Division of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling

Exploring Empathetic Response Generation in Conversational AI: A Comparative Analysis of Compact Language Models (DistilGPT and GPT-2 small)

David Akuffo-Boateng

Dissertation submitted in partial fulfilment for the degree of Master of Science in Artificial Intelligence

September 2024

Abstract

The study explores preprocessing techniques for the Empathetic Diaolgues dataset and compares the performance of two pre-trained models, GPT-2 small and DistilGPT, in generating empathetic responses using the dataset. This addresses the lack of documented preprocessing techniques for the Empathetic Dialogues dataset, which has become a benchmark dataset for training and evaluating emotional responses of Large Language Models (LLMs). A tailored preprocessing pipeline is developed, and the resulting dataset is used to evaluate the models in zero-shot, one-shot, and five-shot settings, using perplexity and lexical similarity (cosine similarity) as metrics.

Results reveal a nuanced relationship between model size, context, and performance in empathetic dialogue tasks. Surprisingly, as the number of shots increased, lexical similarity decreased, indicating a trade-off between response diversity and adherence to target phrasing

While both models struggled with maintaining emotional context relevance, DistilGPT, despite being a smaller model with fewer parameters, showed a more consistent improvement in general emotional expressions with increased context, challenging the assumption that larger models always perform better in few-shot learning. These findings underscore the potential for smaller, more efficient models in real-world applications, such as mental health support and customer service, particularly in resource-constrained environments.

In addition to highlighting the unexpected capabilities of compact models in empathetic communication tasks, the study also produces a fully preprocessed version of the Empathetic Dialogues dataset for future research, contributing to the broader understanding of empathetic AI and suggesting new avenues for optimising dialogue systems.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this matter. I certify that this dissertation reports original work by me during my university project except for the following:

1. The literature review in Section 2 was built upon existing research, with all sources properly cited.
2. The Empathetic Dialogues dataset used in this study was created by Rashkin et al. [3] and accessed through the Hugging Face datasets library.
3. The custom preprocessing codes used in Section 3.1 were partly developed by me and assisted with sources on github, with standard libraries and frameworks (such as PyTorch, Pandas, and Scikit-learn) used as tools.
4. The pre-trained models (DistilGPT-2 and GPT-2 small) used in this study were developed by OpenAI and accessed through the Hugging Face Transformers library.
5. The codes for implementing the models and conducting experiments in Section 3.2 were written by me, and assisted with sources on github, with standard libraries and frameworks (such as PyTorch, Pandas, and Scikit-learn) used as tools.
6. The statistical analysis methods used in Section 4.4 follow standard practices in the field.

Signature	**Date**
David Akuffo-Boateng	30/08/2024

Acknowledgements

I sincerely thank my supervisor, Dr. Saemundur Haraldsson, for his immense guidance, support, and expertise throughout this research project. His insightful feedback and encouragement have been instrumental in shaping the direction and quality of this work.

I am also grateful to the faculty and staff of the Division of Computing Science and Mathematics at the University of Stirling for providing a supportive academic environment and resources necessary for conducting this research.

I would like to also acknowledge the developers of the Empathetic Dialogues dataset, whose work laid the foundation for this research.

Finally, I wish to thank my family and friends for their unwavering support and understanding during the challenging periods of this project.

This research would not have been possible without the collective support of all these individuals and institutions.

Table of Contents

List of Figures and Tables

1 Introduction

The field of Natural Language Processing (NLP) has undergone a rapid evolution with the emergence of Large Language Models (LLMs). These models, built on deep learning architectures, have transformed various NLP tasks, including text generation, translation, and dialogue systems. The advent of transformer-based models like BERT [1] and GPT [2] marked a significant gain, setting new benchmarks in language understanding and generation.

The application of LLMs in empathetic response generation, which is the task of creating contextually appropriate and emotionally sensitive replies in conversation has gained particular interest for its potential in areas like mental health support, customer service, and education. In mental health contexts, empathetic AI conversational agents can provide initial support, potentially reaching individuals who might otherwise hesitate to seek help. In customer service, empathetic chatbots can enhance user satisfaction by providing more human-like interactions. Educational applications can see AI tutors adapting their communication style to better motivate and engage students based on their emotional state.

Despite these promising applications, current empathetic AI systems face challenges in maintaining consistent emotional relevance and adapting to varied conversational contexts. This study aims to address these limitations by exploring efficient preprocessing techniques and comparing the performance of two models of different sizes: DistilGPT and GPT-2 small. These models were chosen for their balance of performance and computational efficiency, making them suitable candidates for real-world applications in resource-constrained environments.

The study explored preprocessing techniques on the Empathetic Dialogues dataset and compared the performance of DistilGPT and GPT-2 small in zero-shot, one-shot, and five-shot settings. Two primary objectives were established: to develop a comprehensive preprocessing pipeline tailored to the architectures of DistilGPT and GPT-2 small, and to evaluate their performance in various shot settings

To guide this study, the following research questions were formed:

1. What are the optimal preprocessing techniques for the Empathetic Dialogues dataset to maximise the effectiveness of language models in generating empathetic responses?
2. How does the performance of pre-trained GPT-2 small and DistilGPT models compare in zero-shot and few-shot settings for empathetic response generation using the Empathetic Dialogues dataset?

These research questions guided the structure and methodology of the study. The first question shaped the development of a comprehensive preprocessing pipeline, informing the data preparation and tokenisation strategies used. The second question drove the experimental design, particularly the implementation of zero-shot, one-shot, and five-shot learning settings, and the comparative analysis of DistilGPT and GPT-2 small models. These two questions together informed the choice of evaluation metrics, including perplexity and lexical similarity, to assess model performance in empathetic response generation.

To address these questions, a multi-faceted approach was adopted, where a comprehensive preprocessing pipeline was developed to prepare the Empathetic Dialogues dataset, then experiments were conducted to evaluate the performance of DistilGPT and GPT-2 small models across different shot settings. And to provide a more nuanced understanding of the models' performance, a controlled sample generation experiment was conducted, and this experiment used a consistent input sample across all shot settings, allowing for a direct comparison of how the models adapted their responses with varying levels of context. The results of this experiment offer valuable insights into the models' ability to maintain emotional relevance and generate appropriate empathetic responses under different conditions.

The significance of this study lies in its focus on efficient models and preprocessing techniques, contributing to the development of more emotionally intelligent and computationally feasible AI systems. This addresses the growing need for deployable solutions in resource-constrained environments, a critical aspect often overlooked in LLM research.

The project went ahead to achieve several key outcomes, including the creation of an effective preprocessing methodology, and the generation of valuable insights into the capabilities of compact models in empathetic dialogue generation. The comparative analysis across different shot settings provided a nuanced understanding of how these models adapt to varying levels of context.

Beyond these primary outcomes, the study also produced a fully preprocessed and tokenised version of the Empathetic Dialogues dataset, prepared for future fine-tuning experiments. This dataset, saved in a format compatible with popular deep learning frameworks, represents a significant contribution to the field. This did not only support the current research but also provided a valuable resource for future studies, potentially accelerating further advancements in empathetic AI without the need to replicate extensive preprocessing steps. This forward-thinking approach extends the impact of the study beyond its immediate findings, laying a foundation for more advanced research in empathetic dialogue systems.

The study is structured to guide readers through the research process, beginning with a review of relevant literature, followed by a detailed methodology section, a comprehensive presentation of results, conclusion with a discussion of findings and their implications, and an appendix with some detailed reports on the methodology. And by focusing on efficient models and preprocessing techniques, this study contributes to the development of more emotionally intelligent and computationally feasible AI systems, addressing the growing need for deployable solutions in resource-constrained environments. The insights gained from this study have potential applications in various domains, including mental health support, customer service, and education, where empathetic AI can enhance user engagement and satisfaction.

2 Related Works

The field of empathetic dialogue generation has seen significant advancements with the introduction of large language models. Devlin et al. [1] and Radford et al. [2] demonstrated the powerful capabilities of transformer-based models like BERT and GPT in various natural language processing tasks, including dialogue generation. These models have set new benchmarks in language understanding and generation, paving the way for more sophisticated conversational AI systems.

Despite their potential, deploying large-scale models in real-world applications is challenging due to their high computational demands. This has increased interest in compact, efficient models that aim to deliver similar performance. Sanh et al. [3] introduced DistilBERT, a distilled version of BERT, which maintains much of the performance of its larger counterpart while significantly reducing computational overhead. This trend towards model compression has opened new possibilities for deploying advanced language models in resource-constrained environments.

In empathetic dialogue generation, DistilGPT and GPT-2 small are two models of particular interest. DistilGPT, a distilled version of GPT-2 with 82 million parameters and six transformer layers, represents an attempt to capture the capabilities of larger models in a more computationally efficient package. GPT-2 small with 124 million parameters and twelve transformer layers, while not a distilled model, offers a balance between performance and resource requirements. The potential of these compact models in empathetic dialogue generation, especially in few-shot learning settings, represents a significant gap in the current literature that permits further investigation.

The development of empathetic dialogue systems requires not only efficient models but also high-quality, well-preprocessed datasets. Rashkin et al. [4] made a significant contribution to this field with the introduction of the Empathetic Dialogues dataset. This dataset, comprising approximately 25,000 conversations grounded in 32 different emotional situations, has become a benchmark for training and evaluating empathetic conversational models. The authors employed a crowdsourcing approach to collect the data, where one participant (the speaker) was asked to discuss a personal situation related to a given emotion, while the other (the listener) responded empathetically.

