselines to assess the impact of Knowledge Distillation in double and single finetuned models respectively. Table 5.3 reports the F1-scores of the different models on the three `EmpatheticPersonas` test sets (ZH (translated), ZH (native) and EN).

Our results indicate that double finetuning yields better performance over single finetuning as well in distilled models. Moreover, Knowledge Distillation is beneficial to overall model performance, whereby higher averages and lower standard deviations in performance across all 3 test sets are obtained when compared to their respective baselines. More specifically, the model distilled at each finetuning stage yielded the best results. Hence, the paper will proceed with the **2-finetuned 2-teachers** distillation pipeline (illustrated by Model 3 in Figure 5.3).

[2] Original code on GitHub: https://github.com/microsoft/unilm/tree/master/minilm
Model on Hugging Face: https://huggingface.co/nreimers/mMiniLMv2-L6-H384-distilled-from-XLMR-Large

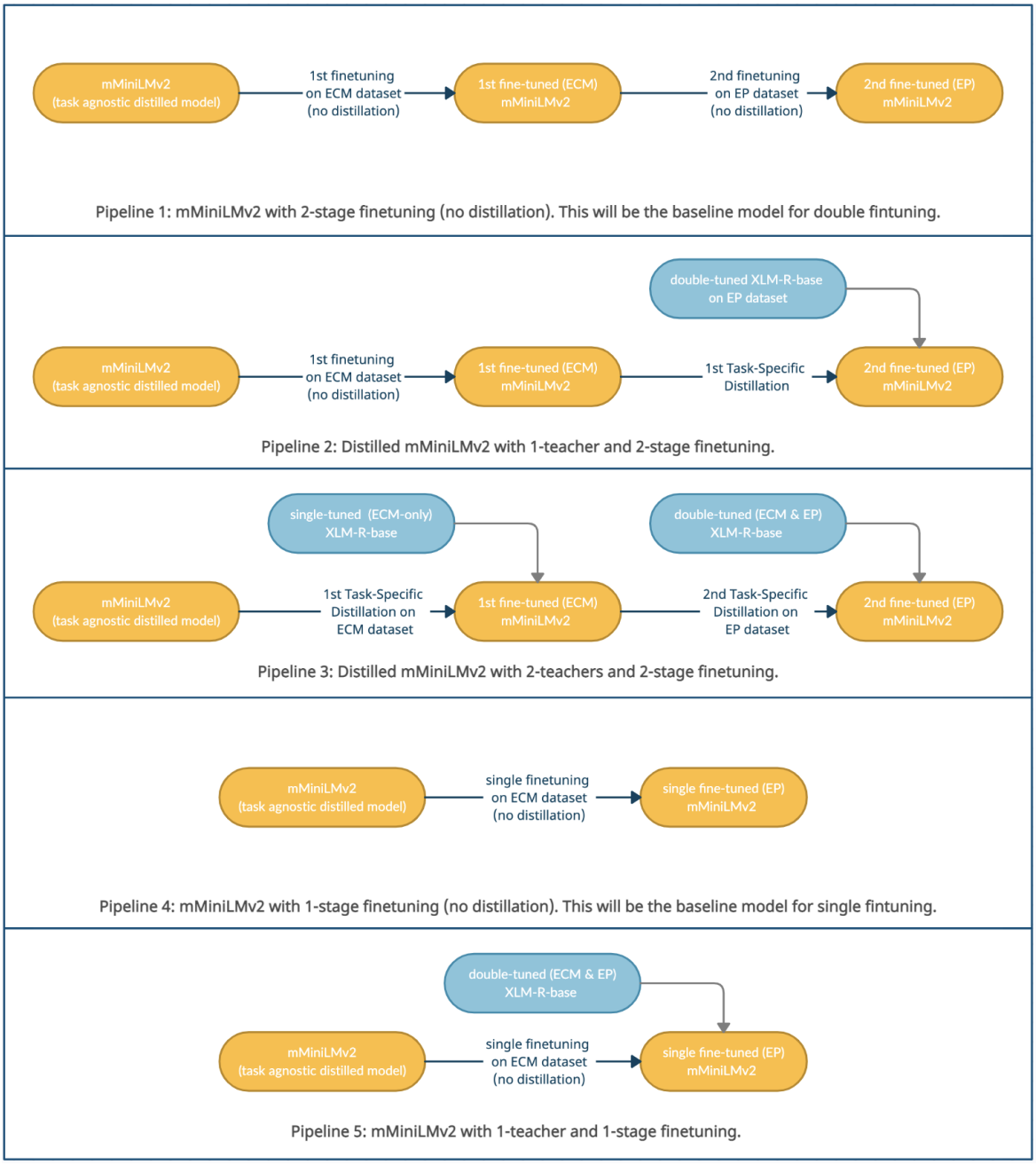


Figure 5.3: Distillation pipelines investigated. Pipeline 1 and 4 do not involve Knowledge Distillation and will be taken as baselines to assess its effectiveness in the other pipelines. Teacher models are represented in blue and student models in yellow.

	Config	ZH Translate	ZH Native	EN	Average	Std Dev
		%	%	%	%	%
2 tune	Baseline (Model 1)	89.64	73.80	84.67	82.70	8.10
	1-teacher (Model 2)	90.44	73.53	85.30	83.09	8.67
	2-teachers (Model 3)	**88.70**	**77.57**	**83.72**	**83.33**	**5.58**
1 tune	Baseline (Model 4)	89.41	72.47	80.98	80.95	8.47
	1-teacher (Model 5)	85.22	78.95	81.74	81.97	3.14

Table 5.3: F1-Scores of various distillation pipelines. Model names correspond to those shown in Figure 5.3. Baseline models are mMiniLMv2 models finetuned without Knowledge Distillation. The average and standard deviation of across the 3 `EmpatheticPersonas` test sets have been tabulated, with **Model 3** attaining the best performance overall.

Following hyperparameter tuning (refer to Appendix A), the F1-Scores of the final distilled model are tabulated in Table 5.4. A thorough analysis of the model's performance will be discussed in the next section. Given its overall performance, in the final *SATbot* implementation, we adopted the **distilled mMiniLMv2 model** as our emotion classifier.

Model	ZH Translate		ZH Native		EN	
	Accuracy	F1-Score	Accuracy	F1-Score	Accuracy	F1-Score
	%	%	%	%	%	%
XLM-R-base	93.86	93.86	90.00	90.09	91.23	91.4
mMiniLMv2	**91.22**	**91.31**	**80.83**	**80.85**	**85.09**	**85.39**
% of XLM-R	97.19	97.28	89.81	89.74	93.27	93.42

Table 5.4: Performance of the final distilled model (2-finetuned 2-teacher) on the 3 `EmpatheticPersonas` test sets, given also as a percentage of the XLM-R-base teacher model. The XLM-R-base performance is also included for ease of comparison. Despite mMiniLMv2 having only 50% of the XLM-R-base teacher model's capacity, it is able to retain a significant proportion of the teacher model's performance (approx. 90% in the worst case and up to 97%).

5.4 Evaluation against Past Works

Paper	Model	ZH Translate	ZH Native	EN
		%	%	%
Alazraki (2021)	roberta-base	-	-	95.1
Hu (2022)	infoxlm-large	96.64	93.33	91.06
Law (present)	xlm-roberta-base	93.86	90.09	91.4
Law (present)	mMiniLMv2	91.31	80.85	85.39

Table 5.5: Model results against different `EmpatheticPersonas` test sets. *Law* refers to the results of the present paper, while *Alazraki* [6] and *Hu* [17] are past works (see Section 1.3). This table includes only results of their best models.

In this section, we will compare our best performing models against previous *SAT-bots*', namely Alazraki's [6] and Hu's [17]. Given that accuracy scores are similar to F1, and that F1 is the main metric of concern, we compare only F1-Scores for simplicity. We note also the differences in models used (eg. size, mono-/multilingual) across papers which makes direct comparisons difficult, and hence provide more qualitative analysis. Results are tabulated in Table 5.5.

Comparison against Hu [17] - Multilingual InfoXLM-large

When comparing model performance against Hu's multilingual model, it should be noted that Hu uses a *BERT-large* sized architecture, which is twice the size of the BERT-base sized architecture and four times the size of the mMiniLMv2 architecture used in this study.

Considering that on average, XLM-R-large models outperform XLM-R-base models by 3.5% [38] due to their increased capacity, the current paper's *BERT-base sized model* performance which differs with Hu's only by -2.78% and -3.24% in the `EmpatheticPersonasZH` translated and native test sets respectively indicate that the model is actually highly competitive. Moreover, the XLM-R-base model of the present study even outperforms Hu's InfoXLM-large model on the English test set by +0.34%.

With regards to the present paper's *distilled model*, the distilled mMiniLMv2 is able to achieve 95%, 87% and 94% of Hu's performance on the `EmpatheticPersonas` ZH(Translated), ZH(native) and EN respectively with only 25% of Hu's model capacity. We also note that in the human trial, our emotion classifier evaluation results, where we used the distilled model, actually outperforms Hu's (see Chapter 8).

The distilled model's biggest benefit, however, is with regards to inference time and capacity saved. Table 5.6 measures the time taken to perform one pass in Hu's model

Model	Inference Speed (s)
InfoXLM-large (Hu)	94.15
mMiniLMv2 (Law)	5.63

Table 5.6: Time taken to compute one inference.

Model	Layers	Hidden Dim.	No. of Parameters
InfoXLM-large	24	1024	355 mil
L6 x H3834 mMiniLMv2	6	384	107 mil

Table 5.7: Model size comparisons.

and the distilled model (mMiniLMv2). As shown, the distilled model runs 16 times faster than Hu's large model during inference. It is also 4 times smaller, making the application much cheaper to host and deploy. Considering the computational advantages and minimal performance trade-off, this result illustrates the potential of distilled models, whereby model performance can be largely recreated with a significantly smaller and efficient model.

Comparison against Alazraki [6] - Monolingual RoBERTa-base

While the present paper's XLM-R model is the same size as Alazraki's (i.e. both are BERT-base sized), Alazraki's model is a *monolingual* model in English, a high resource language, and ours a *multilingual* model. As such, a reduction in performance is expected due to the effect of capacity dilution in multilingual models, whereby limited model capacity is shared between languages [36].

Moreover, publicly available training data in Mandarin is significantly limited in comparison to English, an issue common in NLP research conducted in non-English languages [36]. For the first finetuning, Alazraki's model was trained on a significantly larger Emotions [57] dataset (approximately 190,000 data points were used). However, the largest corpus the present paper was able to obtain in Mandarin was the ECM dataset, from which only 11323 data points could be used (approx. 6% of the Emotions dataset).

Despite so, the present study's XLM-R model still performs relatively well in comparison, differing only by -3.7% on the English test set. As expected, with regards to the distilled mMiniLMv2 model, the performance difference is more significant at -9.71%. Despite so, we would again like to highlight that in the human trials (see Section 8), our model performance surprisingly outperforms Alazraki's models significantly. Ultimately, taking into account the countless advantages of multilingual models, such as ease of future scalability into other languages and cross lingual transfer benefits, the present study believes such a trade-off is marginal.

Chapter 6

Empathetic Rewritings

Empathetic Rewritings refer to the *rewriting* of base utterances to contain high levels of *empathy*. As per previous works, we focus specifically on Geoffrey T. Barrett-Lenard's second phase of empathetic exchange, which is to express empathy as a display of compassion towards the user [58]. This is crucial for the *SATbot*, as empathy is noted to be instrumental in the provision of mental health support and hence, success of such platforms [59]. Moreover, as the chatbot is rule-based and has a fixed conversational flow (refer to Chapter 2), rewritings can bring about greater diversity and variation in conversation and lead to greater engagement.

