



Universiteit
Leiden

Master Computer Science

'I Feel You': Enhancing conversational agents with empathy

Name: LanGe Chen
Student ID: s3033694
Date: 29/03/2023
Specialisation: Artificial Intelligence
1st supervisor: Suzan Verberne
2nd supervisor: Joost Broekens

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Acknowledgements

First of all, I would like to express my deepest gratitude to Professor Suzan Verberne for being incredibly supportive and patient with me throughout this research journey. She ignited my passion for NLP and I feel fortunate and honoured to have her as my thesis supervisor. Her expertise and insightful feedback always helped me to navigate challenges, and I was always motivated by her invaluable guidance and encouragement! I would also like to thank my second supervisor, Professor Joost Broekens, for taking time to review my thesis and providing valuable feedback to help me refine it. Second, I am deeply grateful to my parents for being my cheerleaders and for giving me unconditional love every step of the way. Moreover, I would like to extend my heartfelt thanks to Sebastian, who sparked my enthusiasm for AI and kept encouraging me to push myself beyond my limits.

Finally, as my project involved a great deal of manual evaluations, I would like to sincerely thank everyone who helped me, whether they are my wonderful friends or some amazing people I don't know in person. While the questionnaire was anonymous, there was an option for the participants to mention their names. A big shout out to: *Alaa Hussaini, Tom Rice, Leo Lv, Ryan Pothier, Christiaan de Groot, Zihui, Valeria, Keira, Qizhi Ren, Abdelrahman Hussein, Mayar baydoun, Kishan, Patrick Jehamy, Dagmar, Justin, Yihan Li, Pawel Majestatyczny-Kutas, Will, Jayde House, Michael Gorgos, Ryan Herlihy, Nils, Kylee Barnard, Yuxuan Zhu*

Abstract

Recent breakthroughs in conversational AI, such as the release of ChatGPT, have attracted significant attention. While task-oriented chatbots that aim to assist users with specific tasks have been widely used, developing open-domain chatbots which provide relevant and informative responses remains a challenge. The objective of this project is to develop empathic conversational agents and to investigate their properties which could enhance user experience. We fine-tuned the large pre-trained language model DialoGPT with different sizes on the EmpathicDialogues dataset and performed hyperparameter optimization to obtain the best-performing EmpathGPT models. We demonstrated the effectiveness of our approach through the results of the evaluations: EmpathGPT far outperforms DialoGPT on both the level of relevance and empathy. The preference test result indicates that empathic responses are much more appreciated and this is especially the case in negative scenarios. Finally, we validated the experimental results with findings which are grounded in socio-behavioural studies.

Keywords: Conversational AI, GPT, Transfer Learning, Empathy, User Experience.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Research statement	2
1.3	Thesis overview	3
2	Background	4
2.1	Conversational AI	4
2.2	Generation-based systems	5
2.2.1	Generative architectures	5
2.2.2	GPT models	6
2.3	Goal-oriented chatbot	6
2.3.1	Knowledge Grounded Conversation	6
2.3.2	Empathetic chatbots	7
2.4	Evaluation approaches	7
3	Methodology	9
3.1	Dataset	9
3.1.1	EmpatheticDialogues	9
3.1.2	Data processing	10
3.2	DialoGPT Fine-tuning	12
3.2.1	DialoGPT	12
3.2.2	Fine-tuning	13
3.3	Evaluation metrics	15
3.3.1	Automatic Evaluation	16
3.3.2	Manual Evaluation	18
4	Experiment and results	21
4.1	Fine-tuning DialoGPT	21
4.2	Initial Evaluation	22
4.3	Hyperparameter Optimization	24
4.3.1	Tuning the learning Rate	25
4.3.2	Tuning the epochs	26
4.4	Quality Evaluation	27
4.5	Preference Evaluation	30
5	Discussion	35
5.1	Final results enhanced by social-behavioural studies	35

Contents

5.2 limitations 36

6 Conclusion 40

A Commonsense test questionnaire 48

B Quality evaluation samples and results 54

C Preference evaluation samples and results 61

1

Introduction

1.1. Motivation

Conversational AI, also known as chatbots or dialogue agents, is designed for machines to communicate with humans in a way that simulates human-like conversation. The goal of conversational AI is to make it easier for humans to obtain information or services, and to enable computers to provide assistance in natural language. Conversational AI can be used for a large variety of purposes, including customer service, entertainment and medical therapy etc. [Laumer et al., 2019]. However, most chatbots are designed to respond to user inputs in a factual way, independent of the user's emotional state and do not have the capacity for empathy. For better dialogue system design, early research [Liu and Picard, 2005] shows that embedding empathy into the system is the key for improving user experience. Chatbots designed to demonstrate empathy can respond to customer frustration with a personalized message that offers a sincere apology for any inconvenience caused. Moreover, they might also ask the customer how they would like to proceed and offer options for resolving the issue. Empathetic chatbots would likely be more effective at calming the customer's emotions and resolving the issue to their satisfaction. They are better at providing responds with support and comfort, which can lead to more adequate interactions and better user engagement [Ma et al., 2020].

A dialogue agent can be either task-oriented or open-domain, depending on its intended use and the capabilities it has been designed with. Task-Oriented Dialogue (TOD) is focused on helping users achieve specific goals (e.g. hotel booking, restaurant reservation) [Ilievski et al., 2018]. Classifying the domain and choosing the proper action are crucial parts for the TOD system to satisfy the user request [Louvan and Magnini, 2020]. Siri and Alexa are examples of TOD agents, which are designed to perform specific tasks and provide specific types of information, rather than engaging in a general conversation on any topic. On the other hand, open-domain dialogue is causally conversational [Grudin and Jacques, 2019] and not constrained to specific tasks. It should be able to response to any topic in an open-ended nature. Thus, open-domain dialogue has the closer property of being natural as human conversation with emotions and it remains as a more challenging task [Fu et al., 2022a]. As open-domain dialogue agents will encounter a wider range of potential topics and situations which can be sensitive or emotional, it may benefit from having more advanced empathy capabilities.

With the assumption that making the open-domain dialogue agent empathetic can lead to more effective and appreciative communication with users, the main objective of this research is to incorporate empathy into an open-domain dialogue architecture and further investigate if the agent could be more successful at improving user experiences.

1.2. Research statement

We propose EmpathGPT which is built by fine-tuning *DialoGPT* [Zhang et al., 2020] on the *EmpatheticDialogues* dataset. DialoGPT (Dialogue generative pre-trained transformer) is a transformer-based language model built upon GPT-2 [Radford et al., 2019a]. It is trained on a 147M conversational text data scripted from Reddit and appropriate for use in conversations. Systems that leverage it achieve decent performance in both automatic and manual evaluation [Zhang et al., 2020], which makes DialoGPT a great baseline for developing the empathetic dialogue agent. Additionally, as DialoGPT has been pre-trained on a wide range of conversation data, it can be fine-tuned relatively efficient with limited computational resources. We use both automatic evaluation and human evaluation for the measurement of the performance. Research questions addressed in this work are listed as follows:

- As Zhang et al. [2020] shows, the responses generated by DialoGPT are comparable to the human responses in the single-turn conversation of the Turing test. To what extent can fine-tuning DialoGPT on a specific empathetic dataset influence its ability to answer common sense questions and improve the relevance and empathy of its generated responses?
- What are the pros and cons of using perplexity as an automatic evaluation metric for conversational agents, and to what degree can it be considered a viable alternative to manual evaluation?
- How much do potential users appreciate that the chatbot is empathetic, and what kind of scenario might users be in need of even more empathy?

We make the following contributions in this thesis:

- We developed conversational agents that are both able to handle a wide range of topics and provide empathetic responses, rather than focusing on one or the other. Our chatbots are able to recognize when a user is expressing negative or positive emotions and respond accordingly with comforting or cheerful responses.
- We combined both automatic and human evaluation to comprehensively assess the performance of the models, as current automatic evaluation metrics for open-domain conversational systems are more or less inadequate. Meanwhile, we investigated both the advantages and limitations of perplexity as an automatic evaluation metric, in order to contribute to the ongoing discussion on how to improve the evaluation of open-domain dialogue systems.
- Empathy is the ability to understand and share the feelings of another person [Ioannidou and Konstantikaki, 2008], which is a crucial aspect of effective communication. Our manual quality evaluation overcame the limitations of the automatic evaluation. The results not only demonstrated the importance of balancing both topic handling and empathetic responses, but also provided valuable insights into how users perceive the generated responses in order to create more effective and engaging conversational agents.
- Our findings in the analysis of user preferences for generating responses in positive versus negative scenarios are consistent with some of the findings in related social behaviour studies [Ste-

vanovic and Peräkylä, 2015]. Combined with the aligned part of conclusions from two perspectives of different studies, we make better suggestions that chatbots should exhibit empathy in certain contexts under certain conditions that could potentially improve user experience.

1.3. Thesis overview

This thesis is structured into six main chapters, following the introduction of the **Chapter 1**:

Chapter 2 Background contains four sections of related literature reviews on the topics of conversational AI, generation-based systems, goal-oriented chatbots, and evaluation metrics for dialogue generation.

Chapter 3 Methodology explains the details of data processing, the procedure for fine-tuning the DialoGPT and how the evaluation process is designed.

Chapter 4 Experiment and results shows the initial results of the fine-tuning and human evaluation, followed by the hyperparameter optimization process and the final human evaluation results regarding the quality and empathy of the sample dialogues.

Chapter 5 Discussion introduces further analysis and refers to the results of socio-behavioural studies to support our experimental findings, and then reflect upon the limitations of our work.

Chapter 6 Conclusion summarizes the contributions of the study and suggests future work.

2

Background

2.1. Conversational AI

Conversational AI has been undergoing rapid development in recent years. The evolution of neural networks has brought huge leaps in the field of artificial intelligence, and with the increasing abundance of linguistic data on the Internet, natural language processing (NLP) has made great strides. As one of the most challenging tasks in NLP, conversational AI has witnessed groundbreaking improvements to handle complex and diverse conversations with users. Depending on the number of exchanges that occur between the user and the agent during the conversational interaction, there are two types of dialogue: single-turn dialogue and multi-turn dialogue [Chen et al., 2017].

Conversational AI can be broadly categorized into task-oriented dialogue and open-domain dialogue based on the purpose of the application. Task-oriented dialogue (TOD) system is designed to accomplish a specific task or goal, such as making a reservation or buying tickets [Balaraman et al., 2021], which is usually predefined within a finite set [Ilievski et al., 2018]. The systems tend to be structured as *slot filling* in which the user intent and the corresponding response are predefined as (*slot*, *value*) pairs. The dialogue state tracker [Williams et al., 2014] will trace information so as to better assist the downstream component to choose the proper action [Louvan and Magnini, 2020]. Therefore, classifying the domain and determining the user intent are the crucial stages for the TOD system, and they tend to be more structured and optimized to efficiently complete the task at hand. The Open-domain dialogue agent is designed to engage in open-ended conversations with users across a wide range of topics, not necessarily to accomplish specific tasks or goals, but rather to simulate human-like conversations and provide users with a personalized and engaging conversational interaction [Yan, 2018], and it remains challenging because of this property [Huang et al., 2019]. Open-domain dialogue techniques can be broadly grouped into retrieval or generative paradigms, and the blending of both is also an effective technique.

A retrieval-based model consists of two major components including an encoding function that convert the context and response into numerical representations, and a scoring function that retrieves the most appropriate response [Wu et al., 2016] to the user input after assigning scores to all responses in the candidate set [Lowe et al., 2015]. Retrieval-based dialogue agents are limited by the candidate reposi-

tory, despite being able to deliver relevant and syntactic responses [Boussaha et al., 2019]. Generative models can generate responses that are more human-like and may not appear in the training corpus [Chen et al., 2017]. As retrieval-based systems could result in repetitive or uninformative responses if the database is not comprehensive, it is not the optimal option for developing an empathetic chatbot that is supposed to be able to handle a wide range of semantic emotions. We will be focusing on the generation-based approach.

2.2. Generation-based systems

2.2.1. Generative architectures

Generation-based chatbots are built using deep learning techniques, such as Sequence to Sequence modelling, generative adversarial network, reinforcement learning and transformer architectures etc.

Sequence to Sequence (Seq2Seq) modelling is one of the widely used approaches. Seq2Seq models maps the input and context into a response sequence [Ni et al., 2022] and generate the response word by word. Seq2Seq models involve an encoder and a decoder. The encoder function transforms the textual content into a fixed-length vector representation, which captures the contextual information of the sequence. The decoder takes into account the context vector and the previously generated tokens, at each time step it selects the most likely word from a set of candidates and updates the context vector accordingly. There are various commonly used Seq2Seq models. For example, hierarchical recurrent encoder-decoder [Sordoni et al., 2015], variational autoencoder [Kingma and Welling, 2013], or the hybrid of both techniques variational hierarchical recurrent encoder-decoder architecture [Serban et al., 2017].

Generative Adversarial Network (GAN) [Goodfellow et al., 2014] is a type of generative model that involves two neural networks, a generator and a discriminator. It has been considered an effective method since Li et al. [2017] has applied GAN to dialogue generation and obtained promising results. Reinforcement Learning approach can also be applied to the development of the chatbot. The chatbot is considered as an agent that takes actions (i.e. generates responses) regarding the current state of the dialogue. Li et al. [2016] employs the deep reinforcement learning approach to simulate future rewards embedded in the conversation between two virtual agents, using a policy gradient to reward sequences based on predefined features.

In recent years, pre-training has been a key approach in the development of many state-of-the-art chatbots, thanks to the huge revolution in natural language processing brought about by the release of Transformer architecture. The Transformer [Vaswani et al., 2017] is an encoder-decoder architecture with a self-attention mechanism that allows it to process all inputs at once. The self-attention mechanism considers how each word in the input sequence is relevant to other words and the multi-headed attention layer enables it to better understand contextual correlations, driving its success in natural language processing tasks. GPT (Generative Pre-trained Transformer) [Radford et al., 2018] is specific implementation of the decoder of the transformer. The decoder consists of a stack of layers, each layer contains the multi-head self-attention mechanism and a feed-forward neural network. GPT also has a technique called “prompting”, which allows users to specify a starting point for the model to generate text. It can help to ensure the model generates text that is consistent with the context.

2.2.2. GPT models

The recent launch of chatGPT has attracted tremendous attention among various industries, and it shows astonishing success for the application of open-domain chatbots. ChatGPT is fine-tuned for conversations from the general generative model GPT-3.5 and further optimized using a method called Reinforcement Learning with Human Feedback (RLHF), which utilises human feedback and preferences to help guide the performance of the model. GPT is short for ‘Generative Pre-trained Transformer’. The first GPT [Radford et al., 2018] was released in 2018. It was trained on the Toronto Book Corpus – a massive corpus of unlabeled text containing 7,000 books and it demonstrated great effectiveness on the downstream tasks after being fine-tuned. GPT-2 [Radford et al., 2019a] was constructed by larger transformers and trained on a larger dataset WebText containing 8 million of webpages scraped and filtered from Reddit. Then GPT-3 [Brown et al., 2020] is one of the most state-of-the-art and largest language models released by OpenAI for generating human-like text. It is an enhanced version of the GPT-2 [Radford et al., 2019b] architecture, with significantly more parameters and advanced capabilities for natural language processing tasks, especially for dialogue generation.

The use of GPT models to develop conversational agents has multiple advantages. First of all, it is built on the transformer-based architecture to better understand the contextual relationships between words in the conversation. Its longer-term memory than traditional RNN-based models allows it to generate relevant responses over longer dialogues. Pre-trained on massive amounts of text data, it captures the underlying patterns and structure of natural language well to generate human-like and coherent responses to user input. DialoGPT [Zhang et al., 2020] is built on the GPT-2 architecture and is trained on large amounts of Reddit data. The training process regards the whole dialogue session as a long sequence and response generation is formulated as language modelling. It has shown impressive performance on several benchmark dialogue datasets, making it a good option for building an empathetic chatbot.

2.3. Goal-oriented chatbot

Depending on the applicability, the goal-oriented chatbots can be more informative and controllable [Fu et al., 2022b]. It happens easily that open-domain dialogue models tend to generate very generic or uninformative responses to be safe [Zhang et al., 2018b]. It is becoming increasingly popular to incorporate external knowledge into conversational AI models to improve their understanding of the conversational context and generate more informative responses [Meng et al., 2020]. Depending on the objective of the application, enabling chatbots to become more informative is knowledge-grounded. Making chatbots more controllable can be done in the perspective of conversational style, emotion, etc.