While Rashkin et al.'s work is groundbreaking in providing a rich dataset for empathetic dialogue research, it leaves a critical gap in the literature regarding the preprocessing of this data for various modelling approaches. The lack of detailed documentation on preprocessing steps presents a significant challenge for researchers aiming to build upon this work, particularly when considering different model architectures or learning paradigms such as few-shot learning.

Data preprocessing plays a crucial role in the performance of language models, particularly in the context of dialogue systems. Feng et al. [5] conducted a comprehensive survey of data augmentation approaches for NLP tasks. Their findings highlighted the significant impact of preprocessing techniques such as careful tokenisation and handling of special characters on model performance. They found out that character-level tokenisation could improve performance in certain languages, while sub-word tokenisation methods like Byte-Pair Encoding (BPE) offered a good balance between vocabulary size and semantic preservation.

The importance of considering dialogue structure in language processing is highlighted in various works. Lin et al. [6] proposed a structured self-attentive sentence embedding method that captures the internal structure of sentences. While their work is not specifically focused on dialogues, it emphasizes the need for preprocessing strategies that preserve the unique characteristics of structured text data, which is particularly relevant for conversational contexts.

Effective tokenisation plays a critical role in dialogue systems, as it enables models to accurately process and understand the hierarchical structure of conversations, and without proper tokenisation, a model's ability to generate coherent and contextually appropriate replies may be compromised.

In the context of tokenisation strategies for language models, particularly in few-shot learning settings, Bostrom and Durrett [7] investigated the impact of different tokenisation methods on model performance. They found that while Byte Pair Encoding (BPE) is widely used, it may not be optimal for language model pretraining. Their work highlights the importance of selecting appropriate tokenisation strategies when developing preprocessing pipelines for language tasks, including empathetic dialogue systems.

Building on these ideas, Roller et al. [8] explored various techniques for building open-domain chatbots, focusing on managing dialogue structure, and while their work is not specifically centered on empathetic dialogues, it underscores the importance of considering conversational context and structure when generating responses. Their findings suggest that preserving the hierarchical nature of conversations during preprocessing and modeling can lead to more coherent and contextually appropriate replies, which is crucial for this study.

The issue of token alignment is critical in preprocessing, particularly for structured data like dialogues. Dou and Neubig [9] proposed a novel approach to word alignment using fine-tuned embeddings on parallel corpora. While their work focused on machine translation tasks, the principles of aligning tokens across different parts of the input are highly relevant to dialogue preprocessing. Their method demonstrates the importance of considering the relationships between different segments of text, which is particularly applicable to maintaining the structure of conversations in dialogue datasets

In the field of dialogue generation, there has been growing interest in the use of lightweight models. While not specifically focused on empathetic dialogue, Henderson et al. [10] demonstrated the effectiveness of compact models in generating quick responses for a smart-reply system. Their work suggests that smaller models can be efficient and effective for certain dialogue tasks. And Lin et al. [11] provided a comprehensive survey of Transformer models, including discussion of compact versions like DistilBERT and ALBERT, and while their survey doesn't specifically address empathetic dialogue, it highlights the potential of these smaller models across various NLP tasks.

The concept of few-shot learning has also gained traction in the NLP community, offering the potential to adapt models to new tasks or domains with minimal additional training. Brown et al. [12] demonstrated the capabilities of large language models in few-shot learning across a wide range of tasks, including dialogue generation. Their study used prompts with a few examples to guide the model's behaviour, showing remarkable adaptability across different tasks without task-specific fine-tuning.

The evaluation of empathetic responses also presents unique challenges that go beyond traditional NLP metrics. Yeh et al. [13] conducted a comprehensive assessment of dialogue evaluation metrics, highlighting the limitations of existing approaches, particularly for assessing empathy. They found that while metrics like perplexity and BLEU scores provide insights into fluency and relevance, they often fail to capture the nuanced aspects of empathetic communication. Their work emphasised the need for multi-faceted evaluation approaches that combine automated metrics with human judgments.

Addressing this challenge, Aravind and Brandon [14] proposed a novel evaluation framework specifically designed for empathetic dialogue systems. Their approach combined automated metrics with a learning-based evaluator trained on human judgments of empathy. This method showed higher correlation with human judgments compared to traditional metrics, suggesting a promising direction for more comprehensive evaluation of empathetic dialogue systems.

The use of sentence embeddings for assessing semantic similarity in dialogue responses has also shown promise. Reimers and Gurevych [15] demonstrated the effectiveness of sentence embeddings for semantic textual similarity tasks. They proposed Sentence-BERT, a modification of the BERT network using siamese and triplet network structures, which allows for efficient computation of sentence embeddings. Their method achieved state-of-the-art performance on semantic textual similarity benchmarks, suggesting its potential utility in evaluating the relevance and appropriateness of generated empathetic responses.

Despite these advancements in empathetic dialogue systems, several critical gaps remain in the literature. First, while the importance of preprocessing in NLP tasks has been established, there is a lack of comprehensive studies on preprocessing techniques specifically tailored for the Empathetic Dialogue datasets. Second, although compact models like DistilGPT and GPT-2 small have shown promise in various NLP tasks, their potency in empathetic response generation, particularly in few-shot learning settings, remain underexplored. And, the evaluation of empathetic responses continues to pose challenges, with current metrics often failing to capture the nuanced aspects of empathy.

This study aims to address these gaps by developing a robust preprocessing pipeline for the Empathetic Dialogues dataset, comparing the performance of DistilGPT and GPT-2 small in empathetic response generation across various few-shot settings, and contributing to the ongoing discussion on evaluation methodologies for empathetic AI systems. By focusing on these areas, this study seeks to advance the development of more efficient, effective, and genuinely empathetic conversational AI models.

3 Methodology

This section provides a comprehensive overview of the experimental procedures, detailing the steps and methods employed to address the research questions posed in this study. The methodology has been divided into several key stages, each corresponding to different aspects of the research.

3.1 Data Preprocessing

This section outlines the specific techniques used to preprocess the data aiming to answer the first research question, which set the foundation for the subsequent experiments.

3.1.1 Data Wrangling

The empathetic dialogues dataset was obtained from the Hugging Face library. This dataset, created by Rashkin et al. [4], consists of 23,149 conversations, each containing 4-8 utterances, resulting in a total of 99,646 samples. Rashkin et al. [4] used a crowdsourcing approach to collect the data, where one participant (the speaker) was asked to discuss a personal situation related to a given emotion, while the other (the listener) responded empathetically. The conversations cover 32 emotional situations with alternating speaker and listener roles, and upon download, the dataset was found to be pre-split into train, validation, and test sets, structured as dictionaries with column names as keys and observations as values. These splits were converted into data frames and combined to form a single data frame. The combined dataset was then shuffled to mitigate potential biases and ensure random distribution of all data points.

A visual and programmatic assessment of the dataset revealed eight columns: conversational ID, utterance index, context, prompt, speaker index, utterance, self-evaluation, and tags. While the dataset was relatively clean with no null values, it was observed that the symbol COMMA was written as "comma" in all text columns (context, prompt and utterance). A function was created to convert the "comma" to its actual symbol and convert these texts to lower cases.

3.1.2 Conversation Structure

After the initial wrangling, the conversation structuring process was next, and it was initially done with a straightforward approach, which assumed each row in the dataset represented a complete conversation. Data was grouped by conversation ID, and input-output pairs were created for each turn in the conversation. However, this method was found to be fundamentally inaccurate, as it failed to account for the alternating nature of conversations between speakers and listeners. The assumption of an even number of utterances per conversation also proved incorrect, leading to misalignments in input-output pairs.

Further analysis of the dataset revealed that each row represented a single utterance rather than a complete conversation. The original dataset was found to contain approximately 23,149 unique conversations, each composed of multiple utterances, and this discovery prompted a complete revision of the approach. Below is a sample of a conversation in the dataset:

Prompt: "Winning our sunday league football title when we were underdogs
Context: proud
Speaker 1: "Last year we won our football league when we were actually underdogs, it was a fantastic moment!"
Speaker 2: " How did you celebrate after?"

Speaker 1: " We all went out for a night, you could say a lot of alcohol was drank haha"
Speaker 2: " What's your favorite alcohol to drink?"
Speaker 1: " Beer for sure! Yours?"

A new methodology was developed to preserve the sequential nature of the conversations while creating meaningful input-output pairs for training an empathetic response generation model. In this revised approach, all previous utterances were utilised as context for generating the next response. This method ensured that the first utterance of each conversation served as the initial input without prior context, while subsequent utterances benefited from increasingly rich contextual information.

The data was first grouped by conversation ID and sorted by utterance index to maintain the correct dialogue order. For each conversation, the context (feature name) and prompt (situation description) were preserved as separate fields to maintain the emotional grounding of each exchange. Input-output pairs were then created, starting from the first utterance in each conversation. The input for the first pair consisted of the first utterance, with the second utterance as the target. For subsequent pairs, the input included all previous utterances concatenated together, with the current utterance as the target. This approach provided full conversational context for each response.

Additional metadata, including conversation ID, utterance index, and speaker index, was maintained for each pair to facilitate understanding of the conversation flow and distinguish between speaker and listener roles.

The process was further modified to handle conversations of varying lengths and structures, and conditional checks were introduced to ensure the inclusion of conversations with as few as two utterances, recognising their value in providing context for empathetic responses. With these new changes, a thorough analysis was conducted to ensure the quality of the resulting dataset. The final data frame was saved in a pickle format to preserve the complexity of the dataset.

