Beyond Information: Is ChatGPT Empathetic Enough?

Ahmed Belkhir
Université du Québec à Montréal
belkhir.ahmed@courrier.uqam.ca

Fatiha Sadat
Université du Québec à Montréal
sadat.fatiha@uqam.ca

Abstract

This paper aims to explore and enhance ChatGPT's abilities to generate more human-like conversations by taking into account the emotional state of the user. To achieve this goal, a prompt-driven Emotional Intelligence is used through the *empathetic dialogue* dataset in order to propose a more empathetic conversational language model. We propose two altered versions of ChatGPT as follows: (1) an emotion-infused version which takes the user's emotion as input before generating responses using an emotion classifier based on ELECTRA (Clark et al., 2020); and (2) the emotion adapting version that tries to accommodate for how the user feels without any external component.

By analyzing responses of the two proposed altered versions and comparing them to the standard version of ChatGPT, we find that using the external emotion classifier leads to more frequent and pronounced use of positive emotions compared to the standard version. On the other hand, using simple prompt engineering to take the user emotion into consideration, does the opposite. Finally, comparisons with state-of-the-art models highlight the potential of prompt engineering to enhance the emotional abilities of chatbots based on large language models.

1 Introduction

Conversational agents have become increasingly popular in recent years, with applications ranging from customer service (Ando and Zhang, 2005) to mental health therapy (Abd-Alrazaq et al., 2019). However, while these agents have the potential to provide information in natural language, their current abilities to generate human-like and empathetic conversations are limited (Rapp et al., 2021; Belainine et al., 2020b,a).

To address this challenge, this study explores the emotional abilities of ChatGPT in generating empathetic responses. Specifically, we investigate the effectiveness of incorporating external emotion classifiers using prompt engineering to take the user's emotional state into account when generating responses. This study is motivated by the fact that emotions play a crucial role in human communication, and empathetic responses are essential for building rapport and trust in human-machine interactions (Chen et al., 2021). In customer service for instance, it was shown that up to 40% of consumers' requests are rather emotional without specific informational intents (Xu et al., 2017). Thus, we compare standard ChatGPT that generates responses to simple conversation prompts from the *Empathetic Dialogues dataset* (Ma et al., 2020) to two slightly modified versions with prompt engineering. The first one is an emotion-infused version that takes the user emotion as an additional input before generating responses using an ELECTRA-based emotion classifier; while the second, emotion adapting version tries to consider how the user feels without any external component.

Our study adds to the expanding literature on conversational agents and emotional intelligence and its results have implications for the design and development of conversational agents that can provide personalized and effective support to users. In the following sections, we provide a brief review of the related work (Section 2) then present some relevant preliminary information (Section 3). Section 4 contains a detailed description of our experimental design while our evaluations and results will be presented in section 5. Finally, section 6 concludes this paper and gives some future perspectives.

2 Related Work

According to Allouch et al. (2021), a conversational agent can be defined as *"a dialogue system that can also understand and generate natural language content, using text, voice, or hand gestures, such*

Proceedings of Recent Advances in Natural Language Processing, pages 159–169
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_018

as sign language". Even though the first chatbot in the literature dates back to 1966, with the Rogerian psychotherapist chatbot Eliza developed by Joseph Weizenbaum (Weizenbaum, 1966), chatbot development has only exploded over the past several years (Adamopoulou and Moussiades, 2020). Their applications have been appearing across a variety of industries thanks to huge data sources, machine learning advancements (Grudin and Jacques, 2019), and Large Language Models (LLMs).

Conversational agents can be classified based on the response generation method: rule-based systems choose responses from hand-crafted predefined rules but suffer from dull and repetitive responses (Prendinger and Ishizuka, 2005). Retrieval-based methods use techniques such as keyword matching to find the most approprite response from a fairly large corpus but don't seem very natural (Grudin and Jacques, 2019), and generative chatbots provide more diverse conversations but require massive training data (Sutskever et al., 2014).

Despite all the advancements in the conversational agents research, it appears that people still prefer natural communication to machine-like interactions and feel that a human can understand them better (Rapp et al., 2021). In fact, it was shown in recent studies that customers still prefer interacting with humans over machines (Adam et al., 2021) because generating empathetic and human-like responses is a challenging task for chatbots, as it requires an understanding of the user's emotional state and the ability to respond appropriately.

