A Multilingual Virtual Guide for Self-Attachment Technique

Alicia Jiayun Law[1,†], Ruoyu Hu[1,2,†], Lisa Alazraki[1], Anandha Gopalan[1], Neophytos Polydorou[1,2], Abbas Edalat[1]
[1]*Department of Computing, Imperial College London, UK*
[2]*UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, UK*
Email : {alicia.law15, ruoyu.hu18, lisa.alazraki20, a.gopalan, neophytos.polydorou19, a.edalat}@imperial.ac.uk

***Abstract*—In this work, we propose a computational framework that leverages existing out-of-language data to create a conversational agent for the delivery of Self-Attachment Technique (SAT) in Mandarin. Our framework does not require large-scale human translations, yet it achieves a comparable performance whilst also maintaining safety and reliability. We propose two different methods of augmenting available response data through empathetic rewriting. We evaluate our chatbot against a previous, English-only SAT chatbot through non-clinical human trials ($N = 42$), each lasting five days, and quantitatively show that we are able to attain a comparable level of performance to the English SAT chatbot. We provide qualitative analysis on the limitations of our study and suggestions with the aim of guiding future improvements.**

***Index Terms*—digital psychotherapy, chatbots, attachment theory, Mandarin**

I. Introduction

According to the 2022 Global Burden of Disease study, mental disorders have been ranked amongst the top ten leading causes of burden[1] worldwide since 1990 [1]. With the onset of the COVID-19 pandemic, there has been significant negative impact on the mental health condition of the global population from a variety of environmental stimuli [2], with, for example, the effect in the UK most severe among the 18-34 demographic group but visible in all age demographics [3]. Cases of patients suffering mental health issues associated with a range of negative emotions such as defeat, entrapment and loneliness increased significantly from pre-pandemic levels [4].

As such, provision of mental health support has become more imperative in addressing mental health concerns arising as a result of public health emergencies. Yet, 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 [5], (ii) unaffordable treatment and (iii) limited and unequal distribution of mental healthcare resources [6]. It is therefore desirable to incorporate and supplement existing methods with digital technologies and novel techniques.

The Self-Attachment Technique (SAT) is a self-administrable intervention introduced in [7], [8] and [9]. In SAT, the user enacts both the role of the care-seeker, conceptualised as their childhood or emotional self and represented by the user's favourite childhood photo or VR avatar created from the photo, and that of the care-giver, conceptualised as their adult or thinking self. The adult self establishes an imaginative compassionate relation and then an affectional bonding with the childhood self using the photo or the avatar and their favourite jolly and love songs. Subsequently, for the bulk of the SAT intervention, the adult self re-parents the childhood self to emotional and social maturity by emulating the optimal parent-child interactions whenever the user experiences strong negative emotions, which are projected and externalised onto the childhood self. SAT has had promising results in its pilot study [10].

Prior works have incorporated technologies into the delivery of SAT protocols, with the most recent producing a chatbot assistant [11] aimed at guiding users proficient with practising SAT protocols through protocol recommendation. Conversational agents [12] have significant potential in their application to psychotherapy [13], [14], as recent advancements in the field of Natural Language Processing with large neural pretrained language models [15], [16] using a transformer-based architecture [17] have achieved state-of-the-art results in a variety of tasks that facilitate greater capability of human-computer interaction.

However, prior works are limited only to English, a situation emblematic of much of the recent progress in the application of machine learning models to Natural Language Processing [18], [19]. Monolingual NLP for certain languages can encounter the problem of resource availability, where there is a lower volume of available task-specific data to train a model to the same level of performance as higher-resource languages such as English.

In this paper, we present a computational framework for the delivery of SAT protocols in a Mandarin setting in order to gauge the feasibility of deploying existing English psychotherapeutic intervention into non-English languages, with the aim to contribute to achieving equitable access to mental healthcare for non-English speaking communities in the future. We summarise our contributions as follows:

- We introduce a translation pipeline, leveraging machine translation and post-editing to produce language-specific data from existing task-specific English data.
- We introduce transformer reinforcement learning via Proximal Policy Optimisation (PPO) to train an empathetic, fluent and accurate generation model to produce

† Equal contribution.

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

quality responses via empathetic rewriting.

- We introduce an alternate, supervised learning method for empathetic rewriting and provide quantitative comparison against the previous methods.
- We introduce a multilingual emotion recognition component, and apply knowledge distillation to reduce inference latency.
- We fully integrate the Mandarin version of the chatbot with previous [11] English versions to deploy a fully bilingual application.
- We formally evaluate the chatbot performance in multiple non-clinical trials, and provide qualitative analysis aimed at guiding future work.

II. Background

A. Self-Attachment Technique

Self-Attachment Technique (SAT) [7] is a new psychotherapeutic treatment informed by John Bowlby's Attachment Theory and has shown promise in early pilot studies [10]. 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 attachment (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 [7].

SAT is comprised of 20 self-administered protocols aimed at developing new secure attachment. The protocols invite individuals to envisage 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 [8]. The aims of the 20 protocols can be collated into eight groups:

- Compassion toward the childhood self.
- Affectional bonding with the childhood self; Vowing to care for the childhood self.
- Rebuilding the childhood self's emotional world; Loving the childhood self, zest for life; Bonding with Nature.
- Self-regulation of strong emotions; Reducing negative emotions.
- (Re)-learning to laugh and being playful.
- Learning to change perspective and laugh.
- Socialising the childhood self.
- Enhancing tolerance and resilience.

The previous English SAT chatbot [11] deduces the user's emotional state from open conversation, yet allows the user to select a different emotion whenever they feel that the one inferred is inaccurate. The chatbot then pursues a series of questions depending on the user's emotion to further refine protocol recommendation based on the user's past experience and current state. Protocols deemed unsuitable or those with which the user had previous adverse reactions, particularly protocols aimed at tackling negative emotions, are eliminated from recommendation. Users are encouraged to select and practise a protocol from a list of recommendations. Afterwards, they are prompted to give feedback on changes to their emotional state and undertake further protocols that are selected based on the feedback.