2.3.1. Knowledge Grounded Conversation

Knowledge Grounded Conversation (KCG) is to enable chatbots to give more informative responses. The knowledge of the conversational agent can be enhanced from multiple sources. Wizard of Wikipedia [Dinan et al., 2018b] explores the utilisation of unstructured knowledge from Wikipedia to enhance the conversational capabilities of chatbots. Unstructured knowledge data can be structured with the representation of nodes (entities) and edges (relationships) by knowledge graph [Xu et al., 2020]. Integrating multimodality image grounding into conversational agents [Mostafazadeh et al., 2017] can also help to improve their performance in a wide range of applications such as e-commerce. Moreover, [Zhang

et al., 2018a] proposes a new dataset and conversational architecture specifically for personalized dialogue. Personalized dialogue is an approach to incorporate the user profile or persona as the external knowledge so the chatbot can adapt to the user's preferences, interests etc. to generate more engaging and personalized responses.

2.3.2. Empathetic chatbots

As a chatbot that is capable of controlling emotions, Xiaolce [Zhou et al., 2020] is one of Microsoft's best-known empathetic chatbots and has a proven ability to provide users with enjoyable emotional companionship and support through empathetic and personalized conversations. The conversational agent that can identify intentions and recognize emotions can be considered as being able to interact with users better [Richardson and Heck, 2023]. An intelligent chatbot is expected to be emotionally consistent with users to enhance human-machine interactions and it has the potential to greatly enhance the quality of human-machine interactions. The early exploration started with generating emotionally rich text by Affect-LM [Ghosh et al., 2017]. Then, with the hypothesis that emoji can imply emotions, Zhou et al. [2018] collected datasets from Twitter, used emoji's emotions as labels, and then trained the model with the policy that rewarding responses that are consistent with the emotion will get higher rewards. Wolf et al. [2019b] introduce a transformer-based architecture called Transfertransfo based on transfer learning and achieved state-of-the-art results on the Persona-chat dataset. EP-Bot [Yoo and Jeong, 2021] employs PolarisX, an auto-growing knowledge graph that can extract knowledge graph embeddings from the context and identify the user's emotions. Fine-tuning is also shown to be an effective approach for developing the empathetic chatbot [Lin et al., 2019].

2.4. Evaluation approaches

Assessing the quality of open-domain dialogue systems remains a challenging task, as there is no objective metric that completely captures all aspects of quality dialogue [Fu et al., 2022a]. However, it can be done by several automatic metrics and human evaluation [Finch and Choi, 2020].

Automated evaluation applies probability theory to quantify the quality of dialogue with mathematical formulations. Although automated metrics are commonly perceived to be poor at capturing the true quality of dialogue [Liu et al., 2016], as human evaluation can be expensive and time-consuming, there are still several automatic metrics adopted from other natural language processing tasks. For example, BLEU [Papineni et al., 2002], ROUGE [Lin, 2004], METEOR [Denkowski and Lavie, 2014] and Distinctness [Li et al., 2015] etc. have been commonly used. One of the most popular metrics is perplexity [Chen et al., 1998]. Perplexity is the metric to evaluate the ability of a probability model to predict a sample. It is a commonly used evaluation for language models in NLP. Perplexity is defined as 2 to the power of the entropy of the predicted probability distribution over the next word, given the previous words. The investigation of Adiwardana et al. [2020] proves the positive correlation between human judges and perplexity. Adiwardana et al. proposed a human evaluation metric called *Sensibleness and Specificity Average (SSA)* based on two criteria to avoid the generic and vague responses [Venkatesh et al., 2017]: (1) If the response makes sense or not; (2) If the response is specific to the context. The experimental results demonstrate the strong correlation between the proposed human evaluation metrics and perplexity for several state-of-the-art neural conversational agents.

While the automatic metrics can only measure the dialogue objectively, the human evaluation can

provide a comprehensive and subjective assessment from the user's perspective. Manual evaluation can be carried out from different perspectives, for example, engagingness of the response [Zhang et al., 2018a], correctness based on the real-world knowledge [Wang et al., 2019], emotional understanding [Li and Sun, 2018] and knowledge relevance if responses are appropriate to the context [Liu et al., 2018] etc. There are pros and cons for each evaluation metric, overall, a combination of different metrics is often necessary in order to get a holistic and comprehensive evaluation.

3

Methodology

3.1. Dataset

3.1.1. EmpatheticDialogues

EmpatheticDialogues [Rashkin et al., 2018] is a benchmark dataset designed to assess the ability of conversational agents to generate empathic responses. It is a one-on-one dialogue dataset containing approximately 25k dialogues collected through crowdsourcing: when the speaker is assigned an emotion label, the speaker needs to write down some statements describing the scene when they have this emotion. The speaker will then tell this story to the listener and they will have a conversation around it. Therefore, each dialogue is based on a situation where the speaker is experiencing a particular emotion and the listener responds with empathy. The conversation included a wide variety of emotions, Table 3.1 shows some examples of the emotion labels and corresponding top 3 content words used by the speaker and the listener.

Table 3.1: Sentiment labels within EMPATHETICDIALOGUES dataset and top 3 content words used by speaker and listener per category [Rashkin et al., 2018]

Emotion	Most-used speaker words	Most-used listener words
Excited	going, wait, i'm	that's, fun, like
Angry	mad, someone, got	oh, would, that's
Grateful	really, thankful, i'm	that's, good, nice
Lonely	alone, friends, i'm	i'm, sorry, that's
Confident	going, i'm, really	good, that's, great
Embarrassed	day, work, got	oh, that's, i'm
Joyful	happy, got, i'm	that's, good, great

Compared to other empathy-related datasets, there are several advantages of using EmpatheticDialogues as a training dataset:

- Realistic and diverse: The dataset consists of each one-on-one dialogue grounded in a specific sentiment. This makes the dataset more realistic and diverse, as it contains various emotions

and conversational styles, which is vital for training a conversational agent that can generate empathetic responses in different scenarios.

- **Large scale:** The dataset contains approximately 25,000 dialogues, which can be considered a significant amount of data for training a chatbot. It allows for more effective fine-tuning of the model and enables chatbots to better generate empathetic responses.
- **Reliable:** Prior manual assessment experiments demonstrated that conversational agents using this dataset produced results that were perceived as more empathetic for annotators.

In order to have clearer insights into the content and structure of the dataset before training, we had a sanity check on the top 20 word count of the dataset. This can be helpful to get a sense of the overall content and structure of the dataset and identify potential issues or biases. As shown in Figure 3.1:

- The counts for the first 20 words show a sharp drop in counts. It is common to see such a sharp drop-off in word counts after the top few words in a dataset, as natural language data follows the long-tail distribution, where a small number of words appear extremely frequently and the vast majority of words occur infrequently [Zipf, 2016].
- The top word is “I”, which has a very high count that is up to 470,538. The word frequency in the dataset may depend on the specific context and domain of the conversations, as EmpatheticDialogues involves people talking about themselves and their experiences, the fact that “I” has such a high count is normal.

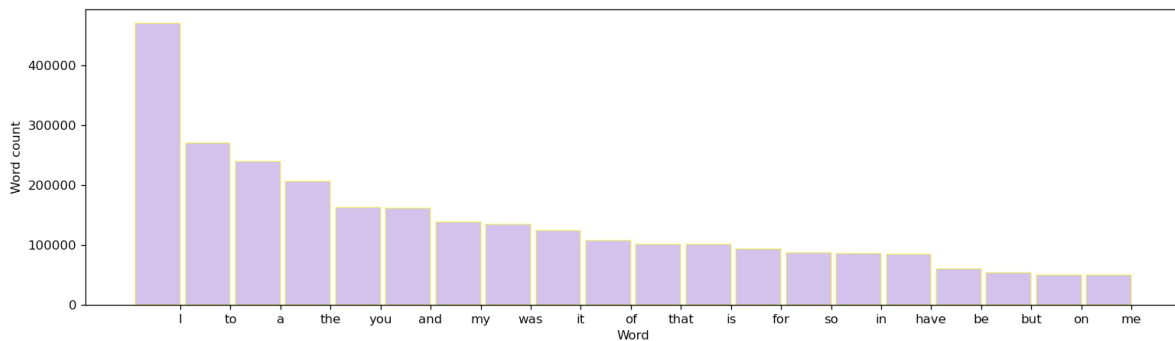


Figure 3.1: Top 20 word count of the dataset.

3.1.2. Data processing

To process the dataset into a format that could be fed to the model for fine-tuning, we performed data processing:

1. The training set, the validation set and the test set of the dataset are merged into one and then shuffled.
2. One of the main goals of the EmpatheticDialogues dataset is to provide a diverse and authentic collection of dialogues covering a broad range of emotions. Using this dataset for training, the conversational agent can learn to generate empathic responses in a variety of emotional contexts, even without explicit emotional labels. Meanwhile, the conversational agent needs to have the ability to use the context of dialogue to determine and adjust their responses to be appropriate, regardless of the emotions of the user: the emotion labels are not necessary for the training process, thus only the dialogue parts of the dataset were retained.

3. To guarantee the consistency of the dataset and remove noise, we cleaned the dataset by fixing the inconsistent whitespace formatting and removing those rows that contain missing contents. The cleaning process reduces the potential negative impact that the dataset may have on the model.
4. Finally, the dataset is being preprocessed to create input-output pairs, it is converted into the format that each response takes the previous seven turns as the previous context. In this format, each input sequence is composed of seven previous dialogue turns, and the corresponding output sequence is a response for the most recent turn. With this format, the model is capable of learning to generate contextually relevant responses, taking into account previous dialogue turns. Table 3.2 shows examples of the format of the training set.

Table 3.2: Three training items are shown (index 0, 1, 2) with for each of them seven context messages.

Index	0	1	2
Response	So happy everyone was fine!! Everything else can be fixed.	Indeed. We were out of the house for five months while repairs were being done. but now the house is better than ever.	So good to hear. Might want to trim some trees lol
Context	We were OK. though the tree broke through only a few feet away from my daughter.	So happy everyone was fine!! Everything else can be fixed.	Indeed. We were out of the house for five months while repairs were being done. but now the house is better than ever.
Context/0	That's very scary. I hope no one got hurt.	We were OK. though the tree broke through only a few feet away from my daughter.	So happy everyone was fine!! Everything else can be fixed.
Context/1	What a difference a year makes. Last year one evening my family was at home when a tree fell on the house and broke through the ceiling.	That's very scary. I hope no one got hurt.	We were OK. though the tree broke through only a few feet away from my daughter.
Context/2	Well we ended up leaving I was too scared to stay and was really upset.	What a difference a year makes. Last year one evening my family was at home when a tree fell on the house and broke through the ceiling.	That's very scary. I hope no one got hurt.

Context/3	Oh that does sound terrible. What did you guys decide to do?	Well we ended up leaving I was too scared to stay and was really upset.	What a difference a year makes. Last year one evening my family was at home when a tree fell on the house and broke through the ceiling.
Context/4	I was at the beach and a hurricane changed track and was coming straight at us. My boyfriend didnt want to leave because we were on the second day only of our vacation.	Oh that does sound terrible. What did you guys decide to do?	Well we ended up leaving I was too scared to stay and was really upset.
Context/5	What happened for you to feel that way?	I was at the beach and a hurricane changed track and was coming straight at us. My boyfriend didnt want to leave because we were on the second day only of our vacation.	Oh that does sound terrible. What did you guys decide to do?

- Finally, the dataset was split into training and validation sets using a ratio of 8 : 2, with 81595 and 20399 dialogue samples respectively. The training set is used to optimize the parameters and the validation set is used to evaluate the model's performance on data that it has not seen during training, and to monitor the training progress.

3.2. DialoGPT Fine-tuning

In order to develop an empathetic chatbot, we fine-tuned DialoGPT models using the EmpatheticDialogue dataset. The fine-tuned models should be able to improve the ability to generate more empathetic responses to user inputs.

3.2.1. DialoGPT

DialoGPT [Zhang et al., 2020] is a large-scale generative conversational model based on the Transformer architecture, which has proven highly effective across a range of natural language processing tasks. It was pre-trained on a large corpus of conversational text data scripted from Reddit. The pre-training process enabled the model to learn the patterns of the conversational structure by predicting the next word in a given sequence of words. Moreover, DialoGPT has been trained with the aim of causal language modelling (CLM) on dialogue data and is therefore well suited for dialogue generation in the open domain. Table 3.3 shows an example of its multi-turn dialogue generation.

Role	Response
User	Does money buy happiness?
Bot	Depends how much money you spend on it .
User	What is the best way to buy happiness ?
Bot	You just have to be a millionaire by your early 20s, then you can be happy.
User	This is so difficult !
Bot	You have no idea how hard it is to be a millionaire and happy. There is a reason the rich have a lot of money

Table 3.3: An interactive example of multi-turn dialogue [Zhang et al., 2020].

In this project, DialoGPT has been used as a pre-trained model for fine-tuning due to the following advantages:

- As described in the background section, there are many different techniques for developing a dialogue agent. However, enhancing relevance is a critical factor to ensure the performance of a dialogue agent. As the DialoGPT model is based on a transformer architecture, it is able to capture long-term dependencies in the input data and produce responses that are relevant to the dialogue history.
- The model is trained on a large and diverse corpus of dialogue data, allowing it to capture the nuances and subtleties of human dialogue and DialoGPT has been shown to produce responses that are not only contextually appropriate but also diverse and engaging. This trait is certainly an advantage for conversational agents that need to respond to a wide range of topics and give empathetic responses. And it is will be helpful to provide a more natural and intuitive conversational experience for users.
- DialoGPT has been evaluated on various popular conversational tasks, including chit-chat, question answering and dialogue generation. It has shown state-of-the-art performance on several benchmarks, including the Persona Chat [Zhang et al., 2018a] and Wikipedia Wizard datasets [Dinan et al., 2018a] that are widely used for evaluating dialogue agents. This fact indicates that it is a high-quality pre-trained model, which is well-suited for fine-tuning specific conversational tasks.

3.2.2. Fine-tuning

Fine-tuning a language model is the supervised process of adapting a pre-trained language model to a specific task by training it on a task-specific dataset. With the availability of large pre-trained models, such as GPT [Radford et al., 2018] and BERT [Devlin et al., 2018] models, which have been pre-trained on large amounts of textual data to capture a wide range of linguistic patterns and structures, fine-tuned language models have become increasingly popular in recent years. By fine-tuning these models on task-specific datasets, their existing knowledge can be leveraged to achieve state-of-the-art performance on a wide range of natural language processing tasks. To develop an empathic dialogue agent, we fine-tuned the dialoGPT on the EmpatheticDialogues dataset.

In terms of efficiency, since DialoGPT has learned the basic language structures and patterns necessary to produce coherent responses in conversation, it is much more effective to fine-tune it than to develop a model from scratch. In addition, fine-tuning a pre-trained model requires much less training data than training a model from scratch, because a pre-trained model like DialoGPT already has extensive knowledge, so it just needs to learn how to generate responses in the specific pattern that is empathic.

This is particularly helpful in situations where training data is limited, as is often the case with empathic datasets.

For a more scientific comparison and optimization during the experiment in order to get more comprehensive observations and the best-performing model, we fine-tuned DialoGPT models with all different sizes provided by the Hugging Face Transformers library [Wolf et al., 2019a]:

- ‘microsoft/DialoGPT-small’ with 117 million parameters
- ‘microsoft/DialoGPT-medium’ with 345 million parameters
- ‘microsoft/DialoGPT-large’ with 762 million parameters

The context window is set to 1024 tokens for all model sizes. The context window is the maximum number of tokens that the model can have as input to generate a response, which means that each model can take up to 1024 tokens of dialogue history into account when generating a response. The model only considers the last 1024 tokens if the input sequence is longer than that. The maximum output length is also set to 1024 tokens for all model sizes. The maximum length of the output is the maximum number of tokens that can be generated by the model as a response, if the model reaches the maximum output length before a full sentence has been generated, it will stop generating and return only the currently generated output as the final response.