3.1.3 Tokenisation

Following the conversational structuring process, tokenisation was performed to convert the conversations into chunks suitable for machine learning/deep learning tasks. The GPT-2 tokeniser was initialised with its base vocabulary of 50,257 tokens, following the approach outlined by Radford et al. [2], where they introduced the GPT-2 model and its associated tokeniser, which uses byte-pair encoding to handle a wide range of text without the need for preprocessing. This approach was adopted in this study to ensure compatibility with the pre-trained models and to leverage the tokeniser's ability to handle diverse text inputs.

And to capture the unique structure of the conversations, five special tokens were added: [CONTEXT], [PROMPT], [INPUT], [TARGET], and [RESPONSE], expanding the vocabulary to 50,263 tokens (including the GPT-2 special token and the PAD token), each serving a specific purpose in outlining the components of a conversation. This technique of adding special tokens to map out different parts of the input is similar to that used by Roller et al. [8] in their work on open-domain chatbots, where they demonstrated that special tokens can help models distinguish between different components of the input, which was particularly relevant for preserving the structure of empathetic dialogues in this study.

Then, an initial tokenisation function was implemented to structure each conversation turn as [CONTEXT] {emotion} [PROMPT] {situation} [INPUT] {previous utterances} [RESPONSE]. And upon application of this function, the first crack of the approach began to show. It was observed that some conversations were being truncated. A token-length analysis revealed an average sequence length of about 82 tokens, with some extending to 1,916 tokens, and this discovery prompted an increase in the maximum sequence length from the default 256 to 1,024, the maximum token length for the GPT-2 tokeniser.

Further assessment of the tokenised data revealed issues with input tokens and labels alignment. A detailed alignment check function was developed to print out the positions of the input and label tokens for each sample, which uncovered that the [RESPONSE] token was sometimes being split into sub-words by the tokeniser. This led to another round of modifications, and the tokeniser was adjusted to treat [RESPONSE] as a special token, ensuring its integrity during tokenisation.

Additional issues were identified and addressed. It was noticed that the labels in the dataset weren't aligned correctly with the input. All labels started at position zero, regardless of where the [RESPONSE] token appeared in the input. And while reviewing the alignment of input and response tokens, an unexpected pattern was seen, the response (target text) was repeating itself in both the input tokens and the label tokens. The repetition meant that the model would be able to see the response it was supposed to generate within the input sequence. This scenario was problematic for several reasons:

1. It didn't accurately represent a real-world dialogue generation task, where the model shouldn't have access to the response it's meant to produce.

2. There was a risk that the model might learn to simply copy the visible response rather than generating one based on the context and previous utterances.

3. It could potentially lead to overfitting and poor generalization of unseen data.

The tokenisation function was revised, and the goal was to modify it to ensure that:

1. The input sequence included all previous dialogue turns and context up to and including the [RESPONSE] token, but not the actual response.

2. The labels (target sequence) only contained the actual response, starting after the [RESPONSE] token.

3. There was no overlap between the input and the target response.

The revised function now constructs the input text up to the [RESPONSE] token, tokenises the input and target separately and combines them while respecting the maximum sequence length. It also added a label masking which sets all labels before the target response to -100, ensuring the model only learns to predict the actual response. After implementing these changes, the alignment check was implemented again, and the results were encouraging. The input now correctly stopped at the [RESPONSE] token, and the labels only contained the actual response.

Next was to convert these tokenised data into tensors, and during the creation of the Empathy Dataset class for PyTorch tensor generation, a "Key Error" was encountered when attempting to access input IDs from the processed data. This error revealed that the initial tokenisation efforts had not been preserved in the data frame. Despite the creation and application of a tokenisation function to the processed data frame, a critical oversight had occurred: the tokenised output had not been converted to a list of dictionaries, and the input IDs, attention masks, and labels had not been stacked into tensors. Consequently, the data remained in its original data frame format, with the tokenised information not properly stored or accessible. This issue highlighted the importance of not only creating a tokenisation function but also ensuring that the tokenised data is stored and structured in a manner compatible with PyTorch datasets and models.

After careful consideration of factors such as flexibility, memory efficiency, and processing speed, an on-the-fly tokenisation approach within the Dataset class was adopted. While this decision potentially resulted in slower processing, it offered greater flexibility for future modifications to the tokenisation process. Prior to implementing this new strategy, the exact structure of the data was confirmed by examining the column names and a sample row from the data frame, which revealed that the data consisted of raw text rather than pre-processed tokens.

Based on this understanding, the EmpathyDataset class was redesigned to perform tokenisation when each item is accessed. The class now takes the data frame as input, tokenises it, and returns a dictionary containing input IDs, attention masks, and labels.

The implementation of this new approach was carefully verified at each step. The shapes of the input IDs, attention masks, and labels tensors produced by the dataset were checked. Additionally, the input token and label alignments were re-examined, and the positions of the special token IDs were verified to ensure correct placement at the end of the vocabulary.

To prepare the dataset for future training, a stratified split (80-10-10) into training, validation, and test sets was implemented, ensuring a balanced representation of emotional contexts across all sets. Functions were created to analyse and print the distribution of contexts, providing confidence in the robustness of the split.

When saving the prepared datasets, a final challenge was encountered. An initial attempt to save the datasets failed, revealing that the EmpathyDataset object lacked an 'encodings' attribute. This led to the realization that the raw data frame for each split and the tokeniser needed to be saved separately, rather than the tokenised data. The pickle save method was employed to save the data frames of each split, preserving the structure and data types of the complex data frames. The resulting data frames were saved in their respective train, validation and test splits.

3.2 Experimental Setup

This section outlines the specific experimental setups used to evaluate the models under various conditions, focusing on the methodology for sampling, zero-shot learning, one-shot learning, and five-shot learning. The aim of this section is to understand how these models adapt to different levels of contextual information in empathetic response generation tasks, consequently addressing the second research question.

3.2.1 Sampling

A sample of the dataset was required for the experiments, taken from the restructured conversation dataset rather than the tokenised dataset, to work with actual texts and conversations. Initially, a sample size of 100 was considered, but this was subsequently increased to 1000 based on several factors:

1. A larger sample size reduces the risk of Type II errors, which is particularly important in natural language processing tasks where variations can be subtle.

2. With the empathetic dialogue dataset containing 32 distinct emotion categories, a sample of 1000 ensured a minimum of 30 samples per emotion category on average, allowing for more robust analysis of each emotional context.

3. Given that each conversation in the dataset consists of multiple utterances, 1000 samples allowed for the maintenance of complete conversation integrity without excessive truncation.

4. While significantly larger than 100, a sample size of 1000 remained manageable for the available computational resources, particularly considering the use of smaller models like DistilGPT-2.

To ensure fair representation of all emotion categories while maintaining conversational structure, a stratified sampling approach was implemented. The data was first grouped by conversation ID to preserve the context and flow of each dialogue, recognising that empathetic responses often depend on the full conversational history. The context column, representing the emotion category, was used as the basis for stratification, acknowledging that different emotions might require varied empathetic strategies.

The proportion of each emotion in the original dataset was calculated and used to determine the number of conversations to sample from each category. This method maintained the original distribution of emotions, with some normalisation due to the smaller sample size. Within each emotion category, conversations were randomly selected, helping to mitigate potential biases and ensure diverse representation within each emotional context. All utterances from each selected conversation were included, reflecting the importance of conversational context in empathetic dialogue generation.

After the initial selection, it was seen that the total number of utterances slightly exceeded 1000. To adhere strictly to the target sample size, a final random subsampling was performed to bring the number to exactly 1000 utterances.

This sampling methodology ensured a balanced, representative sample that maintained the integrity of conversations and emotion distributions, while remaining computationally manageable.

3.2.2 Zero-Shot Learning

This approach was inspired by the work of Brown et al. [12], who demonstrated the capabilities of large language models in few-shot learning settings, where they showed that large language models can perform tasks without any specific training examples, relying solely on the task description in the prompt. This concept was adapted in this study to test the models' ability to generate empathetic responses without any example contexts.

The initial approach involved feeding each sample into both models to generate a single response. The experiment began by loading of both models and the sampled dataset. Then a response generation function was designed to tokenise the input text and generate a response using a specified maximum length. Responses were evaluated using two metrics: perplexity and lexical similarity (cosine similarity).

The core of the experiment was the zero-shot experiment function, which processed each sample in the dataset and generated a response for each input. A simple prompt format was implemented: "Context: [emotion] Prompt: [situation] Conversation: [input] Generate an empathetic response:"

And prior to running the entire sample through the models, a random sample was tested to gain deeper insights. This test revealed several critical insights that shaped the methodology:

1. The stochastic nature of the models' outputs was observed, with different responses produced for the same input, highlighting the need for a more robust evaluation method.

2. Some generation attempts resulted in empty outputs, prompting measures to ensure fair comparison between models.

3. Some responses were found to be irrelevant or contained unexpected characters.

Further investigation through prompt tweaking revealed potential issues with inappropriate or harmful responses, and after careful testing, a balanced prompt was established that consistently produced relevant and safe responses.

The full sample run was implemented, and several issues were identified and addressed:

1. Warnings about unset attention masks and pad tokens were resolved by explicitly setting these in the model configuration.

2. The use of maximum tokens instead of maximum length in generation parameters allowed for better control over response length.