Several studies have explored the use of different techniques to improve the emotional abilities of conversational agents. For example, Asghar et al. (2018) used a heuristic search technique in order to ensure variety and emotional relevance in the generated replies. Other research aimed to identify the emotion of the input message by embedding each input word in a three-dimensional emotion embedding space which dimentsion are Valence, Arousal, and Dominance (VAD) (Warriner et al., 2013). To address the relevance of the emotional responses, Lin et al. (2019) proposed the empathy hypothesis stating that the type of generated emotion should be consistent with the contextual emotional state of the user, while Wei et al. (2019) argued that we can't assume that the output emotion has to match the input emotion. They claimed that using a predefined label to train the response generator results in poor response quality. Zhang et al. (2018) proposed to generate multiple responses for six emotional categories and the best response is then selected with a ranking algorithm.

In recent years, the field of natural language processing has witnessed an unprecedented race to develop new LLMs based on the transformer architecture (Vaswani et al., 2017) which showed a great potential at capturing complex patterns in language data. For instance, GPT-3 by OpenAI (Brown et al., 2020) has proven its capacity to produce coherent and human-like language, PALM by Google (Chowdhery et al., 2022) has contributed to reducing the computational requirements for training large models and PaLM 2 promised advanced reasoning and general capabilities compared to the current state of the art of language models (PaLM2).

Recently, ChatGPT has demonstrated its remarkable ability to understand and converse with humans fluidly. Since its release in November 2022 with impressive language abilities, there has been a growing interest in evaluating the conversational language model for different aspects of Natural Language Understanding (NLU). For instance, Bang et al. (2023) evaluates the multilingual performance of ChatGPT on three tasks of language identification, sentiment analysis, and machine translation. Lai et al. (2023) evaluates the performance of ChatGPT, beyond English on many Natural Language Processing (NLP) tasks such as NER, NMT, POS, NLI, QA, CSR. Kocoń et al. (2023) tried to evaluate ChatGPT on 25 different NLP tasks and found that it did very well in most of them, but didn't outperform the state of the art in any particular task. However, to the best of our knowledge, there is still no work on the evaluation of ChatGPT on the emotional intelligence level.

The exceptional performance of LLMs on a variety of tasks, even with zero-shot or few-shot settings, has inspired NLP academics to reevaluate the predominant training paradigms from previous years. For example, prompt engineering is a relatively new promising technique that appears to improve LLMs' performance on downstream tasks. For example, in the context of zero-shot mathematical reasoning, Kojima et al. (2022) found that simply prompting GPT-3 with "Let's think step by step" quadrupled the accuracy on the MultiArith arithmetic dataset, from 18% to 79%!

In this paper, we focus on the potential of prompt engineering and external emotion classifiers to enhance the emotional abilities of ChatGPT. Our study builds upon previous research on prompt engineering and explores the effectiveness of external emotion classifiers in improving ChatGPT's ability to generate empathetic responses.

3 Preliminaries

In this section, we introduce some important notions that would be important to understand the design and implementation sections.

3.1 Problem formulation

A multi-turn dialogue defined as $D = \{U_1, ..., U_M\}$ consists of M alternate utterances of two interlocutors (Belainine et al., 2022). Each utterance U_i can be associated with an emotion label E_i. Given a dialogue D, we aim to generate the next utterance U_{M+1} that would be coherent, not only with the previous semantics, but also with the previous emotional state(s).

3.2 Emotion classification

Emotions are states of feelings resulting from internal or external changes in our lives and depend on the speaker's attitude and personality (Al-Omari et al., 2020). They can be classified into 6 basic categories according to Ekman (1992) or 8 classes according to Plutchik (1980). However, a recent study showed that using 27 emotion labels in addition to a neutral label can be effective for fine-grained emotion classification (Demszky et al., 2020). Using Principal Preserved Component (PPCA) Analysis (Cowen et al., 2019), they showed that the 28 used labels are highly significant.

One of the most challenging problems in the automated understanding of language is emotion recognition & classification. However, transfer learning can leverage the effectiveness of pre-trained LLMs to tackle such a task more effectively (Chronopoulou et al., 2019; Belainine et al., 2020b,a). By re-training (or fine-tuning) the pre-trained model on a smaller dataset that is tailored to the new task (emotion classification for example) while keeping some or all of the pretrained weights unchanged, the model we obtain is adapted to the new task. Compared to training the model from scratch, this method can result in faster convergence and greater performance using a fraction of the processing power (Pan and Yang, 2010).