Hu [17] carried out Empathetic Rewritings in Mandarin using generative language models. While it yielded utterances that were generally fluent, it was noted that they were less native-sounding/ lacked colloquialism, influenced by the English dataset that the training data was translated from (eg. utterances like "我拿错了棍子的一端" which literally translates to "*I got the wrong end of the stick*" that are not a typical of the Mandarin language were found).

The present paper hypothesizes that this is a result of extensive finetuning on the translated dataset. Specifically, Hu performed supervised training on Chinese GPT-2 using the machine translated `EmpatheticPersonasZH` as a parallel dataset over 450 epochs. As such, we believe this eroded the model's native linguistic knowledge.

Progressing in the same vein as Hu [17], this paper will adopt the use of generative language models to generate empathetic rewritings in Mandarin. As Hu's findings have indicated that multilingual generative models significantly underperforms monolingual Chinese GPT-2 for Chinese text generation, our implementation will use **Chinese GPT-2** only.

However, an alternative training methodology: *reinforcement learning using proximal policy optimisation* [60], will be adopted in aims to tackle the limitations faced in his implementation. Rather than the supervised training methodology which uses the `EmpatheticPersonasZH` dataset directly to train text generation, this methodology will use the dataset only to train classifiers. These classifiers will then subsequently be used to generate rewards for the model to encourage empathetic text generation

that is fluent and semantically accurate during training. As the use of the dataset is more *indirect*, we hypothesize that this will diminish the impact of the translated dataset on the LM. The following sections describes the training setup in detail.

6.1 Supervised Warm-Start

Before commencing reinforcement learning, literature has shown that introducing LMs to the task via **supervised learning** prior leads to more effective learning [59, 61], and this has been proven true in our initial experimentation. By performing supervised warm-starts, we expose the model to:

- empathetic, psychotherapeutic-style speech, and
- the rewriting task (rather than text continuation, which the LM was originally trained for)

prior to reinforcement learning.

Inspired by [59, 62], we adopt a "reverse engineering" strategy for the supervised learning, which utilises parallel dataset of low-high empathy pairs and [HIGH]/[LOW] tokens to control transformations. The prompts are as such:

[HIGH] emotion [SEP] base utterance [REWRITE] empathetic rewriting

[LOW] emotion [SEP] empathetic rewriting [REWRITE] base utterance

whereby:

- *emotion*
  The emotion attached to the base utterance. Either:
  愤怒(Anger)，焦虑(Anxious)，快乐(Happy)，悲伤(Sadness)，所有情绪(All Emotions)
- *base utterance*
  The base utterance to be rewritten
- *empathetic rewriting*
  Rewritings from the `EmpatheticPersonasZH` dataset (see Section 3.1)
- *[HIGH]/[LOW] token*
  low-to-high/ high-to-low empathy transformation respectively

The dataset was also upsampled to achieve an approximately balanced distribution across base utterances, without which, we noted poor utterance generation for the minority utterances. This yields a training dataset with 5673 data points. We perform supervised warm-start for only 1 epoch with learning rate 1e-05.

6.2 Reinforcement Learning with Proximal Policy Optimisation (PPO)

The reinforcement learning (RL) framework is defined as follows:

- **States ($\mathcal{S}$)**
  The model prompts. This is similar to that in supervised warm-start (Section 6.1), however using only [HIGH] transformation token and omission of the "empathetic rewritings" after the [REWRITE] token as this is the action to be taken by the agent:

  [HIGH] emotion [SEP] base utterance [REWRITE]

- **Actions ($\mathcal{A}$)**
  The possible empathetic rewritings generated by the LM.

- **Policy (π)**
  This is defined by the generative language model (i.e. Chinese GPT-2).

- **Rewards ($\mathcal{R}$)**
  The reward is given as:

$$R(x, y) = r(x, y) - \beta log \frac{\pi(y|x)}{\rho(y|x)} \tag{6.1}$$

  whereby:

  - $r(x, y)$: reward model (see Section 6.3)
  - $\beta log \frac{\pi(y|x)}{\rho(y|x)}$: adaptive Kullback-Leibler (KL) Divergence penalty, included as part of PPO (see below)

To train the policy, we adopt **Proximal Policy Optimisation (PPO)**[60], with reference to the implementation in [61].

Proximal Policy Optimisation (PPO)

PPO [60] is a Trust Region Policy Optimisation (TRPO) Method [63]. In trust region methods, complex objective functions are approximated using simpler quadratic surrogate functions, enabling more efficient implementation. Trust regions are defined as the bounds within the surrogate function where the surrogate is a good approximation of the original objective function. By executing optimisation within the trust region, we can ensure good convergence and monotonic improvement in the policy [64].

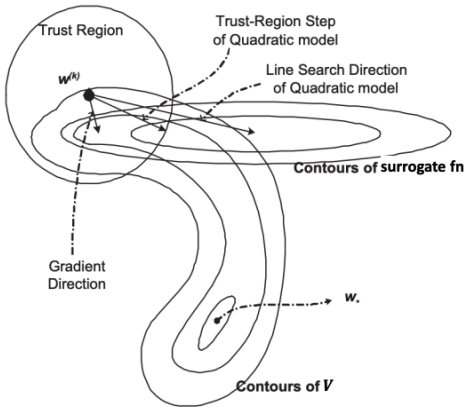


Figure 6.1: Illustration of the trust region method taken from [64]. V represents the original objective function that is highly complex while the surrogate function is an approximation of the original objective. The trust region then encompasses the region where the surrogate and V overlaps (i.e. where the surrogate is a good approximation of the original function).

PPO further simplifies TRPO to enable more efficient optimisation. There are 2 ways to implement PPO. In this paper we use the adaptive Kullback-Leibler (KL) Divergence penalty (i.e. the second term in the reward function). The term "adaptive" is used as β is varied dynamically to achieve a target KL value (KL_{target}) as such:

$$e_t = clip(\frac{KL(\pi_t, \rho) - KL_{target}}{KL_{target}}, -0.2, 0.2) \tag{6.2}$$

$$\beta_{t+1} = \beta_t(1 + K_\beta e_t) \tag{6.3}$$

In PPO, KL Divergence is used to penalise parameter updates that result in large deviations in the active policy (π) from the reference policy (ρ), thereby preventing destructively large policy updates. This also serves as an entropy bonus to encourage model exploration. Finally, a study done by the OpenAI team using PPO to finetune transformers for human preferences [61] has also found that it encourages coherence and semantic preservation during text generation.

Figure 6.2 illustrates how training is performed using PPO. At each step of the training, the model outputs a rewriting based on the prompt. We then compute the semantic, fluency and empathy rewards for the rewriting, as well as the KL Divergence Penalty for the current policy. The policy is then updated based on the rewards and advantages computed [1]. For details on parameters used during training, see Appendix B.

[1]Policy updates are implemented using the *trl* library: https://github.com/lvwerra/trl

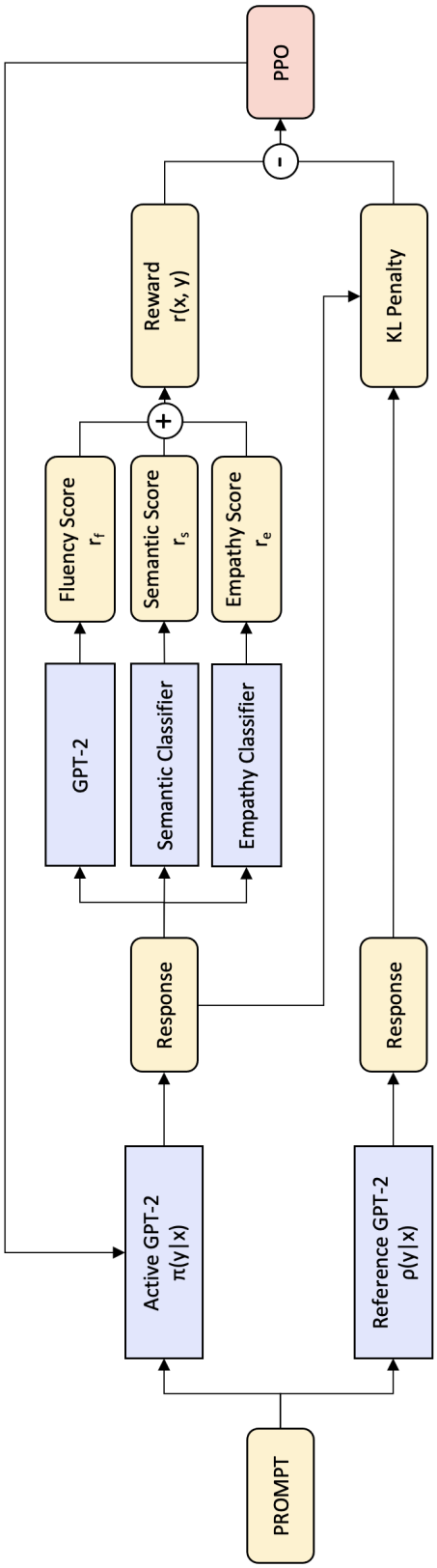


Figure 6.2: The reinforcement learning training set-up, inspired by [59] and [61]. The active GPT-2 (π) constantly undergoes PPO updates while the reference GPT-2 (ρ) is only used for monitoring policy divergence.

6.3 Reward Model, $r(x, y)$

The reward model is comprised of reward metrics we seek to optimise. More specifically, for the task of empathetic rewritings, we seek to reward utterances that are first and foremost **empathetic**, but also **fluent** and **semantically-relevant**. The following subsection describes each component in greater detail.

6.3.1 Empathy Reward, r_e

Empathy reward is the key component to the empathetic rewritings task. This component seeks to reward rewritings that contain high levels of empathy and penalise rewritings with low empathy. To quantify the amount of empathy contained within an utterance, an **empathy classifier** was developed. The *logits to the highly empathetic class* computed by the empathy classifier is then taken as the empathy reward.

Empathy Classifier

An empathy classifier was developed using XLM-R [2] trained on the `EmpatheticPersonasZH` 1100 empathy-annotated dataset (see Section 3.1). While the dataset was originally annotated for 3 classes (non-, slightly- and highly-empathetic), we modify the dataset to contain only 2 classes, merging the non- and slightly-empathetic classes into a single class, and use this to train a binary empathy classifier. The rationale behind this was due to the following:

- The empathetic rewriting task seeks to transform our utterances into highly-empathetic rewritings and hence, we are only concerned with the highly empathetic class.
- Previous studies indicated that empathy classifiers performed poorly at classifying "slightly-empathetic" utterances, given the subjectivity of the task [6, 17]. Hence, we eliminate this class to prevent "confusing" the empathy classifier.

Following hyperparameter tuning (see Appendix A), the binary empathy classifier has an overall accuracy and F1-Score of **90%** with the following class-based performance:

Empathy Class	Precision	Recall	F1-Score
	%	%	%
Non/Slightly	91%	95%	93%
Highly	**89%**	80%	84%

Table 6.1: Empathy Classifier performance on the `EmpatheticPersonasZH` empathy test dataset.

[2] XLM-R was used as per the Emotion Classifier, following findings in Section 5.1.

Overall, the empathy classifier displays strong performance. Most importantly, it has a high precision [3] of 89% for classifying texts in the highly-empathetic class. This is crucial as it indicates low likelihood of misclassifying rewritings (i.e. false positives), ensuring trained utterances are truly highly empathetic when classified as such and suited for our mental health application.

6.3.2 Semantic Reward, r_s

The **semantic reward** seeks to reward rewritings that carry the same semantic meaning as the base utterance. Without this component, sentences that are highly empathetic but do not carry the correct semantic meaning will be generated as the model seeks to exploit the empathy reward. To measure semantic similarity of the rewriting to its base utterance, we devise a semantic classifier. The semantic reward is the *logit for the semantic class corresponding to the base utterance* (eg. if it was a rewriting for base utterance "这是由特别事件引起的吗", the semantic reward is the logits obtained for class 0).