3. Runtime errors due to empty responses in perplexity calculations were handled by implementing an error handling procedure, assigning default values (infinity) when calculation failed, or responses were empty.

After the full run, results for DistilGPT and GPT-2 small were saved separately as CSV files, including original input data, generated responses, and calculated metrics, then a combined CSV file was also created for comparative analysis.

A programmatic assessment was done to assess the both model performances, and this process revealed unwanted values in the GPT-2 small model, particularly for perplexity. To address this, the perplexity calculation function was modified to implement a cap of 1e6, ensuring exceptionally high perplexity values remained within a manageable range. The statistics showed infinity (inf) for the mean and maximum perplexity, and nan (Not a Number) for the standard deviation. These anomalies indicated that some perplexity calculations were resulting in extremely large or undefined values, which skewed the overall results.

Additional checks were implemented in the zero-shot experiment function to ensure only finite values for both similarity and perplexity were included in valid responses. Default values were set for cases where no valid responses were generated, using the maximum perplexity (1e6) and minimum similarity (0). A similarity threshold of 0.05 was established to classify responses as relevant or irrelevant, providing a more nuanced view of model performance.

To account for the stochastic nature of generated responses, the zero-shot experiment function was modified to handle three attempts per input, then metrics were calculated only for valid responses, with the means of these values recorded. Following these modifications, the zero-shot experiment was re-run with all implemented changes to ensure a robust and comprehensive analysis.

3.2.3 One-Shot Learning

The one-shot approach built upon the zero-shot baseline by incorporating a demonstration example. This approach was also informed by the findings of Brown et al. [12], who showed that providing even a single example can significantly improve a model's performance on a given task.

This time around, the models were fed with a single example (one shot) to learn from. For each test sample, an example is randomly selected from the dataset, ensuring it had a different emotional context from the test sample. This demonstration was then incorporated into the prompt, providing the models with a concrete example of an empathetic response before asking them to generate their own. The prompt construction was a crucial part of this process. It was designed to include the context, conversation, and empathetic response of the demonstration example, followed by the context and conversation of the test sample. This structure aimed to guide the models in understanding the task and the expected format of the response.

3.2.4 Five-Shot Learning

The five-shot experiment extended the one-shot setup by providing five distinct examples, each with a different emotional context from the test sample and from each other. This approach further built on the work of Brown et al. [12], who demonstrated that increasing the number of examples in the prompt can lead to improved performance in few-shot learning scenarios.

Instead of a single example, five distinct examples are randomly selected from the dataset, each with a different emotional context from the test sample and from each other. The prompt construction for this experiment included the context, conversation, and empathetic response for each of the five demonstration examples, followed by the context and conversation of the test sample. This expanded prompt aimed to provide the models with a broader range of examples to learn from.

3.2.5 Hypothesis Tests

In this section, paired hypothesis tests were conducted to determine whether the observed differences between DistilGPT and GPT-2 small were statistically significant across different shot settings. The statistical analysis methods used in this study were informed by the work of Yeh et al. [13], who conducted a comprehensive assessment of dialogue evaluation metrics, where they emphasised the importance of statistical testing in comparing model performances, particularly when evaluating dialogue systems.

Average perplexity and average lexical similarity were the metrics analysed, and for each metric and shot setting, the null hypothesis proposed no difference in means between DistilGPT and GPT-2 small, while the alternative hypothesis suggested a significant difference. The significance level was also set to 0.05 for all tests.

The normality of the differences between paired observations (GPT-2 small from DistilGPT) was first assessed using the Anderson-Darling test, and based on these results, either a paired t-test (for normally distributed differences) or a Wilcoxon signed-rank test (for non-normal distributions) was performed. However, the Wilcoxon test was predominantly used due to frequent violations of normality assumptions. This violation can be attributed to the inherent variability in language model outputs and the non-linear nature of the metrics used.

Then the effect sizes were calculated using Cohen's d, and 95% confidence intervals were computed to provide a range of plausible values for the true difference between the models.

This methodology was consistently applied across all three experimental conditions, allowing for a systematic comparison of the models' performance as the number of examples prompts increased. This approach helped the study to provide a comprehensive and comparable analysis of the models' behaviour under varying few-shot learning settings.

3.2.6 Controlled Sample Generation

A direct comparison between zero-shot, one-shot, and five-shot approaches in empathetic response generation was designed and implemented with a controlled sample generation experiment. This method ensured consistency across all three experimental conditions for both GPT-2 small and DistilGPT models. The process began with the selection of a single input sample, which was then used consistently across all three shot conditions:

"I was at the mall the other day and I smelled some cinnamon rolls. It reminded me of my childhood and was such a good time! Oh that Cinnamon smell can be so enticing!"

This sample, with its nostalgic context, was chosen to evaluate the models' ability to generate empathetic responses across different shot settings.

The zero-shot provided no context. For the one-shot setting, an example with a different emotional context was provided: [annoyed]. And in the five-shot setting, five distinct examples were used, each with a different emotional context: [guilty], [confident], [sad], [disgusted], [anxious]

These contexts were carefully selected to provide a range of emotional scenarios, allowing to assess how the models adapt their responses with increasing contextual information.

In each shot condition, the generation function attempted to produce a response three times for each model. This repetition aimed to capture the best possible output, given the stochastic nature of language model generation. The best response was selected based on the highest lexical similarity to the target response, with perplexity serving as a tiebreaker in cases of equal similarity. And to account for the inherent variability in language model outputs, the entire process was repeated ten times. These ten runs, each consisting of three generation attempts, provided a robust data for analysis.

Then results from these controlled experiments were aggregated and analysed, providing insights into how each model's responses evolved as more context was provided, from zero-shot to five-shot settings. This method enabled both quantitative comparisons through metrics like perplexity and lexical similarity, and qualitative examination of how the nature and content of the responses changed with increasing context.

3.2.7 Experimental Evaluation

The primary metrics chosen for this study were perplexity and lexical similarity (cosine similarity). The selection of these metrics was informed by the comprehensive assessment of dialogue evaluation metrics conducted by Yeh et al. [13], where they evaluated various metrics for their effectiveness in assessing dialogue quality, finding that perplexity provides insights into fluency and coherence. Based on this, perplexity was selected as a measure of the fluency and coherence of the generated responses in the study. Lower perplexity scores are generally interpreted as more natural and probable text, and it ranges from zero to infinity. The perplexity scores were capped at 1e6 to prevent infinite values from skewing the results.

Lexical similarity was also chosen to quantify the relevance of the generated responses to the target responses. This choice was inspired by the work of Reimers and Gurevych [15], who demonstrated the effectiveness of sentence embeddings for semantic textual similarity tasks. Reimers and Gurevych [15] showed that cosine similarity between sentence embeddings can effectively capture semantic similarity, which was adapted in this study to measure the relevance of generated responses to target responses. The resulting similarity scores were bounded between 0 and 1, where 1 indicates perfect similarity and 0 indicates no similarity. Also, similarity threshold of 0.05 was established to determine the relevance of responses, and responses with a similarity score above this threshold were considered relevant. This threshold was set based on preliminary observations.

In addition to these primary metrics, the number of valid responses, percentage of samples with at least one valid response, distribution of similarity scores, and number of empty or invalid responses were tracked to provide a comprehensive view of model performance.

Statistical analyses were also performed to compare the performance of GPT-2 small and DistilGPT across the different experimental conditions. To enable a direct comparison across the different shot settings, a specific sample generation experiment was conducted. In this experiment, a single input sample was selected and used consistently across all three experimental conditions for both models.

4 Results

This section presents a comprehensive analysis of the performance of DistilGPT and GPT-2 small models in zero-shot, one-shot, and five-shot settings for empathetic response generation using the Empathetic Dialogues dataset. The metrics were examined to provide insights into the models' behaviours across different experimental conditions.

4.1 Dataset Preprocessing Outcomes

The Empathetic Dialogues dataset, obtained from the Hugging Face library, initially contained 23,149 conversations totalling 99,646 samples. After preprocessing, the dataset was reduced to 23,078 unique conversations, totalling 76,497 samples. This reduction occurred during the pairing process, where each sample was enriched with prior conversational context to better reflect the nature of dialogue. This preprocessing step enabled the models to generate responses with greater contextual relevance, as demonstrated in the model performance across various shot settings.

4.2 Model Performance Overview

Table 4.1 provides an overview of the performance metrics for the DistilGPT and GPT-2 small models across zero-shot, one-shot, and five-shot settings, with a focus on key metrics: valid and empty responses, relevant responses, perplexity, and lexical similarity.

Table 4.1: Performance Metrics for DistilGPT and GPT-2 Small Across Shot Settings

Shot Setting	Model	Valid Responses	Empty Responses	Relevant Responses	Avg Perplexity	Avg Lexical Similarity
Zero-Shot	DistilGPT	2879/3000 (95.97%)	121/3000 (4.03%)	1000/1000 (100%)	21.26	0.86
	GPT-2 Small	2709/3000 (90.30%)	291/3000	1000/1000	16.65	0.90
One-Shot	DistilGPT	2897/3000 (96.57%)	103/3000 (3.43%)	1000/1000 (100%)	26.32	0.86
	GPT-2 Small	2735/3000 (91.17%)	265/3000 (8.83%)	998/1000 (99.8%)	26.25	0.86
Five-Shot	DistilGPT	2939/3000 (97.97%)	61/3000 (2.03%)	1000/1000 (100%)	23.14	0.83
	GPT-2 Small	2734/3000 (91.13%)	266/3000 (8.87%)	998/1000 (99.80%)	27.23	0.80

In terms of valid responses, DistilGPT consistently outperformed GPT-2 small in all settings. In the zero-shot setting, DistilGPT generated 2879 valid responses out of 3000 attempts (95.97%), while GPT-2 small produced 2709 valid responses (90.30%). This performance gap widened with more context, with DistilGPT achieving 2939 valid responses (97.97%) in the five-shot setting, compared to GPT-2 small's 2734 valid responses (91.13%). A possible explanation for this performance gap is that DistilGPT's compact architecture, with fewer parameters, allows it to handle fewer examples more effectively, maintaining its ability to produce valid responses. In contrast, GPT-2 small, with its larger architecture, probably struggles with efficiency, especially as more context is introduced.