B. Empathy in Digital Psychotherapy

According to psychotherapy research, an important component in the efficacy of psychotherapeutic interventions is the capability of the therapist to engage in an empathetic manner with the patient [21]. Similarly to prior works [11], [22], we focus on Godfrey T. Barrett-Lennard's second phase of empathetic dialogue with the aim of producing empathetic responses demonstrating compassion towards the user.

Prior works on empathetic dialogue systems [14], [22], [23] have highlighted the importance of empathy in digital mental health support. However, most prior works focus on English deployment, and at the time of writing, there is little open-domain, language-specific and task-specific data for open empathetic response generation in Mandarin.

C. Related works

Applications of digital psychotherapeutic interventions, such as *Cognitive-Behavioural Therapy* (CBT) [24] remain largely monolingual, though with increasing monolingual adoption in non-English languages in recent years, such as French in the case of Lopez et al. [25]. Works such as those by Bakker et al. [26] and Weaver et al. [27] present English-only digital platforms for CBT and cite the adaptation to more languages as a future research direction to increase impact.

Previous SAT chatbots [11] make use of crowd-sourced English data for their responses, and a fixed conversation flow that determines the appropriate response type at each conversation step. Responses are retrieved from a pool of possible responses ranked by a weighted metric combining sentence fluency, novelty to previous conversation, and perceived empathy. We aim to leverage the existing task-specific data produced by [11] through translation.

Several approaches exist to facilitate multilinguality in chatbot response generation, though we focus broadly on two translation approaches:

- **Inference-time** translations [28], [29], wherein the semantics of the output utterance is determined prior to performing translation on a selected response. This has a relatively low data footprint, and changes to the translation system can be deployed immediately. However, it requires higher computational demand at inference-time. Translation mechanisms can be embedded within the response generation step such as in Graça et al. [30] and Dimitra et al. [31] to provide customer support and public administration functionalities respectively. Ralston et al. [28] apply this approach to the provision of mental health support to students by wrapping conversation logic within source-target and target-source translation steps using multiple external APIs. Lin et al. [29] identify the high

[2] Affect dysregulation is defined as the "impaired ability to regulate and/or tolerate negative emotional states" [20].

cost of this approach, as well as the susceptibility to noisy data at inference-time, and propose learning language-agnostic representations to allow for better multilingual adaptation through training on translated data.

- **Pre-computed** [32] translations, wherein the responses are saved and retrieved at the relevant stages of the conversation. This approach allows for all the responses to be known prior to the retrieval step, allowing for finer control over the responses presented to the user. Due to the sensitive nature of conversations in mental health contexts, this feature is inherently beneficial for ensuring the safety of the produced responses. However, in order to deploy to multiple languages, this approach requires the creation of language files for each supported language, as seen in the mental health support chatbot from Nieminen et al. [32]. *Retrieval-based* methods are nonetheless commonly used in conversational agents for mental health in a monolingual context, such as in Alazraki et al. [11], where a retrieval model allows the English SAT chatbot to guide users through carrying out self-attachment therapy protocols, and in Vaira et al. [33], where a similar model is used to provide support to new mothers.

For the purpose of this paper, we focus on pre-computing responses in order to ensure the safety and reliability of the conversation provided to the user, as well as to prevent translation errors that may negatively impact user experience.

D. Patient Safety

The SAT chatbot is a mental health application targeted at patients suffering from mental health conditions. In the interest of their safety, we take the following measures:

1) *Safe & Non-Toxic Chatbot Conversations*
Empathetic rewritings are produced from generative language models trained using a controlled dataset that has been vetted for safety. Utterances are also scored for empathy via an empathy classifier (the scores assigned are discrete labels) and only those that have attained a high empathy score are selected. Finally, as an additional precaution, all utterances are manually vetted for toxic content before being included in the dataset.
2) *Terminating Therapy*
SAT Protocols involve users interacting with their childhood self, which can inadvertently trigger strong emotions in 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 the SAT chatbot, while a mental health application, is not equipped, without a concurrent intervention by a human psychotherapist, to treat patients suffering from serious mental health conditions such as severe depression. Participants are only permitted to take part in the human evaluation trials once they have been thoroughly informed of the associated risks, and clear consent has been received.

E. Data Protection

While patients do not need to provide personal information to interact with the chatbot, the *contents* that patients discuss with the chatbot are themselves considered personal data under the UK's Data Protection Act (DPA) [34] and General Data Protection Regulations (GDPR) [35]. To ensure adherence to data protection laws, the SAT chatbot does not store user interactions beyond the treatment sessions. We also do not store metadata from user's devices (e.g. geolocation, IP/MAC addresses, IMEI codes etc.).

Data compliance was 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 were informed on how their personal information would be handled, and were required to provide consent before participating. Moreover, participant responses are anonymised, and all responses collected are used strictly for the purposes of the current study.

III. Dataset

A. Data Analysis

The `EmpatheticPersonas` dataset [11] (EP) is a crowd-sourced dataset intended for the development of the SAT chatbot. This dataset is comprised of two main components. Firstly, it contains 1,181 written expressions of **emotion**, aimed at training an emotion classifier. The examples in the dataset are approximately evenly distributed across four emotion classes: there are 284 examples relating to Fear/Anxiety, 297 relating to Anger, 300 relating to Sadness and 300 relating to Joy/Contentment. Secondly, the dataset contains 2,144 **empathetic rewritings** of 45 base utterances. 1,100 of these have also been annotated for empathy using a discrete scale from 0 to 2 (where 0 represents non-empathetic utterances, 1 represents slightly empathetic ones and 2 corresponds to highly empathetic utterances). The annotated rewritings are aimed at training an empathy classifier.

We produced an additional native Mandarin dataset consisting of 120 emotional utterances balanced across four emotion classes for testing purposes.

B. Dataset Translation

As crowd-sourcing data is a time consuming and costly process, we leveraged publicly available machine translation tools to aid the translation process of the existing EP dataset into Mandarin. We formulated our translation pipeline as follows:

1) We used a publicly available machine translation tool (Google Translate) to obtain a base English(EN)-Mandarin(ZH) translation of the EP dataset.
2) We performed post-editing (`v1`) on the translated dataset to remedy major translation errors affecting sentence semantics (Fig. 1).
3) An additional post-editing step (`v2`) was introduced after early trials identified a need to inject language-specific terms and colloquialisms to further improve the localisation quality of the translations (Fig. 2).