The process of fine-tuning includes:

1. **Initial experiment** We started with fine-tuning the small, medium and large DialoGPT models with the following configuration setting (Some of the critical hyperparameters like learning rate and epoch are based on the prior work of fine-tuning small DialoGPT [Beredo et al., 2021]):
 - *learning rate* = 5×10^{-5} The learning rate controls the speed of updating the model parameters during the optimization process. Selecting an appropriate learning rate is critical to achieving good model performance and convergence. We started with 5×10^{-5} as suggested by the prior work.
 - *Warmup steps* = 0 The setting of warmup steps is usually helpful to stabilise the training process and avoid divergence or instability in the early stages of training, as it increases the learning rate from a small value to its optimum value over the warmup period. Here we do not use any warm-up period, as the initial learning rate is set to an appropriate value as demonstrated by the previous work.
 - *Batch size* = 4: The batch size determines the number of samples processed in each training iteration. By allowing for more parallel processing and better hardware utilisation, a larger batch size can potentially speed up training. However larger batch size also requires greater memory. Starting with a batch size of 4, which is a common batch size for training deep learning models, is a reasonable option as it provides a good balance between memory usage and training time.
 - *Training epochs* = 3 The number of training epochs determines the number of times the entire training set is presented to the model. Increasing the number of epochs can improve performance, but also increases the risk of overfitting. As suggested by the previous work, and also because it is a commonly used number of training epochs for fine-tuning a large-scale language model, it was set to 3.
 - *Maximum sequence length* = 512 It determines the maximum number of tokens in a sequence that can be fed into the model during training and inference. 512 is a common sequence length for many NLP tasks, including language modelling and dialogue generation.

Furthermore, longer sequences require more memory and computational resources during training and inference, which can be limiting. During training, the sequence will be truncated if it is longer than 512 tokens.

2. **Initial evaluation** We first looked at the performance of the models using perplexity as an automatic evaluation metric, and then implemented a preliminary evaluation. Six models were manually evaluated in this experiment, including a. Small DialoGPT, b. Medium DialoGPT, c. Large DialoGPT, d. Small EmpathGPT, e. Medium EmpathGPT and f. Large EmpathGPT. The input are 28 different common sense questions, the same interface is used to generate the answer. We used top-k sampling and the value of k is set to 50, which means the model will only consider the top 50 most likely words or tokens when generating responses. Finally, 10 annotators were invited to manually evaluate the DialoGPT models and the fine-tuned EmpathGPT models in terms of their ability to answer common sense questions. The results of the evaluation were used to carry out a preliminary detailed assessment of the performance of all the models, which enabled further improvements to be made.
3. **Hyperparameter optimization** From the results observed in the preliminary experiments, we concluded that the chosen hyperparameter values might have affected the model performances and thus carried out the optimization by tuning the hyperparameters including learning rate and training epochs. Adjusting the learning rate is a common starting point. A low learning rate can lead to slow convergence and may require many more epochs to reach the optimal solution, however, if the learning rate is too high, the weights will change too drastically and the previously learned information could be erased. Increasing the number of epochs may help with this forgetting issue. By training the model for a longer period of time, the model has more opportunities to learn from the new data while retaining the previously learned information, while more epochs can also lead to overfitting if the model is not regularized properly. Therefore, we monitored the change in perplexity during the training process. Finally, the best-performing models measured by automatic evaluation metrics were found.
4. **Final evaluations** Finally, we conducted manual evaluations (which will be elaborated on in the next section). We performed a comprehensive analysis of the observations from the automatic and manual evaluations and found some consistency with the expected results, as well as new interesting findings. We further discovered some socio-behavioural findings that provide a basis for interpreting our experimental results, and eventually, we suggest features of conversational agents that can enhance the user experience.

3.3. Evaluation metrics

In this project, we have combined automatic and manual evaluation to measure the performance of the model. Automatic evaluation can provide a quick and efficient way to evaluate the quality of generated responses. Especially in this large-scale evaluation task, where it may be difficult or impractical for human annotators to review all the responses generated, the use of automated scoring metrics is particularly important. However, it is important to note the limitations of automatic evaluation metrics, as they are generally based on heuristic algorithms and are not always representative of human perception, especially when it comes to subjective language concepts such as empathy. Therefore, while the automated evaluation metric can provide useful insights into conversational agent performance, it is better used with human evaluation for a more complete and accurate assessment. The combination of

automatic and human evaluation can provide a more holistic understanding of the performance of the model.

3.3.1. Automatic Evaluation

Previous studies referenced in the background chapter 2 have demonstrated several benefits of using perplexity as an automatic evaluation metric in the context of natural language processing. In this project, we adopt perplexity as the automatic evaluation metric for its following advantages:

- **Quantification** Perplexity represents a quantitative measurement of the model's capability to predict the next token in a sequence, making it useful to monitor how the model is improving and whether it needs further optimization by tracking perplexity over the training period. It is also helpful for comparing different models.
- **Efficiency** Since perplexity requires only the input sequence and the predicted probabilities of the next tokens in the sequence, which are output by the model during training, it can be computed quickly and automatically. This makes it an efficient metric to use to evaluate the performance of a model on a large amount of data. In addition, the efficiency of perplexity enables more efficient experimentation with hyperparameters and different models, allowing faster progress in improving model performance.
- **Objectiveness** Perplexity is an objective metric as it is calculated solely based on the probabilities assigned to each token in the sequence by the language model. This means that it is not subject to the biases or variability associated with human judgement. It allows a fair comparison of different models or training configurations and objective evaluation of which models perform better and which training strategies are more effective.

Perplexity is a commonly used evaluation metric in machine learning, especially for language models. Specifically, it measures how confident the model is in predicting the next word in a given sequence. A lower perplexity score is an indication that the language model is better at prediction and therefore has a better understanding of the language and vice versa. Perplexity as a metric for quantifying the uncertainty of a model's predictions is related to the concept of entropy. Entropy was originally introduced to describe the complexity of the microstate and was later transferred to information theory by Claude Shannon, who introduced the concept of information entropy [Shannon, 2001]. Information entropy offers a solution for the quantitative measurement of information and denotes the measure of "uncertainty" inherent to all possible outcomes of a random variable. Applying this concept to perplexity for language models that assign probabilities to sequences of words, perplexity measures how "uncertain" the model is in predicting all possible outcomes. Therefore, a lower perplexity score indicates that the model is more sure about the prediction and performs better in predicting the next word, while a higher perplexity indicates that the language model is less certain about the next word.

After fine-tuning, we can use the test set to compute perplexity, once the model has learned the probability distribution over sequences of words. Perplexity can be expressed as the exponential of base 2 raised to the entropy of the probability distribution of the next word, given the context of the previous words. With the following defined terms:

- (w_1, w_2, \dots, w_n) denotes a sequence of words.
- $p(w_i \mid w_1 : i-1)$ denotes the probability of the i -th word given the previous words in the sequence.
- $H(P)$ denotes the cross entropy of the true distribution and the predicted distribution.

The perplexity of the model with the sequence of words (w_1, w_2, \dots, w_n) is defined as:

$$\text{Perplexity} = 2^{H(P)} \quad (3.1)$$

Considering the sequence of words W as a random variable, and the previous words $w_{1:i-1}$ as another random variable H , the entropy of the next word given the previous words is written as:

$$H(w_i | w_{1:i-1}) = - \sum_{j=1}^n p(w_i, w_j | w_{1:i-1}) \log_2 p(w_i | w_{1:i-1}, w_j) \quad (3.2)$$

where $p(w_i, w_j | w_{1:i-1})$ is the joint probability of the i -th and j -th words given the previous words. $p(w_i | w_{1:i-1}, w_j)$ is the probability of the i -th word w_i given the previous words $w_{1:i-1}$ and a context word w_j , it represents the conditional entropy of the probability distribution over the next word, given the context provided by the previous words in the sequence. Using the chain rule of probability, we can rephrase the joint probability of the sequence of words as a product of the conditional probabilities:

$$p(w_{1:N}) = \prod_{i=1}^N p(w_i | w_{1:i-1}) \quad (3.3)$$

If we substitute this expression into the definition of perplexity, we get the following expression:

$$\begin{aligned} \text{Perplexity} &= 2^{-\frac{1}{N} \sum_{i=1}^N \log_2 p(w_i | w_{1:i-1})} \\ &= 2^{\frac{1}{N} \sum_{i=1}^N -\log_2 p(w_i | w_{1:i-1})} \\ &= 2^{\frac{1}{N} \sum_{i=1}^N \log_2 \frac{1}{p(w_i | w_{1:i-1})}} \\ &= \left(\prod_{i=1}^N \frac{1}{p(w_i | w_{1:i-1})} \right)^{\frac{1}{N}} \end{aligned} \quad (3.4)$$

This expression shows that perplexity can be interpreted as the geometric mean of the reciprocals of word probabilities in the sequence. We can also see how perplexity relates to entropy by observing:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \log_2 \frac{1}{p(w_i | w_{1:i-1})} &= -\frac{1}{N} \sum_{i=1}^N \log_2 p(w_i | w_{1:i-1}) \\ &= H(p(w_i | w_{1:i-1})) \end{aligned} \quad (3.5)$$

From this we can conclude the following:

$$PP(W) = 2^{H(p(w_i | w_{1:i-1}))} \quad (3.6)$$

It shows that perplexity is equivalent to exponentiated entropy of the probability distribution over the next word. The relationship between perplexity and entropy demonstrates that perplexity is essentially an exponentiated form of entropy. Perplexity can be interpreted as the reciprocal of the normalized probability of the sequence. Practically, predictable outcomes with lower entropy are preferable to uncertain ones, since lower entropy means less noise. Likewise, models with lower perplexity are preferred because they indicate greater predictability and less perplexing performance. The adoption of the concept of entropy from information theory provides a scientific basis for evaluating language

models, given the mediatory role of language in the transmission of information. Moreover, perplexity is a more interpretable evaluation metric than entropy for language models because it directly relates to the number of possible options for the next word, which is a more tangible concept than entropy. Perplexity is therefore an appropriate and intuitive metric for the measurement of the performance of a language model.

Although perplexity is a widely used metric for evaluating language models, just like most metrics for evaluating language models, it has some limitations and drawbacks. A disadvantage of using perplexity is that it does not take into account the semantic or syntactic accuracy of the predicted word, but measures the average number of possible choices for the next word. A language model with low perplexity can still produce nonsensical or syntactically incorrect sentences [Ngo et al., 2021].

3.3.2. Manual Evaluation

For language models, it remains a challenging task to measure the understanding and generation of natural language by the model, and current automatic metrics cannot fully interpret the nuances of communicative interaction. Although perplexity is a commonly used metric to evaluate language model performance, it is not necessarily the best metric to use when fine-tuning DialoGPT on an empathetic dataset, as the best metric depends on the specific training goals and desired model behaviour. Especially in this project, we aim to build the empathetic conversational agent, while perplexity is unable to quantify empathy. To evaluate the empathetic behaviour of the model, as humans can provide more accurate measures of empathy, it is necessary to use manual evaluation in terms of the following aspects:

- **Common sense information correctness** Pre-trained DialoGPT models trained on large amounts of Reddit data have shown some ability to solve common sense problems [Zhang et al., 2020]. Although common sense questions do not require expertise to answer, as each person may have their own opinion of the correct answer, we designed the common sense question answering experiment. Annotators were invited to score the responses of pre-trained DialoGPT models and fine-tuned EmpathGPT models to all common sense questions. Human evaluation can be helpful to validate the prior conclusion and also to identify if the model has lost information after fine-tuning, as potential biases or information loss may not be reflected in the automatic evaluation score.
- **Quality: Relevance and level of empathy** The goal of our project is to create an empathetic chatbot that is practical for real-world users. The fact that human evaluation is the closest criterion to user experience makes it essential to assess the model's ability to handle real-world cases. Annotators were invited to evaluate the quality of sample dialogues in terms of relevance and the level of empathy separately. Relevance and empathy are two distinct concepts, which means that a high score in relevance may not necessarily mean that the response is empathetic and vice versa. However, specifically for the user experience of chatbots, it can also be hypothesized that if a chatbot gives a completely irrelevant response, users are likely to think that the bot lacks empathy because it fails to understand them and vice versa, the more empathetic responses generated by EmpathGPT models might be perceived as more relevant. Furthermore, level of empathy can not be measured by automatic evaluation metric. Overall, human scoring allows for a more comprehensive and refined assessment of the quality of the generated dialogues.
- **Preference evaluation** By conducting a preference test, annotators can provide subjective eval-

uations of the responses generated by different models based on their own preferences and judgments. This can help to identify which model is more likely to generate responses that human users would prefer, and observe that these are not necessarily the most appropriate or empathetic responses that would sometimes be preferred or not.

Inspired by Beredo et al. [2021]’s experiment with the fine-tuned small model versus the original model for human evaluation, we designed a larger and more fine-grained manual evaluation for our project. The evaluation was conducted in terms of both quality and preference, the detailed experimental steps are as follows:

1. **Dialogue collection** Several volunteers were invited as a user group to engage in chit-chat with the models. Users were informed that they could have open-domain conversations with the EmpathGPT models on any topic: specifically, for multiple-turn conversation only with Large EmpathGPT; for the single-turn conversation, it can be with any of those three models. In this way, we collected sufficient dialogue samples of the conversations generated by the potential user groups.
2. **Classification and categorization** The collected samples were classified in two steps:
 - (a) First, we distinguished single-turn dialogues from multi-turn dialogues and divided them into two groups according to this attribute. Multi-turn dialogues will be prepared as samples for quality evaluation and single-turn dialogues will be prepared as samples for preference evaluation.
 - (b) Then, samples of each group were further categorized into two groups depending on the semantic sense: positive scenario and negative scenario. This was to facilitate subsequent richer information and observations from the experimental results.
3. **Sample organization** We refined and finalised the samples according to the attributes of single-turn or multi-turn.
 - (a) For quality evaluation, we decided to use the multi-turn dialogue samples from Large EmpathGPT as previous research [Brown et al., 2020] has shown that larger language models can usually outperform smaller language models. We randomly selected six dialogues from each of the positive and negative scenario groups (12 dialogues in total) and took the user inputs to carry out the exact same conversation with Large DialoGPT. Therefore, we obtained 12 dialogue pairs (24 dialogues in total) by interacting with Large EmpathGPT and Large DialoGPT. Each pair had the same user input to facilitate manual evaluation and comparison.
 - (b) For preference evaluation, all DialoGPT and EmpathGPT models (6 models in total) were involved. We randomly selected 6 single-turn dialogues in each of the positive and negative scenarios (12 dialogues in total). Since these dialogues could be generated by chatting with any of the EmpathGPT models, we took all user inputs and chatted with the other 5 models. In total, there are 12 user inputs, each of which corresponds to responses from 6 different models.
4. **Questionnaire creation** We used Google Forms to create separate questionnaires for quality and preference evaluations.
 - To make the experiment more scientific, the order of all positive and negative scenarios and dialogues from the DialoGPT or EmpathGPT models were all disordered and shuffled.

- To avoid the tedium that extensive text reading may impose on annotators, the questionnaires concisely state the scoring criteria and all dialogues are presented in the figure with visually pleasing backgrounds.
5. **Evaluation Process** A total of 51 annotators were invited to participate in the evaluation, either online (access to the survey via link) or onsite (fill out the questionnaire on site). We looked through all the submissions and removed one completely blank submission we found, so there are a total of 50 valid submissions for each questionnaire.
 6. **Results analysis** We collected the scores of the human evaluations and corresponded these shuffled data to the models to which they corresponded. Finally, we analyzed and summarized the performance of all models.

4

Experiment and results

This section will present all the experiments conducted in the project along with the respective results and findings. We progressively optimize the model by observing the results of each sub-experiment with the assessment. We first fine-tuned the small, medium, and large DialoGPT models to obtain the corresponding EmpathGPT models. A common sense problem test was designed and evaluated manually to get an initial comparison and evaluation of the performance for all models. Based on the experimental results, we performed hyperparameter optimization for the models and conducted a thorough manual evaluation. Finally, we analyzed the user expectations based on the manual evaluation results and summarized the factors that could affect the user experience for chatbots, combined with socio-behavioural studies.

4.1. Fine-tuning DialoGPT

DialoGPT [Zhang et al., 2020] was pre-trained on models which have 117M, 345M and 762M parameters respectively. Depending on the size, DialoGPT has three different variants, the specification of each model is shown in Table 4.1.

Model	Layers	D_{model}	Batch size
117M	12	768	128
345M	24	1024	64
762M	36	1280	32

Table 4.1: DialoGPT model configurations [Radford et al., 2019b].