3.3 Prompt engineering

Prompt Engineering can be defined as the design of instructions (prompts) in a way that improves the quality of the results from existing language models without further training on new datasets (Liu et al., 2023). As mentioned earlier, this technique has shown promising results in steering Large Language Models and improving their results without retraining or even fine-tuning (Kojima et al., 2022).

3.4 The ELECTRA model

The ELECTRA (*Efficiently Learning an Encoder that Classifies Token Replacements Accurately*) model is a type of neural network architecture that was introduced by researchers at Google (Clark et al., 2020). It has been shown to outperform other pre-trained language models such as BERT (Devlin et al., 2018) on several NLP benchmarks, including sentiment analysis (Mala et al., 2023).

The main innovation behind the ELECTRA model is *replaced token detection* instead of *masked token prediction*. In fact, for popular LLMs like BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019), the pre-training job is masking a portion of the unlabeled input and then training the network to retrieve this original input. This method works well, but its data efficiency is limited because it only learns from a fraction of the tokens. Researchers from Stanford University and Google Brain proposed replacing certain tokens with plausible substitutes produced by a small language model as an alternative to masking then trying to determine if each token is an original or a replacement using the pre-trained discriminator. This resulted in a significantly more computationally efficient model thanks to learning from the entire set of input tokens. Studies such as B et al. (2023) have shown that the proposed method greatly speeds up training and improves performance on downstream NLP tasks (Clark et al., 2020).

3.5 ChatGPT

ChatGPT is a Large Language Model based on the GPT-3.5 architecture and developed by OpenAI (Ouyang et al., 2022). It was trained on massive textual corpora and can provide human-like replies to a variety of natural language cues, from straightforward queries to more complicated dialogues. Using a transformer-based design, the model is able to capture long-range relationships in the input data and produce output that is incredibly fluent and coherent (Guo et al., 2023). It was originally trained based on InstructGPT (Ouyang et al., 2022) but it is also continuously improved using RLHF (Stiennon et al., 2020).

4 Experimental Design

In this section, we present a detailed description of the three ChatGPT versions used in these evaluations as well as the dataset used and the ELECTRA-based emotion classifier that we will need for the emotion infused version and for the evaluation part.

4.1 The Emotion Classifier

4.1.1 Datasets

For emotion classification, we used the GoEmotions dataset (Demszky et al., 2020), a large dataset of over 58k Reddit comments manually annotated with 28 fine-grained emotion labels by up to five different human annotators. It includes basic emotions like joy and anger but also more complicated ones like nervousness and caring. The authors argue that the chosen emotion labels are highly significant according to the Principal Preserved Component Analysis (PPCA) (Demszky et al., 2020). Figure 1 shows that the distribution of emotion labels is not balanced. We should keep this in mind when choosing appropriate evaluation metrics.

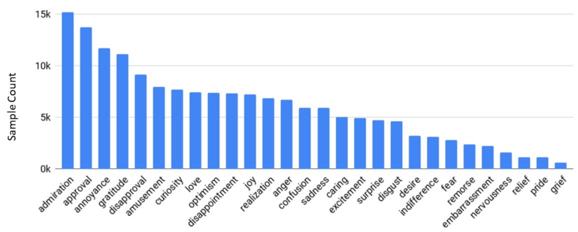

Figure 1: Labels distribution in GoEmotions dataset.

To analyze the dialogue performance of the chatbot systems, we will be using the Empathetic Dialogues dataset (Rashkin et al., 2018). This is a large-scale dataset made up of over 25,000 human-to-human dialogues designed to elicit sympathetic reactions. It was constructed by asking the participants to share personal tales and then to respond sympathetically to the stories of others. The dataset is all about emotionally grounded personal situations and therefore it is rich in terms of emotions.

4.1.2 Fine-tuning

Thanks to its impressive perfomance on the sentiment analysis task (Mala et al., 2023), which is similar to the emotion classification task, we chose the ELECTRA pre-trained model to build our emotion classifier. We fine-tuned it on the GoEmotions dataset using the PyTorch framework by adding a three-layer classification head consisting of:

- A fully connected layer used to reduce the feature dimensionality.
- A dropout layer to prevent overfitting.
- A fully connected layer used to map the reduced feature space to the number of emotion labels in the dataset (28).