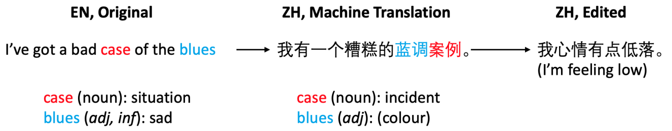


Fig. 1. Example of major translation errors targeted in post-editing. Most translation errors stemmed from literal translations of EN colloquialisms into ZH.

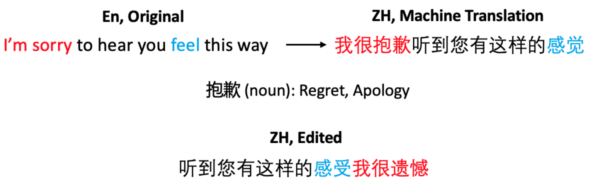

En, Original
ZH, Machine Translation
I'm sorry to hear you feel this way
我很抱歉听到您有这样的感觉
抱歉 (noun): Regret, Apology
ZH, Edited
听到您有这样的感受我很遗憾

Fig. 2. Example of minor translation errors edited to increase localisation accuracy. Note that the original translation maintained the coherence of the source sentence, but used words that were not the most appropriate to the context.

It is worth noting that the introduction of post-editing steps allows screening of candidate responses for potentially harmful or dangerous utterances in addition to remedying errors.

Reference-based sentence evaluation metrics such as BLEU [36] and ROUGE [37] evaluate the quality of a translation against reference target sentences. As human-translated target sentences were not available, we instead evaluated the efficacy of our translation using the reference-free [38] sentence fluency metrics SLOR [39] and PRSIM-SRC[3] [40] along with sentence perplexity (PPL), with results shown in Table I.

We observe from Table I that both post-edit revisions yield better fluency scores across all three metrics, suggesting that the inclusion of post-edits did improve the fluency of the utterances over base machine translated text, and improved translation quality with respect to the source sentence. We note that `v2` yields a slightly higher improvement over the `base` version in SLOR (3.92 vs 3.84) than `v1` (3.87 vs 3.84), whilst `v1` scores higher on PRISM-SRC (34.71 vs 35.04) and Perplexity (18.08 vs 18.83). We hypothesise that this may be due to the fact that the edits made in `v2` are of a finer-grain nature, using 'rarer' tokens that are more commonly associated with colloquialisms.

We also compare the quality of our utterances to the English version through human evaluation in Section V-C, where we show that our post-edited utterances attain a comparable level of user experience to the English version.

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

Revision	**SLOR**	**PRISM-SRC**	**PPL**
`Base`	3.84	39.05	19.77
`v1`	3.87	**34.71**	**18.08**
`v2`	**3.92**	35.04	18.83

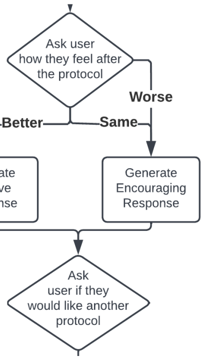

Ask user how they feel after the protocol
Better
Same
Worse
Generate Positive Response
Generate Encouraging Response
Ask user if they would like another protocol

Fig. 3. Updated conversation flow component for producing empathetic responses based on the user's change in emotion (after practising a protocol) and their original emotional state.

IV. Implementation

Our chatbot uses a rule-based conversation flow as in [11], where the chatbot first works to recognise the user's emotional state and from this it guides the next stages of the conversation to establish a context for protocol recommendation, with previously ineffective protocols removed from the set of potential recommendations. We introduce an addition to the conversation flow as shown in Fig. 3, to account for the user's change in emotional state after carrying out a protocol and produce appropriate responses. The chatbot contains two core components: *emotion recognition* and *empathetic rewritings*.

A. Emotion Recognition

As the conversation flow is dictated by the user's emotional state, the chatbot needs to correctly identify the user's emotions. We developed an emotion classifier capable of identifying four emotions: sadness, anger, fear/anxiety and joy/contentment. We double finetuned the pretrained language model (PLM) XLM-R[4] [41] using an emotion dataset in native Mandarin (NLPCC) [42], followed by the EP emotion data. Our model's results are shown in Table II, where at least 90% accuracy and F1-scores are attained across all test sets. The first finetuning was introduced to enhance performance in native Mandarin. However, our findings show that the model performs sufficiently well even without finetuning on the NLPCC dataset (see 'single' results in Table II). This is beneficial for low resource languages where there may be a lack of native in-domain data.

B. 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 operate. Hence, adopting such models in real time applications becomes highly impractical due to cost and latency issues [43].

[3] We show NLL in our work as opposed to the Log-likelihood shown in the original work [40].

[4] Available at https://huggingface.co/xlm-roberta-base

TABLE II
EMOTION CLASSIFIER RESULTS AGAINST DIFFERENT EMPATHETICPERSONAS TEST SETS WITH SINGLE AND DOUBLE FINETUNING.

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

To optimise the emotion classifier for runtime efficiency, we performed Knowledge Distillation [44] as a compression technique to reduce the size of the model while maintaining its performance. Using the double finetuned model (Section IV-A) as the teacher model, we performed Knowledge Distillation on a L6xH384 mMiniLMv2 student model[5] [45], a task-agnostic model distilled from an XLM-R-large model. We performed double finetuning, with distillation occurring at each stage, inspired by the multi-stage distillation framework shown in [46].

Distillation was performed using the Triple Loss method [43], which incorporates distillation loss (L_{dist}) and cosine embedding loss (L_{cos}), in addition to the classic supervised training loss (L_{ce}), during training.