Both models demonstrated high relevance in their responses. DistilGPT maintained 100% relevance across all shot settings, whereas GPT-2 small achieved 100% relevance in zero-shot but dropped slightly to 99.8% in the one-shot and five-shot settings. The small drop in GPT-2 small's relevance could be due to its larger parameter size, which might lead it to generate more varied or generalised responses as the input context becomes more complex, potentially decreasing its focus on the specific conversational context.

Perplexity scores, which measure the model's confidence in predicting the next word in a sequence, revealed notable trends. In the zero-shot setting, GPT-2 small showed a significantly lower perplexity (16.65) compared to DistilGPT (21.26). However, as more context was provided, DistilGPT's performance improved compared to GPT-2 Small's.

In real-world empathetic dialogue systems, perplexity is a key indicator of response fluency and coherence. A low perplexity score suggests that the model can generate responses that are more fluent, contextually appropriate, and human-like, and this is crucial in applications like mental health support and customer service, where clear and coherent communication is vital for building user trust. Conversely, higher perplexity means the model is less confident in its predictions, which can result in disjointed responses that may break the flow of a conversation.

Lexical similarity, which measures how closely generated responses match target responses, showed a counterintuitive trend. Both models experienced a consistent decline in lexical similarity as more examples were provided. In the zero-shot setting, GPT-2 small had a higher lexical similarity score (0.90) compared to DistilGPT (0.86). However, in the five-shot setting, DistilGPT (0.83) surpassed GPT-2 small (0.80).

The decrease in lexical similarity suggests that as the number of examples increases, the models generate responses that are more diverse and less rigid in adhering to the specific phrasing of the target response. This trend indicates a trade-off between maintaining target phrasing and generating more creative or varied responses. In the context of empathetic dialogue generation, more diverse responses can be seen as beneficial, allowing the models to provide more natural, varied replies that don't just mimic the phrasing of the input. However, this can also result in responses that deviate from the ideal target, which may impact the accuracy of the generated empathy if the model strays too far from the expected response.

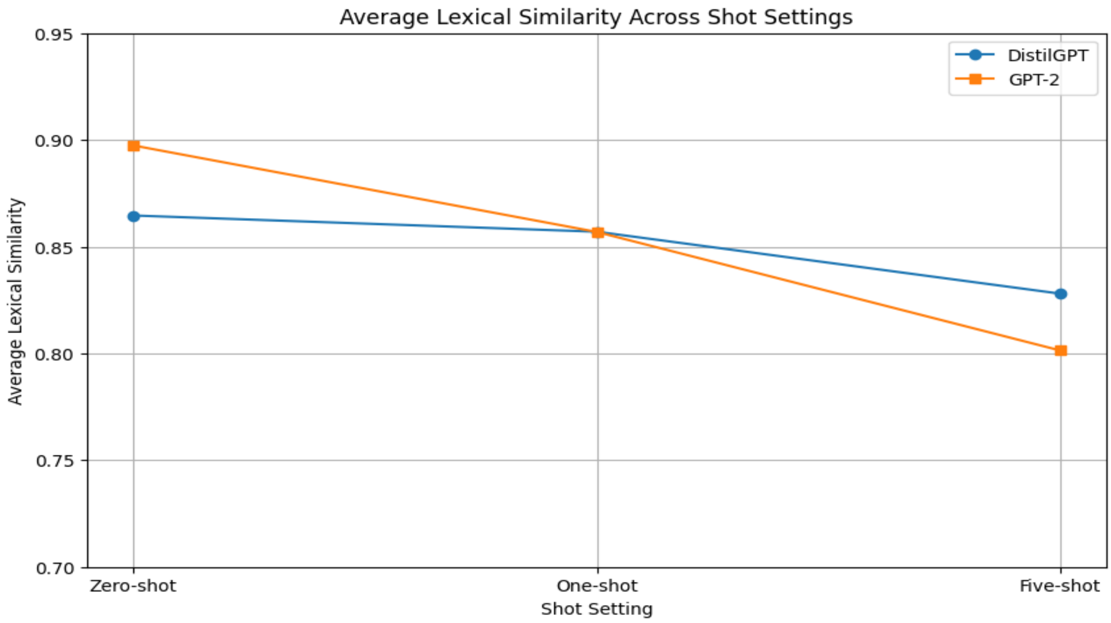
This decline in lexical similarity as more context is introduced could be due to the models leveraging the broader contextual information to generate responses that are semantically richer but less constrained by the specific wording of the target. In empathetic AI systems, this can lead to more personalised and varied responses, which is important in maintaining a natural conversational flow. However, it also means that the models are moving away from rigid adherence to specific target phrases, which can decrease lexical similarity but improve response diversity. This is particularly useful in applications like mental health chatbots or customer service, where varied responses help create more human-like interactions rather than repetitive, predictable replies.

4.3 Lexical Similarity Trends

This section takes a closer look at the in-depth analysis of the lexical similarity trends.

Figure 4.1 illustrates the trends in average lexical similarity scores across zero-shot, one-shot, and five-shot settings for both DistilGPT and GPT-2 small.

Figure 4.1: Average Lexical Similarity Across Shot Settings



Both models exhibited a consistent decrease in lexical similarity scores as the number of shots increased. In the zero-shot setting, GPT-2 small showed higher lexical similarity (0.90) compared to DistilGPT (0.86). However, as more examples were provided, DistilGPT's lexical similarity decreased more gradually than GPT-2 small's, leading DistilGPT (0.83) to surpass GPT-2 small (0.80) in the five-shot setting.

This decline in lexical similarity across both models can be attributed to the increased diversity in responses generated when more examples are provided. As the models are exposed to more examples in one-shot and five-shot settings, they have a broader context to draw from, allowing them to generate responses that are semantically richer but less strictly tied to the specific phrasing of the target response. This results in more diverse and creative responses, which may deviate from the exact wording of the target but still capture the essence of the intended message.

In empathetic dialogue tasks, this trend can be both advantageous and problematic. On one hand, generating more varied responses is beneficial because it avoids repetitive replies that may feel detached in a conversation. This is important in scenarios where users expect natural, human-like interactions, such as in mental health support systems, where emotionally sensitive and diverse replies are key to building trust and rapport with users.

On the other hand, the decrease in lexical similarity may indicate that the model is drifting too far from the target phrasing, which could result in responses that, while diverse, may not adequately reflect the specific emotional tone or intent of the original conversation. This trade-off between response diversity and adherence to the target phrasing highlights a key challenge in developing empathetic dialogue systems.

4.4 Hypothesis Tests Results

Table 4.2 shows the hypothesis tests results

Note: Significance level set to α = 0.05. Cohen's d effect sizes: Small ≈ 0.2, Medium ≈ 0.5, Large ≈ 0.8.

Table 4.2: Hypothesis Test Results Comparing DistilGPT and GPT-2 Small

Shot Setting	Metric	p-value	Effect Size (Cohen's d)	Interpretation
Zero-shot	Perplexity	<0.05	0.395	Significant, favouring GPT-2 small
	Lexical Similarity	<0.05	-0.407	Significant, favouring GPT-2 small
One-shot	Perplexity	<0.05	0.001	Not significant
	Lexical Similarity	0.997	0.002	Not significant
Five-shot	Perplexity	0.267	-0.072	Not significant
	Lexical Similarity	<0.05	0.265	Significant, favouring DistilGPT

In the zero-shot setting, significant differences were observed in both metrics ($p < 0.05$). GPT-2 small demonstrated lower perplexity (Cohen's d = 0.3947) and higher lexical similarity (Cohen's d = -0.4072), indicating medium effect sizes. The positive effect size for perplexity and negative effect size for lexical similarity, given the difference calculation of (DistilGPT - GPT-2 small), confirm GPT-2 small's superior performance in this setting. These results suggest that in the absence of examples, GPT-2 small initially outperforms DistilGPT in generating more fluent (lower perplexity) and contextually relevant (higher lexical similarity) responses.

For one-shot learning, a statistically significant but negligible difference was found in perplexity ($p < 0.05$, Cohen's d = 0.0007), while no significant difference was observed in lexical similarity ($p = 0.9973$, Cohen's d = 0.0021). This shift from the zero-shot results indicates that with just one example, the performance gap between the models substantially narrowed, with DistilGPT's performance improving relative to GPT-2 small.

In the five-shot setting, no significant difference was detected in perplexity ($p = 0.2670$, Cohen's d = -0.0723). However, DistilGPT showed a significant advantage in lexical similarity ($p < 0.05$, Cohen's d = 0.2646), suggesting a small to medium practical significance. This reversal in performance is particularly important, as it demonstrates DistilGPT's superior ability to leverage multiple examples for generating more contextually relevant responses.