Semantic Classifier

A semantic classifier was developed using XLM-R [4] trained on the 2144 `EmpatheticPersonasZH` empathetic rewritings dataset (see Section 3.1). Following hyperparameter tuning, we yield a semantic classifier with an accuracy, macro-F1, macro-precision and macro-recall of **96%**. For more details on the Semantic Classifier (i.e. utterance-to-class mapping, hyperparameter tuning, class-based performance), see Appendix A.

6.3.3 Fluency Reward, r_f

The **fluency reward** was included to prevent rewritings that are highly empathetic but are incoherent/grammatically incorrect. This is computed as:

$$r_f(er) = \frac{1}{PPL(er)} - RP(er) \tag{6.4}$$

whereby:

- er: empathetic rewriting
- $\frac{1}{PPL(er)}$: inverse perplexity (computed by Chinese GPT-2)
- $RP(er)$: repetition penalty

The repetition penalty was included following observations of the model repeating keywords/empathetic terms in attempts to exploit the semantic and empathy reward.

[3] Precision = true positives / predicted positive = TP/(TP+FP)

[4] XLM-R was used as per the Emotion Classifier, following findings in Section 5.1.

6.3.4 Total Reward, r

The total reward is implement as a multi-objective function comprised of the weighted sum of the empathy, fluency and semantic rewards, written as:

$$r = w_e r_e + w_f r_f + w_s r_s \tag{6.5}$$

6.4 Evaluation

For each of the 45 base utterance-emotion pairs, 25 rewritings were generated (without duplication), yielding a total of 1125 utterances. Inference was conducted using top-p (nucleus) sampling, with parameters "temperature" and "repetition penalty" varied to obtain more variation and diversity in utterances [65]. Table 6.2 are some examples of empathetic rewritings generated for each emotion category.

We evaluate the rewritings for fluency and empathy and compare these against Hu's [17]. Fluency is measured as the inverse perplexity, while empathy is measured as the percentage of rewritings classified as being highly empathetic by the empathy classifier in Section 6.3.1. The results are as follows:

	Fluency		**Empathy**
	Mean	**Std Dev.**	**%**
Hu (2022)	18.46	8.80	93.07
Law (present)	17.86	6.88	89.68

Table 6.3: Empathetic Rewritings evaluation. Perplexity measures the fluency of the rewritings generated (lower better), computed using GPT-2. Empathy measures the percentage of rewritings classified as being highly empathetic by the empathy classifier in Section 6.3.1.

With regards to **empathy**, the present paper reports a lower percentage of utterances (-3.39%) as being high empathy compared to Hu's [17]. However, considering almost 90% of utterances are classified as high empathy, this performance is comparable.

Regarding fluency, the new methodology is shown to yield *improved performance* on this aspect, with a lower **perplexity** score (-0.6 mean, -1.92 standard deviation) when compared to Hu's [17], indicating better alignment to native Chinese speech. In the introduction of this chapter, the paper also highlighted a key concern of using translated `EmpatheticPersonasZH` dataset towards the empathetic rewritings generated, which is the presence of non-native phrases derived from English speech. While there is no measurable metric at present to measure this aspect, we requested feedback from participants regarding this during the human trial and note improvement in participant sentiment during the evaluation over Hu's trial (see Section 8).

Emotion	Base Utterance	Empathetic Rewriting
焦虑 *Anxiety*	你认为你应该成为别人的救星吗? *Do you believe that you should be the saviour of someone else?*	你是否认同你可能正在尝试拯救他人? 如果是这样，也许是因为这个事件而导致你感到一些不愉快的情绪。 *Do you agree that you might be trying to save others? If this is the case, perhaps this might be why you are feeling unhappy.*
悲伤 *Sadness*	这是由特别事件引起的吗? *Was this caused by a specific event/s?*	我很遗憾听到你感觉不舒服。是不是有什么特定的事情让你有这样的感觉? *I'm sorry to hear that you're feeling unwell. Was it a particular event that made you feel this way*?
快乐 *Joy*	那很好！让我推荐一个你可以尝试的练习。 *That's Good! Let me recommend a protocol you can attempt.*	我很高兴你心情很好。我可以向你推荐一个练习，你可以尝试让自己感觉更好！你怎么认为? *I'm so happy that you're feeling good. Could I recommend you a protocol, you can try it to help yourself feel even better! What do you think?*
愤怒 *Anger*	这是由最近或遥远的事件（或多个事件）引起的吗? *Was this caused by a recent or distant event(s)?*	你能告诉我更多关于这件事发生的具体时刻，以便我们进一步了解情况。如果你不介意，请问这事是最近或很久以前发生的? *Could you please tell me more about this event's timing, so we can better understand the situation. If you don't mind, may I ask if this was a recent or distant event?*
所有情绪 *All Emotions*	感谢你的参与。再见。 *Thank you for taking part. See you soon.*	我非常感谢你的参与，希望我们能很快再次见面。 *I am very grateful for your participation, hope we will see each other again soon.*

Table 6.2: Examples of empathetic rewritings generated for each emotion category.

6.5 Limitations of Methodology

While the methodology yields successful performance overall, we would like to acknowledge that it is **extremely sensitive**. As PPO is a stochastic policy method, its actions are drawn from a probability distribution. This means that actions vary each time, resulting in starkly different outcomes between different runs of training. Moreover, performance varies significantly based on the weights attached to the reward components (i.e. w_e, w_s and w_f), which makes hyperparameter tuning difficult. In comparison, a supervised training methodology might yield more stable and predictable performance outcomes.

Secondly, we note that rewritings generated using this training methodology contained certain random words (eg. "简介:" [*introduction:*] and "mobile") at the beginning and end of rewritings, and it is not clear why this occurs. We clean the rewritings of these instances prior to being used by the chatbot.

6.6 Preventative Measures: Toxicity

As discussed in Section 4.2, GPT is known to display concerning toxic behaviour [45]. Given that *SATbot* is a mental-health application, such texts are unacceptable and it is crucial empathetic rewritings are vetted for such instances.

One of the precautions the present paper has taken was the **removal of personas**. In previous *SATbot*s, 5 personas of different ages and sexes were included. However, the inclusion of personas was observed to lead to sexist remarks from being generated during training. Moreover, we note little variation in the persona data (eg. rewritings of a 18-39 female are indistinguishable from the rewritings of a 40-69 male) which makes training of different personas difficult. Similar comments were also noted in previous works [6, 17]. Hence, after careful consideration, we have omitted them in this paper for safety.

Utterances have also been pre-generated and **manually inspected** for any toxic speech or distressing content (eg. violence and self harm) before use in the chatbot. It is however promising to note that no negative content was found in utterances generated by the final trained model. In the future, a hate-speech detector could be devised to automate this inspection process.

Chapter 7

Web Platform and User Interface

The *SATbot* has been developed as a web application to enable ease of access by users. In this chapter, we discuss the architecture of the *SATbot* application and our contributions.

7.1 Architecture

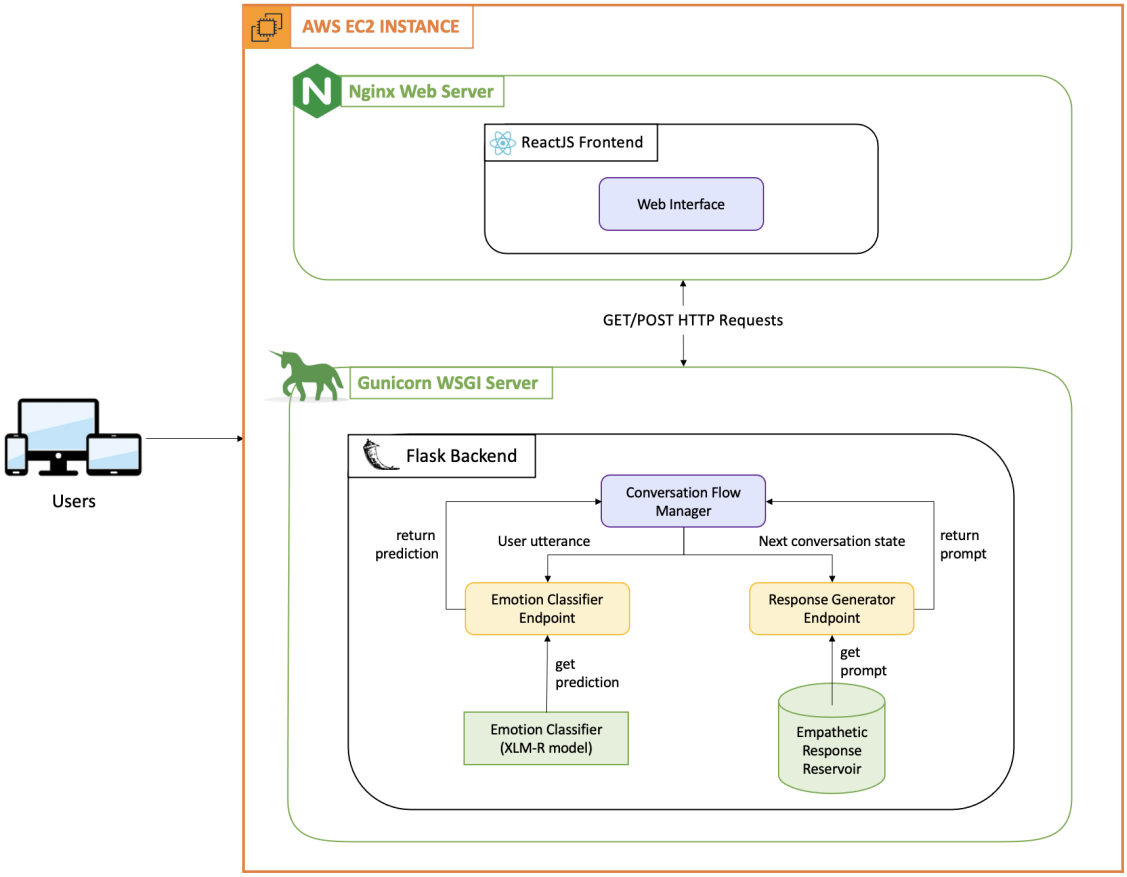


Figure 7.1: *SATbot* web application architecture.

Figure 7.1 illustrates the architecture of the *SATbot* web application, which is largely based upon previous works [6, 13, 17]. It involves a **ReactJS** frontend, which utilises the **react-chatbot-kit** library [66] for the chatbot components. The backend API is developed using the **Flask** web development framework in Python3. Communication between the frontend and backend API is done via GET and POST HTTP requests.

For deployment, we utilise **Nginx** to serve the React frontend and **Gunicorn** to serve the Flask backend. The application is deployed on an AWS EC2 instance in the cloud and is accessible using web browsers.

7.2 Key Contributions

Our key contributions to the *SATbot* include:

1. **Bilingual (EN-ZH) Language Support**
   Motivated by the aims of this paper to extend SAT to increased users, we developed a fully bilingual chatbot capable of conversing in **both** English and Mandarin [1]. Protocols in both languages are also shown in the protocol viewer.

   Previous versions: Only single languages were supported (either English-only [6] or Mandarin-only [17]).

2. **Updated User Interface**
   While maintaining the core components of previous *SATbots* (the chatbot and protocol viewer), we updated the user interface, including colours and a more modern, minimalist aesthetic (see Figure 7.2) for better clarity and engagement.

Figure 7.2: *SATbot* user interface.