We used cross-entropy as a loss function which includes the softmax function in its computation to calculate the probability distribution over the predicted classes according to equation 1:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j}) \qquad (1)$$

where N is the batch size (set to 128), M is the number of classes (28 in this case), $y_{i,j}$ is the binary label for the i-th example and j-th class, and $p_{i,j}$ is the predicted probability of the i-th example belonging to the j-th class.
During fine-tuning, the weights of the pre-trained ELECTRA model are frozen, and only the weights of the added classification head are optimized.

4.2 ChatGPT and emotions

We used three versions of ChatGPT to evaluate the impact of incorporating emotions in the generation process. Each version is fed with the first $n-1$ user utterances containing the context of the conversation, while the last, n^{th} utterance in any given conversation from the Empathetic Dialogues is predicted by the different chatbot models. We ran the experiments using the ChatGPT API.

4.2.1 ChatGPT-A: Regular ChatGPT

This is the basic version of ChatGPT and it was used without any modification. It is trained to generate responses to conversation prompts using only the text prompt as input. This model serves as a baseline to compare the performance of the other two versions. It will be denoted ChatGPT-A.

4.2.2 ChatGPT-B: Emotion-Infused ChatGPT

This version of ChatGPT takes into account the emotional state of the user in generating responses. We used the ELECTRA-based emotion classifier to classify the emotion of the user input. We use the emotion class as an additional input along with the text prompt to generate a response by adding the following sentence with the "assistant" role: *"Looks like you're feeling [emotion]"* just before asking ChatGPT to reply. We will be calling this model ChatGPT-B and figure 2 illustrates the idea behind injecting emotions into ChatGPT.

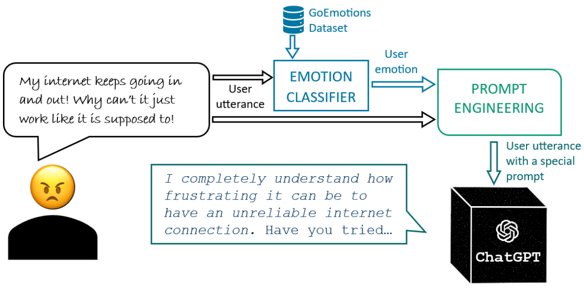

Figure 2: Illustration of emotion-infused ChatGPT.

4.2.3 ChatGPT-C: the Emotion-adapting ChatGPT

This version of ChatGPT uses simple prompt engineering to take the user's emotion into consideration. Specifically, we add the following sentence: *"Try to understand how I'm feeling"* in the "user" role before asking ChatGPT to continue the conversation. This chatbot version is going to be called ChatGPT-C in the following sections.

By comparing the performance of these three versions, we aim to assess the effectiveness of the proposed methods for enhancing the emotional abilities of ChatGPT in generating more human-like responses in empathetic conversations.
To illustrate the difference between the three versions of ChatGPT, here is a sample dialogue:

USER: My internet keeps going in and out and it is really bothering me! Why can't it just work like it is supposed to!
ASSISTANT: Are you on wireless, or a wired network?
USER: It is wireless. I keep on having to unplug and then plug back in the router and modem! Very frustrating!
CHATGPT-A: Some possible solutions include resetting your router, updating the firmware, checking for interference from other devices, and contacting your Internet Service Provider (ISP) for assistance.
CHATGPT-B: I completely understand how frustrating it can be to have an unreliable internet connection. Have you tried contacting your internet service provider to see if there is an issue with your service?
CHATGPT-C: I can imagine how frustrating it must be for you to have unstable internet.

5 Evaluations and Results

5.1 Emotion classification results

During fine-tuning of the ELECTRA-based classifier on the GoEmotions dataset, we experimented with different values for the hyperparameters in order to get the best results. The hyperparameters include a batch size of 128, a learning rate of 1e-4 and a number of 10 training epochs.

We divided the GoEmotions dataset into train, validation and test sets with ratios of 80|10|10 and we achieved excellent results in terms of the different metrics used. We got an AUROC score of up to 98.54%, an accuracy of 86.92% and an F1-score of 84.48%, indicating very reliable performance across all classes, despite the dataset being unbalanced. The full classification results per emotion label are illustrated in the table 1.