1) **Classic Supervised Training Loss, L_{ce}**
This is the cross-entropy loss between the student model's predicted distribution (c_i) and the target training labels (q_i) which is in the form of a one-hot vector.

$$L_{ce} = \sum_i q_i * log(c_i) \tag{1}$$

2) **Distillation Loss, L_{dist}**
This is the cross-entropy loss between the student model's *softened* predicted distribution (s_i) and the teacher's *softened* predicted distribution (t_i) [47].

$$L_{dist} = \sum_i t_i * log(s_i) \tag{2}$$

These softened predictions are also known as the softmax-temperature probability distribution, given by:

$$p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{3}$$

where T denotes temperature and z_i denotes 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 to improve performance [43] . The loss is as follows:

$$L_{cos} = 1 - cos(T(x), S(x)) \tag{4}$$

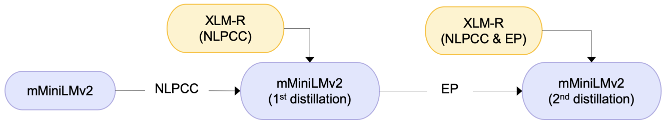


Fig. 4. Model training pipeline for the emotion classifier. It involves a two stage finetuning, with distillation occurring at each stage. Teacher models (XLM-R) are represented in yellow and student models (mMiniLMv2) in blue.

TABLE III
ACCURACY AND F1-SCORES OF THE DISTILLED STUDENT MODEL (MMINILMV2) ON THE 3 EP TEST SETS. WE ALSO INCLUDE THE PERFORMANCE OF THE TEACHER MODEL (XLM-R) AND THE ENGLISH EMOTION CLASSIFIER FROM [11] FOR COMPARISON.

	ZH Translate		ZH Native		EN	
Model	Acc.	F1	Acc.	F1	Acc.	F1
	%	%	%	%	%	%
RoBERTa-base [11]	-	-	-	-	94.96	95.10
XLM-R-base	93.86	93.86	90.00	90.09	91.23	91.40
mMiniLMv2	91.22	91.31	80.83	80.85	85.09	85.39

Thus, the final training loss is taken as the average of the three losses:

$$L_{total} = \frac{L_{ce} + L_{dist} + L_{cos}}{3} \tag{5}$$

The emotion classifier training pipeline is illustrated in Fig. 4. Following hyperparameter tuning, we obtained a performance (accuracy and F1) of ∼81% and ∼85% on the native Mandarin and English test sets respectively (Table III). Considering that the mMiniLMv2 has only 40% of the XLM-R-base teacher model's capacity (see Table IV), the model performs extremely well, retaining a significant proportion of its teacher model's performance (approx. 90% in the worst case and up to 97% in the best case.). We also note a significant reduction in average inference time using the distilled model (Table V) compared to the base XLM-R model. This had a noticeable impact on trial participant feedback (see Section V-C).

Additionally, we compared the distilled model's performance against the previous SAT chatbot emotion classifier (see Table III). When comparing against the English RoBERTa-base classifier deployed in [11], the distilled model has 90% of its performance at 40% of its capacity (Table IV). On the other hand, the non-clinical trial results point to a significant improvement in participant sentiment toward our emotion classifier compared to the one from [11].

Overall, considering the computational advantages and minimal performance trade-off, the above results illustrate the

[5] Available at https://github.com/microsoft/unilm/tree/master/minilm

TABLE IV
MODEL SIZE COMPARISONS.

Model	Layers	Hidden Dim.	No. Params.
XLM-R [41]	12	768	270M
mMiniLMv2 [45]	6	384	107M

TABLE V
AVERAGE INFERENCE TIME.

Model	Inference Time (s)
XLM-R-base	0.1877
mMiniLMv2	0.0308

potential of performing Knowledge Distillation, whereby classification performance can be largely recreated with a significantly smaller and more efficient model.

C. Empathetic Rewriting

In order to increase the level of empathy expressed in the chatbot's responses, we augmented the existing translated responses by having lower-empathy utterances rewritten to be more empathetic. As the chatbot has a rule-based conversational flow, rewriting has the additional benefit of boosting diversity in its conversation, thus potentially leading to greater user engagement.

We adopted the generative language model Chinese GPT-2[6] to generate the empathetic rewritings in Mandarin. This model was trained using reinforcement learning (RL) with proximal policy optimisation [48], based on [49]. Prior to training, we performed a supervised warm-start since literature has shown that it leads to more effective learning [23], [49].

The training setup is illustrated in Fig. 5. To facilitate training, we devised a reward model to reward utterances that are first and foremost empathetic, but also fluent and semantically relevant.

The **empathy reward** r_e is the key component of the empathetic rewriting task. This component aims to reward rewritings that convey a high degree of empathy, and penalise rewritings conveying low empathy. To quantify the degree of empathy conveyed by an utterance, we developed an empathy classifier using an XLM-R model trained and evaluated on the empathy-annotated EP data, obtaining an overall accuracy and F1-score of 90%. The logit of the highly empathetic class computed by the classifier is then taken as the empathy reward.

The **semantic reward** r_s aims to reward rewritings that deliver the same semantic meaning as the base utterance. Without this component, utterances that are highly empathetic but do not carry the correct semantic information may be generated as the model seeks to exploit the empathy reward. To measure the semantic similarity of a rewriting to its base utterance, we trained and evaluated an XLM-R model on the empathetic rewritings in the EP dataset, obtaining an overall accuracy and F1-score of 96%. The semantic reward is the logit of the semantic class corresponding to the base utterance.

The **fluency reward** r_f was adapted from the fluency function in [11] and included to prevent rewritings that are highly empathetic but are incoherent/grammatically incorrect. This is computed as:

$$r_f(er) = \frac{1}{PPL(er)} - RP(er) \tag{6}$$

where er denotes the empathetic rewriting, $\frac{1}{PPL(er)}$ is the inverse of the perplexity (computed by a GPT-2 model) and $RP(er)$ denotes a cumulative penalty for every repeated word within that rewriting (excluding stop words). Attempting to remove the repetition penalty term resulted in the model seeking to exploit the semantic and empathy reward by repeating keywords/empathetic terms.

The final reward was implemented as a multi-objective function comprised of the weighted sum of the empathy, fluency and semantic rewards, written as:

$$r = w_e r_e + w_f r_f + w_s r_s \tag{7}$$

Similarly to related works [32], we pre-generated and manually inspected empathetic responses for any toxic speech or distressing content (e.g. relating to violence or self harm) before approving them to be used by the chatbot. It is worth noting, however, that no problematic content was found in the utterances generated by the final trained model. In the future, a hate-speech detector could be devised to automate this inspection process.