[1]Note that, as the present paper only carried out Empathetic Rewritings in Mandarin, English utterances are from Alazraki's work [6]

Previous versions:

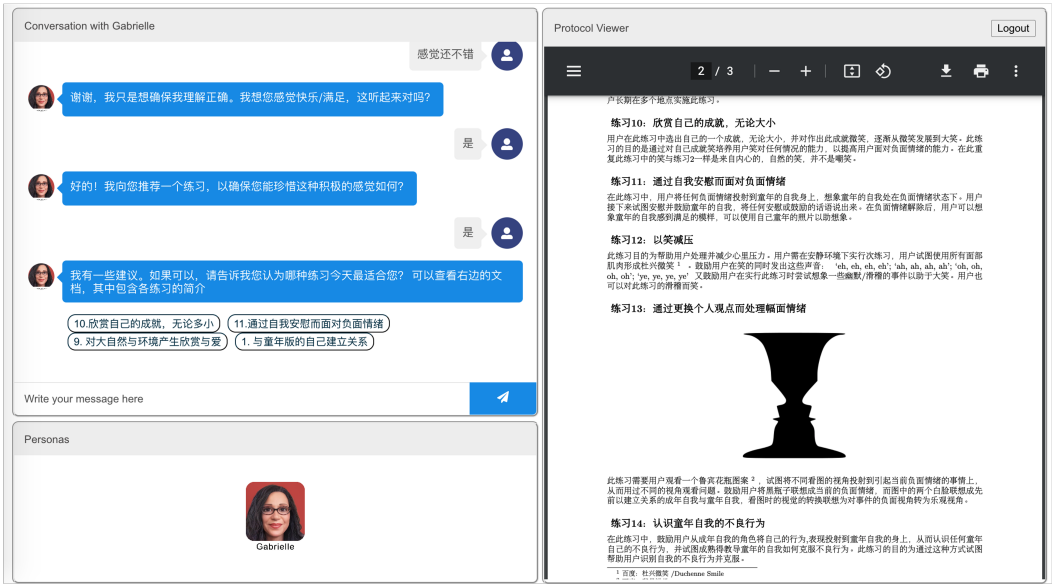


Figure 7.3: Previous *SATbot* versions' user interface, which comprises of a chatbot conversation pane and a PDF viewer for protocols.

3. **Increased Interaction**
   To make navigating the long list of protocols less overwhelming for users, we created an interactive protocol viewer which displays only a single protocol at a time. To view protocols, users can interact with the protocol pane on the right to select and view their protocols of interest.

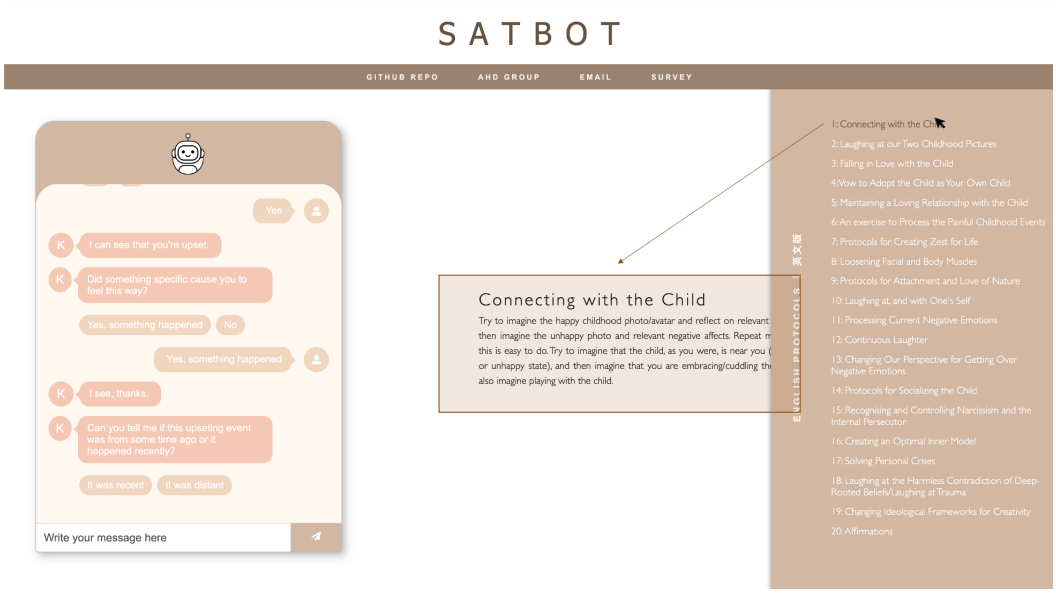


Figure 7.4: Protocol appears in the protocol viewer when user selects from the protocol pane.

Protocols are also displayed in the viewer when users select a protocol suggested by the chatbot.

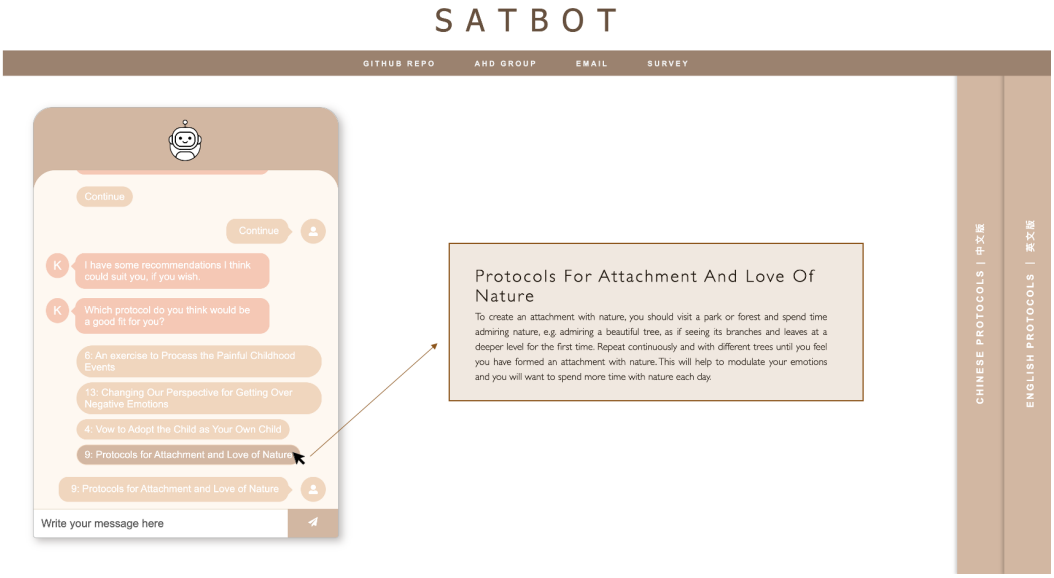


Figure 7.5: Protocol appears in the protocol viewer when user selects from the list of protocol suggestions provided by the chatbot.

Previous versions: A non-interactive PDF viewer which displayed all protocols (see Figure 7.3).

4. **Mobile-compatible**
   Following feedback from users (see next Chapter), the website is viewable on all screen sizes, even mobile phones.
   Previous versions: Not mobile compatible.

Figure 7.6: *SATbot* on mobile.

Chapter 8

Non-Clinical Trial

A **non-clinical human trial** was carried out for the formal evaluation of the *SATbot*. As the chatbot is bilingual, participants were required to be fluent in both Simplified Chinese and English. Given the limited participant pool, knowledge in SAT protocols or psychotherapy were beneficial but not required. In total, 27 participants (10 female, 17 male) between the ages of 25-60 consented to the trial.

At the start of the trial, participants are provided with the link and login credentials to access the web platform. Throughout the trial, participants are instructed to abide by the following:

1. To interact with the application once per day across the period of 5 days.
2. To have a minimum of 3 interactions in Chinese and 1 in English [1]
3. To note down any unnatural sounding utterances generated by the chatbot when using the Chinese setting.

At the end of the trial, an anonymous feedback questionnaire is issued to each participant for them to evaluate their experience using the *SATbot*.

8.1 Questionnaire Responses

The questionnaire is primarily comprised of 5-point Likert Scale questions (ranging from strongly disagree to strongly agree), and contains also open-ended text responses to allow participants to provide additional comments and/or elaborate on the reasoning behind their choices.

The questionnaire aims to evaluate each of the chatbot components that have been designed, these include the chatbot's:

- Emotion prediction capabilities (relates to Chapter 5)
- Quality of empathetic rewritings (relates to Chapter 6)

[1] We emphasise Chinese interactions as it is the main focus of the present paper.

- User interface and user experience (relates to Chapter 7)
- Overall experience

We summarise the findings from the questionnaire in the subsections below. We include also comparisons with previous work (where applicable). While comparing against previous chatbots, we would like to highlight the difference in participant pools between studies (Hu [17] and Alazraki [6] had 14 and 16 participants respectively, while the present paper has 27), and hence emphasise that results are purely indicative. A full compilation of participant responses can be found in Appendix C.

8.1.1 Responses Regarding Emotion Classification

As the emotion classifier developed in Chapter 5 is multilingual, we had participants assess the chatbot's emotion predictions in both English and Mandarin.

"The chatbot was good at guessing my emotions in *English*."

89% of participants agreed to this statement, out of which most participants **strongly** agreed. This is a significant increase over Alazraki's [6] where only 63% of participants agreed. This finding is surprising as classifier training was primarily Mandarin-focused, and highlights the immense capabilities of multilingual models.

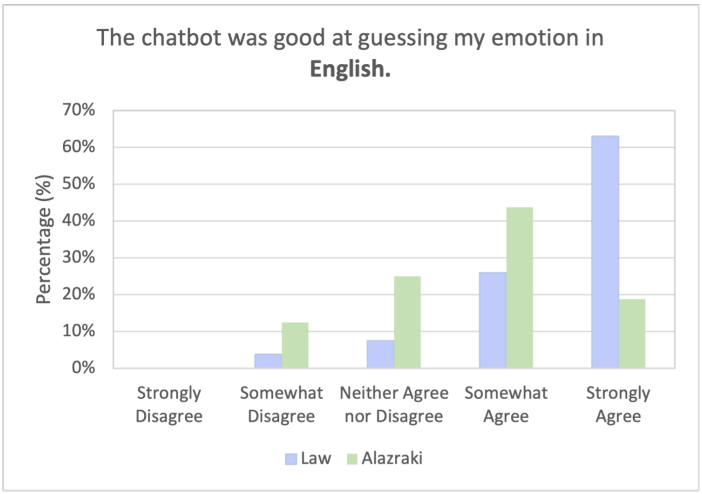


Figure 8.1: Participant evaluation of emotion classification performance in English.

"The chatbot was good at guessing my emotions in *Mandarin*."

A stronger performance in Mandarin emotion classification over English was observed, with 93% of participants agreeing that the classifier correctly guessed their emotions in Mandarin. This is as expected given that model training was Mandarin-focused, as previously mentioned.

The model performance in Mandarin is slightly better than Hu's [17], where 89% of participants agreed. However, our evaluation finds a larger percentage of participants who **strongly** agreed to the statement. Taking into account that the emotion classifier used in this paper contains only 25% of Hu's model's capacity and runs 16 times faster at inference, this performance highlights the success of Knowledge Distillation in the current *SATbot*. We note also several comments from participants commending the speed of the chatbot.

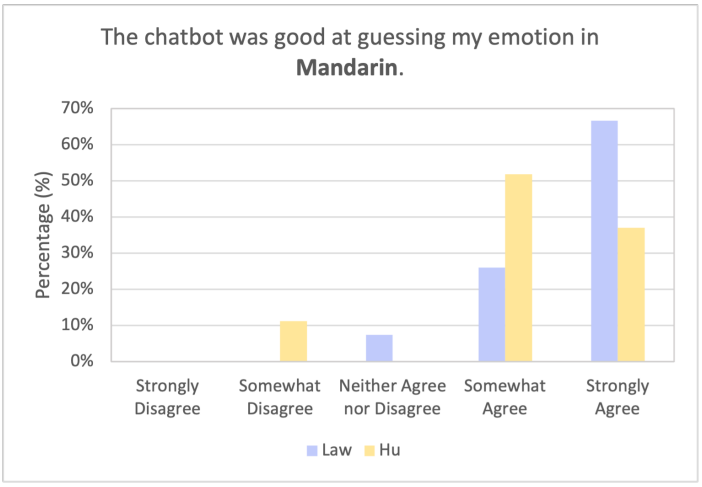


Figure 8.2: Participant evaluation of emotion classification performance in Mandarin.