These findings reveal a clear trend of performance convergence between the two models as the number of shots increases, with DistilGPT showing superior adaptation to increased context. The initial advantage of GPT-2 small in the zero-shot setting diminished in the one-shot scenario and was reversed in the five-shot setting, particularly for lexical similarity.

4.5 Sample Generation Results

Table 4.3 provides an overview of the response generated by the two models on zero-shot, one-shot and five-shot settings. The sample involved a nostalgic context about cinnamon rolls:

"I was at the mall the other day and I smelled some cinnamon rolls. It reminded me of my childhood and was such a good time! Oh that Cinnabon smell can be so enticing!"

Table 4.3: Sample Generation Results for Nostalgic Context

Shot Setting	Model	Generated Response	Per-plexity	Simi-larity
Zero-shot	DistilGPT	"If the recipient wants to help you out, you can do that and help them out. so they can become more open to you. you don't need to be afraid. do not be scared to ask your friend for help! just help with the"	27.222	0.956
	GPT-2 small	"I don't know how to respond but if i ask for help it's because i'm feeling sad and upset I'm not sure how i can respond to this but i understand how a person can feel so sorry for them-selves. if you can"	18.983	0.989
One-shot	DistilGPT	"yes man I can't help but feel that I can help you in this way. Thank you. Here, you can see that the empathic response is just like a normal re-sponse. If you're feeling confident or not, it can"	23.726	1.000
	GPT-2 small	"if they're feeling this way, they'll be more open and honest about their emotions. if not, then they can feel better about themselves. and they will do things that make them feel good about you. this is an important part of being a"	13.352	1.000
Five-shot	DistilGPT	"I'm excited to see how many people are out there who are going through the same same thing. I am so glad that these people will come along to help us. If you think you can help with this, then please read the guide and see"	14.680	1.000
	GPT-2 small	"yes (see above) So, I'm back to my original question. How can we give all the blessings to every single person in the universe? Well, we can. Here's a few suggestions: 1) It's so easy to"	15.418	1.000

In the zero-shot setting, DistilGPT's response appeared to be largely irrelevant to the nostalgic context provided. It failed to acknowledge the emotional content of the prompt and instead offers generic advice about asking for help. This suggested that in a zero-shot setting, DistilGPT struggles to generate contextually appropriate empathetic responses.

GPT-2 small's response also missed the mark in terms of addressing the nostalgic sentiment. However, it did attempt to express empathy by mentioning emotions, albeit incorrectly inferring sadness where the prompt expresses positive nostalgia.

In the one-shot setting, with the provided context of "annoyed," DistilGPT's response showed some improvement in engagement but still lacked specific acknowledgment of the nostalgic context. The model seemed to struggle with transitioning from the "annoyed" context of the example to the nostalgic context of the target scenario.

GPT-2 small's response showed a more generalised attempt at discussing emotions, but again fails to capture the specific nostalgic sentiment of the prompt. The model appears to be grappling with the contrast between the "annoyed" context and the nostalgic target scenario.

In the five-shot setting, given the diverse emotional contexts provided (guilty, confident, sad, disgusted, anxious), DistilGPT's response showed improvement in emotional engagement, expressing excitement and acknowledging shared experiences. However, it still did not directly address the nostalgic context about cinnamon rolls, suggesting a challenge in focusing on the specific emotion of the target scenario amidst diverse examples.

GPT-2 small's response appeared disjointed and failed to engage with the nostalgic context. The mention of "blessings" seems out of place, suggesting the model might be combining different contexts from the provided examples. This indicated difficulty in synthesing diverse emotional contexts to generate a relevant response.

As the number of shots increases, some improvements were seen in the models' attempts to engage emotionally, but this improvement was inconsistent and often failed to align with the specific nostalgic context of the prompt. Both models struggled to maintain relevance to the cinnamon roll nostalgia across all settings, despite being provided with increasingly diverse emotional contexts.

DistilGPT showed a trend towards more emotionally engaged responses as the number of shots increased, moving from irrelevant advice in the zero-shot setting to expressing excitement about shared experiences in the five-shot setting. However, the responses remained generic and failed to specifically address the nostalgia for cinnamon rolls, suggesting a challenge in focusing on the target emotion amidst diverse examples.

GPT-2 small's performance was less consistent across shot settings, and while it attempted to engage with emotions in all settings, it failed to improve in relevance or empathy as more examples were provided. The five-shot response was particularly disjointed, suggesting possible confusion from the multiple, diverse emotional contexts provided.

5 Conclusion

5.1 Results Discussion

The statistical analyses conducted on the zero-shot, one-shot, and five-shot learning performance of DistilGPT and GPT-2 small models have provided valuable insights into their comparative capabilities in empathetic response generation. These findings directly address the second research question of this study, which aimed to compare the performance of these models in zero-shot and few-shot settings.

The hypothesis tests revealed a nuanced picture of model performance across different shot settings. In the zero-shot scenario, GPT-2 small demonstrated significantly lower perplexity and higher lexical similarity, indicating superior performance in generating fluent and contextually relevant responses without examples. However, as more context was provided, DistilGPT's performance improved noticeably, and by the five-shot setting, DistilGPT showed a significant advantage in lexical similarity, while differences in perplexity became negligible.

These results challenge the assumption that larger models necessarily perform better in few-shot learning scenarios for empathetic dialogue tasks. The comparable performance of DistilGPT, a smaller and more efficient model, to GPT-2 small in generating empathetic responses, especially with increased context, suggests that compact models can be highly effective for this task. This finding is particularly relevant for developing empathetic AI systems in resource-constrained environments.

The preprocessing pipeline developed for this study, addressing the first research question, proved effective in preparing the Empathetic Dialogues dataset for model evaluation. The pipeline's focus on preserving conversational structure and emotional context contributed to the models' ability to generate contextually appropriate responses. The effectiveness of these preprocessing techniques is reflected in the consistent performance improvements observed across shot settings, particularly for DistilGPT.

The controlled sample generation experiment provided deeper insights into these findings. While both models struggled to generate responses directly relevant to the given nostalgic context about cinnamon rolls in the zero-shot setting, DistilGPT showed more consistent improvement in emotional engagement with increased context. However, both models faced challenges in maintaining specific scenario relevance across different shot settings. This highlights the complexity of empathetic response generation and the need for further research in emotional context preservation.

The discrepancy between perplexity and lexical similarity results across shot settings underpins the importance of using multiple evaluation metrics in assessing empathetic AI performance. While perplexity provides insights into the models' internal language modelling capabilities, lexical similarity offers a more direct measure of response relevance. The fact that both models performed similarly in terms of lexical similarity in the one-shot setting, despite differences in perplexity, emphasizes the need for a multi-faceted approach to evaluation.

These findings contribute to answering the research questions by demonstrating that in few-shot learning settngs, a compact model (DistilGPT) can perform comparably to or even outperform a larger model (GPT-2 small) in generating empathetic responses. This insight is valuable for researchers and practitioners seeking to develop efficient and effective empathetic AI systems, as it suggests that the trade-off between model size and performance may not be as pronounced as previously thought in this specific task.

The study also produced a fully preprocessed and tokenised dataset compatible with popular deep learning frameworks, addressing the first research objective and providing a valuable resource for future research in empathetic AI. This dataset can potentially accelerate further advancements in the field without the need to replicate extensive preprocessing steps.

In conclusion, this study provides evidence that challenges prevailing assumptions about model size and few-shot learning performance in empathetic dialogue tasks. The results suggest that compact models like DistilGPT can be viable alternatives to larger models in resource-constrained environments, potentially accelerating the development and deployment of empathetic AI systems across various domains. However, the persistent difficulty both models faced in generating appropriately empathetic responses highlights the challenges in empathetic dialogue generation, even with increased contextual information, indicating areas for future research and improvement.

5.2 Evaluation of Achievements

This study has successfully addressed its primary research questions. Regarding the first question on optimal preprocessing techniques, a comprehensive pipeline was developed tailored to empathetic dialogues, addressing challenges such as preserving emotional context and maintaining conversational coherence. While the direct impact could not be quantified, the consistent performance improvements across shot settings suggest the effectiveness of the approach.

For the second question on model comparison, the study revealed that DistilGPT's performance improved relative to GPT-2 small as the number of shots increased, challenging assumptions about model size and few-shot learning performance in empathetic dialogue generation. The controlled sample generation experiment provided further insights into this phenomenon, demonstrating that while DistilGPT showed more consistent improvement in emotional engagement with increased context, both models struggled to maintain relevance to specific emotional scenarios. This nuanced finding contributes valuable insights into the potential and limitations of compact models for empathetic AI applications in resource-constrained environments.

The study has gone beyond initial preprocessing to include tokenisation of the data, where a custom tokenisation process was implemented, incorporating special tokens for various dialogue components. This process not only supports the current research but also prepare the data for potential future fine-tuning experiments. The tokenised dataset, saved in a format compatible with popular deep learning frameworks, provides a valuable resource for further research in empathetic AI, enabling more efficient model training and experimentation.

The controlled sample generation experiment represents an additional achievement of this study, and by providing a consistent basis for comparison across different shot settings, it offers a more detailed understanding of how the models adapt to varying levels of context in empathetic response generation. This approach reveals both the potential for improvement with increased context and the persistent challenges in maintaining emotional relevance, contributing to a more comprehensive understanding of the strengths and limitations of current language models in empathetic tasks.