	Precision	Recall	F1-score
Admiration	0.91	0.91	0.91
Amusement	0.95	0.87	0.91
Anger	0.86	0.88	0.87
Annoyance	0.86	0.76	0.80
Approval	0.80	0.86	0.83
Caring	0.80	0.81	0.81
Confusion	0.88	0.84	0.86
Curiosity	0.77	0.94	0.85
Desire	0.80	0.86	0.83
Disappointment	0.80	0.81	0.80
Disapproval	0.76	0.87	0.81
Disgust	0.87	0.84	0.86
Embarrassment	0.90	0.87	0.89
Excitement	0.72	0.92	0.80
Fear	0.93	0.88	0.90
Gratitude	0.96	0.93	0.94
Grief	0.86	0.86	0.86
Joy	0.84	0.88	0.86
Love	0.90	0.95	0.92
Nervousness	0.69	0.75	0.72
Optimism	0.90	0.83	0.86
Pride	0.88	0.78	0.82
Realization	0.80	0.88	0.84
Relief	0.82	0.90	0.86
Remorse	0.64	0.86	0.73
Sadness	0.80	0.78	0.79
Surprise	0.74	0.93	0.82
Neutral	0.92	0.87	0.90
AUC			0.99
Accuracy			0.87
Macro avg	0.83	0.86	0.84
Weighted avg	0.87	0.87	0.87

Table 1: The detailed emotion classification results.

By examining table 1, we can see that almost all emotion labels achieved above 80% in precision, recall and F1-score. The lowest scores correspond to labels with the least training examples (such as pride that has less than 10 examples). This is expected since the labels with the most examples would be easier for the model to classify (such as admiration that has over 300 training examples). Overall, despite the big number of classed to choose from, our emotion classifier achieves impressive results, especially compared to the BERT-based model in the (Demszky et al., 2020) paper which only reached 40% 63% and 46% in precision, recall and F1-score, respectively. We can therefore assume that our model can be reliably used to predict the user and chatbot emotional expressions.

5.2 ChatGPT-B vs. ChatGPT-A

To compare the performance of the emotion-infused ChatGPT (ChatGPT-B) to the regular ChatGPT (ChatGPT-A), we ask both models to predict the last reply of each conversation as described in section 4.2. We then give an emotion label to each reply of both chatbots using our ELECTRA-based emotion classifier.

When examining results, we found that in 45% of the conversations from Empathetic Dialogues, both ChatGPT versions' replies were given the same emotion label. However, if we can use the probability of each emotion as an indication of the emotion intensity, we can plot the change in percentage of each emotion label probability in figure 3 and see some interesting results, even for replies with the same emotion label.

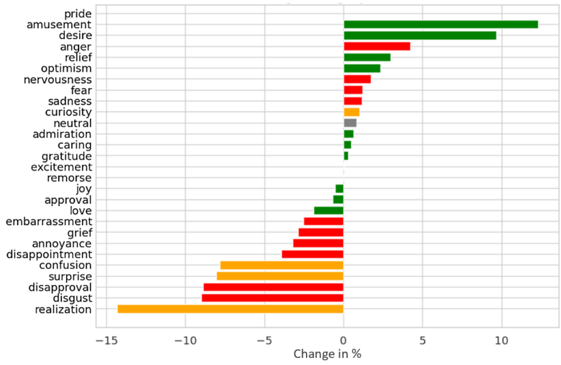

Figure 3: ChatGPT-A vs. ChatGPT-B emotion intensity.

As we can see in figure 3, positive emotions (with the green color) tend to be more pronounced in the emotion-infused ChatGPT, while negative (red-colored) and ambiguous (orange-colored) emotions were less intense overall. This indicates that when giving the user emotion as an input to ChatGPT, the chatbot tends to use more empathetic language. The "anger" emotion seems to be the exception here. This means that the replies that express this negative emotion are more pronounced with the emotion-infused ChatGPT. This can be explained by the fact that the chatbot tries to be empathetic by expressing anger about the same thing that the user was angry about

We also analyzed the replies of which the emotion label changed according to the emotion classification model, representing 55% of the conversations we tested. We plot the frequency change in percentage in the horizontal bar chart of figure 4.