Overall Remarks

In both cases, a small population of participants selected "somewhat disagree" and "neither agree nor disagree" (11% for English and 7% for Mandarin). A general feedback was that the participant's emotions were not classified under any of the 4 emotion categories that the model was trained to identify. This has been a prominent issue brought up in previous works and is currently being addressed by another student.

8.1.2 Responses Regarding Empathetic Rewritings

"The chatbot sounded very empathetic throughout the conversation."

85% of participants agreed to this statement. This is closely aligned with previous chatbots, with Hu's at 86% and Alazraki's at 88%. The trend across the choices is also similar, with most participants stating they **somewhat** agree.

The main discrepancy with previous models is that a few participants disagreed with this statement in the present trial (1 participant strongly disagreed and 1 somewhat disagreed). In the open text response, participants explained that they felt the some of the empathetic texts were very forced upon and felt "pandering". The use of a fixed script by the rule-based chatbot also made the chatbot appear too "standard-

ised" and "generic" for true display of empathy.

In essence, while the chatbot is overall empathetic in conversation, it fails to capture the small nuances necessary for empathetic connection. To tackle this, we provide suggestions for future improvements in Chapter 9.

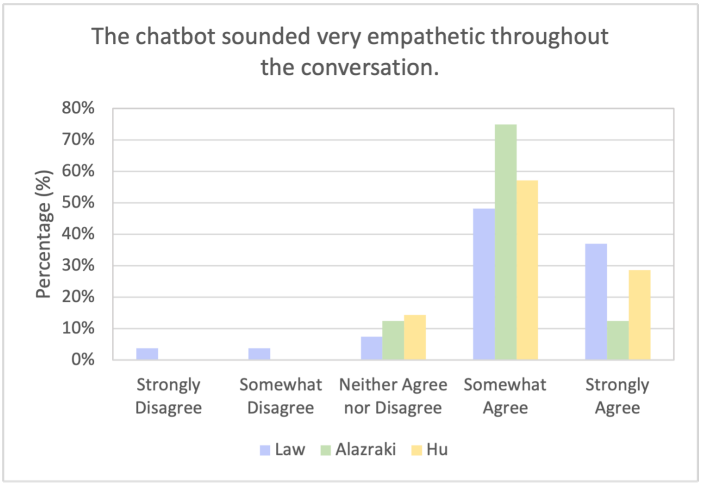


Figure 8.3: Participant evaluation of the chatbot's display of empathy during conversation.

"The chatbot provided fluent and natural-sounding responses in Mandarin."

96% of participants agreed to this statement, majority of whom **somewhat** agreed. 1 participant also selected neither agree nor disagree. Overall, participants commented that they were able to understand the utterances generated by the chatbot well. Explanations provided by participants that did not **strongly** agree to the statement primarily fall under the following:

- **Utterances felt too formal.**
  The Mandarin language contains polite and respectful terminology that are not commonly used in day-to-day speech. For example, the characters "您" and "你" both mean "you", but the former is used to refer to people politely (eg. addressing elderly, customers, senior officials) while the latter is more commonly used. This occurred as we adopted polite terminology in aims to enhance our empathy objective during text generation, unaware of how it would impact user's experience. We have now rectified instances highlighted by users.

- **Presence of English-style figures of speech not common to Mandarin.**
  This was one of the key points we emphasized in Chapter 6, that was also present in Hu's [17] chatbot. While a new training methodology was adopted to address this, it appears that the issue is still present in some cases (see Appendix C for the instances). This is expected as the use of the translated `EmpatheticPersonasZH` dataset, while reduced, is still involved in the generative model training to some extent (both in terms of the classifiers and supervised warm start-up). However, we do note **fewer** participant feedback

regarding this compared to Hu's for a larger participant pool, suggesting potential improvement in this aspect.

Comparing against Hu's chatbot where 77% of participants agreed to this statement, the current *SATbot* appears to yield significant improvement in this aspect, indicating the success of our adopted training methodology (RL with PPO) in training generative LMs to yield better quality outputs.

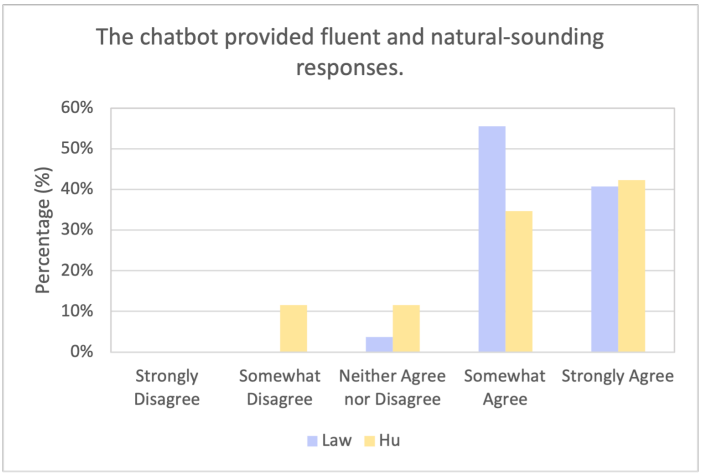


Figure 8.4: Participant evaluation of the chatbot's fluency and naturalness in conversation.

"The chatbot conversations were engaging."

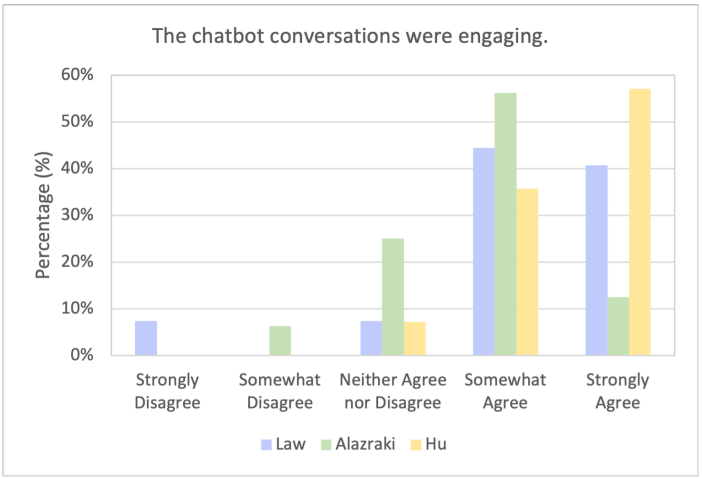


Figure 8.5: Participant evaluation of the chatbot's engagement.

85% of participants agreed to this statement. This is similar to Hu's trial which reported 93% agreement and much higher than the 69% reported in Alazraki's trial.

It was also noted that several participants disagreed with the statement, more specifically 2 participants strongly disagreed and 2 participants neither agreed nor disagreed. Participants attributed their decisions to the rigid nature of the rule-based chatbot, making it appear to "follow a predetermined script". Participants suggested the chatbot to be "more fluid" and include the use of more "free text input" (i.e. open dialogue). This response is similar to those received in Hu's trial, and is currently a project undertaken by another student.

8.1.3 User Interface

"I am satisfied with the chatbot user interface."

89% of participants agreed with this statement, out of which majority participants **strongly** agreed. Users noted the interface to be 'light-hearted', 'aesthetically pleasing' and 'calming' as a psychotherapeutic chatbot.

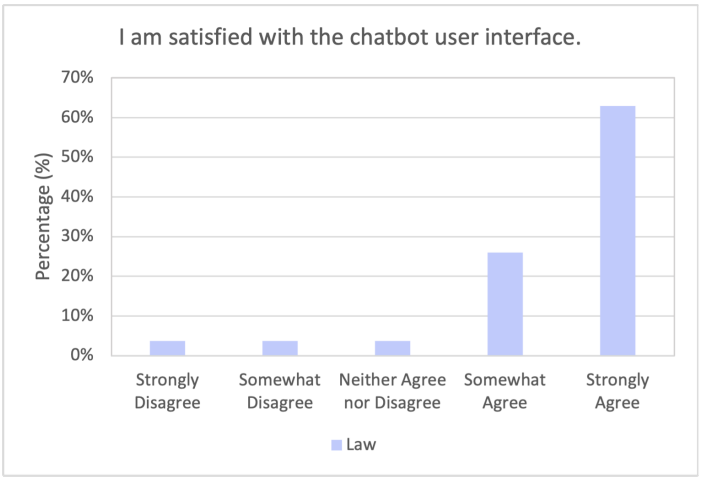


Figure 8.6: Participant evaluation of the chatbot's user interface.

As previous works did not carry out an assessment regarding user interface during their trials, no clear comparisons can be drawn. However, several participants familiar with the previous chatbot's interface noted that the improved features, such as an interactive protocol viewer and bilingual language support, helped "increase engagement" and improved their overall experience.

In our trial, we also note that some participants disagreed with the statement (1 strongly disagree, 1 somewhat disagree, 1 neither agree nor disagree). Participants explained that parts of the chatbot was not visible on their desktop device and that the application was also not mobile-compatible. Following this feedback, we have addressed this issue and reached out to participants who have confirmed that this issue is resolved.

8.1.4 Overall

"Overall, the platform was useful."

89% of participants agreed that the platform was useful overall. This percentage of participants in agreement is similar to the sentiments found in previous work, at 92% for Alazraki's and 93% for Hu's.

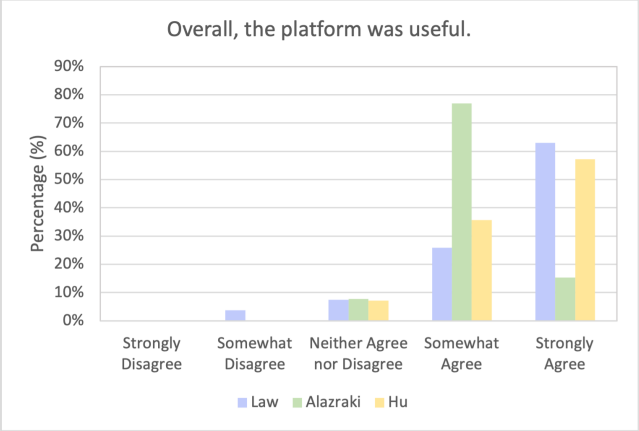


Figure 8.7: Participant evaluation of the chatbot's usefulness.

8.2 Study Limitations

The key limitation faced in this study is that majority of participants do not possess **knowledge of SAT protocols** or possess **background in psychology**. This was a necessary trade-off as participants were required to be English-Mandarin bilingual speakers, which significantly limited the size of the participant pool. As such, results from the trial evaluation, such as those pertaining empathy and the usefulness of the chatbot, may not be entirely reflective of the chatbot's performance from a psychology-specific point of view. In the future, trial participants should undergo training in SAT where they can be informed about the treatment and its protocols for better understanding before providing feedback. If possible, having Mandarin-speaking clinicians to participant in trials will also be extremely valuable.

Another limitation was with regards to the **study size**. While the present trial has recruited more participants compared to previous studies, this sample size is still relatively small. Moreover, the demographic is skewed towards males (63% male, 37% female). This is once again due to the stringent requirements of the trial screening which limited the participant pool. In future trials, recruitment should continue to increase the trial sample size and focus on balancing demographics.