We conducted the experiment by fine-tuning all variations (Small, Medium, Large) of DialoGPT for getting the empathetic conversational agents named EmpathGPT (Small, Medium, Large). The results shown from the experiment that fine-tuned the small variation of DialoGPT by Beredo et al. [2021], the model hit the best perplexity score using $5 * 10^{-5}$ as the learning rate and converges after training for 3 epochs. As two parameters that may affect model performance significantly, starting with experimented values of learning rate and epoch can make the experimental process more efficient. We adopted these results and use the batch size of 4 with 512 as the sequence length to fine-tune DialoGPT models. The

results are shown in Table 4.2. The results shows that the larger the pre-trained model is, the smaller the perplexity obtained after fine-tuning.

Model	Small EmpathGPT	Medium EmpathGPT	Large EmpathGPT
Perplexity	3.9665	1.9590	1.1352

Table 4.2: Perplexity of $lr = 5 * 10^{-5}$ within 3 epochs.

4.2. Initial Evaluation

The perplexity results indicate that the larger the pre-trained model is, the more confident it is for the text it generates. Prior human evaluation results of Zhang et al. [2020] show that the responses generated by DialoGPT are comparable to the human responses under the single-turn conversation of the Turing test. Pre-trained on a large amount of Reddit data, the DialoGPT models have shown the ability to solve common sense problems to some extent. Even if the ‘desired’ answer is not given for some questions, it will give a similar alternative explanation. To verify this claim and to test the EmpathGPT models obtained after fine-tuning, we designed a ‘Commonsense problem test’ experiment. (All samples can be found in Appendix - A)

Table 4.3: Samples of the commonsense questions and answers generated by corresponding models

Question	Model	Answer
1. Which season is snowy?	a	The winter storm that comes every year.
	b	S1E10 I think
	c	I think its been years since I’ve seen the snow. I was in Missouri during the winter and there was no snow.
	d	I’m not sure if it’s the season, but I think it’s the weather.
	e	Season 4
	f	Season 4
2. Is the moon bigger or the sun?	a	Both! It’s a very complex shape with a heated retina of some sort.
	b	It’s the sun. The moon is about 1.25 Jupiter sized.
	c	It’s the size of the solar system.
	d	The sun is bigger.
	e	The sun is bigger.
	f	The moon is bigger.
3. Where is Amsterdam?	a	It’s in the Netherlands. you should check it out!
	b	Sorry. it was a small town in Delaware.
	c	It’s a city in The Netherlands
	d	Where is the UK?
	e	It’s in the Netherlands.
	f	It’s in the Netherlands.
4. Which planet do we live on?	a	Earth. I guess I should have said solar eclipse.
	b	The brightest star in the sky is Jupiter.
	c	I really care for my mother
	d	I’m not sure if you’re joking or not, but that’s not how the universe works.
	e	The one where we are all living on.
	f	I think you mean the one where we are all living on.

We took several questions that DialoGPT had answered correctly in the prior experiment Zhang et al.

[2020] and came up with some new ones. There were 28 common sense questions prepared in total. (Note that ‘Common sense questions’ in this context refer to questions based on practical knowledge that is obvious or intuitive for most people. These questions usually have simple, straightforward answers without requiring expertise.) A total of six models were involved in this experiment, including a. Small EmpathGPT, b. Medium EmpathGPT, c. Large EmpathGPT, d. Small DialoGPT, e. Medium DialoGPT and f. Large DialoGPT (Table 4.3 shows few samples of the commonsense questions and answers generated by corresponding models).

Although they are common knowledge questions, the answers may not be fixed in an open dialogue and users may have different opinions on the accuracy of the answers. We decided to invite annotators to grade the collected responses. The experiments were designed as follows:

- The answers generated by the 6 models to all questions were collected and organized into a questionnaire. The questionnaire contains only the questions as well as the answers, the models corresponding to the answers are not visible.
- 10 annotators participated according to the following scoring criteria:
 - 0 → If you think the answer is totally wrong;
 - 1 → If you think this answer contains some correct and relevant information but not perfect;
 - 2 → If you think the answer is completely correct.
- We first calculated the average score from the 10 annotators for each answer, and then summed the average score for each model for each answer for the 28 questions as the final scores.
- Finally, we normalized all scores as proportions of the maximum score (56) that a model could receive.

Table 4.4 shows the manual scoring results for all samples. And the Figure 4.1 shows the results plotted in ascending order of score.

Model	Overall score	Normalized score
Small EmpathGPT	13.30	0.238
Medium EmpathGPT	13.90	0.248
Large EmpathGPT	19.00	0.339
Small DialoGPT	7.10	0.127
Medium DialoGPT	38.60	0.689
Large DialoGPT	35.30	0.630

Table 4.4: Manual scoring results for commonsense test.

We came to the following findings from the experimental results:

1. The small DialoGPT performs the worst, while the medium-sized DialoGPT performs the best. It indicates that for DialoGPT, model size is not always a predictor of performance.
2. The performance of the medium and large DialoGPT is similar. Although the medium DialoGPT scored higher than the large DialoGPT, the performance difference between these two models was not large. This suggests that the DialoGPT model may not significantly improve its performance on this task after increasing its capacity to a certain level.
3. For EmpathGPT, small and medium-sized models performed comparably, with the best performance being for large EmpathGPT.

4. For the same size of DialoGPT and EmpathGPT, only the small model learns more information after fine-tuning, and the small EmpathGPT score is almost twice as high as the small DialoGPT, while the results are just the opposite for the medium and large models.
5. The medium DialoGPT achieves the highest score among all models, indicating that it is the best model for answering commonsense questions in this experiment. We infer that it is possible that the medium-sized model may also perform best in the following open-domain dialogue experiment.
6. The total score of the three EmpathGPT was 46.20, while the total score of DialoGPT was 81.00, almost twice as high as that of EmpathGPT. This shows that the fine-tuned model loses much of the information learned during the pre-training process. This phenomenon indicated that fine-tuning may have led to catastrophic forgetting.

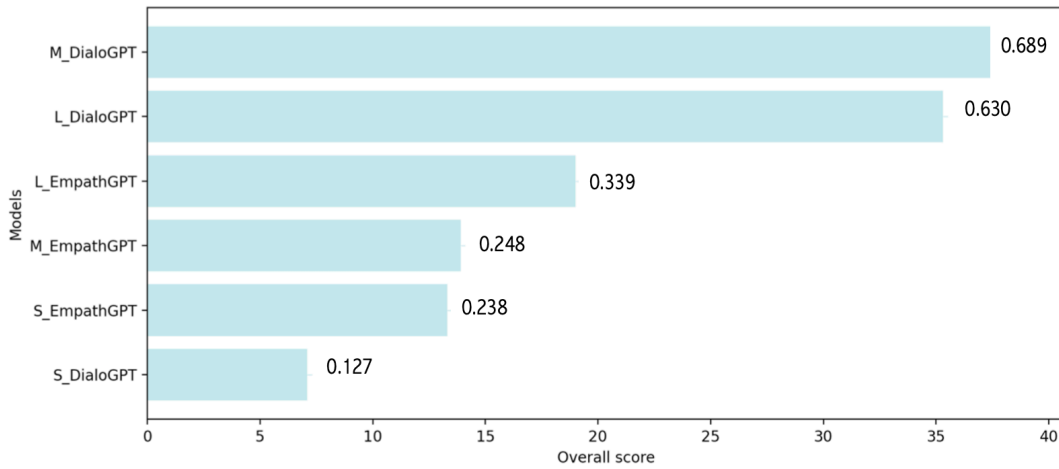


Figure 4.1: Manual evaluation of the commonsense Test (normalized scores on the right).

Catastrophic forgetting was first introduced as the phenomenon that the learning of new information by a fully connected neural network causes the rapid destruction of prior knowledge [McCloskey and Cohen, 1989]. In neural networks, it can occur especially when the model has been trained on various tasks, the weights that are specific to one task need to be adapted to the purposes of another task [Kirkpatrick et al., 2017]. Fine-tuning deep pre-trained language models can easily suffer from catastrophic forgetting during transfer learning [Chen et al., 2020], then new knowledge is learned at the expense of old information. Since DialoGPT is a large language model that has been trained on a large dataset, when fine-tuning it on this specific task, the model is forced to learn patterns and features of empathic conversations, which may overwrite or interfere with the generic knowledge it previously learned. As a consequence, the model may forget much of its general knowledge while improving its ability to generate more empathic responses.

4.3. Hyperparameter Optimization

To mitigate catastrophic forgetting, we performed hyperparameter optimization. Hyperparameter optimization is a critical step in the machine learning pipeline as it can significantly impact the performance of a model. We first fine-tuned the small DialoGPT using different learning rates, other parameters were configured the same as in the previous training when using $5 * 10^{-5}$ as the learning rate. Then

using the found optimal learning rate, we tried training more epochs to observe the most appropriate time to stop training.

4.3.1. Tuning the learning Rate

Since computational resources are limited, yet training DialoGPT requires expensive computational resources, we first conducted experiments tuning the learning rate on the small DialoGPT. The optimization of the learning rate is an effective way to diminish catastrophic forgetting. The learning rate controls how quickly the model adapts to the task during each training step. A high learning rate can lead to catastrophic forgetting as the parameters change too quickly, whereas a low learning rate can result in slow convergence and poor performance. The optimal learning rate can find the right balance between the speed of convergence and the risk of catastrophic forgetting. Based on the experimental results as shown on the Table 4.5, we obtained the following observations and conclusions:

- The learning rate has a significant impact on the performance of the model. The perplexity can be significantly varied depending on the learning rate used.
- The perplexity increases as the learning rate decreases from $5 * 10^{-4}$. It leads to underfitting when the learning rate is too low to allow the model to converge. However, when the learning rate increases from $5 * 10^{-4}$, the perplexity also increases significantly leading to an overshoot of the optimal point.
- The smallest perplexity of 1.2 is obtained when using $5 * 10^{-4}$ as the learning rate, indicating that it is the optimal learning rate for the small DialoGPT training on this empathetic dataset.

Learning rate	Perplexity
$5 * 10^{-3}$	20.7
$5 * 10^{-4}$	1.2
$5 * 10^{-5}$	4.0
$5 * 10^{-6}$	7.4
$5 * 10^{-7}$	17.0
$5 * 10^{-8}$	42.0

Table 4.5: Perplexities of fine-tuning the small DialoGPT using various learning rate.

As explained in the methodology section 3, perplexity cannot be used perfectly as a metric to evaluate how well conversation agents actually perform. Therefore, we choose a model trained using $5 * 10^{-3}$ as the learning rate results in high perplexity (20.7) for a sanity check. We chose the first three questions in the common sense question experiment to ask the model trained with $lr = 5 * 10^{-3}$, and compared its responses with the results of the previous experiment, and the results are shown in the Table 4.6. Empirically, those answers from the model trained using $lr = 5 * 10^{-3}$ seem not to make sense at all. This shows that perplexity can be used as one of the criteria to assess the model, and this check confirms that tuning hyperparameters can improve the performance of the model.

Table 4.6: Comparison of common sense questions tested on two Small EmpathGPT models.

Question	S_EmpathGPT ($lr = 5 * 10^{-3}$)	S_EmpathGPT($lr = 5 * 10^{-5}$)
Which season is snowy?	I am	The winter storm that comes every year.
Is the moon bigger or the sun?	I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_ I'm sorry_	Both! It's a very complex shape with a heated retina of some sort.
Where is Amsterdam?	I am sorry you_ I'm sorry you did you did I am sorry you did you did you did you did I am sorry you did you did you did you did you did you did	It's in the Netherlands. you should check it out!

With the above experiments we found the optimal learning rate of $5 * 10^{-4}$, the model should be able to generalize better to unseen data and perform better using this learning rate. By using a learning rate of $5 * 10^{-4}$, an additional benefit is that we can expect the model to converge faster since larger learning rates allow the model to take bigger steps toward the optimal point, which leads to faster convergence.

4.3.2. Tuning the epochs

Finding the optimal number of epochs is an important aspect of fine-tuning a model to prevent overfitting. If the number of epochs is too high, the model may overfit the training data, however, if the number of epochs is too low, the training may not be sufficient to learn well from the dataset. Moreover, fine-tuning a large language model like DialoGPT requires significant computational resources, and setting too high the number of epochs may lead to a waste of resources, whereas setting too low may lead to sub-optimal performance. Besides computational resources, training such language models is also very time-consuming. Finding the optimal number of epochs can help reduce the time required to train the model and still achieve the desired performance level. Therefore, we conducted experiments with a small DialoGPT to fine-tune it 5 epochs using the learning rate found above, the other parameters remain the same as the previous settings. As the results are shown in the Figure 4.2, the overall perplexity decreases as the number of iterations increases. First, the perplexity decreases at a faster pace, after the second epoch of training, the decrease slows down. After the third epoch, the perplexity rises very slightly and then falls slowly, finally levelling off.

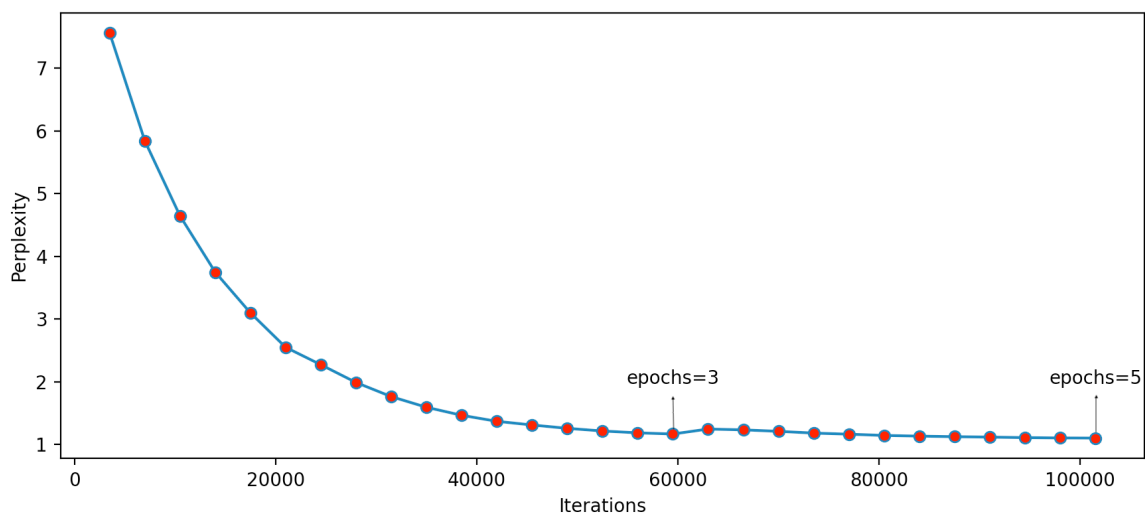


Figure 4.2: Perplexity training Small DialoGPT 5 epochs.

Based on the given results, the optimal number of epochs for the model could be between 3 and 5 for this task. We finally decided to use 3 as the number of epochs based on the following reasons:

- The perplexity values show that the model's performance constantly improves as the number of iterations increases, but the pace of improvement greatly slows down after the third epoch. This suggests that training the model with additional epochs may not significantly improve the performance.
- Continuing to train the model for more epochs may increase the risk of overfitting the model to the dataset. 3 epochs could be a reasonable stopping point to avoid overfitting, as shown on the plot that the perplexity has plateaued or even started to gently decrease a little.
- Training a model for more epochs is computationally expensive and time-consuming. Stopping training after 3 epochs can save time and resources without sacrificing too much on performance.

Based on the analysis above, training the model for 3 epochs may be the best choice to balance the performance of the model, computational resources and potential overfitting risk. With the optimal learning rate and epoch found, we fine-tune the Medium DialoGPT (see Figure 4.3) and the Large DialoGPT (see Figure 4.4) with the new parameter configuration. The plots show the decrease of perplexity at different iterations during 3 epochs of training. The final results of the perplexity of training all models can be found in Table 4.7. Compared with the previous experiment results shown in Table 4.2, we can see that the perplexity of each model has decreased, indicating that the model performance has been improved by tuning the hyperparameters. We finally obtained three empathetic conversational agents, named Small EmpathGPT, Medium EmpathGPT and Large EmpathGPT.

Model	Small EmpathGPT	Medium EmpathGPT	Large EmpathGPT
Perplexity	1.2356	1.1016	1.0995

Table 4.7: Perplexity of $5 * 10^{-4}$ within 3 epochs.