Overall, the emotion-infused ChatGPT-B tends to use positive emotions more frequently whereas

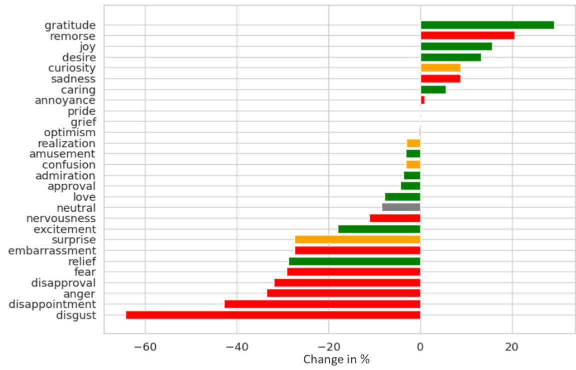

Figure 4: ChatGPT-A vs. ChatGPT-B emotion frequency.

negative and ambiguous emotions were used more rarely compared to regular ChatGPT. There are few exceptions out of the 28 emotion labels, though: remorse and sadness are used more, which shows more empathy towards the user, and relief and excitement are less often used, showing more understanding of the user request and less asking for elaboration. More importantly, the negative emotions like disgust, disappointment, anger etc., saw the biggest drop in use by ChatGPT-B. We can also notice that the neutral emotion is used less often, indicating more emotional replies. To analyze the results further, we created a confusion matrix to see the frequency change in each emotion label per user emotion to see which emotion labels were becoming what. This matrix is in the figure 5.

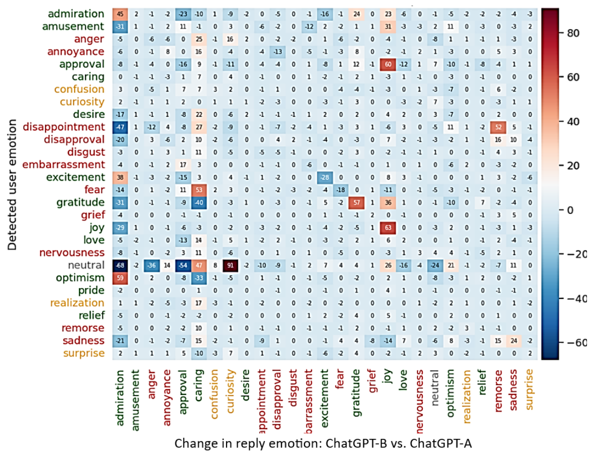

Figure 5: ChatGPT-B: Reponse emotion per user emotion.

Examining the heatmap, we can see that ChatGPT-B uses "caring" and "joy" emotions more often compared to the regular ChatGPT. The most noticeable change however is the use of the "curiosity" emotion. In fact, it is used much more often when it detects that the user is neutral. This indicates that the chatbot expresses interest in what the user is saying and that it is making inquiries in an attempt to learn more about the issue of the human.

5.3 ChatGPT-C vs ChatGPT-A

The emotion adapting version of ChatGPT, ChatGPT-C, which used the prompt *”try to understand how I’m feeling”* at the end of the user’s utterance shows different results. In figure 6, we can see that the chatbot tends to use negative emotions more often and positive emotions less often. This is likely due to the fact that this particular prompt is associated with negative emotions. In fact, a person wouldn’t say ”try to understant how I feel” when expressing joy or excitement, but rathen when he feels sad or annoyed; and ChatGPT tries then to match the emotion of the user in this case. To confirm that, we can examine the emotion frequency change per user emotion illustrated in the heatmap of the figure 7. The most noticeable changes are in the following situations:

- When the user is neutral, the chatbot expresses admiration much less often and instead tries to mimic either the ”caring” emotion or the ”anger” and ”sadness” emotions.
- When the user appears to be sad, the chatbot expresses ”approval”, ”joy” noticeably less often and expresses more ”caring” and ”sadness” instead.
- If the chatbot finds that the user is fearing something, it expresses the ”fear” emotion instead of ”approval” or ”curiosity”.

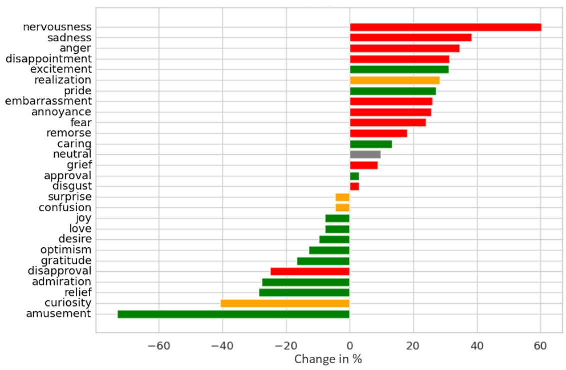

Figure 6: ChatGPT-C vs ChatGPT-A emotion frequency.