Finally, given that the present paper focused on improving the *SATbot* user interface, a **second trial** was not executed due to the lack of time. While user feedback has been updated in the chatbot (where applicable), these could not be evaluated in a further trial to assess improvement.

Chapter 9

Future Work

Following the work done in the present paper, as well as through feedback from the human trial, we note several areas for improvement, as well as interesting opportunities for new exploration in future works. We discuss these according to the *SATbot*'s key components: emotion classification, empathetic text generation and user interface.

9.1 Emotion Classification

Code-Switching

Code-Switching, sometimes referred also as Code-Mixing, is a phenomena prevalent in multilingual communities, whereby individuals alternate between two or more languages within the same conversation [67]. As a potential input to the chatbot could contain code-switched text, it would be interesting to see how model performance can be optimised for such inputs.

To provide a baseline for future works, we developed a small English-Mandarin code-switched test dataset, `EmpatheticPersonasCS` (see Section 3.1) and assessed our present model's performance on code-switched data. The results are as follows:

Model	Accuracy	F1-Score
	%	%
mMiniLMv2	76.00	76.22

Table 9.1: Emotion Classifier performance on the `EmpatheticPersonasCS` code-switched test dataset.

During the human trial, we also encouraged participants to experiment code-switching with the chatbot. 16 out of 27 participants attempted this, out of which 84% of participants agreed that the chatbot was able to guess their emotions when doing so.

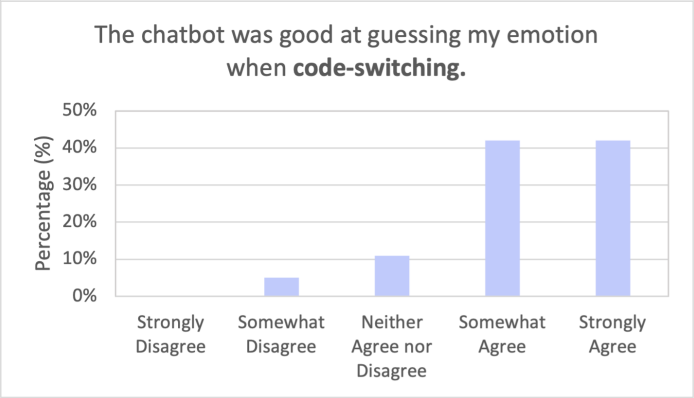


Figure 9.1: Participant evaluation of emotion classification performance when using code-switching.

Considering that code-switching was not an explicit training objective in the present study, the above performance is commendable. However, it still does not meet the performance as when we use only single-language inputs (see Section 5.4). Given the promising results, future work could investigate methods to improve the emotion classifiers' performance when using code-switching to reach comparable results.

Increasing Languages

Multilingual models are extremely powerful. To highlight this, we had participants converse with the emotion classifier in other languages (beyond English and Mandarin) to assess how far its multilingual capabilities extend. 18 out of 27 participants attempted this, and it was reported that the model was able to classify 13 participant instances in Malay, French, Cantonese and Korean despite only being trained in English and Mandarin (see Appendix C).

With the model already yielding positive zero-shot performance, it will be exciting to see how the model performs when being specifically trained on these other languages. We hope future works will continue to push the bounds of multilingual models, incorporating more languages within a single model. As there are currently 2 other students working on this same project in Russian and Cantonese, future work can combine all projects as a starting point to develop a 4-language (English, Mandarin, Russian, Cantonese) multilingual emotion classifier.

Moreover, seeing that our experience using machine translated data was able to yield successful results for emotion classification, future works can leverage this fact to easily obtain data in other languages too to train the classifier for these languages.

Learning from Mistakes (Continuous Learning Framework)

As a way to improve on the *SATbot*'s emotion classification abilities, we would like to propose a "continuous learning pipeline" to be incorporated into the chatbot in the future. When users identify the chatbot's predictions to be incorrect, it would

good to have a system in place to store these utterances in the database along with their corrected emotion, and use these to allow the model to continually learn from its mistakes. This will serve as a targeted approach to deal with the model's weaknesses. A user consent option should also be included to ensure compliance with data protection laws.

9.2 Empathetic Text Generation

Advanced Empathy

In our trial, several participants reported that the chatbot was overly empathetic, which resulted users to find the chatbot disingenuous. As a potential solution, rather than choosing high empathy utterances throughout conversation, we can use dynamically varying empathy levels.

Some participants also noted that the display of empathy is superficial. At present, we note that empathy is displayed very commonly using phrases such as "我很抱歉听到（*I am sorry to hear that*）", "如果你不介意*(I hope you don't mind)*". However, true expressions of empathy is comprised of the following [59]:

1. *Emotional Reactions*
   Expressing compassion.

2. *Interpretations*
   Display of understanding towards one's predicament/issue.

3. *Explorations*
   Improving understanding by exploring one's emotions and situation.

Hence, we propose that future works look into creating new empathetic datasets based on this theoretically-grounded framework for empathy, to enhance empathy on a deeper level and better capture its nuances.

Open Dialogue

A popular comment that was observed not only in this paper, but also past works, is that the rule-based model is too rigid, which affects perceived empathy and engagement when using the chatbot. Hence, the present paper believes the next step forward is to make the *SATbot* an open dialogue chatbot. While a student is currently working on this, the project is done in English and we would look like to see this extend to Mandarin and other languages.

The present paper's positive experience using Chinese GPT-2 highlights its strong text generation capabilities. Future work can leverage empathetic dialogue datasets such as PsyQA [68] and Chinese GPT-2 to develop an open dialogue language model in Mandarin.

9.3 User Interface/ Additional Features

The following are a list of curated participant responses regarding possible feature improvements in the chatbot taken from the trial:

1. Extending the *SATbot* to include a mobile application for greater ease of access.
2. Include speech detection rather than purely typed responses to more closely emulate a therapy session.
3. Participant profile to track long term progress (eg. past protocols practiced, emotion monitoring, feeling history).
4. Graphics/ illustrations to be shown with protocols for easier understanding.

Chapter 10

Conclusion

The framework outlined in this paper, built upon previous works [6, 13, 17], serves as a foundational proof-of-concept for existing mental health conversational agents to expand into other languages. We show that it is possible to use machine translated data, coupled with improved training methodologies, to yield safe and quality outputs with large pretrained language models that are suitable for mental health applications.

This is particularly so for classification tasks. Using machine translated data and a multilingual XLM-R-base [38] model, we successfully developed 3 highly performant classifiers for this study. These include:

1. An emotion classifier capable of achieving 90.00% accuracy and 90.09% F1-Score in monolingual Mandarin, and 91.23% accuracy and 91.4% F1-Score in monolingual English,

2. A binary empathy classifier capable of achieving 90.00% accuracy and F1-Score, and

3. A semantic classifier capable of achieving 96.00% accuracy and F1-Score.

We developed also a training pipeline for text generation to attain empathetic rewritings that are fluent and semantically correct using Chinese GPT-2 [44]. Our methodology yields 89.68% empathetic utterances with an improved fluency score over previous works (-0.6 mean, -1.92 standard deviation) [1].

Moreover, we looked into the use of Knowledge Distillation to distill large pretrained models such that they can be cheaper, quicker, yet still performant for practical deployment in applications. We yield a 25% smaller model that runs 16 time faster while attaining (on average) 92% the performance of the SOTA Mandarin model in emotion classification [17].

A human trial comprised of N=27 bilingual (English-Mandarin) participants was conducted to formally evaluate our framework components. The evaluation results

[1] Note that the lower the fluency score, the better.

show that the methodology in this paper yields improved performance in terms of emotion classification and utterance quality over previous works, and scored comparatively in all other aspects. We take into account participant feedback regarding potential areas for improvement and provide suggestions on how to resolve these, as well as discuss interesting opportunities for future exploration.

Overall, our work has shown promising potential for extending psychotherapeutic applications to non-English languages. We hope this finding helps encourage and enable works in other languages, increasing the reach of mental health treatment to wider communities in the future.

Bibliography

[1] Alize Ferrari, Damian Santomauro, Ana Herrera, Jamile Shadid, Charlie Ashbaugh, Holly Erskine, Fiona Charlson, Louisa Degenhardt, James Scott, John McGrath, Peter Allebeck, Corina Benjet, Nicholas Breitborde, Traolach Brugha, Xiaochen Dai, Lalit Dandona, Rakhi Dandona, Florian Fischer, Juanita Haagsma, and Harvey Whiteford. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry 2022*, 9:137–150, 01 2022. doi: 10.1016/S2215-0366(21)00395-3. pages 1

[2] Michael Daly, Angelina R. Sutin, and Eric Robinson. Longitudinal changes in mental health and the covid-19 pandemic: evidence from the uk household longitudinal study. *Psychological Medicine*, page 1–10, 2020. doi: 10.1017/S0033291720004432. pages 1

[3] World Health Organisation. Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide, March 2022. URL `https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-` Accessed 4th May 2022. pages 1

[4] Xuezheng Qin and Chee-Ruey Hsieh. Understanding and addressing the treatment gap in mental healthcare: Economic perspectives and evidence from china. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 57, 2020. doi: 10.1177/0046958020950566. pages 1

[5] Abbas Edalat. Introduction to self-attachment and its neural basis. In *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015. doi: 10.1109/IJCNN.2015.7280780. pages 1, 2

[6] Lisa Alazraki. A deep-learning assisted empathetic guide for self-attachment therapy. Master's thesis, Imperial College London, 2021. pages 1, 2, 3, 4, 5, 6, 7, 9, 10, 19, 25, 26, 32, 36, 38, 42, 52

[7] Algorithmic Human Development. Algorithmic human development. URL `http://humandevelopment.doc.ic.ac.uk/`. Accessed 4th May 2022. pages 1

[8] Kate E. Murray, Chantelle J. Musumeci, and Elija Cassidy. Crossing the digital divide: A content analysis of mainstream australian mental health web-

sites for languages other than english. *Health & Social Care in the Community*. doi: https://doi.org/10.1111/hsc.13890. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/hsc.13890. pages 1

[9] University of California, Irvine. Non-english-speaking, medically compromised patients may not benefit from telemedicine. URL https://medicalxpress.com/news/2022-02-non-english-speaking-medically-compromised-patients-benefit.html. Accessed 4th May 2022. pages 1

[10] Milken Institute School of Public Health - The George Washington University. Limited english proficiency and health care: How to support non-english speakers. URL https://onlinepublichealth.gwu.edu/resources/limited-english-proficiency-health-care-how-to-support/. Accessed 4th May 2022. pages 1

[11] Yael Dvir, Julian D. Ford, Michael Hill, and Jean A. Frazier. Childhood maltreatment, emotional dysregulation, and psychiatric comorbidities. *Harvard review of psychiatry*, 2014. doi: 10.1097/HRP.0000000000000014. pages 2

[12] Abbas Edalat. Self-attachment: A self-administrable intervention for chronic anxiety and depression, 03 2017. pages 2

[13] Ali Ghachem. Evaluation of a virtual agent in guiding users from the non-clinical population in self-attachment intervention. Master's thesis, Imperial College London, 2021. pages 2, 3, 4, 6, 38, 52

[14] Pralabh Saxena. Rule-based vs ai-based chatbots, 2021. URL https://medium.com/predict/rule-based-vs-ai-based-chatbots-28613db3fe2c. Accessed 7th May 2022. pages 2

[15] Asbjørn Følstad and Petter Bae Brandtzaeg. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience 5*, 2020. doi: https://doi.org/10.1007/s41233-020-00033-2. pages 2

[16] Alaa A Abd-alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. An overview of the features of chatbots in mental health: A scoping review. *International journal of medical informatics (Shannon, Ireland)*, 132, 2019. ISSN 1386-5056. pages 2