D. Supervised Empathetic Rewriting

While the RL-based methodology yielded overall quality responses, we should note that this method can be 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 response to this, we also introduce a simpler, supervised learning (SL) approach to empathetic rewriting. We fine-tuned a GPT-2 model by prompting it with the user's emotional state $\mathcal{S}_e$ and a basic, low-empathy utterance $\mathcal{S}_L$, and used a high-empathy utterance $\mathcal{S}_H$ as the learning target. We employed the empathy classifier (EC) used in the RL method to form an additional binary classification learning objective over all the utterances X_g generated at each training step:

$$L_{EC} = CrossEntropyLoss(EC(X_g), 1) \tag{8}$$

Where 1 is the label for highly empathetic sentences. We then updated the model using the combined loss

$$L_{Total} = L_{LM} + L_{EC} \tag{9}$$

where L_{LM} is the language modelling loss produced by the GPT-2 model.

[6] Available at https://huggingface.co/uer/gpt2-chinese-cluecorpussmall

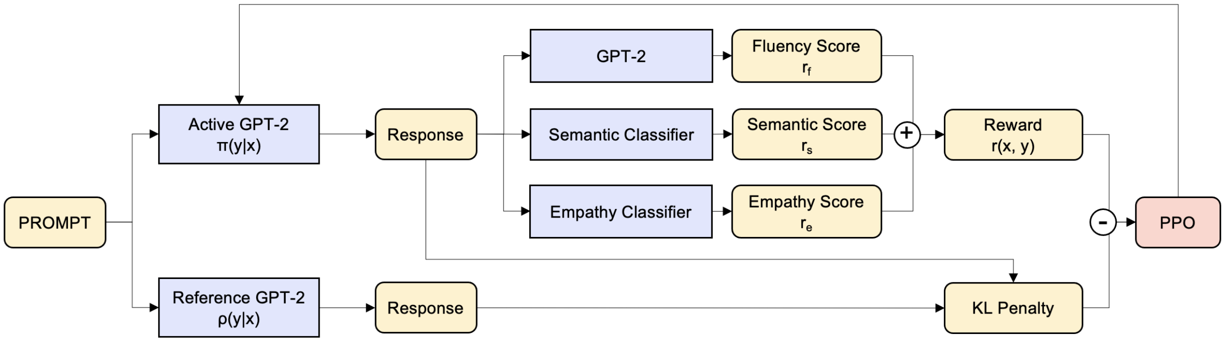


Fig. 5. Reinforcement learning setup for empathetic rewriting.

We then pre-generated and manually inspected responses in the same manner as the RL approach. We compare the responses generated by the two approaches through non-clinical human trials against the English SAT chatbot [11] in Section V-C.

V. Non-Clinical Trial

A. Study Design

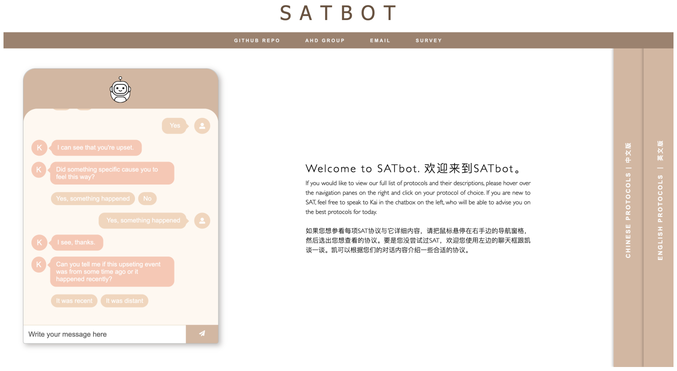
Formal evaluation of the SAT chatbot was carried out through non-clinical human trials. Separate trials were conducted on chatbots using responses generated via the reinforcement learning approach and the supervised learning method. For the purpose of the trial, participants were required to be fluent in both Simplified Mandarin and English in order to fully experience the bilingual chatbot. We note that users fluent in either language are nonetheless able to utilise the chatbot to practise SAT. Given the limited participant pool, knowledge of SAT protocols or psychotherapy was beneficial but not required. Participants were nonetheless provided with information detailing the SAT protocols prior to the trials. In total, 42 participants (20 female, 22 male) aged 25 to 60 consented to and participated across three trials.

Throughout each trial, participants were instructed to interact with the application once per day over a period of five days. There had to be a minimum of three interactions in Mandarin and one in English. Participants were also asked to note down any unnatural sounding utterances generated by the chatbot when using the Mandarin setting.

At the end of each trial, an anonymous feedback questionnaire was issued to each participant, aimed at evaluating their experience using the chatbot. The questionnaire sought to collect user feedback on: (i) the chatbot's emotion recognition capabilities, (ii) the quality of and the empathy conveyed by the chatbot's responses, (iii) the overall experience of using the chatbot and (iv) the perceived usefulness of the chatbot. Participants were asked to evaluate each aspect by providing their level of agreement with a particular statement on a Likert scale.

B. User Interface

Fig. 6 shows the user interface of the web platform that was deployed for the non-clinical trial. Protocols were available to view on the platform upon selection in both English and Mandarin.



K I can see that you're upset.

K Did something specific cause you to feel this way?

K I see, thanks.

K Can you tell me if this upseting event was from some time ago or it happened recently?

Welcome to SATbot. 欢迎来到SATbot。

If you would like to view our full list of protocols and their descriptions, please hover over the navigation panes on the right and click on your protocol of choice. If you are new to SAT, feel free to speak to Kai in the chatbox on the left, who will be able to advise you on the best protocols for today.

如果您想参看每项SAT协议与它详细内容，请把鼠标悬停在右手边的导航窗格，然后选出您想查看的协议。要是您没尝试过SAT，欢迎您使用左边的聊天框跟凯谈一谈。凯可以根据您们的对话内容介绍一些合适的协议。

Fig. 6. Trial platform web interface.