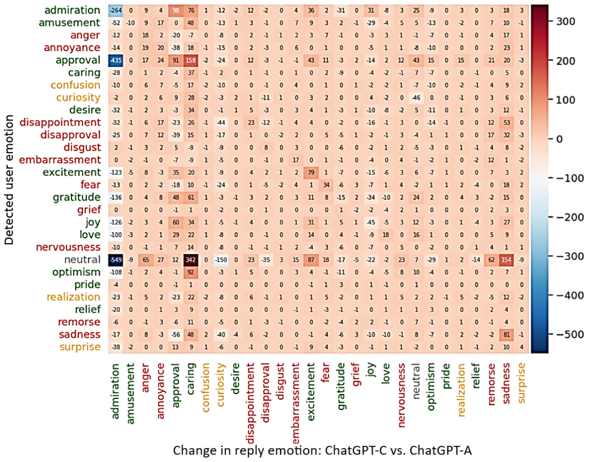

Figure 7: ChatGPT-C: Reponse emotion per user emotion.

5.4 Comparaisons to the SOTA Models

We also compared our three ChatGPT versions with other emotion-aware chatbot models as proposed in the literature (SOTA). The original transformer model (Vaswani et al., 2017) proposed in 2017 uses a self-attention mechanism to generate responses to user input. Emoprepend (Rashkin et al., 2018) is an improved version of the transformer with an emotion classifier trained to minimize the joint generation and classification loss. EmoEL (Lin et al., 2019) is a transformer-based model that combines responses from different decoders corresponding to different emotions. EmpDG (Li et al., 2019) is another chatbot model that uses both coarse-grained dialogue-level and fine-grained token-level emotions to generate empathetic responses.

	Response emotion choice				Perplexity
	Acc.	*Prec.*	*Recall*	*F1*	
Transformer	15.04	17.74	7.2	5.46	89.89
Emoprepend	15.04	**18.15**	7.72	5.58	94.77
MoEl	15.59	16.96	7.52	5.69	112.3
EmpDG	15.19	14.92	7.87	6.02	111.9
ChatGPT-A	28.64	13.19	**14.14**	**12.4**	30.63
ChatGPT-B	**39.55**	14.1	13.35	12.34	28.38
ChatGPT-C	22.71	10.03	12.23	9.4	**26.35**

Table 2: Benchmarking with other empathetic chatbots.

We compared our proposed ChatGPT versions with these models, in terms of the output emotion compared to the emotion label of the Empathetic Dialogues considered as ground truth response emotion. We also compared the fluency of the responses generated by each chatbot model to see which ones produce the most coherent replies.

The results in table 2 show that our proposed ChatGPT-B, the emotion-infused ChatGPT, outperformed both EmoEL and EmpDG in terms of emo-

Source	Utterance
Context	**Yeah about 10 years ago i had a horrifying experience . it was 100 % their fault but they hit the water barrels and survived . they had no injuries but they almost ran me off the road .**
Reference	*Did you suffer any injuries ?*
Transformer	Oh no , i am so sorry to hear that .
emoprepend	Oh no , did they catch them ?
moel	Oh no ! i hate spiders !
emp_DG	Oh no ! i am so sorry . i am so sorry . i am so sorry . i am so sorry to hear that .
ChatGPT-A	Wow, that sounds really scary! Have you been able to cope with the experience since then?
ChatGPT-B	I'm sorry to hear about your scary experience. It's alarming to think about what could have happened, but I'm glad you're okay.
ChatGPT-C	That sounds really scary and traumatic, and it's understandable that you would still remember it vividly.

Table 3: Reply examples from the different chatbot models.

tion response accuracy. Specifically, ChatGPT-B, with zero-shot, gave the highest emotion accuracy of up to 39.5%, while the regular ChatGPT gave the highest recall, and F1-score of 14.14% and 12.4%, respectively. These scores might appear to be on the low side, but we need to keep in mind that neither of the different ChatGPT versions were ever trained on the Empathetic Dialogues dataset, unlike the other models, and nevertheless produce impressive zero-shot results. Furthermore, we used a large number of emotion labels (28 fine-grained labels) which makes it harder to match the reference emotion exactly. In fact, a conversational agent can appear empathetic and emotional with several classes of emotions. For example, when looking at the answers from chatbots, we find that sometimes in the reference the answer to something like "I had an accident" is a question like "are you okay now?" which expresses the emotion 'curiosity' while the chatbot says "I hope you are okay now" which represents the emotion "caring". Moreover, in the reference, 25% of the answers are questions (expressing the "curiosity" emotion) while our chatbot responds are dominated by emotion "caring." Despite this, the ChatGPT versions perform the best overall with no prior training on the Empathetic Dialogues, in contrast to the other models.