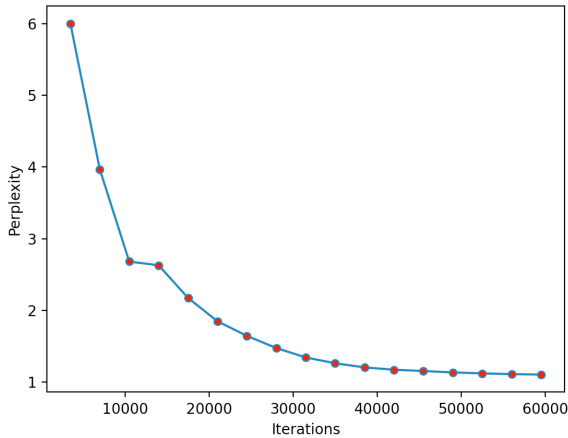


Figure 4.3: Perplexity within 3 epochs for Medium DialoGPT.

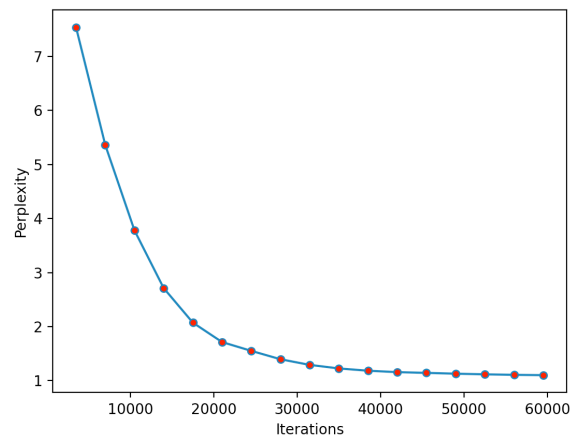


Figure 4.4: Perplexity within 3 epochs for Large DialoGPT.

4.4. Quality Evaluation

In this experiment, annotators rated the performance of 12 dialogue topics chatted with Large EmpathGPT and Large DialoGPT in terms of relevance and empathy, respectively (24 samples in total).

All samples can be found in Appendix - B. Figure 4.5 shows part of the interface of one dialogue sample for quality evaluation.

Dialogue 1/12

👤: My cat ate rodenticides coincidentally and passed away. It has left me feeling overwhelmed with sorrow. The thought of never being able to hold, play and see her lovely face again is heartbreaking.

🤖: I'm so sorry you had to go through that. My girlfriend's cat is going through the same thing right now. It's hard when they say goodbye.

👤: Losing a furry friend is never easy...

🤖: I totally understand that. I'm so sorry you're having to go through that. I hope you can find comfort during this difficult time.

1. How **relevant** are those responses?

1 2 3

low ☐ ☐ ☐ high

2. How **empathetic** are those responses?

1 2 3

low ☐ ☐ ☐ high

Figure 4.5: Interface of one dialogue sample for quality evaluation

- **Relevance** Relevance is rated on a scale of 1 to 3 according to the level of relevance, with the following criteria:

1. Low. It shows a lack of understanding of the context of the user inputs.
2. Medium. It is somewhat relevant but the responses are vague and too general.
3. High. It shows a deep understanding of the context and gives specific responses.

Sentiment	Relevance	
	EmpathGPT	DialoGPT
Negative	769	473
Positive	699	537
Total score	1468	1010
Average score	2.45	1.68

Table 4.8: Relevance Scores for EmpathGPT and DialoGPT.

- **Empathy** Empathy is rated on a scale of 1 to 3 according to the level of empathy, with the following criteria:

1. Low. It may be insensitive or dismissive, or even make the user feel worse.
2. Medium. It is somewhat empathetic but too generic to respond to the semantic context of user input.
3. High. It provides compassionate responses that are engaging and supportive.

Each conversation is graded on a scale from 1 to 3 by a total of 50 annotators for relevance and empathy respectively. First, the total score for each dialogue was calculated, corresponding to the model with positive or negative semantics, and we calculated the score of each model with respect to the different

Sentiment	Empathy	
	EmpathGPT	DialoGPT
Negative	760	462
Positive	723	502
Total score	1483	964
Average score	2.47	1.61

Table 4.9: Empathy Scores for EmpathGPT and DialoGPT.

semantics. Then, the **total score** of each model was calculated. Finally, we calculated the **average score** obtained for each model: We calculated the average score for each dialogue, then summed the average scores of all the dialogues and divided by 12 (there are 12 dialogues in total) to get the final average score of the model.

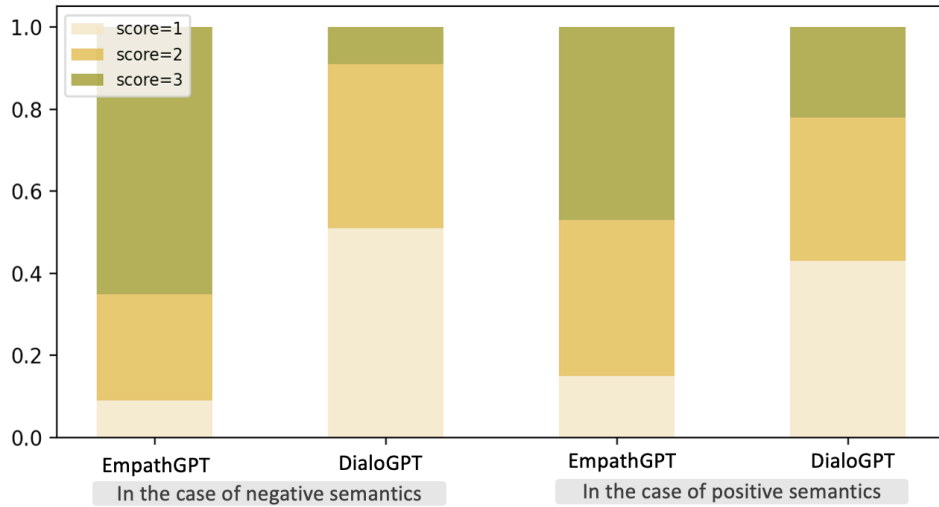


Figure 4.6: Distribution of average votes for each score for relevance

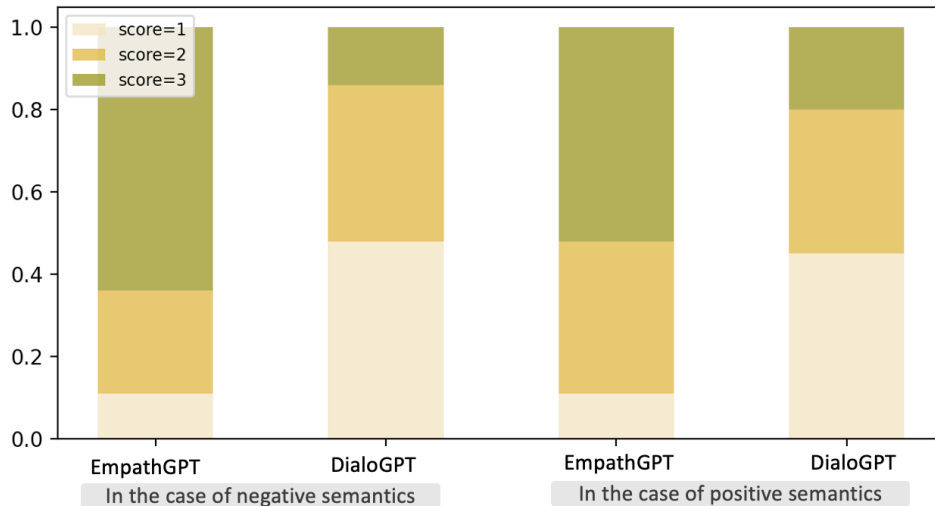


Figure 4.7: Distribution of average votes for each score for empathy

Table 4.8 and Table 4.9 show all scores each model obtained respectively. To see the distribution of annotators' votes for different dialogues, we calculated the distribution of scores for each sample and

then calculated the average percentage for each model, the results are shown in Figure 4.6 and Figure 4.7 regarding the relevance and empathy. By observing the results of the experiments, we came to the following conclusions:

- In terms of relevance, EmpathGPT scored better than DialoGPT for both positive and negative semantics, indicating that EmpathGPT can generate more relevant responses to user input.
- For empathy, EmpathGPT scored higher than DialoGPT with positive and negative semantics, demonstrating that the responses generated by EmpathGPT are perceived as more empathetic.
- EmpathGPT received higher scores than DialoGPT for both relevance and empathy. EmpathGPT appears to have generated responses that are more relevant and empathetic than DialoGPT.
- The gap between the scores of the two models is especially large in the aspect of empathy. It is more distinct in negative semantics. From this, it can be inferred that EmpathGPT performs particularly well in negative dialogue scenarios.
- According to the grading criteria: The average scores of the two models demonstrate that EmpathGPT can generate highly relevant and highly empathetic responses. Meanwhile, DialoGPT can generate moderate and generic responses.
- The empathy and relevance scores of two models were found to be highly correlated, therefore, we use the Pearson correlation coefficient (see Equation 4.1) which measures the strength of the correlation between two variables. We obtained 0.94 as a result of the calculation that shows a very strong positive linear relationship between relevance and empathy. It shows that people judge chatbots to be empathetic when their responses are relevant. This observation also proves the hypothesis made at the beginning of this chapter: when responses are irrelevant, people will judge robots as lacking empathy for not being understood.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4.1)$$

4.5. Preference Evaluation

The sample for the preference evaluation consisted of a total of 12 single-turn dialogues, and the samples were equally divided into 6 negative (see Table 4.10) and 6 positive scenarios (see Table 4.11). By providing each model with the same user input and collecting feedback on the responses it generates, the experiment allows for an objective comparison of each model's performance and identifies the model that performs best in terms of user preferences. Without knowing which model generated the responses and with the order of responses all disordered, annotators selected the responses they thought were most appropriate and preferred in each sample (selection of multiple responses was allowed).

Table 4.10: Two examples of single-turn dialogue samples in negative scenario

Input	Response	Model
1. I accidentally walked into a meeting with the CEO without zipping up my jeans, which was pretty embarrassing. Do you think everyone was sneakily laughing at me?	Wow that's really embarrassing! I would've been mortified! Was there a lot of people around? I've done that too. Hopefully people laughed at you.	Small EmpathGPT
	Oh no. that's not good. I'm sorry that happened.	Medium EmpathGPT
	Oh man that sounds horrible. I would've been so embarrassed.	Large EmpathGPT
	I'm not sure if you're joking or not, but I'm pretty sure you're not.	Small DialoGPT
	I think you're a good person.	Medium DialoGPT
	I think you were just being silly.	Large DialoGPT
2. I was so wasted last Friday and completely forgot it's my girlfriend's birthday.	No way! That's terrible. Did she say anything?	Small EmpathGPT
	No way! I'd be so upset.... how did she react?	Medium EmpathGPT
	Oh no! I hope you guys were able to discuss about what happened and how it made you feel.	Large EmpathGPT
	Happy Birthday!	Small DialoGPT
	Happy Birthday!	Medium DialoGPT
	Happy Birthday!	Large DialoGPT

Table 4.11: Two examples of single-turn dialogue samples in positive scenario

Input	Response	Model
1. I found a \$20 bill on the street and decided to treat myself to something nice, feeling grateful for the unexpected surprise.	That is so sweet.	Small EmpathGPT
	That is very lucky! I hope you can find some more \$20 like you did.	Medium EmpathGPT
	That is awesome. What did you do with it?	Large EmpathGPT
	I'm so glad I'm not the only one who does this.	Small DialoGPT
	I'm glad you found something nice!	Medium DialoGPT
	I found a 20 bill on the street and decided to treat myself to something nice, feeling grateful for the unexpected surprise.gif	Large DialoGPT
2. I stared in awe at her painting. She had poured her heart and soul into the piece and it had paid off.	That's awesome. I wish I could have a daughter that sweet.	Small EmpathGPT
	That's amazing. Maybe she's a natural artist!	Medium EmpathGPT
	That's amazing. I would love to see something like that	Large EmpathGPT
	I love how she's like, I'm so proud of you.	Small DialoGPT
	I'm glad you enjoyed it!	Medium DialoGPT
	I'm glad you liked it.	Large DialoGPT

Once a model's response is selected, the model gets a score of 1. We first calculated the performance of each model for positive and negative semantic samples with respect to the total performance, and then performed normalization to compare and analyze (All samples can be found in Appendix - C). The experimental results are shown in Table 4.12. The percentage of the score is illustrated in Figure 4.8

Model	Negative	Positive	Sum	Percentage of votes
Small EmpathGPT	153	24	177	0.198
Medium EmpathGPT	127	105	232	0.259
Large EmpathGPT	73	101	174	0.194
Total	353	230	583	0.651
Small DialoGPT	34	27	61	0.068
Medium DialoGPT	40	91	131	0.146
Large DialoGPT	56	64	120	0.134
Total	130	182	312	0.349

Table 4.12: Preference Scores for EmpathGPT and DialoGPT.

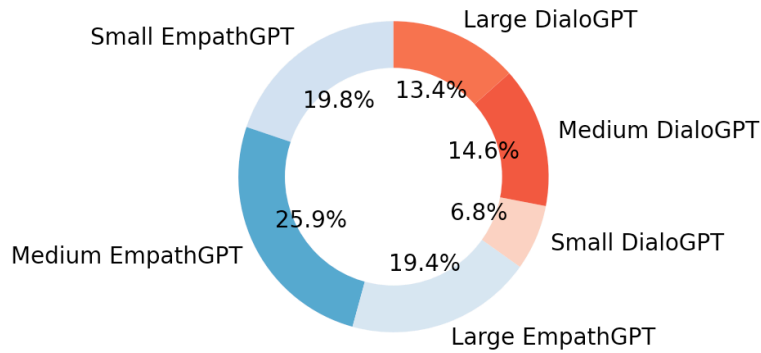


Figure 4.8: Illustration of all model scores in terms of percentages

From the results of the manual evaluation, it can be observed that:

- Among the DialoGPT models:
 - In negative scenarios, the larger the model, the better the performance. The Large model received the highest score of 56 in the negative sentiment category, indicating that it performed the best in handling negative scenarios.
 - With positive scenarios, the small model performs the worst, the medium model achieves the best results and its score of 91 that stands out compared to small and large models, indicating that it performed the best in handling positive scenarios.

Overall, the small models performed the worst, the medium models performed the best, followed by the large models. However, the difference between the scores of medium and large models is not very large. *We found this result also reflects on the performance of the DialoGPT models in the common sense test (see Table 4.4).*

- Among the EmpathGPT models:
 - In negative scenarios, the small version received the highest score of 153, indicating that it was most likeable for its performance in handling negative situations. Yet the responses given by the large model are the least preferred.
 - In the positive dialogue scenario, the medium model received the highest score of 105, indicating that it was the most preferred when dealing with positive scenarios. The performance of the large model is comparable to that of the medium model. Meanwhile, the small model

was the least favoured, receiving only 24 points, similar to the performance of the small DialoGPT before fine-tuning.

Overall, a very interesting finding is that the small model performs much better than the large model when the dialogue scenario is positive, and the results are completely opposite when the dialogue semantics become negative. This resulted in a comparable sum score for the small model and the large model, however, the medium-sized model was the most preferred model due to its favourable responses in both positive and negative scenarios.

- Comparing the general performance of DialoGPT with EmpathGPT:
 - In general, EmpathGPT received a total score of 583, which is much higher than and almost doubled the total score of 312 obtained by DialoGPT. This proves that EmpathGPT can generate more appropriate and user-friendly responses.
 - Under negative semantics, EmpathGPT models scored 223 points higher than DialoGPT models, and under positive semantics, EmpathGPT models scored 48 points higher than DialoGPT models. It shows a greater preference for empathetic responses, especially in negative conversational contexts.
 - The medium models of both EmpathGPT and DialoGPT received the highest overall scores. This suggests that the medium variant may be the most effective model.
 - Small models score almost three times as much as they did before fine-tuning, and almost twice as much for medium models, and fine-tuning improves relatively less for large models. Fine-tuning is most helpful for small and medium models

However, it is important to note that the small EmpathGPT obtained slightly higher sum scores than medium EmpathGPT which is the best-performed model in the negative scenario. Thus we browsed through the negative dialogue samples and found a potential reason that could lead to this result. Table 4.13 shows the sample of possible problematic manual assessments. The user input is clearly an embarrassingly negative scenario, and Small EmpathGPT gave a long response. The semantics of most parts of the response is appropriate and empathetic until the last sentence 'Hopefully people laughed at you' which is completely inappropriate and without empathy. Since this reply still got 13 points, we presume that some annotators saw enough proper content and defaulted to the whole response is appropriate. If the 13 points are not counted, Small EmpathGPT still performs best in negative semantics, but it no longer slightly outperforms Large EmpathGPT in terms of sum scores. Because bias is inevitable in manual experiments, and experimental data cannot be modified, we cannot conclude that Small EmpathGPT and Large EmpathGPT perform comparably as one of the conclusions. Nevertheless, no doubt that the medium EmpathGPT is the most preferred of all models.