C. Evaluation

Participants were first asked to evaluate the emotion classifier's capabilities. When assessing whether the chatbot was good at guessing emotions, 89% and 93% of participants agreed with this statement for the emotion classification in English and Mandarin respectively compared to previous works [11], where only 63% of participants agreed (see Fig. 7). Taking into account that the distilled model is half the size of the one used in [11], this highlights the success of Knowledge Distillation at achieving performant yet compact models.

Participants were also asked to evaluate the quality of the chatbot's utterances. When asked to gauge whether the chatbot came across as highly empathetic throughout the conversation, 85% of the participants that had interacted with the RL chatbot agreed, while this proportion was 86% for the participants that had interacted with the SL chatbot (see Fig. 8). This result is consistent with the previous English-only implementation, where 88% of participants had agreed with this statement.

When asked if they found that the chatbot provided fluent and natural-sounding responses, 96% of participants that had used the RL chatbot agreed, while this proportion was 77% for the participants that had engaged with the SL chatbot (see Fig. 9).

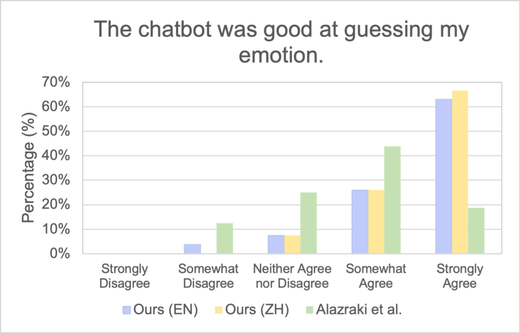


Fig. 7. Participant evaluation of emotion classification performance.

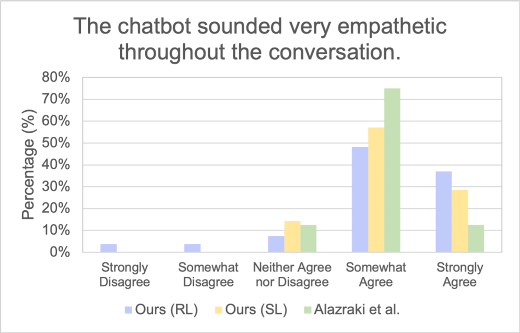


Fig. 8. Participant evaluation of the chatbot's display of empathy during conversation.

With regards to the level of engagement when conversing with the chatbot, 85% of participants agreed that they were engaged when the chatbot used RL-trained utterances, while 93% agreed when the chatbot used SL-trained utterances (see Fig. 10). The perceived user engagement of our platform is thus significantly higher than the previous English-only implementation, where only 69% of trial participants had agreed with the above statement.

Finally, participants were asked to evaluate the chatbot's usefulness. 89% of participants agreed that the platform was useful. This is roughly consistent with the proportion of users who had agreed that the English-only platform was useful, which was 92% (see Fig. 11).

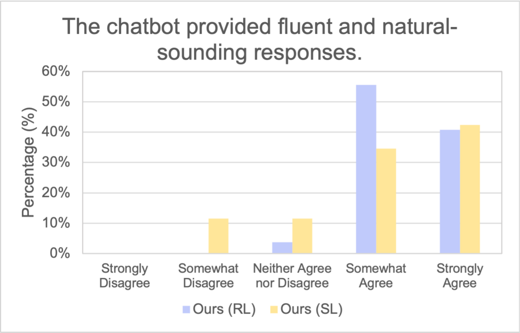


Fig. 9. Participant evaluation of the chatbot's fluency and naturalness in conversation.

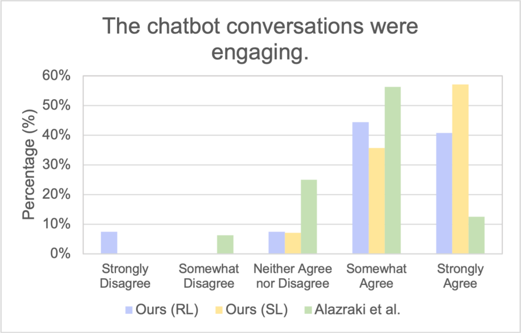


Fig. 10. Participant evaluation of the chatbot's engagement.

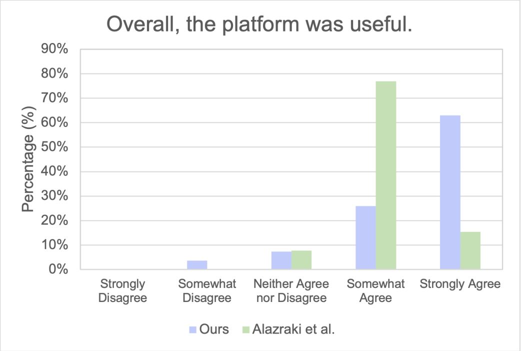


Fig. 11. Participant evaluation of the chatbot's usefulness.

VI. Discussion

A. Results

Our framework builds upon previous work to provide a foundation for future multilingual development and for the deployment of computational methods for mental health support to more languages, with a more scalable pipeline and less reliance on in-domain, in-language data. The human trials of our platform have shown promising results with respect to the perceived empathy, usefulness, user engagement, quality of responses and ability to correctly identify users' emotions.

B. Study Limitations

The key limitation faced in this study is that the majority of participants did not possess knowledge of SAT protocols prior to the trial. 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. Efforts were made to inform participants on how to carry out SAT protocols prior to the start of the human trial through information documents, and we received positive feedback from some participants regarding SAT. Nonetheless, the results from the trial evaluation, such as those pertaining to empathy and 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 receive information on the treatment and its protocols, for a better understanding before providing feedback. If possible, having Mandarin-speaking clinicians participating in the trials would also be extremely valuable.

Another notable limitation was the study size. Whilst our trial recruited more participants than the previous study [11], the participant sample was still relatively small. Moreover, the study sizes across the three trials conducted were inconsistent (13, 14 and 27), with imbalanced demographics across the sexes. 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.