On the perplexity front, it's clear that GPT-3.5-based ChatGPT models outperform the other chatbot models. In fact, since a lower perplexity generally means a more coherent expression (Bahl et al., 1983), we can see that ChatGPT-based models are vastly superior on this level. Specifically, the emotion-adapting ChatGPT-C has the lowest perplexity score of 26.35 while the emotion-infused ChatGPT-B has a slightly higher perplexity score of 28.38. The emotion-aware versions of ChatGPT are slightly more coherent when compared to the regular ChatGPT-A that got a perplexity of 30.63, likely thanks to the responses being more emotionally informed. While this is the worst score out of the three ChatGPT models, it is still well ahead of all the other models that have a perplexity score of more than 89.89. To see why this is the case, we can examine some examples in table 3. We can clearly see that ChatGPT models' responses are more natural and coherent compared to other models. For example, while emp_DG's reply does express remorse, it does so in a repetitive and unnatural sentence structure: *"oh no ! i am so sorry . i am so sorry . i am so sorry . i am so sorry to hear that ."* which explains the bad perplexity score for this model.

6 Conclusions and Future Work

In this study, we looked at how ChatGPT may elicit emotional reactions. Our findings imply that using prompt engineering and external emotion classifiers to augment conversational bots' emotional intelligence can be successful.

Our research adds to the expanding pool of knowledge regarding conversational agents and their emotional intelligence. The findings suggest that external knowledge sources, such as emotion classifiers, can provide a more nuanced understanding of the user's emotional state, and can lead to more effective and natural responses. Additionally, our study highlights the potential of prompt engineering to steer existing language models to produce outcomes tailored to our preferences without re-training or even fine-tuning. Future research might examine how well ChatGPT performs with other prompt designs. Other datasets can also be examined to see how that impacts the generated replies. We can also conduct a cross-lingual study to explore the benefits and limits of prompt engineering in generative AI.

References

Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

Martin Adam, Michael Wessel, and Alexander Benlian. 2021. Ai-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, 31(2):427–445.

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Hani Al-Omari, Malak A Abdullah, and Samira Shaikh. 2020. Emodet2: Emotion detection in english textual dialogue using bert and bilstm models. In *2020 11th International Conference on Information and Communication Systems (ICICS)*, pages 226–232. IEEE.

Merav Allouch, Amos Azaria, and Rina Azoulay. 2021. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24):8448.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 154–166. Springer.

Mala J B, Anisha Angel S J, Alex Raj S M, and Rajeev Rajan. 2023. Efficacy of electra-based language model in sentiment analysis. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pages 682–687.

Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, (2):179–190.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Billal Belainine, Fatiha Sadat, and Mounir Boukadoum. 2022. End-to-end dialog generation using a single encoder and a decoder cascade with a multi-dimension attention mechanism. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*.

Billal Belainine, Fatiha Sadat, Mounir Boukadoum, and Hakim Lounis. 2020a. Towards a multi-dataset for complex emotions learning based on deep neural networks. *Workshop on Linguistic and Neurocognitive Resources (LiNCr2020), Language Resources and Evaluation Conference (LREC 2020)*, pages 50–58.

Billal Belainine, Fatiha Sadat, Hakim Lounis, and Mounir Boukadoum. 2020b. Towards an emotionally driven natural language generation. *Montreal AI Symposium 2020*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

ChatGPT. Introducing chatgpt [online].

Ja-Shen Chen, Tran-Thien-Y Le, and Devina Florence. 2021. Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing. *International Journal of Retail & Distribution Management*, 49(11):1512–1531.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. *arXiv preprint arXiv:1902.10547*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Alan S Cowen, Petri Laukka, Hillary Anger Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature human behaviour*, 3(4):369–382.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Jonathan Grudin and Richard Jacques. 2019. Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–11.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2019. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys. Vol. 55, No. 9*.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

JB Mala, Anisha Angel SJ, Alex Raj SM, and Rajeev Rajan. 2023. Efficacy of electra-based language model in sentiment analysis. In *2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS)*, pages 682–687. IEEE.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

PaLM2. Palm 2 technical report [online].

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, 19(3-4):267–285.

Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. 2019. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Rui Zhang, Zhenyu Wang, and Dongcheng Mai. 2018. Building emotional conversation systems using multi-task seq2seq learning. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 612–621. Springer.