[17] Ruoyu Hu. A multi-language virtual psychotherapy chatbot. Master's thesis, Imperial College London, 2022. pages 3, 4, 7, 9, 10, 15, 20, 25, 27, 32, 34, 36, 38, 42, 43, 44, 52

[18] Media Department for Digital, Culture and Sport. The data protection act 2018 keeling schedule. 2020. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/969513/20201102_-_DPA_-__MASTER__Keeling_Schedule__with_changes_highlighted__V4.pdf. Accessed 29th May 2022. pages 4

[19] Media Department for Digital, Culture and Sport. General data protection regulation keeling schedule. 2020. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/969514/20201102_-_GDPR_-__MASTER__Keeling_Schedule__with_changes_highlighted__V4.pdf. Accessed 29th May 2022. pages 4

[20] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11325. URL https://ojs.aaai.org/index.php/AAAI/article/view/11325. pages 10

[21] Lucia Specia. Natural language processing - language modelling [lecture]. Imperial College London. Feb 2022. pages 11

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762. pages 11

[23] Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2021.08.002. URL https://www.sciencedirect.com/science/article/pii/S2666651021000231. pages 11

[24] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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*, pages 483–498. Association for Computational Linguistics, June 2021. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41. pages 11

[25] Colin Wei, Sang Michael Xie, and Tengyu Ma. Why do pretrained language models help in downstream tasks? an analysis of head and prompt tuning, 2021. URL https://arxiv.org/abs/2106.09226. pages 11

[26] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to bert embeddings during fine-tuning?, 2020. URL https://arxiv.org/abs/2004.14448. pages 11

[27] Samuel Flender. What exactly happens when we fine-tune bert?, 2021. URL https://towardsdatascience.com/what-exactly-happens-when-we-fine-tune-bert-f5dc32885d76. Accessed 20th May 2022. pages 11

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692. pages 11, 12

[29] Ke Tran. From english to foreign languages: Transferring pre-trained language models, 2020. URL https://arxiv.org/abs/2002.07306. pages 11, 12

[30] Maayan Shvo, Andrew C. Li, Rodrigo Toro Icarte, and Sheila A. McIlraith. Interpretable sequence classification via discrete optimization, 2020. URL https://arxiv.org/abs/2010.02819. pages 11

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL https://arxiv.org/abs/1810.04805. pages 12

[32] Eunchan Lee, Changhyeon Lee, and Sangtae Ahn. Comparative study of multiclass text classification in research proposals using pretrained language models. *Applied Sciences*, 12(9), 2022. ISSN 2076-3417. doi: 10.3390/app12094522. URL https://www.mdpi.com/2076-3417/12/9/4522. pages 12

[33] Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. Multilingual bert, 2019. URL https://github.com/google-research/bert/blob/master/multilingual.md. Accessed 20th May 2022. pages 13, 16

[34] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.419. URL https://aclanthology.org/2020.coling-main.419. pages 13

[35] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021. doi: 10.1109/TASLP.2021.3124365. pages 13

[36] Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. A primer on pretrained multilingual language models. *CoRR*, abs/2107.00676, 2021. URL https://arxiv.org/abs/2107.00676. pages 13, 16, 17, 26

[37] Peifeng Wang. Cross-lingual transfer learning [lecture notes]. University of Southern California. pages 13

[38] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2019. URL https://arxiv.org/abs/1911.02116. pages 13, 16, 17, 25, 52

[39] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Association for Computational Linguistics, Jul 2019. doi: 10.18653/v1/P19-1493. URL https://aclanthology.org/P19-1493. pages 13, 15

[40] Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr. pages 13

[41] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training, 2020. URL https://arxiv.org/abs/2007.07834. pages 13, 14, 16

[42] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019. URL https://arxiv.org/abs/1901.07291. pages 14

[43] Milan Gritta and Ignacio Iacobacci. XeroAlign: Zero-shot cross-lingual transformer alignment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 371–381. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.findings-acl.32. URL https://aclanthology.org/2021.findings-acl.32. pages 14

[44] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. pages 14, 52

[45] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301. pages 14, 36

[46] Zhe Zhao, Hui Chen, Jinbin Zhang, Xin Zhao, Tao Liu, Wei Lu, Xi Chen, Haotang Deng, Qi Ju, and Xiaoyong Du. Uer: An open-source toolkit for pre-training models. *EMNLP-IJCNLP 2019*, page 241, 2019. pages 14

[47] Liang Xu, Xuanwei Zhang, and Qianqian Dong. Cluecorpus2020: A large-scale chinese corpus for pre-training language model, 2020. URL https://arxiv.org/abs/2003.01355. pages 14

[48] Cindy Wang and Michele Banko. Practical transformer-based multilingual text classification. In *NAACL*, 2021. pages 15, 16

[49] Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946. Association for Computational Linguistics, August 2021. doi: 10.18653/v1/2021.findings-acl.437. URL https://aclanthology.org/2021.findings-acl.437. pages 16

[50] Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. Multilingual is not enough: Bert for finnish, 2019. URL https://arxiv.org/abs/1912.07076. pages 16

[51] Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. Xlm-e: Cross-lingual language model pre-training via electra, 2022. pages 16

[52] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019. URL https://arxiv.org/abs/1910.01108. pages 20, 21

[53] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2019. URL https://arxiv.org/abs/1909.10351. pages 20, 22

[54] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices, 2020. URL https://arxiv.org/abs/2004.02984. pages 20

[55] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531. pages 20, 21

[56] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers, 2020. URL https://arxiv.org/abs/2012.15828. pages 22

[57] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1404. URL https://aclanthology.org/D18-1404. pages 26

[58] Geoffrey T. Barrett-Lenard's. The empathy cycle: Refinement of a nuclear concept. *Journal of Counseling Psychology*, 28, 1981. doi: 10.1037/0022-0167.28.2.91. pages 27

[59] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach, 2021. URL https://arxiv.org/abs/2101.07714. pages 27, 28, 31, 50

[60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347. pages 27, 29

[61] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2019. URL https://arxiv.org/abs/1909.08593. pages 28, 29, 30, 31, 67

[62] Robert West and Eric Horvitz. Reverse-engineering satire, or "paper on computational humor accepted despite making serious advances". *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7265–7272, Jul. 2019. doi: 10.1609/aaai.v33i01.33017265. URL https://ojs.aaai.org/index.php/AAAI/article/view/4712. pages 28

[63] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2015. URL https://arxiv.org/abs/1502.05477. pages 29

[64] Heeyoul Choi, Sookjeong Kim, and Seungjin Choi. Trust-region learning for ica. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 1, pages 41–46, 2004. doi: 10.1109/IJCNN.2004.1379867. pages 29, 30

[65] Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*, 2019. pages 34

[66] Fredrik Oseberg. React-chatbot-kit, technical documentation. 2021. URL https://fredrikoseberg.github.io/react-chatbot-kit-docs/. Accessed 23rd July 2022. pages 38

[67] Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. BERTologiCoMix: How does code-mixing interact with multilingual BERT? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121, Kyiv, Ukraine, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.adaptnlp-1.12. pages 48

[68] Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support.

In *Findings of the Association for Computational Linguistics: ACL 2021*, 2021. pages 50

Appendices

A Classifier Hyperparameter Tuning

Parameter	Values Considered
Adam Epsilon	1e-04, 1e-06, 1e-08
Scheduler	Constant, Constant with Warmup, Linear with Warmup, Cosine with Warmup, Polynomial Decay with Warmup
Learning Rate	9e-06 to 9e-05
Batch Size	1, 2, 4, 8, 16, 32

Table 1: Model Parameters investigated for XLM-R Classifiers.

Table 1 shows the various parameters investigated during hyperparameter tuning of classifiers used in this paper. Tuning for each parameter was done one at a time. In each run, epoch was set to a large value of 20 but with early stopping applied using the following parameters to prevent overfitting:

- early stopping metric: evaluation loss
- early stopping delta [2]: 0.0001
- early stopping patience: 10
- evaluation steps: approx. 4 times per epoch

For the following classifiers, Adam Epsilon of 1e-08, Linear Scheduler with Warmup and Batch Size of 8 achieved the best performance in all cases. The key parameter that varied was **learning rate** and the final selected values are listed in the following sections.

Emotion Classifier (XLM-R Base)

Tuning	Learning Rate
First Tuning (ECM)	9.00e-06
Second Tuning (EP)	2.00e-05

Table 2: Best learning rates for the emotion classifier.

Emotion Classifier (Distilled MiniLMv2)

For the distilled emotion classifier, an additional parameter needed to be tuned: temperature. Typically, temperatures within the range of [2,5] are evaluated and

[2]The value in which evaluation loss must be reduced to be considered an improved checkpoint

these were the values considered during tuning. The following were the optimal learning rates and temperatures used for the distilled model:

Tuning	Learning Rate	Temperature
First Tuning (ECM)	3.00e-05	4
Second Tuning (EP)	1.00e-05	5

Table 3: Best learning rates and temperature for the distilled emotion classifier.

Empathy Classifier

Tuning	Learning Rate
Single Tuning (EP)	4.00e-05

Table 4: Best learning rate for the empathy classifier.

Semantic Classifier

The semantic classifier has 20 classes, each pertaining to a base utterance. The labelling is as such:

1. 这是由特别事件引起的吗?
   Was this caused by a specific event/s?

2. 这是由最近或遥远的事件（或多个事件）引起的吗?
   Was this caused by a recent or distant event (or events)?

3. 您最近是否尝试过练习6，并发现由于旧事件而重新点燃了无法控制的情绪?
   Have you recently attempted protocol 6 and found this reignited unmanageable emotions as a result of old events?

4. 您最近是否尝试过练习11，并发现由于旧事件而重新点燃了无法控制的情绪?
   Have you recently attempted protocol 11 and found this reignited unmanageable emotions as a result of old events?

5. 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.

6. 您是否对某人有强烈的感受或表达以下任何情绪:
   Have you strongly felt or expressed any of the following emotions towards someone:

7. 您认为您应该成为别人的救星吗?
   Do you believe that you should be the saviour of someone else?

8. 您是否将自己视为受害者，将自己的负面情绪归咎于他人?
   Do you see yourself as the victim, blaming someone else for how negative you feel?
9. 您觉得您在试图控制某人吗?
   Do you feel that you are trying to control someone?
10. 当出现问题时，您是否总是责怪和指责自己?
    Are you always blaming and accusing yourself for when something goes wrong?
11. 在之前的对话中，您是否考虑过其他提出的观点?
    In previous conversations, have you considered other viewpoints presented?
12. 您是否正在经历个人危机（与亲人相处有困难，例如与朋友闹翻）?
    Are you undergoing a personal crisis (experiencing difficulties with loved ones e.g. falling out with friends)?
13. 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.
14. 根据您所说的，我相信您正在感受。这个对吗?
    From what you have said I believe you are feeling . Is this correct?
15. 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 am sorry. Please select from the emotions below the one that best reflects what you are feeling:
16. 感谢您的参与。再见
    Thank you for taking part. See you soon.
17. 这是我的建议，请选择您想尝试的练习
    Here are my recommendations, please select the protocol that you would like to attempt.
18. 请现在尝试通过此练习。完成后，按"继续"
    Please try to go through this protocol now. When you finish, press 'continue'
19. 采取此练习后，您感觉更好还是更糟?
    Do you feel better or worse after having taken this protocol?
20. 您想尝试另一种练习吗?
    Would you like to attempt another protocol?

The following learning rates were used:

Tuning	Learning Rate
Single Tuning (EP)	4.00e-05

Table 5: Best learning rate for the semantic classifier.