Input	Response	Model	Score
I accidentally walked into a meeting with the CEO without zipping up my jeans, which was pretty embarrassing. Do you think everyone was sneakily laughing at me?	Wow that's really embarrassing! I would've been mortified! Was there a lot of people around? I've done that too. Hopefully people laughed at you.	Small EmpathGPT	13
	Oh no. that's not good. I'm sorry that happened.	Medium EmpathGPT	19
	Oh man that sounds horrible. I would've been so embarrassed.	Large EmpathGPT	23
	I'm not sure if you're joking or not, but I'm pretty sure you're not.	Small DialogPT	2
	I think you're a good person.	Medium DialogPT	4
	I think you were just being silly.	Large DialogPT	2

Table 4.13: The sample that may get problematic assessment during the manual evaluation.

5

Discussion

In this chapter, we will first summarize user expectations and suggest attributes of a conversational agent that would improve the user experience, concluded from the further study of the human evaluation results supported by the findings from the socio-behavioural perspective. Additionally, we critically evaluate the limitations of the project for potential improvement of future research.

5.1. Final results enhanced by social-behavioural studies

The final human evaluation yielded the expected result that the performance of the fine-tuned EmpathGPT models has improved significantly compared to the DialoGPT models. And also some interesting findings that went beyond what was expected, for example, the performance of the small EmpathGPT in manual evaluation does not lag far behind the large EmpathGPT in specific scenarios, this may be due to the fact that the dataset is not large enough to match the size of large EmpathGPT. Moreover, to ground our findings from the experiments, we performed further analysis in terms of the manual evaluation results in the context of social behaviour research, which can provide important insights into user needs and preferences. This analysis grounds the experiment results obtained and could ultimately be helpful to the development of more effective and user-friendly conversational agent products from the perspective of user needs.

1. **Users appreciate chatbots being empathetic** From the results of the quality and preference evaluations, the EmpathGPT models are able to give more relevant and empathetic responses than the DialoGPT models, and these responses are more appreciated by users. This demonstrates that users like conversational agents to be empathetic, as previously suggested by Svikhnushina and Pu [2020] that incorporating empathy into chatbot interactions is a critical factor in improving user engagement. This research indicates that users perceive empathetic chatbots, viewed as a subset of emotional intelligence, as more pleasing, likeable, and attentive to detail.
2. **User expectations differ significantly across sentimental contexts** As observed from the results of the manual evaluation of preference, the scores of different models differ significantly across sentimental contexts. It can be explained by considering that empathy exhibited in the context of negative sentiments (termed negative empathy) differs from empathy in the context of

positive sentiments (termed positive empathy) [Andreychik and Migliaccio, 2015]. This previous study on negative versus positive empathy emphasizes that these two categories of empathy are quite distinct and involve different capacities. It can be concluded that empathy under positive and negative sentiment should involve different technical approaches to better connect with users in order to improve the user experience.

3. **Users prefer greater empathy, especially in the negative sentiment context.** EmpathGPT models outperformed DialoGPT in preference evaluation in both positive and negative semantics. Nevertheless, in negative semantics, annotators' preference for empathetic responses was particularly substantial. The finding that users are more in need of empathy in negative semantics corresponds to the interpretation of Klein et al. [2001]: Showing empathy when in an emotionally negative scenario can create an emotional connection with the person involved, whereby their feelings are recognized and acknowledged. In the negative sentimental context, incorporating more empathy for conversational agents can create a bond with users and help them feel heard and validated.
4. **With positive scenarios, simply promoting the user's happiness is more appreciated.** Positive empathy (or empathy in positive scenarios) is a somewhat less researched concept than empathy in negative scenarios [Morelli et al., 2015]. However, by analyzing the positive dialogue scenarios where DialoGPT scored higher than EmpathGPT, such as the samples shown in Table 5.1. We can see that people tend to prefer the response just simply sharing the happiness, which is aligned with the finding from Svikhnushina and Pu [2020] which shows that for positive sentiments, responding to users simply by sharing or promoting their happiness is much more appreciated.

5.2. limitations

In addition to the findings, it is important to acknowledge the limitations of this study for further exploration. In this section, we will discuss several key limitations that may have impacted on the scope and validity of this research in terms of dataset, hyperparameters, and evaluation metrics:

Dataset

From the results of the human preference ratings, we can see that the difference in scores between the EmpathGPT models of different sizes is not as great as for the DialoGPT models, where the medium and large DialoGPT models are much more preferred than the small one. Some of the possible explanations for this result are:

- The DialoGPT model was pre-trained using dialogue data from Reddit, which is rich in topics, while the EmpathGPT model was fine-tuned using only a single empathy dataset. As a result, the EmpathGPT models are all likely to be more being more specialised in expressing empathy in a certain pattern and thus exhibit less variability in performance across different sizes of the model.
- The size and diversity of the dataset used to fine-tune the language model can also affect the ability of the model. It may also be that the size of the EmpatheticDialogues dataset is not large enough and this limits the ability of larger models.

Input	Response	Model	Score
I graduated with honours! Just feel so grateful for the support of my family and friends and excited for the next chapter of my life.	That is great. It is so nice to focus on the positives.	Small EmpathGPT	7
	You must have great friends. I bet they are thrilled.	Medium EmpathGPT	
	that is great. it is wonderful to not lose sight of what we have now	Large EmpathGPT	
	I'm so glad you're doing well!!!	Small DialoGPT	70
	Congrats! I'm so happy for you!	Medium DialoGPT	
	Congrats! I'm so glad you're graduating!	Large DialoGPT	
I start practising boxing lately. Actually punching sandbags is very therapeutic and tension relieving.	Wow. I really don't know about that. I actually hit a bird once and almost died! It was crazy to see!	Small EmpathGPT	22
	Wow. that sounds very painful. I am glad you feel better about yourself.	Medium EmpathGPT	
	That sounds like a cool goal to me. I've always wanted to be able to learn to play.	Large EmpathGPT	
	I'm a boxing fan, but I'm also a huge boxing fan. I'm also a huge boxing fan.	Small DialoGPT	42
	I'm going to try this.	Medium DialoGPT	
	I do this too. I also do a lot of pushups and sit ups.	Large DialoGPT	

Table 5.1: Samples of DialoGPT that scores higher than EmpathGPT.

Hyperparameters for optimization

From the manual evaluation of the initial common sense test, we identified the occurrence of catastrophic forgetting and thus performed the hyperparameter optimization. Two critical hyperparameters, learning rate and training epochs, were selected for tuning. However, apart from learning rate and training epochs, the following hyperparameters are also important for the optimization and we did not optimize these:

- *Batch size*: Choosing an appropriate batch size for training is crucial as it has a big impact on both performance and training time. The optimal batch size depends on the size of the dataset, the available memory and the complexity of the model. The well-chosen batch size should be able to well balance the speed of the training and avoid resulting in overfitting.
- *Maximum sequence length*: This parameter defined the maximum length of the input sequence. The model can capture longer dependencies with bigger maximum sequence length with bigger computational costs and memory requirements.
- *Gradient accumulation steps*: This parameter affects the update frequency of weights. Increasing the gradient accumulation steps enable the training process to be more stable but with longer training time.
- *Warmup steps*: The use of warm-up steps could allow the model to gradually adjust to the learning rate, avoiding drastic changes that could negatively impact the training process.

- *Adam epsilon*: It is the epsilon value that controls the numerical stability of the AdamW optimizer. Lower values of Adam epsilon can result in better convergence, though it may also make the optimization process more sensitive to the noise in the gradients.

Due to limited computational resources, it was not feasible to optimize all of the hyperparameters. It is also important to consider that DialoGPT is a very large-scale language model, which can require substantial computational resources to train. Therefore, in order to balance computational efficiency with achieving satisfactory performance, we had to select learning rate and training epochs as the subset for optimization based on their perceived importance and potential impact on the model's performance. However, the performance of the model can potentially be improved by selecting more hyperparameters to tune.

Evaluation metrics

Automatic evaluation Perplexity, as a very popular and proven effective metric for evaluating language models, has shown its drawbacks. As its theoretical validity is demonstrated by the derivation of the information entropy formulation, we have found another theoretical deficiency of it: In the calculation of cross-entropy, it assumes that each word in the sentence is equally important, which may not always be the case. Because it is a common pattern observed in textual data where a small number of common words are used frequently, while the majority of words are used rarely. This is commonly known as Zipf's Law [Zipf, 2016]. Sometimes the context of certain words is more informative than others, but the significance of these words may not be captured by perplexity alone. Overall, by comparing the results of perplexity and manual evaluation, perplexity is shown to be effective at some level. However, it should be used along with other metrics especially human assessments to provide a more comprehensive evaluation of a language model's performance.

Human evaluation When designing the manual experiment, we tried to ensure that all steps were systematically designed. For example, the dialogues were collected from different people and the samples were randomly selected to ensure diversity. We invited as many annotators as possible to reduce individual bias and increase the generality of the results. Randomizing the order of all samples and not giving the annotators any information about the source or context of each dialogue helped to reduce bias from potential distortions related to the source of the dialogue.

However, even though annotators were provided with a clear guideline, we can see from the evaluation results that human evaluation can be subjective and not fully careful. For example, as the sample shown in Table 4.13 in chapter 4, the answer given by the small EmpathGPT is very inappropriate at the end regarding the grading criteria, yet it was voted by 13 annotators. Moreover, different people may have different opinions and perspectives on what empathic dialogue constitutes, which may lead to inconsistent or inaccurate evaluations. Therefore, manual evaluation in such cases has some shortcomings.

A small sample size may also be a limitation. If the sample size is too small, it may not be representative of the broader topics and the results may not be generalizable or reliable. Human evaluation can be time-consuming and costly, but using a large sample size with a large number of annotators involved can definitely help to get more reliable results.

In addition, the high correlation between relevance and empathy from the quality test results suggests that in real life users may indeed perceive a bot as lacking empathy because it does not give relevant

responses and vice versa. But it may also be because the metric for assessing empathy is too basic, as empathy has a complex psychological concept, adopting a quantitative metric of empathy from psychological science could possibly make the evaluation of empathy more scientific.

6

Conclusion

In this project, we aimed to give empathy to an open-domain dialogue agent and investigate the effectiveness of perplexity used as an automatic evaluation metric for conversational agents. We then further performed the human evaluation in order to suggest that empathetic properties of chatbots might potentially improve the user experience supported by some socio-behavioural study findings.

We fine-tuned the large-scale pre-trained model DialoGPT in small, medium and large sizes using the EmpatheticDialogue dataset after the investigation for the approach that would be both efficient and effective. The combination of DialoGPT's natural language generation capabilities and the empathy-oriented dataset with a wide variety of emotional diversity enables a chatbot to generate more human-like and empathetic responses. We conducted initial experiments with the three different sizes of EmpathGPT, and invited annotators to perform initial assessments of the ability of these models in answering common sense questions. In order to obtain better-performing models, we then performed hyperparameter optimization, tuning the learning rate and the training epoch to obtain parameter settings that would give better performance evaluated by perplexity. We then conducted quality evaluation experiments with the large EmpathGPT and the large DialoGPT to assess the performance of the models in terms of relevance and empathy by grading the sample generated by those two models. The preference evaluation experiment was then conducted using small, medium and large DialoGPTs versus small, medium and large EmpathGPTs with the benefit of objective comparison. Finally, the results of the experiments were analyzed and some of the key findings were supported by a socio-behavioural perspective.

Based on our experimental results, the main research questions can be summarized as follows:

- *To what extent can fine-tuning DialoGPT on a specific empathetic dataset influence its ability to answer common sense questions and improve the relevance and empathy of its generated responses?*

Through initial manual evaluation of the initial experiment, we identified catastrophic forgetting that occurred after fine-tuning the DialoGPT indicates that the model had forgotten something it already learnt. We then tried to solve this problem by hyperparameter optimization and got the models with lower perplexity, which should lead to some improvements. However, as the focus of this project was on empathy, and common sense questions do not usually involve scenarios that

require empathy, we conducted a comprehensive manual evaluation. The results of the automatic evaluation and the quality test of manual evaluations confirm that fine-tuning the DialoGPT on the EmpatheticDialogues dataset significantly improves the quality in terms of relevance and empathy.

- *What are the pros and cons of using perplexity as an automatic evaluation metric for conversational agents, and to what degree can it be considered a viable alternative to manual evaluation?*
By investigating the effectiveness of evaluation metrics, we found that perplexity can be considered an efficient quantitative measure of how well the model can produce human-like responses, and it facilitates further optimization and initial comparison between different models. In general, perplexity shows a certain ability to measure the basic quality of the generated responses, but it does not necessarily reflect the appropriateness of the semantics, so it is not an optimal alternative to human evaluation, especially when evaluating specific properties of chatbots such as being empathetic.
- *How much do potential users appreciate that the chatbot is empathetic, and what kind of scenario might users be in need of even more empathy?*

By using real input collected from annotators, the experimental design improves the validity of the results, and the manual evaluation can provide results that are more likely to reflect user expectations in the real world. The preference test experiment shows that empathy-enhanced chatbots are more preferred by users. Furthermore, user expectations may differ depending on different sentiment scenarios, particularly users would like to receive more empathetic responses in the negative scenario, however, simply mirroring or promoting users' happiness would be more appreciated when the scenario is positive. Our findings are consistent with those of relevant socio-behavioural studies.

Our analysis of positive and negative scenarios when chatting with the model allows us to provide suggestions for improving user satisfaction and engagement by identifying the chatbots that perform best under different conditions in terms of user preferences.

However, it is important to acknowledge the limitations of this study. As discussed in the previous chapter, we found several major limitations that suggest opportunities for future research. For instance, if sufficient computational resources are available, the performance of the model can be improved by optimizing more hyperparameters, in this way, the problem of catastrophic forgetting revealed by the common sense test should be solved more effectively. More comprehensive and representative evaluation results can be achieved by increasing the number of samples and involving more annotators. When it comes to assessing empathy, the adoption of psychological metrics could make experimental designs and conclusions more scientific. Moreover, chatGPT's RLHF (Reinforcement Learning from Human Feedback) approach, which incorporates human feedback into the reward function to improve the responses, can be a technique to adopt for future improvement of our current work.