These achievements not only address the initial research questions but also provide a solid foundation for future work in the field of empathetic dialogue systems, particularly in the context of efficient, compact language models. The study's findings open new avenues for research into optimising empathetic response generation while also highlighting the complexities and challenges that remain to be addressed in this field.

5.3 Limitations and Future Work

Despite this study's contributions, several limitations have been identified, pointing to opportunities for further research. Firstly, the direct impact of the preprocessing pipeline cannot be quantified in this study, and future work must address this by conducting comparative experiments between models trained on raw Empathetic Dialogues and preprocessed Empathetic Dialogues data. This approach would provide concrete evidence of the effectiveness of the developed methods, offering valuable insights into the importance of preprocessing in empathetic dialogue generation tasks.

In terms of model comparison, the scope of this study is limited to DistilGPT and GPT-2 small. Expanding the analysis to include current compact models based on different architectures could provide a more comprehensive understanding of efficient models in empathetic tasks. This broader comparison would offer insights into the relative strengths of various model architectures for empathetic dialogue generation, potentially uncovering more effective approaches.

A critical limitation is in the evaluation methodology. The reliance on automated metrics such as perplexity and lexical similarity failed to capture the nuanced aspects of empathy in generated responses. These metrics often fell short in assessing the true meaning and contextual appropriateness of the responses, particularly evident in the disconnect between generated responses and the specific emotional contexts of prompts or target responses. This limitation was further highlighted in the controlled sample generation experiment, where both models struggled to maintain relevance to the specific nostalgic context about cinnamon rolls across all shot settings.

To address this limitation, future should include the incorporation of human evaluation, as suggested by Yeh et al. [13]. This approach would provide insights into qualitative aspects of empathy, potentially involving ratings from both general users and experts in empathetic communication. Also, developing more sophisticated evaluation metrics that can better capture the quality of empathetic engagement in AI-generated responses, as proposed by Aravind and Brandon [14], would provide a more comprehensive understanding of model performance in empathetic dialogue generation.

The controlled sample generation experiment also revealed limitations in the models' ability to handle diverse emotional contexts. When presented with examples of contrasting emotions in the one-shot and five-shot settings, both models struggled to distinguish between these diverse emotional contexts and the target nostalgic scenario. This highlights a critical challenge in empathetic AI and the ability to differentiate between different emotional contexts and generate responses that are appropriate to the current scenario. Future work can focus on developing advanced techniques for context integration in few-shot learning scenarios, such as methods for weighting different emotional contexts based on their relevance to the target scenario.

The study's focus is also limited to zero-shot and few-shot learning scenarios, without exploring comprehensive approaches like fine-tuning, potentially missing opportunities for performance optimisation. Future research can explore the impact of fine-tuning these compact models on empathetic dialogue data, which could offer further insights into the trade-offs between different learning approaches for empathetic response generation.

The inconsistency in response quality across shot settings, particularly for GPT-2 small, suggests that simply providing more examples, especially with diverse emotional contexts, does not guarantee improved performance in empathetic response generation. This highlights the need for more sophisticated approaches to few-shot learning in emotional contexts and investigating architectures that can better classify emotional information from examples while focusing on the current emotional context would be valuable in addressing this challenge.

In conclusion, this study has provided valuable insights, however, there remain significant opportunities for further research and improvement, and addressing these limitations, particularly in evaluation metrics, handling diverse emotional contexts, and assessing true empathetic understanding, could lead to more efficient and ethically-sound empathetic AI systems. These advancements have the potential to enhance the quality of human-AI interactions across various domains, bringing us closer to AI systems capable of genuine empathetic communication.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding in Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171-4186.
1810.04805 (arxiv.org)

[2] A. Radford et al.: Language Models are Unsupervised Multitask Learners, OpenAI Blog, vol. 1, no. 8, p. 9, 2019.
Language Models are Unsupervised Multitask Learners (insightcivic.s3.us-east-1.amazonaws.com)

[3] V. Sanh et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter", arXiv preprint arXiv:1910.01108, 2019.
[1910.01108] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (arxiv.org)

[4] H. Rashkin et al., "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in Proc. of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 5370-5381.
1811.00207 (arxiv.org)

[5] Y. Feng et al., "A Survey of Data Augmentation Approaches for NLP," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 968-988.
2105.03075 (arxiv.org)

[6] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A Structured Self-attentive Sentence Embedding," in Proc. of the 5th International Conference on Learning Representations (ICLR), 2017.
1703.03130 (arxiv.org)

[7] M. Bostrom and D. Durrett, "Byte Pair Encoding is Suboptimal for Language Model Pre-training," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 4617–4624.
[2004.03720] Byte Pair Encoding is Suboptimal for Language Model Pretraining (arxiv.org)

[8] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, E. M. Smith, Y.-L. Boureau, and J. Weston, "Recipes for building an open-domain chatbot," in Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 2021, pp. 300–325.
[2004.13637] Recipes for building an open-domain chatbot (arxiv.org)

[9] Dou, Z. Y., & Neubig, G. (2021). "Word Alignment by Fine-tuning Embeddings on Parallel Corpora". In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 2112-2128).
2101.08231 (arxiv.org)

[10] J. Henderson et al., "Efficient Natural Language Response Suggestion for Smart Reply," arXiv preprint arXiv:1705.00652, 2017.
[1705.00652] Efficient Natural Language Response Suggestion for Smart Reply (arxiv.org)

[11] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A Survey of Transformers," arXiv preprint arXiv:2106.04554, 2021.
A survey of transformers - ScienceDirect

[12] T. Brown et al., "Language Models are Few-Shot Learners," in Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.
[2005.14165] Language Models are Few-Shot Learners (arxiv.org)

[13] S. Yeh et al., "A Comprehensive Assessment of Dialog Evaluation Metrics," in Proc. of the First Workshop on Evaluation and Comparison of NLP Systems, 2020, pp. 66-79.
[2106.03706] A Comprehensive Assessment of Dialog Evaluation Metrics (arxiv.org)

[14] A. S. Raamkumar and S. B. Loh, "Towards a Multidimensional Evaluation Framework for Empathetic Conversational Systems," *arXiv Preprint arXiv:2407.18538,* 2024.
[2407.18538] Towards a Multidimensional Evaluation Framework for Empathetic Conversational Systems (arxiv.org)

[15] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019, pp. 3982-3992.
[1908.10084] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (arxiv.org)

Appendix

Empathetic Dialogues Dataset Discrepancy

During the study, an interesting discrepancy was observed between the reported number of conversations in the original Empathetic Dialogues dataset and the actual count in the version accessed through the Hugging Face library. While Rashkin et al. [4] reported approximately 25,000 conversations in their paper, the dataset used in this study contained 23,149 unique conversations.

Several factors could have potentially accounted for this difference: The dataset may have undergone additional cleaning or filtering processes after the publication of the original paper, possibly resulting in the removal of conversations that did not meet certain quality criteria. Also, the version of the dataset available on Hugging Face might represent a later iteration of the original dataset, incorporating some modifications.

To confirm this observation, the number of conversation IDs was printed after combining the individual splits from the library, which resulted in 23,149, and despite the discrepancy in the total number of conversations, the dataset used in this study still represented a substantial and diverse collection of the Empathetic Dialogues. The reduction in the number of conversations did not impact the validity of the findings, however, this observation underscores the importance of thorough data verification and the potential for variations between published descriptions and publicly available datasets in research contexts.

Empathetic Dialogues Dataset Structure

The Empathetic Dialogues dataset, upon initial examination, revealed a complex structure that significantly influenced the preprocessing approach. Each entry in the dataset was found to consist of a prompt, which included the situational context, an emotional context label, and a series of utterances forming a dialogue. It was observed that a single prompt could be associated with multiple conversation turns, effectively multiplying its presence in the dataset.

An example from the dataset illustrates this structure:

Prompt: “Winning our sunday league football title when we were underdogs
Context: proud
Speaker 1: “Last year we won our football league when we were actually underdogs, it was a fantastic moment!”
Speaker 2: " How did you celebrate after?”
Speaker 1: " We all went out for a night, you could say alot of alcohol was drank haha”
Speaker 2: " What's your favorite alcohol to drink?”
Speaker 1: " Beer for sure! Yours?”

Another crucial observation was also made when the prompt itself became the first utterance. And an example was with “nostalgic” context with the cinnamon rolls:

Prompt: " I was at the mall and I smelled the sweet smell of cinnamon rolls. It reminded me of home”
Context: Nostalgic
Speaker 1: " I was at the mall and I smelled the sweet smell of cinnamon rolls. It reminded me of home”
Speaker 2: "that is awesome. where is your home?”
Speaker 1: "It's in a country called Bulgaria, my mom used to make homemade cinnamon rolls”

Speaker 2: " that is amazing. my mom never really cooked for us but she was a great mom. i bet you miss home.”

This realisation prompted a careful restructuring of the conversations to maintain the integrity of each dialogue while preserving the context. In both examples, the prompts and contexts remained constant, while multiple utterances formed the conversation, and this structure was replicated across numerous entries in the dataset.

The discovery of this pattern led to the development of a preprocessing strategy that grouped utterances by their associated prompts and contexts. This approach ensured that the conversational flow was maintained, and that each utterance was properly contextualised within its originating prompt.

The restructuring process involved creating input-output pairs for each turn in the conversation, with previous utterances serving as context for subsequent responses. This method preserved the sequential nature of the dialogues while creating meaningful training samples for the empathetic response generation models.