C. Future Work

We note that the human trial was conducted with the purpose of quantifying the efficacy of using the multilingual chatbot, and was not aimed at determining the therapeutic effects of SAT on a Chinese-speaking population. An 8-week psychological intervention can be conducted in the future, where participants are exposed, step by step, to the SAT protocols through weekly sessions, and in which a Mandarin-capable SAT chatbot can be used in guiding users through carrying out SAT protocols on a daily basis. Future work could also investigate the application of this translation-based method to the delivery of other rule-based psychotherapy methods, such as CBT, on a Chinese-speaking population.

Code-switching, sometimes referred to as code-mixing, is a phenomenon prevalent in multilingual communities, whereby individuals alternate between two or more languages within the same conversation [50]. Since 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.

Moreover, it would be worth investigating the performance of the same model used in this paper on different languages, especially those with low resource availability. An extension could also be designed for expanding the range of emotions recognised by the classifier, and to assess the efficacy of formulating the emotion classification task as a multi-label problem, as human emotions can be complex and are typically not mutually exclusive.

A common reflection amongst trial participants focuses on the inherent rigidity of rule-based conversation. Therefore, future work could investigate the incorporation of open-dialogue to facilitate more natural conversations.

VII. Acknowledgement

Students Ruoyu Hu and Neophytos Polydorou were supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. We appreciate the general support from the Empowered Human Foundation in Canada and the UK.

References

[1] A. J. Ferrari, D. F. Santomauro, A. M. M. Herrera, J. Shadid, C. Ashbaugh, H. E. Erskine, F. J. Charlson, L. Degenhardt, J. G. Scott, J. J. McGrath *et al.*, "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*, vol. 9, pp. 137–150, 01 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215036621003953

[2] B. Pfefferbaum and C. S. North, "Mental health and the COVID-19 pandemic," *New England Journal of Medicine*, vol. 383, pp. 510–512, 2020. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMp2008017

[3] M. Daly, A. R. Sutin, and E. Robinson, "Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK household longitudinal study," *Psychological Medicine*, pp. 1–10, 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33183370/

[4] R. C. O'Connor, K. Wetherall, S. Cleare, H. McClelland, A. J. Melson, C. L. Niedzwiedz, R. E. O'Carroll, D. B. O'Connor, S. Platt, E. Scowcroft, B. Watson, T. Zortea, E. Ferguson, and K. A. Robb, "Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 mental health & wellbeing study," *The British Journal of Psychiatry*, vol. 218, pp. 326–333, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7684009/

[5] J. Shang, S. Wei, J. Jin, and P. Zhang, "Mental health apps in China: Analysis and quality assessment," *JMIR Mhealth Uhealth 2019*, vol. 7, 11 2019. [Online]. Available: https://mhealth.jmir.org/2019/11/e13236

[6] X. Qin and C.-R. 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*, vol. 57, 2020. [Online]. Available: https://doi.org/10.1177/0046958020950566

[7] A. Edalat, "Introduction to self-attachment and its neural basis," *International Joint Converence on Neural Networks (IJCNN)*, pp. 1–8, 2015. [Online]. Available: https://doi.org/10.1109/IJCNN.2015.7280780

[8] ——, "Self-attachment: A holistic approach to computational psychiatry," in *Computational Neurology and Psychiatry*, 2017, pp. 273–314. [Online]. Available: https://spiral.imperial.ac.uk/handle/10044/1/74080

[9] ——, "Self-attachment: A self-administrable intervention for chronic anxiety and depression," Department of Computing, Imperial College London, London, England, Tech. Rep., 2017. [Online]. Available: https://spiral.imperial.ac.uk/handle/10044/1/94925

[10] A. Edalat, M. Farsinezhad, M. Bokharaei, and F. Judy, "A pilot study to evaluate the efficacy of self-attachment to treat chronic anxiety and/or depression in Iranian women," *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 6376*, vol. 19, p. 6376, 5 2022. [Online]. Available: https://www.mdpi.com/1660-4601/19/11/6376

[11] L. Alazraki, A. Ghachem, N. Polydorou, F. Khosmood, and A. Edalat, "An empathetic AI coach for self-attachment therapy," in *2021 IEEE Third International Conference on Cognitive Machine Intelligence (CogMI)*, 2021, pp. 78–87. [Online]. Available: https://ieeexplore.ieee.org/document/9750315

[12] T. Ji, Y. Graham, G. J. F. Jones, C. Lyu, and Q. Liu, "Achieving reliable human assessment of open-domain dialogue systems," *arXiv preprint arXiv:2203.05899*, 3 2022. [Online]. Available: https://arxiv.org/abs/2203.05899v1

[13] H. Sun, Z. Lin, C. Zheng, S. Liu, and M. Huang, "Psyqa: A Chinese dataset for generating long counseling text for mental health support," *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1489–1503, 6 2021. [Online]. Available: https://arxiv.org/abs/2106.01702v1

[14] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, "Towards empathetic open-domain conversation models: a new benchmark and dataset," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 5370–5381, 11 2019. [Online]. Available: https://arxiv.org/abs/1811.00207v5

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, 2018. [Online]. Available: https://arxiv.org/abs/1810.04805v2

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, p. 9, 2019. [Online]. Available: https://openai.com/blog/better-language-models/

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all

you need," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 5999–6009, 2017. [Online]. Available: https://arxiv.org/abs/1706.03762v5

[18] J. Hu, S. Ruder, A. Siddhant, G. Neubig, O. Firat, and M. Johnson, "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization," *37th International Conference on Machine Learning, ICML 2020*, vol. PartF168147-6, pp. 4361–4371, 2020. [Online]. Available: https://arxiv.org/abs/2003.11080v5

[19] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, 2021. [Online]. Available: https://arxiv.org/abs/2010.11934v3

[20] Y. Dvir, J. D. Ford, M. Hill, and J. A. Frazier, "Childhood maltreatment, emotional dysregulation, and psychiatric comorbidities." *Harvard review of psychiatry*, 2014. [Online]. Available: https://doi.org/10.1097/HRP.0000000000000014

[21] R. Elliott, A. C. Bohart, J. C. Watson, and D. Murphy, "Therapist empathy and client outcome: An updated meta-analysis," *Psychotherapy*, vol. 55, pp. 399–410, 12 2018. [Online]. Available: https://psycnet.apa.org/record/2018-51673-006