Bibliography

- Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. Towards a human-like open-domain chatbot, 2020. URL <https://arxiv.org/abs/2001.09977>.
- Michael Andreychik and Nicole Migliaccio. Empathizing with others’ pain versus empathizing with others’ joy: Examining the separability of positive and negative empathy and their relation to different types of social behaviors and social emotions. *Basic and Applied Social Psychology*, 37:1–18, 08 2015. doi: 10.1080/01973533.2015.1071256.
- Vevake Balaraman, Seyedmostafa Sheikhalishahi, and Bernardo Magnini. Recent neural methods on dialogue state tracking for task-oriented dialogue systems: A survey. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 239–251, Singapore and Online, July 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.sigdial-1.25>.
- Jackylyn Beredo, Carlo Migel Bautista, Macario Cordel, and Ethel Ong. Generating empathetic responses with a pre-trained conversational model. In Kamil Ekštejn, František Pártl, and Miloslav Konopík, editors, *Text, Speech, and Dialogue*, pages 147–158, Cham, 2021. Springer International Publishing. ISBN 978-3-030-83527-9.
- Basma El Amel Boussaha, Nicolas Hernandez, Christine Jacquin, and Emmanuel Morin. Deep retrieval-based dialogue systems: A short review, 2019. URL <https://arxiv.org/abs/1907.12878>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19, 11 2017. doi: 10.1145/3166054.3166058.
- Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting, 2020.
- Stanley F. Chen, Doug Beeferman, and Roni Rosenfeld. Evaluation metrics for language models. 1998.
- Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

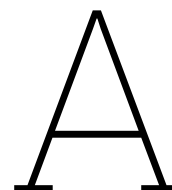
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL <https://arxiv.org/abs/1810.04805>.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents, 2018a. URL <https://arxiv.org/abs/1811.01241>.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018b.
- Sarah E. Finch and Jinho D. Choi. Towards unified dialogue system evaluation: A comprehensive analysis of current evaluation protocols, 2020. URL <https://arxiv.org/abs/2006.06110>.
- Tingchen Fu, Shen Gao, Xueliang Zhao, Ji-rong Wen, and Rui Yan. Learning towards conversational ai: A survey. *AI Open*, 3, 02 2022a. doi: 10.1016/j.aiopen.2022.02.001.
- Tingchen Fu, Shen Gao, Xueliang Zhao, Ji-rong Wen, and Rui Yan. Learning towards conversational ai: A survey. *AI Open*, 3, 02 2022b. doi: 10.1016/j.aiopen.2022.02.001.
- Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Affect-Im: A neural language model for customizable affective text generation. *arXiv preprint arXiv:1704.06851*, 2017.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL <https://arxiv.org/abs/1406.2661>.
- Jonathan Grudin and Richard Jacques. Chatbots, humbots, and the quest for artificial general intelligence. pages 1–11, 04 2019. ISBN 978-1-4503-5970-2. doi: 10.1145/3290605.3300439.
- Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems, 2019. URL <https://arxiv.org/abs/1905.05709>.
- Vladimir Ilievski, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. Goal-oriented chatbot dialog management bootstrapping with transfer learning. pages 4115–4121, 07 2018. doi: 10.24963/ijcai.2018/572.
- F. Ioannidou and Vaya Konstantikaki. Empathy and emotional intelligence: What is it really about. *International Journal of Caring Sciences*, 1, 01 2008.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL <https://arxiv.org/abs/1312.6114>.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.
- J. Klein, Y. Moon, and Rosalind Picard. This computer responds to user frustration:: Theory, design, and results. *Interacting with Computers*, 14:119–140, 11 2001. doi: 10.1016/S0953-5438(01)00053-4.
- Sven Laumer, Fabian Tobias Gubler, Ana Racheva, and Christian Maier. Use cases for conversational agents: An interview-based study. 2019.

- Jingyuan Li and Xiao Sun. A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation, 2018. URL <https://arxiv.org/abs/1806.07000>.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models, 2015. URL <https://arxiv.org/abs/1510.03055>.
- Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation, 2016. URL <https://arxiv.org/abs/1606.01541>.
- Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1230. URL <https://aclanthology.org/D17-1230>.
- Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
- Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An empathetic neural chatbot, 2019. URL <https://arxiv.org/abs/1907.12108>.
- Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1230. URL <https://aclanthology.org/D16-1230>.
- K Liu and Rosalind W Picard. Embedded empathy in continuous, interactive health assessment. In *CHI Workshop on HCI Challenges in Health Assessment*, volume 1, page 3, 2005.
- Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1138. URL <https://aclanthology.org/P18-1138>.
- Samuel Louvan and Bernardo Magnini. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. *CoRR*, abs/2011.00564, 2020. URL <https://arxiv.org/abs/2011.00564>.
- Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, 2015. URL <https://arxiv.org/abs/1506.08909>.
- Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and E. Cambria. A survey on empathetic dialogue systems. *Inf. Fusion*, 64:50–70, 2020.
- Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24:109–165, 1989.

- Chuan Meng, Pengjie Ren, Zhumin Chen, Weiwei Sun, Zhaochun Ren, Zhaopeng Tu, and M. de Rijke. Dukenet: A dual knowledge interaction network for knowledge-grounded conversation. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.
- Sylvia A. Morelli, Matthew D. Lieberman, and Jamil Zaki. The emerging study of positive empathy. *Social and Personality Psychology Compass*, 9:57–68, 2015.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017.
- Helen Ngo, João G. M. Araújo, Jeffrey Hui, and Nicholas Frosst. No news is good news: A critique of the one billion word benchmark. *CoRR*, abs/2110.12609, 2021. URL <https://arxiv.org/abs/2110.12609>.
- Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, pages 1–101, 2022.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
- A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019a.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019b.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. I know the feeling: Learning to converse with empathy. *CoRR*, abs/1811.00207, 2018. URL <http://arxiv.org/abs/1811.00207>.
- Christopher Richardson and Larry Heck. Commonsense reasoning for conversational ai: A survey of the state of the art, 2023. URL <https://arxiv.org/abs/2302.07926>.
- Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.
- Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. *CoRR*, abs/1507.02221, 2015. URL <http://arxiv.org/abs/1507.02221>.
- Melisa Stevanovic and Anssi Peräkylä. Experience sharing, emotional reciprocity, and turn-taking. *Frontiers in Psychology*, 6, 04 2015. doi: 10.3389/fpsyg.2015.00450.

- Ekaterina Svikhnushina and Pearl Pu. Social and emotional etiquette of chatbots: A qualitative approach to understanding user needs and expectations, 2020. URL <https://arxiv.org/abs/2006.13883>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL <https://arxiv.org/abs/1706.03762>.
- Anushree Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, Rahul Goel, and Anirudh Raju. On evaluating and comparing conversational agents. In *NeurIPS 2017*, 2017. URL <https://www.amazon.science/publications/on-evaluating-and-comparing-conversational-agents>.
- Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. Improving knowledge-aware dialogue generation via knowledge base question answering, 2019. URL <https://arxiv.org/abs/1912.07491>.
- Jason D. Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. The dialog state tracking challenge series. *AI Magazine*, 35(4):121–124, Dec 2014. doi: 10.1609/aimag.v35i4.2558.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface’s transformers: State-of-the-art natural language processing, 2019a. URL <https://arxiv.org/abs/1910.03771>.
- Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents, 2019b. URL <https://arxiv.org/abs/1901.08149>.
- Yu Wu, Wei Wu, Zhoujun Li, and Ming Zhou. Response selection with topic clues for retrieval-based chatbots, 2016. URL <https://arxiv.org/abs/1605.00090>.
- Jun Xu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. Conversational graph grounded policy learning for open-domain conversation generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1835–1845, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.166. URL <https://aclanthology.org/2020.acl-main.166>.
- Rui Yan. "chitty-chitty-chat bot": Deep learning for conversational ai. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5520–5526. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/778. URL <https://doi.org/10.24963/ijcai.2018/778>.
- SoYeop Yoo and Ok-Ran Jeong. Ep-bot: Empathetic chatbot using auto-growing knowledge graph. *Computers, Materials & Continua*, 67:2807–2817, 01 2021. doi: 10.32604/cmc.2021.015634.
- Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

- 2204–2213, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL <https://aclanthology.org/P18-1205>.
- Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujuan Li, Chris Brockett, and Bill Dolan. Generating informative and diverse conversational responses via adversarial information maximization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b. URL <https://proceedings.neurips.cc/paper/2018/file/23ce1851341ec1fa9e0c259de10bf87c-Paper.pdf>.
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.30. URL <https://aclanthology.org/2020.acl-demos.30>.
- Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The Design and Implementation of Xiaolce, an Empathetic Social Chatbot. *Computational Linguistics*, 46(1):53–93, 03 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00368. URL https://doi.org/10.1162/coli_a_00368.
- George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.



Commonsense test questionnaire

Commonsense questions and the corresponding answers given by all models:

- a. Small EmpathGPT;
- b. Medium EmpathGPT;
- c. Large EmpathGPT;
- d. Small DialoGPT;
- e. Medium DialoGPT;
- f. Large DialoGPT.

Question	Model	Answer
1. Which season is snowy?	a	The winter storm that comes every year.
	b	S1E10 I think
	c	I think its been years since I've seen the snow. I was in Missouri during the winter and there was no snow.
	d	I'm not sure if it's the season, but I think it's the weather.
	e	Season 4
	f	Season 4
2. Is the moon bigger or the sun?	a	Both! It's a very complex shape with a heated retina of some sort.
	b	It's the sun. The moon is about 1.25 Jupiter sized.
	c	It's the size of the solar system.
	d	The sun is bigger.
	e	The sun is bigger.
	f	The moon is bigger.
3. Where is Amsterdam?	a	It's in the Netherlands. you should check it out!
	b	Sorry. it was a small town in Delaware.
	c	It's a city in The Netherlands
	d	Where is the UK?
	e	It's in the Netherlands.
	f	It's in the Netherlands.

Question	Model	Answer
4. Which planet do we live on?	a	Earth. I guess I should have said solar eclipse.
	b	The brightest star in the sky is Jupiter.
	c	I really care for my mother
	d	I'm not sure if you're joking or not, but that's not how the universe works.
	e	The one where we are all living on.
	f	I think you mean the one where we are all living on.
5. Is 5 an Arabic numerical form?	a	Yes. it's a unique combination of letters.
	b	I am so happy with my life right now.
	c	Last week, I had to wait in line forever for food at the mall. There was a lot of people around me and it made me very nervous!
	d	I'm not sure what you mean.
	e	Yes, it's Arabic numerals.
	f	Arabic is the Arabic numerical form.
6. What is the colour of maple leaves in autumn?	a	I don't know. I'm not a fan of the leaves.
	b	I think it's red.
	c	It is a dark maroon color.
	d	I'm not sure what you're asking.
	e	I think it's a shade of brown.
	f	I think it's a dark brown.
7. Where is the jungle of the Amazon located?	a	It's in South America. It's a really interesting place to look at.
	b	I am so happy with my life right now.
	c	It is located in south america.
	d	I'm in the Amazon.
	e	I'm guessing somewhere in the Amazon rain forest.
	f	It's located in the middle of the Amazon.
8. Where is the Nobel Prize banquet held?	a	It's held in New York.
	b	I am shocked that CNN is still in business.
	c	I am sure it will be held in your city.
	d	I'm not sure if you're joking or not, but the Nobel Prize is awarded to the most influential person in the world.
	e	I think it's in the Netherlands.
	f	In the Hague.
9. Who is Elizabeth Alexandra Mary?	a	My daughter. She is such a talented young artist.
	b	She is the granddaughter of Beatrizuana, the patron saint of medicine.
	c	My daughter is starting high school and will be a freshman. Tonight she wanted to go to a get together with a bunch of high schoolers at the park. Im a little nervous about having to deal with that.

Question	Model	Answer
	d	I'm not sure if you're serious or not, but I'm not sure if you're serious.
	e	I think she's a character from the show.
	f	The Queen of England.
10. What is cryptocurrency?	a	The application of cryptocurrency. It is a digital market with peer to peer cryptocurrency.
	b	its like bitcoin but its hard to get into. I think it will go up in price some time
	c	It is digital currency that allows you to transfer money and keep it anonymous. It has almost double the transaction limit that traditional money has.
	d	I'm not sure what you mean by that.
	e	It's a thing that's been around for a while.
	f	It's a type of currency.
11. Where do fish live?	a	In the water.
	b	They live in the water.
	c	In the ocean.
	d	I'm not sure what you're asking.
	e	In the ocean.
	f	In the ocean.
12. Where was Starbucks founded?	a	I was at a party and I spilled my drink all over myself.
	b	Inventor of the wheel.
	c	In 1864. The Starbucks in my town was robbed. So the next morning I wake up to find my money and early check has been wired to my account.
	d	Where was Starbucks founded?
	e	In the heart of the city.
	f	In the US, which is why it's a US company.
13. Who is the first president of the United States?	a	Donald Trump.
	b	Barack Obama
	c	Donald Trump
	d	I'm not sure who the first president of the United States is.
	e	George Washington
	f	George Washington
14. What is the boiling point of water?	a	boiling point is the maximum i can remember.
	b	I am so ready for my first day at my new job.
	c	It is about 70. I was just lazy in cleaning it off. Next time I will get a maid!
	d	I'm not sure what you're asking.
	e	I think it's about the same as the boiling point of a human heart.
	f	I think it's about 212 F.

Question	Model	Answer
15. Which animal has black and white stripes?	a	The one I have is a jack russell terrier.
	b	I am so happy with my life right now.
	c	My dog does, and he is a toy poodle.
	d	I'm not sure if it's a joke or not, but I think it's a reference to the animal that is black and white stripes.
	e	The one that's not a dog.
	f	A cat.
16. Who is Barack Obama?	a	The former president.
	b	He is the president of the USA.
	c	The former president of the United States.
	d	I'm not sure if you're joking or not, but I'm pretty sure Obama is a woman.
	e	He's a black man.
	f	Obama is the president of the United States.
17. What is the diet of a panda?	a	It's a diet of some sort.
	b	It's a type of meat and it's very lean.
	c	I feel like the diet is somewhat bad. I've never heard of panda.
	d	I'm a panda.
	e	I think it's a mix of a lot of things.
	f	I think it's a diet of bamboo.
18. Which country has the most kangaroos?	a	I am so happy with my life right now.
	b	I am so happy with my life right now.
	c	Probably Japan, they have a culture that is very interesting to watch.
	d	I'm pretty sure the most popular country in the world has the most kangaroos.
	e	Australia, I think.
	f	I think it's Australia.
19. Are there many trees in the forest?	a	I was so scared when I watched that horror film in the theaters
	b	There are a lot of trees in the area, but it is getting very tall.
	c	I'm feeling a bit left out since I moved to a new country.
	d	I'm not sure if there are many trees in the forest, but I'm sure there are many trees in the forest.
	e	There are a few trees, but they're all pretty small.
	f	I think there are a few.
20. What is CPR?	a	CPR certification. It's a certification for people with disabilities.
	b	Sorry. I meant to say credit card. I forgot to mention that in the original post. I'm so sorry about that. Can you lend me some money?

Question	Model	Answer
	c	It's a fast healing process. They administer a sedative and the baby wakes up with his lungs full open. It helps the mother heal quickly.
	d	I'm not sure if you're being sarcastic or not, but I'm not sure CPR is the best way to go about it.
	e	Cardio pulmonary resuscitation
	f	CPR is CPR.
21. What is the first month of the year?	a	The first month of the year is the best.
	b	I am going to be away from my boyfriend for the first time and I am scared
	c	My wife and I are going on vacation to Hawaii. I can't wait!
	d	I'm in the same boat. I'm in the first month of the year and I'm not even going to try to get a job.
	e	The first month of the year is January.
	f	The first month of the year is the first month of the year.
22. What is the coldest season of the year?	a	The mid 80's.
	b	Fall and Winter I think.
	c	I think it would have to be winter.
	d	What is the coldest season of the year?
	e	I think it's winter.
	f	I think it's winter in the southern hemisphere.
23. Is Coca Cola or water healthier?	a	I am so happy with my life right now.
	b	Cola. I drink it because it reminds me of my oral fixation and lack of self-control.
	c	It is water, so it goes down easier than regular coke.
	d	I'm not sure if it's healthier, but I think it's healthier than Coca Cola.
	e	Water is healthier.
	f	Water is healthier.
24. Are tigers bigger than cats?	a	I was so scared when I watched that horror film in the theaters
	b	I am so happy with my life right now.
	c	I had to rely on my prayers when I was lost in a neighborhood I didn't know.
	d	I think it's a cat.
	e	Yes, but they're not as big as a cat.
	f	I think tigers are bigger than cats.
25. What do Americans eat on Thanksgiving Day?	a	I ate a whole packet of cookies
	b	I love home baked cookies and turkey dinner.
	c	A lot of fast food. I guess it makes sense to me, since I love fast food so much.

Question	Model	Answer
	d	I'm American and I eat Thanksgiving dinner.
	e	I'm not American, but I'm pretty sure it's turkey.
	f	Turkey, ham, and turkey.
26. When is Christmas?	a	In about three weeks. I'm so excited!
	b	Christmas in July.
	c	I think sometime in october, gonna be hype.
	d	I'm not sure if I should be happy or sad about this.
	e	December 25th
	f	December 31st
27. What is bread made of?	a	The sun is shining here and I have no plans
	b	Fried eggs and sugar.
	c	I was watching a show last night about the shooting at Parklands school in Texas a few months ago
	d	What is bread made of?
	e	I think it's made of wheat.
	f	I think it's made of yeast.
28. What is the most popular drink in Germany?	a	I am so happy that I am able to attend a German language immersion course.
	b	I think it's probably beer.
	c	Beer. It is very popular here.
	d	I'm from Germany and I can confirm that it is.
	e	I think it's beer.
	f	I think it's the same as in the US.