The following class-specific performance was obtained:

Base Utterance Label	Precision	Recall	F1-Score
1	0.78	0.95	0.86
2	1.00	0.81	0.90
3	0.99	0.88	0.93
4	0.90	0.95	0.92
5	0.97	0.97	0.97
6	1.00	1.00	1.00
7	1.00	1.00	1.00
8	0.91	1.00	0.96
9	1.00	1.00	1.00
10	1.00	1.00	1.00
11	1.00	1.00	1.00
12	1.00	0.93	0.97
13	0.81	1.00	0.89
14	1.00	1.00	1.00
15	1.00	1.00	1.00
16	1.00	1.00	1.00
17	1.00	0.80	0.89
18	1.00	1.00	1.00
19	1.00	0.96	0.98
20	1.00	1.00	1.00

Table 6: Semantic Classifier performance on the `EmpatheticPersonasZH` test dataset.

B Text Generation Hyperparameter Tuning

We perform hyperparameter tuning only on the parameters listed in Table 7 during RL training of the Chinese GPT-2. The remainder of parameters follow those in [61]. As text generation training takes a longer time, we consider parameters in larger intervals. We set the number of training steps to 10000. Monitoring the model performance using wandb, we perform early stopping once the reward function converges to its maximum.

Parameters	Values Considered	Selected
learning rate	1e-06, 5e-06, 1e-05, 5e-05	5e-06
batch size	8, 16, 32	16
empathy weight, w_e	1, 2, 3, 4, 5	4
semantic weight, w_s	0.25, 0.30, 0,35, 0.40, 0.45, 0.50	0.25
fluency weight, w_f	1, 2, 3, 4, 5	1

Table 7: Model parameters investigated during RL training of Chinese GPT-2.

C Human Trial Questionnaire Response

C.1 5-point Likert Scale Responses

Statement	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
The chatbot was good at guessing my emotion in English.	0/27	1/27	2/27	7/27	17/27
The chatbot was good at guessing my emotions in Mandarin.	0/27	0/27	2/27	7/27	18/27
The chatbot sounded very empathetic throughout the conversation.	1/27	1/27	2/27	13/27	10/27
The chatbot provided fluent and natural-sounding responses.	0/27	0/27	1/27	15/27	11/27
I am satisfied with the chatbot user interface.	1/27	1/27	1/27	7/27	17/27
The chatbot conversations were engaging.	2/27	0/27	2/27	12/27	11/27
Overall, the platform was useful.	0/27	1/27	2/27	7/27	17/27

Table 8: Main questions from the trial, supplementary to Section 8.1.

Language	Strongly Disagree	Somewhat Disagree	Neither Agree nor Disagree	Somewhat Agree	Strongly Agree
Code-switching (EN-ZH)	0/19	1/19	2/19	8/19	8/19
Malay	0/11	0/11	2/11	7/11	2/11
Cantonese	0/2	0/2	0/2	0/2	2/2
French	1/2	0/2	0/2	1/2	0/2
Portuguese	0/1	0/1	1/1	0/1	0/1
Hungarian	1/1	0/1	0/1	0/1	0/1
Korean	0/1	0/1	0/1	0/1	1/1

Table 9: Extra questions from the trial regarding the statement "The chatbot was good at guessing my emotion in [language]", supplementary to Section 9.1. Do note that this section is intended only for discussion purposes and is not a true indication of the model's capabilities in these languages.

C.2 Open-text Responses

Please state any incorrect/ unnatural sounding expressions when conversing with the chatbot in Mandarin

(Note: all expressions highlighted here have been resolved from the model.)

- 1. 弊大于利*(More harm than good)* sounds a little unnatural or too formal
  2. 我很感激你分享它*(I am very grateful you shared this)* could be better as '感谢您与我分享' *(thank you for sharing this with me)*
  3. 或其他人对您的负面情绪负责？*(others are responsible for your negative emotions)* could be replaced with 您的负面情绪源于他人*(your negative emotions are the result of others)*
  4. 可能是您... *(Maybe it's that you're...)* can be replaced with 有没有可能是... *(is it possible that...)*
  5. ...将问题归咎于自己的倾向？*(...direct the blame towards yourself)* can be replaced with ...将责任归纳与自己*(...put the blame on yourself)*
  6. 治疗方案*(treatment program)* sounds a little serious.

- 1. 我希望我们能抓住它并把它装进瓶子里。*[literally translates to: I hope we can take hold of this and put it in a bottle]* can be replaced with 我希望我们能继续保持这个心情*(I hope we can hold on to this feeling)*
  2. 我很[欣赏]这可能是新的*[literally translates to: I appreciate that this might be new]* can be replaced with 这练习可能是新的*(This protocol might be new)*

- 您似乎朝着更快乐迈出了一大步！*(You are taking a big step towards happiness)* This feels quite unnatural, appears to be a translation from "you're taking a big step towards happiness" or something of the sort. Other than that, everything else sounded quite natural, albeit slightly formal.

- 1. 感谢您的[开放] *(Thank you for your openness)*
  2. 现在您可以尝试[通过]您选择的协议*(You may now [go through (literally)] your selected protocol)*

- 1. 协议*[direct translation of the word "protocol"]* sounds a bit unnatural to me, maybe 练习*(protocol)* is better.

- It did sound a little too formal for a regular empathetic conversation. Apart from that everything is pretty good.

- 协议*[direct translation of the word "protocol"]*, feel like 练习*(protocol)* is a more accurate term.

- The chatbot switches between 您and 你, would it be better to unify? Some of the sentences feels a bit long

- As a native Chinese speaker, I had no trouble understanding the conversation, but some sentences did not sound very natural.

What did you like/not like about the user interface.

- Having seen previous versions of the chatbot, I found this interface a strong improvement. The webpage is now definitely more interactive and aesthetic. I particularly like the clean protocol navigation pane and how the protocols appear when you select on the chatbot's suggestions. It allows me to focus only on the protocols I want to practice and not get distracted by other options.

- The layout of the chatbot was really nicely designed. The viewing pane for the protocol was a nice touch and made navigating really easy. I enjoyed the modern design compared to other chatbots from certain websites. The colour was also very inviting which made using the chatbot a much more enjoyable experience. Small comment was that the white font was difficult to read against the backdrop.

- Much more light-hearted appearance, really appreciate the integration of multiple languages into a single chatbot. Really like how the protocol information is integrated into the website design as opposed to a pdf in previous version, as well as highlighting the selected protocol.

- Nice colour palette, easy on the eyes. Text on the right is not really eye catching so only read when I was directed to by the chatbot (this could be good/bad depending on if that was purposeful).

- The pink template used set an intimate scene to dive into emotional details of the user. A brighter template would not have the same effect, while a darker template would add to a negative emotion.

- There's no particular dislike with the experience, however I would like to suggest adding illustrations to the protocols. It would be helpful for people who are not as well-read or educated.

- The part where the chatbot displayed the correct protocol based on what I chose was extremely cool. Also, nice colour palette choice.

- I like the colour scheme of the website as it is very aesthetically pleasing and calming for an emotion bot.

- UI is super cute, and the protocol on the right hand side is easy for me to find what I want.

- The chatbot replies really fast and accurately. The layout of the website is neat and tidy.

- Neat and clean UI. Good prompts from the chat bot to move to the next questions.

- Signing in via chat does not really make sense. Design was good though.

- Can't scroll down at all to see the full screen. Parts of it are hidden. *(note: issue resolved)*
- It is very cute and prompting the exercise after selection is useful.
- I liked how easy it is to navigate the chat bot.
- The interface is amazing! I love it!
- The user interface is much nicer.
- Very engaging. Responsive.
- Interface is perfect.
- Like, very lively.

Please describe the overall emotional impact of your experience

- I felt that the language used by the chatbot was really kind and it felt that I was almost speaking to a real person. The chatbot was able to guess the emotions I was portraying from my responses with some accuracy. The suggested protocols were very insightful and creative, and gave a new perspective when dealing certain emotions.
- Really enjoyed interacting with this platform, more so than previous version mainly due to a much nicer user interface. Was very useful in recommending SAT protocols to me in an intuitive manner. The textual responses can sound unnatural at times, but did not encounter sentences that I could not understand.
- Positive! I am new to Self Attachment Therapy and this project was very insightful. Also, most online therapy that I am aware of are primarily in single languages/ English. It was extremely nice to see this chatbot incorporate both languages, to more inclusivity!
- The chatbot is really light-hearted and empathetic. The lovely user interface makes me feel relaxed. And I do feel happier after practicing several protocols.
- Overall, I think it has achieved a general level of emotional impact that is engaging and somewhat empathetic. It is also a pleasant conversation in general.
- It was a good journey to learn about how to improve one's mood, with chatbot responses that come through quickly.
- I have a much greater insight and wanted to learn more about the SAT protocols.
- I did feel understood by the bot, and the protocols made me feel assured and better.

- It does help lifting my mood when I have no one to talk to.
- It was a great experience, I learnt a lot from the chatbot.
- Some prompts felt good, others felt very pandering.
- Pretty good! colour and feature are welcoming too.
- Surprised to have experience such unique chat bot.
- Enjoyable experience with Kai.
- No real emotional impact.
- No impact registered.
- Lights up the mood.
- It helped me a lot.
- Very pleasant.
- It was okay.
- Positive!

Any further suggestions on how the platform can be improved?

- It would be nice if it could track long term progress - e.g. track to see when I practiced the exercises and how it affects me over time rather than comparing to an instant change in emotion. Also a wider range in emotions would be nice for future updates. A mobile app with user personalisation would be nice. Also perhaps graphics with the solutions just to clarify what it's telling me to do.
- Open text input beyond the initial message asking how the user was feeling. This might be much more advanced but might allow the user to express more emotions. It would be great to allow for speech detection.
- There needs to be more emotion categories, sometimes the chatbot does not accurately classify my emotions because of the limited options (which is not its fault but good to consider in the future).
- I found that the protocols suggested were quite limited. In some cases, the protocols listed were not something I am looking for. Perhaps more protocols could be added.
- It would be great if there is an introduction to what Kai provides. An explanation of all the different protocols on the sidebar will be clear as well.
- The utterance of this bot is not very rich compared to other chatbots. If you have time, maybe add some more utterances.

- Subsequent dialogue after user's emotions are predicted needs to be more fluid and better at recognising free text input.
- Probably using less formal language and words while adding engaging illustrations to help with understanding.
- Seeing the success of this Chinese version, I would like to see it extended to more languages.
- Sometimes Kai is unable to detect emotions when not stating the emotion in the statement.
- Make the font slightly bigger or possibly a deeper colour for easier readability.
- More detailed profile, with information of feeling history and emotion patterns.
- I guess if it is a bit more customised in terms of its answers it will be better.
- Mobile version would be nice, that way I can talk to it on the go.
- Think there's a bug in the code, on the restart bit.
  (Note: issue resolved)
- Make it available for phone would be very useful!
- Should continue improving the Chinese responses.
- More languages support of course. Love it.
- None, its perfect to use.
- Need mobile version.
- None.
- No.

Final Comments

- I felt that the chatbot accurately predicted my emotions. it was especially great at predicting when i typed multiple languages together, amazing stuff!! Overall, it's amazing and easy to use, compared with the first version, especially in terms of user experience.
- Well it actually feels pretty mundane like the conversation is very general and standardised. But overall, I'm very impressed with the speed of the chatbot. There was no delay in the response from the chatbot which is highly commendable.
- Does not feel like the bot is advanced enough to produce an engaging conversation that could have any effect on a person. Feels more like all it can do is perform sentiment analysis on my inputs and follow a predetermined script.

- Whilst the platform is largely free of sentence errors, I found there to be a higher concentration of strange-sounding sentences when presented with additional questions in a negative-emotion conversation path.
- I feel as though the chat bot was performing as it should. It was able to pick up sentiments from my responses and suggested appropriate protocols.
- Great idea and would be a useful tool!