[22] A. Sharma, A. S. Miner, D. C. Atkins, T. Althoff, and P. G. Allen, "A computational approach to understanding empathy expressed in text-based mental health support," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5263–5276, 2020. [Online]. Available: https://arxiv.org/abs/2009.08441

[23] A. Sharma, I. W. Lin, A. S. Miner, D. C. Atkins, and T. Althoff, "Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach," 2021. [Online]. Available: https://arxiv.org/abs/2101.07714

[24] S. Wilhelm, H. Weingarden, I. Ladis, V. Braddick, J. Shin, and N. C. Jacobson, "Cognitive-behavioral therapy in the digital age: Presidential address," *Behavior Therapy*, vol. 51, pp. 1–14, 1 2020. [Online]. Available: https://doi.org/10.1016/J.BETH.2019.08.001

[25] R. Lopez, E. Evangelista, L. Barateau, S. Chenini, A. Bosco, M. Billiard, A. D. Bonte, S. Béziat, I. Jaussent, and Y. Dauvilliers, "French language online cognitive behavioral therapy for insomnia disorder: A randomized controlled trial," *Frontiers in Neurology*, vol. 10, p. 1273, 12 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6906191/

[26] D. Bakker, N. Kazantzis, D. Rickwood, and N. Rickard, "Development and pilot evaluation of smartphone-delivered cognitive behavior therapy strategies for mood- and anxiety-related problems: Moodmission," *Cognitive and Behavioral Practice*, vol. 25, pp. 496–514, 11 2018. [Online]. Available: https://doi.org/10.1016/J.CBPRA.2018.07.002

[27] A. Weaver, A. Zhang, X. Xiang, P. Felsman, D. J. Fischer, and J. A. Himle, "Entertain me well: An entertaining, tailorable, online platform delivering CBT for depression," *Cognitive and Behavioral Practice*, 10 2021. [Online]. Available: https://doi.org/10.1016/J.CBPRA.2021.09.003

[28] K. Ralston, Y. Chen, H. Isah, and F. Zulkernine, "A voice interactive multilingual student support system using IBM watson," *18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, pp. 1924–1929, 12 2019. [Online]. Available: https://arxiv.org/abs/2001.00471v1

[29] Z. Lin, Z. Liu, G. I. Winata, S. Cahyawijaya, A. Madotto, Y. Bang, E. Ishii, and P. Fung, "Xpersona: Evaluating multilingual personalized chatbot," *Proceedings of the 3rd workshop on natural language processing for conversational AI*, pp. 102–112, 3 2021. [Online]. Available: https://arxiv.org/abs/2003.07568v2

[30] J. Graça, P. Dimas, H. Moniz, A. Martins, and G. Neubig, "Project MAIA: Multilingual AI agent assistant," *22nd Annual Conference of the European Association for Machine Translation*, p. 495, 2020. [Online]. Available: https://aclanthology.org/2020.eamt-1.68

[31] A. Dimitra, A. Ruge, R. Ion, S. Segărceanu, G. Suciu, O. Pedretti, P. Gratz, and H. Afkari, "A machine translation-powered chatbot for public administration," *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pp. 327–328, 2022. [Online]. Available: https://aclanthology.org/2022.eamt-1.54/

[32] H. Nieminen, L. Kuosmanen, R. Bond, A. K. Vartiainen, M. Mulvenna, C. Potts, and C. Kostenius, "Coproducing multilingual conversational scripts for a mental wellbeing chatbot - where healthcare domain experts become chatbot designers," *European Psychiatry*, vol. 65, pp. 293–293, 2022. [Online]. Available: https://doi.org/10.1192/j.eurpsy.2022.748

[33] L. Vaira, M. A. Bochicchio, M. Conte, F. M. Casaluci, A. Melpignano, L. Vaira, M. A. Bochicchio, M. Conte, and F. M. Casaluci, "Mama bot: A system based on ML and NLP for supporting women and families during pregnancy," *ACM International Conference Proceeding Series*, pp. 273–277, 6 2018. [Online]. Available: https://doi.org/10.1145/3216122.3216173

[34] UK Government, "Data protection act 2018," 2018, accessed 29th May 2022. [Online]. Available: https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[35] European Parliament and the council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," *Official Journal of the European Union*, 2016, accessed 29th May 2022. [Online]. Available: https://gdpr-info.eu/

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. [Online]. Available: https://aclanthology.org/P02-1040

[37] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text summarization branches out*, pp. 74–81, 2004. [Online]. Available: https://aclanthology.org/W04-1013

[38] K. Ethayarajh and D. Sadigh, "Bleu neighbors: A reference-less approach to automatic evaluation," *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 40–50, 2020. [Online]. Available: https://arxiv.org/abs/2004.12726v3

[39] K. Kann, S. Rothe, and K. Filippova, "Sentence-level fluency evaluation: References help, but can be spared!" *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 313–323, 2018. [Online]. Available: https://arxiv.org/abs/1809.08731v1

[40] B. Thompson and M. Post, "Automatic machine translation evaluation in many languages via zero-shot paraphrasing," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 90–121, 2020. [Online]. Available: https://arxiv.org/abs/2004.14564v2

[41] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019. [Online]. Available: https://arxiv.org/abs/1911.02116

[42] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11325

[43] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2019. [Online]. Available: https://arxiv.org/abs/1910.01108

[44] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 3 2015. [Online]. Available: https://arxiv.org/abs/1503.02531v1

[45] W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei, "Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers," 2020. [Online]. Available: https://arxiv.org/abs/2012.15828

[46] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for natural language understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 4163–4174. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.372

[47] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[48] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[49] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," 2019. [Online]. Available: https://arxiv.org/abs/1909.08593

[50] S. Santy, A. Srinivasan, and M. Choudhury, "BERTologiCoMix: How does code-mixing interact with multilingual BERT?" in *Proceedings of the Second Workshop on Domain Adaptation for NLP*. Kyiv, Ukraine: Association for Computational Linguistics, Apr. 2021, pp. 111–121. [Online]. Available: https://aclanthology.org/2021.adaptnlp-1.12