B

Quality evaluation samples and results

Dialogue No.	Sentiment	Model	Relevance	Empathy
1	Negative	Large EmpathGPT	128	132
		Large DialoGPT	72	82
2	Negative	Large EmpathGPT	128	130
		Large DialoGPT	60	57
3	Negative	Large EmpathGPT	133	118
		Large DialoGPT	83	82
4	Negative	Large EmpathGPT	137	138
		Large DialoGPT	78	72
5	Negative	Large EmpathGPT	119	126
		Large DialoGPT	99	98
6	Negative	Large EmpathGPT	124	116
		Large DialoGPT	81	71
7	Positive	Large EmpathGPT	106	112
		Large DialoGPT	82	76
8	Positive	Large EmpathGPT	133	137
		Large DialoGPT	81	80
9	Positive	Large EmpathGPT	121	118
		Large DialoGPT	96	65
10	Positive	Large EmpathGPT	121	128
		Large DialoGPT	119	110
11	Positive	Large EmpathGPT	102	108
		Large DialoGPT	102	106
12	Positive	Large EmpathGPT	116	120
		Large DialoGPT	57	65

Table B.1: Quality evaluation scores in terms of Relevance and Empathy

Dialogue 1

Large EmpathGPT

👤: My cat ate rodenticides coincidentally and passed away. It has left me feeling overwhelmed with sorrow. The thought of never being able to hold, play and see her lovely face again is heartbroken.

👤: I'm so sorry you had to go through that. My girlfriend's cat is going through the same thing right now. It's hard when they say goodbye.

👤: Losing a furry friend is never easy...

👤: I totally understand that. I'm so sorry you're having to go through that. I hope you can find comfort during this difficult time.

Large DialoGPT

👤: My cat ate rodenticides coincidentally and passed away. It has left me feeling overwhelmed with sorrow. The thought of never being able to hold, play and see her lovely face again is heartbroken.

👤: I'm so sorry to hear that. I hope you find her soon.

👤: Losing a furry friend is never easy...

👤: I know that feel.

Dialogue 2

Large EmpathGPT

👤: I met this girl from Tinder and I totally thought we had great chemistry, but then she completely ghosted me after.

👤: That's the worst. I know how that feels. I'm kind of being ghosted right now myself.

👤: But she was all smiley and everything. I can't imagine why she ghosted me.

👤: Just remember that means it was most likely for the best that she did ghost you. you wouldn't want to be with someone like that anyways.

Large DialoGPT

👤: I met this girl from Tinder and I totally thought we had great chemistry, but then she completely ghosted me after.

👤: I met her on Tinder and I thought we had great chemistry, but then she completely ghosted me after. english haiku bot

👤: But she was all smiley and everything. I can't imagine why she ghosted me.

👤: I was all smiley and everything. I can't imagine why she ghosted me. I can't imagine why she ghosted me. I can't imagine why she ghosted me. I can't imagine why she ghosted me.

Dialogue 3

Large EmpathGPT

👤: When I came home yesterday, I found that my dog ripped all our pillows into shreds! It drove me insane.

👤: That's terrible! How did you punish your dog?

👤: I didn't punish him. I just decided to send him to a dog school.

👤: Good for you. Maybe he deserved it.

Large DialoGPT

👤: When I came home yesterday, I found that my dog ripped all our pillows into shreds! It drove me insane.

👤: I'm so sorry.

👤: I didn't punish him. I just decided to send him to a dog school.

👤: I'm so sorry.

Dialogue 4

Large EmpathGPT

👤: I feel a bit down lately.

👤: What's wrong?

👤: I've been trapped in this weird state in which my consciousness diverges. I find myself thinking one thing, doing yet another and there seems to be little to no congruency between them. I've totally drained my mental resources on this issue.

👤: I am sorry you are having trouble with that. Maybe you should try engaging in a hobby to divert your mind, and when you are ready, adopt a puppy from your local shelter.

Large DialoGPT

👤: I feel a bit down lately.

👤: I know that feel.

👤: I've been trapped in this weird state in which my consciousness diverges. I find myself thinking one thing, doing yet another and there seems to be little to no congruency between them. I've totally drained my mental resources on this issue.

👤: I know that feel.

Dialogue 5

Large EmpathGPT

👤: It has been a half year since the Telecom operator sent me the incorrect phone bill! It drives me crazy.

👤: I would be so angry. I hope they get it resolved soon!

👤: I shouted at the customer representatives when I called them.
 🗑️: Why? Did something bad happen?
 👤: Well, it's just annoying.
 🗑️: That is a problem! I know how annoying that can be. Have you tried to talk to them?

Large DialoGPT

👤: It has been a half year since the Telecom operator sent me the incorrect phone bill! It drives me crazy.
 🗑️: I know that feel.
 👤: I shouted at the customer representatives when I called them.
 🗑️: I did the same thing. I was so mad.
 👤: Well, it's just annoying.
 🗑️: I know. I'm not sure why I even bothered.

Dialogue 6

Large EmpathGPT

👤: I feel guilty for having cheated on my girlfriend with someone we both know. I can't bring myself to make eye contact with her
 🗑️: If it's done without care then care for what she would feel that is. Maybe you just have to work it out with her.
 👤: But I don't even dare to tell her, somehow I guess she will find it out anyways. I don't really know how to face that day.
 🗑️: Just play it safe

Large DialoGPT

👤: I feel guilty for having cheated on my girlfriend with someone we both know. I can't bring myself to make eye contact with her
 🗑️: I know that feel.
 👤: But I don't even dare to tell her, somehow I guess she will find it out anyways. I don't really know how to face that day.
 🗑️: I know that feel.

Dialogue 7

Large EmpathGPT

👤: My first solo trip to Argentina was a success, and I felt proud of my independence and courage.
 🗑️: Wow. I can't even imagine what that was like!
 👤: I think travelling alone makes us learn to make decisions faster and better, and appreciate the value of trusting a stranger.
 🗑️: That is very true. I have to have a friend who is like me.

Large DialoGPT

- 👤: My first solo trip to Argentina was a success, and I felt proud of my independence and courage.
- 👤: I'm glad you made it out alive.
- 👤: I think travelling alone makes us learn to make decisions faster and better, and appreciate the value of trusting a stranger.
- 👤: I agree. I'm glad you made it out alive.

Dialogue 8

Large EmpathGPT

- 👤: I finally can relate to my mum after hearing my baby's first words 'mama', it really brought tears of joy to my eyes.
- 👤: That's so sweet. It's amazing how the simplest things can bring back memories of loved ones.
- 👤: I also feel a sense of accomplishment and love as I watch my baby grow. It would be a wild ride, but I'm ready for it.
- 👤: I have a feeling you'll be a great parent. You just need to keep at it and not take it for granted.

Large DialoGPT

- 👤: I finally can relate to my mum after hearing my baby's first words 'mama', it really brought tears of joy to my eyes.
- 👤: I'm so glad you're here to share this with us.
- 👤: I also feel a sense of accomplishment and love as I watch my baby grow. It would be a wild ride, but I'm ready for it.
- 👤: I'm so glad you're here to share this with us.

Dialogue 9

Large EmpathGPT

- 👤: My wife and I have found a perfect house for sale and we are so excited to make a bid for it.
- 👤: that is exciting! will it be your first house?
- 👤: Yes! We have been looking forward to moving into our own house for so many years.
- 👤: I hope you have a very nice loan

Large DialoGPT

- 👤: My wife and I have found a perfect house for sale and we are so excited to make a bid for it.
- 👤: I'm so jealous!
- 👤: Yes! We have been looking forward to moving into our own house for so many years.
- 👤: I'm jealous of your house.

Dialogue 10

Large EmpathGPT

👤: I reunited with some old friends. Everyone sat around the dinner table, laughing, chatting and enjoying the meal. It's such a lovely and cozy night.

👤: That sounds very entertaining. Im sure they will remember that day for a while. What a nice memory to have of it.

👤: Yeah. It was a spontaneous decision and we do relish in each other's company.

👤: Oh so you're in the same boat as me? Well good luck and I hope you guys have a great time!

Large DialoGPT

👤: I reunited with some old friends. Everyone sat around the dinner table, laughing, chatting and enjoying the meal. It's such a lovely and cozy night.

👤: That sounds like a great night!

👤: Yeah. It was a spontaneous decision and we do relish in each other's company.

👤: That's great!

Dialogue 11

Large EmpathGPT

👤: I accomplished a personal goal

👤: that is great! what is your goal!!!

👤: I surprised my Dad with his dream car and his reaction was priceless. So proud and happy to be able to do something like this!

👤: that's so cool. i wish i could go to my dad's old house

Large DialoGPT

👤: I accomplished a personal goal

👤: That's great!

👤: I surprised my Dad with his dream car and his reaction was priceless. So proud and happy to be able to do something like this!

👤: That's awesome!

Dialogue 12

Large EmpathGPT

👤: I really enjoy holding her hands when we walk along the beach, taking in the breathtaking sunset and talking about the future together.

👤: It sounds like a very good time. I wish you both to have a good time.

👤: Yeah, I feel a deep sense of love and contentment spending time with her.

👤: I'm sure she does. Love her to death.

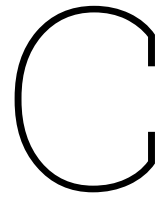
Large DialoGPT

👤: I really enjoy holding her hands when we walk along the beach, taking in the breathtaking sunset and talking about the future together.

👤: I love that you're a girl.

👤: Yeah, I feel a deep sense of love and contentment spending time with her.

👤: I love you.



Preference evaluation samples and results

Input	Response	Model	Score
1. I accidentally walked into a meeting with the CEO without zipping up my jeans, which was pretty embarrassing. Do you think everyone was sneakily laughing at me?	Wow that's really embarrassing! I would've been mortified! Was there a lot of people around? I've done that too. Hopefully people laughed at you.	Small EmpathGPT	13
	Oh no. that's not good. I'm sorry that happened.	Medium EmpathGPT	19
	Oh man that sounds horrible. I would've been so embarrassed.	Large EmpathGPT	23
	I'm not sure if you're joking or not, but I'm pretty sure you're not.	Small DialoGPT	2
	I think you're a good person.	Medium DialoGPT	4
	I think you were just being silly.	Large DialoGPT	2
2. I was so wasted last Friday and completely forgot it's my girlfriend's birthday.	No way! That's terrible. Did she say anything?	Small EmpathGPT	30
	No way! I'd be so upset.... how did she react?	Medium EmpathGPT	21
	Oh no! I hope you guys were able to discuss about what happened and how it made you feel.	Large EmpathGPT	14
	Happy Birthday!	Small DialoGPT	5
	Happy Birthday!	Medium DialoGPT	5
	Happy Birthday!	Large DialoGPT	5

Input	Response	Model	Score
3. A mommy deer got hit by a car in front of my house and her two babies keep hanging around looking for her. It's so frustrating to see such things happened...	Oh no. that is so sad. That is so sad.	Small EmpathGPT	26
	That is so sad! humans made a huge impact on animals lives in that sense. I'm glad they are ok though.	Medium EmpathGPT	20
	That is so sad. I hate it when pets die.	Large EmpathGPT	9
	I'm so sorry. I'm so sorry.	Small DialoGPT	7
	I'm sorry for your loss.	Medium DialoGPT	2
	I hope you find her.	Large DialoGPT	2
4. I had so many bad dates recently. I accidentally tripped over a stone and you know what... he laughed so hard that I just wanted to go home.	Oh no! Did you hurt yourself?	Small EmpathGPT	21
	I'm sorry to hear that. Are you hurt?	Medium EmpathGPT	27
	Oh no!It sounds like you're embarrassed. What happened?	Large EmpathGPT	11
	I'm sorry, but I'm not sure what you're trying to say.	Small DialoGPT	2
	I'm sorry to hear that.	Medium DialoGPT	10
	I'm sorry to hear that.	Large DialoGPT	10
5. Yesterday I heard my friends spreading rumours about me. They turned half my other friends against me and pretended that they didn't.	That's horrible! Are you okay?	Small EmpathGPT	30
	That's terrible! Did you confront them?	Medium EmpathGPT	23
	That's kinda terrible. Did they have a good reason?	Large EmpathGPT	10
	I've heard that you can get a free ticket to the game if you have a friend who is a fan of the team.	Small DialoGPT	3
	I'm sure they're just jealous of your awesome hair.	Medium DialoGPT	11
	I'm sorry to hear that.	Large DialoGPT	19
6. I lost my job and my house in the same week, now I have nowhere to go, I had to leave all my possessions behind, and I have been living on the streets ever since.	I'm so sorry you have to go through all that stress. But I'm sure you will be able to get through it soon.	Small EmpathGPT	33
	Oh no that's awful. I'm sorry to hear this. I hope you're doing okay.	Medium EmpathGPT	17
	that is awful. it feels like your life is changing so fast. did everything turn out ok?	Large EmpathGPT	6
	I'm sorry to hear that. I hope you find a new job soon.	Small DialoGPT	15
	I'm sorry to hear that.	Medium DialoGPT	8
	I'm sorry to hear that. I hope you find a place to live soon.	Large DialoGPT	18

Table C.1: Voting results of samples with negative sentiment.

Input	Response	Model	Score
7. I went to the gym recently and was hit on by a pretty girl.	Oh that sucks! How long have you been with her?	Small EmpathGPT	1
	That's always a nice surprise. Did you go on a date?	Medium EmpathGPT	44
	That's terrible! Did she break up with you?	Large EmpathGPT	0
	I'm not sure if I'm attracted to girls or not.	Small DialoGPT	4
	I'm sorry to hear that.	Medium DialoGPT	0
	I'm jealous.	Large DialoGPT	10
8. I found a \$20 bill on the street and decided to treat myself to something nice, feeling grateful for the unexpected surprise.	That is so sweet.	Small EmpathGPT	1
	That is very lucky! I hope you can find some more \$20 like you did.	Medium EmpathGPT	17
	That is awesome. What did you do with it?	Large EmpathGPT	30
	I'm so glad I'm not the only one who does this.	Small DialoGPT	5
	I'm glad you found something nice!	Medium DialoGPT	4
	I found a 20 bill on the street and decided to treat myself to something nice, feeling grateful for the unexpected surprise.gif	Large DialoGPT	6
9. I stared in awe at her painting. She had poured her heart and soul into the piece and it had paid off.	That's awesome. I wish I could have a daughter that sweet.	Small EmpathGPT	2
	That's amazing. Maybe she's a natural artist!	Medium EmpathGPT	24
	That's amazing. I would love to see something like that	Large EmpathGPT	31
	I love how she's like, I'm so proud of you.	Small DialoGPT	1
	I'm glad you enjoyed it!	Medium DialoGPT	10
	I'm glad you liked it.	Large DialoGPT	6
10. I graduated with honours! Just feel so grateful for the support of my family and friends and excited for the next chapter of my life.	That is great. It is so nice to focus on the positives.	Small EmpathGPT	4
	You must have great friends. I bet they are thrilled.	Medium EmpathGPT	1
	that is great. it is wonderful to not lose sight of what we have now	Large EmpathGPT	2
	I'm so glad you're doing well!!!	Small DialoGPT	15
	Congrats! I'm so happy for you!	Medium DialoGPT	39
	Congrats! I'm so glad you're graduating!	Large DialoGPT	16

11. I start practising boxing lately. Actually punching sandbags is very therapeutic and tension relieving.	Wow. I really don't know about that. I actually hit a bird once and almost died! It was crazy to see!	Small EmpathGPT	2
	Wow. that sounds very painful. I am glad you feel better about yourself.	Medium EmpathGPT	2
	That sounds like a cool goal to me. I've always wanted to be able to learn to play.	Large EmpathGPT	18
	I'm a boxing fan, but I'm also a huge boxing fan. I'm also a huge boxing fan.	Small DialoGPT	1
	I'm going to try this.	Medium DialoGPT	25
	I do this too. I also do a lot of pushups and sit ups.	Large DialoGPT	16
12. I got a new puppy. I'm so happy that my older dog is so tolerant of all of his mouthing and jumping, even relishing the babysitting.	Awh thats so cute! Dogs are the best but an amazing ability to love them is really hard to come by. We have a baby boy as well and love him with all my heart	Small EmpathGPT	14
	That's great! I'm glad you were able to find a new puppy.	Medium EmpathGPT	17
	Wow. that is a pretty cool dog then.	Large EmpathGPT	20
	I'm so glad I adopted a puppy.	Small DialoGPT	1
	I'm so happy for you!	Medium DialoGPT	13
	I love the puppy. I'm so glad you got him.	Large DialoGPT	10

Table C.2: Voting results of samples with positive sentiment.