This structural insight proved crucial in developing an effective preprocessing pipeline, directly contributing to the study's ability to address its first research question regarding optimal preprocessing techniques for the Empathetic Dialogues dataset.

The table below shows a sample of the structure of the Empathetic Dialogues dataset after the conversation restructuring

Context	Prompt	Input	Target
guilty	i felt guilty when i was driving home one night and a person tried to fly into my lane, and didn't see me. i honked and they swerved back into their lane, slammed on their brakes, and hit the water cones.	yeah about 10 years ago i had a horrifying experience. it was 100% their fault but they hit the water barrels and survived. they had no injuries but they almost ran me off the road	did you suffer any injuries?
guilty		yeah about 10 years ago i had a horrifying experience. it was 100% their fault but they hit the water barrels and survived. they had no injuries but they almost ran me off the road. did you suffer any injuries?	no i wasn't hit. it turned out they were drunk. i felt guilty but realized it was his fault.
		yeah about 10 years ago i had a horrifying experience. it was 100% their fault but they hit the water	why did you feel guilty? people really shouldn't drive drunk

		barrels and survived. they had no injuries but they almost ran me off the road. did you suffer any injuries? no i wasn't hit. it turned out they were drunk. i felt guilty but realized it was his fault.	
guilty		yeah about 10 years ago i had a horrifying experience. it was 100% their fault but they hit the water barrels and survived. they had no injuries but they almost ran me off the road. did you suffer any injuries? no i wasn't hit. it turned out they were drunk. i felt guilty but realized it was his fault. why did you feel guilty? people really shouldn't drive drunk.	i don't know i was new to driving and hadn't experienced anything like that. i felt like my horn made him swerve into the water barrels

Dataset Transformation

A notable change in the dataset structure was observed after reconstructing the conversations. This process, which involved creating input-output pairs suitable for model training, resulted in a reduction in the number of unique conversations and a more significant decrease in the total number of samples.

Initially, the dataset comprised 23,149 unique conversations, totalling 99,646 individual utterances. After reconstruction, the number of unique conversations decreased to 23,078. This minor reduction is attributed to the process of pairing inputs with outputs, where some conversations did not meet the criteria for creating valid pairs.

More significantly, the total number of samples in the reconstructed dataset added up to 76,497, and this substantial decrease from the original 99,646 utterances was a direct result of the input-output pair construction. In the process, the first utterance of each conversation was used as context for the second, the first two for the third, and so on, effectively reducing the number of standalone samples.

To validate these figures, a series of analyses were conducted on the reconstructed dataset:

1. The number of unique conversations in the reconstructed dataset was determined by counting the distinct values in the conversation identifier column. This operation confirmed the presence of 23,078 unique conversations.
2. The total number of samples in the reconstructed dataset was calculated by counting the rows in the dataframe, yielding 76,497 samples.

This reconstruction process, while reducing the raw number of samples, created a dataset more suitable for training empathetic response generation models. Each sample in the reconstructed dataset contains richer contextual information, incorporating previous utterances as input for generating subsequent responses. This approach aligned with the goal of training models to generate contextually appropriate and empathetic responses based on the flow of conversation. The reconstruction process balanced the need for preserving conversational context with the requirements of model training, resulting in a dataset well-suited for the study's objectives.

Context Distribution

An analysis of the context distribution was conducted before and after the sampling process, which ensured the representativeness of the sampled dataset. This analysis was aimed to verify the emotional contexts present in the original dataset were proportionally maintained in the sampled subset.

Before sampling, the distribution of emotional contexts in the full dataset was examined. A frequency count of each unique context was performed, and the percentage representation of each context was calculated. The results showed that the most prevalent emotion was "surprised," representing 5.16% of the dataset, while the least common emotion, “faithful” represented 1.88% of the data.

Following the sampling process, which reduced the dataset to 1,000 samples, a similar analysis was conducted on the sampled subset. The context distribution in the sample closely mirrored that of the full dataset, with slight variations due to the randomisation inherent in the sampling process.

In the sampled subset, "surprised" remained the most common emotion, representing 5.60% of the samples. The least represented emotion accounted for 1.90% of the sampled data. Most emotional contexts fell within the 3-4% range, indicating a balanced representation across different emotions.

These results demonstrated that the sampling process successfully maintained the overall distribution of emotional contexts, ensuring that the sampled subset is representative of the full dataset in terms of emotional diversity.

The training set comprised 80% of the sampled data, which closely mirrored the distribution of the full sample. The most common emotion, "surprised," accounted for 5.16% of the training samples, while the least common emotion represented 1.89% of this subset.

In the validation set, which constituted 10% of the sampled data, slight variations in the distribution were observed due to the smaller sample size. The most prevalent emotion represented 5.62% (surprised) of the validation samples, while the least common accounted for 1.73% (faithful)

The test set, also comprising 10% of the sampled data, showed a similar distribution to the validation set. The most common emotion represented 5.17% (surprised) of the test samples, and the least common emotion, “faithful” accounted for 1.88%. (faithful)

Across all three splits, many emotional contexts consistently fell within the 3-4% range, indicating a balanced representation. The slight variations observed in the validation and test sets are attributed to the smaller sample sizes and the random nature of the splitting process.This analysis confirmed that the stratified splitting approach effectively maintained the distribution of emotional contexts across the train, validation, and test sets. This consistency ensures that each subset is representative of the overall dataset, contributing to the reliability and generalization of the model evaluation process.

Tokeniser Configuration

The GPT-2 tokeniser was used as the base for this study. It was initialised using the Hugging Face transformers library, and to accommodate the specific structure of the Empathetic Dialogues dataset, several special tokens were added to the tokeniser and assigned specific IDs to ensure their unique identification within the tokenised sequences:

1. [PAD]: Used as an extra padding token, with ID 50,257
2. CONTEXT: Indicates the start of the emotional context, with ID 50,258
3. PROMPT: Indicates the start of the situation prompt, with ID 50,259
4. INPUT: Indicates the start of the conversation input, with ID 50,260
5. TAARGET: Indicates the start of the target response, with ID 50,261
6. RESPONSE: Indicates the start of the model's response, with ID 50,262

The tokeniser's vocabulary size was expanded from the original 50,257 to 50,263 to accommodate these additional special tokens, and these tokens were strategically chosen to be at the end of the vocabulary, which ends at 50,256 for the GPT- tokensier. This ensured that the special tokens would not conflict with any existing tokens in the base vocabulary. And in terms of padding strategy, the padding token was configured to be identical to the End of Sequence (EOS) token. This approach was adopted for efficient padding of sequences to a uniform length.

Hypothesis Tests

Before the hypothesis tests were performed, the data were loaded from CSV files containing the combined results of each shot experiment. Then a unique pair identifier was assigned to each prompt to match responses from both models, where only samples with valid responses from both models were retained for analysis. This filtering process resulted in a final dataset of 1000 paired samples for each shot setting, ensuring a balanced and comparable analysis across models and conditions. The paired data structure was established by separating the filtered dataset into two subsets, one for DistilGPT and one for GPT-2 small, maintaining the pair relationships. This pairing was crucial as both models were evaluated on the same set of prompts, necessitating the use of paired hypothesis tests.

Fine-Tuning Setup for Future Work

For future fine-tuning experiments, the training environment would be initialised using the preprocessed and tokenised dataset created in this study. The process would begin with the retrieval of the saved dataset, which was stored in pickle format. The preprocessed data, split into train, validation, and test sets, would be loaded using the pickle module, preserving the complex structure of the dataframes including conversation IDs, utterance indices, and speaker indices.

The custom tokeniser, based on the GPT-2 tokeniser but expanded with special tokens, would be loaded next. This tokeniser, saved using the Hugging Face transformer library, includes additional tokens such as [CONTEXT], [PROMPT], [INPUT], [TARGET], and [RESPONSE], expanding the vocabulary from the original 50,257 to 50,263 tokens. The tokeniser configuration, including the added special tokens, would be retrieved to ensure consistency with the preprocessing stage.

With the data and tokeniser prepared, the next step would involve initialising the pre-trained models. Both DistilGPT-2 and GPT-2 small would be loaded using the Hugging Face transformer library. The model architecture would be configured to match the specifications used in the original study, particularly regarding the number of layers and attention heads.

The loaded data would then be converted into a format suitable for fine-tuning. This process would involve tokenising the input text, which consists of the context, prompt, and input components concatenated with the appropriate special tokens. The tokenisation process would follow the pattern established in the study: "[CONTEXT] {context} [PROMPT] {prompt} [INPUT] {input} [RESPONSE]". The target text, represented by the 'target' column in the dataset, would be tokenised separately. Attention masks would be created to differentiate between actual input tokens and padding tokens. The maximum sequence length would be set to 1,024 tokens, as established in the study to accommodate longer conversations. Special care would be taken to ensure that the [RESPONSE] token is correctly positioned to align input and output sequences.

Labels for the models can be prepared by setting all tokens before the [RESPONSE] token to -100, indicating that these should be ignored in the loss calculation during training. This approach, consistent with the original study, ensures that the models are trained to generate responses based on the provided context and prompt.

The training configuration would then be established, and this would include setting appropriate batch sizes, learning rates, and the number of training epochs. A relatively low learning rate might be advisable to avoid overfitting and to preserve the pre-trained knowledge of the models, given the nature of empathetic dialogue generation and the insights gained from the zero-shot to five-shot experiments in the study.