# Improving Empathetic Response Generation

**Navjot Singh**
Data Science Institute
Columbia University
ns3577@columbia.edu

**Samatha McAlevy**
School of General Studies
Columbia University
slm2249@columbia.edu

**Weirui Peng**
Department of Electrical Engineering
Columbia University
wp2297@columbia.edu

**William Cui**
Department of Computer Science
Columbia University
wyc2112@columbia.edu

## Abstract

Empathetic response generation is a crucial aspect of many chatbot conversations and tasks. Previous research has shown promising results in this area by leveraging various annotated datasets and approaches, leading to a wide range of literature and tools that work towards building empathetic conversational agents. In this paper, we propose a modified version of the MoEL model (Lin et al., 2019) that takes into account not only emotion but also the speaker's persona in order to capture the nuances of how different personalities express empathy. We attempt to incorporate methods from different empathetic dialogue paradigms into one model and aim to achieve richer results. Specifically, we adapt our ModMoEL model to work with the PEC dataset (Zhong et al., 2020) while leveraging rich persona information about speakers. We also explore the potential of adding additional auxiliary signals to improve empathetic natural language generation through experimentation with meaningful variations of MoEL.

## 1 Introduction

Empathy, the ability to understand and reflect on another person's emotions, is a natural human trait that has remained a persistent challenge in dialogue systems. Empathetic response generation involves accurately detecting and tracking the user's emotional state and generating an appropriate response while remaining engaging and avoiding generic or repetitive responses. The ability to generate high-quality empathetic dialogue has numerous practical applications, including improving customer service interactions, enhancing the capabilities of open-domain conversational AI, and even positively impacting the emotional states of users.

There has been significant progress in empathetic response generation through various research approaches in recent years. This progress has been partially driven by the availability of large empathy-focused datasets like the EMOTIONAL SUPPORT CONVERSATION (ESConv, Liu et al., 2021) dataset and EMPATHETICDIALOGUES (ED, Rashkin et al., 2018). The ED dataset, in particular, has been instrumental in developing the state-of-the-art empathy approaches such as MoEL (Lin et al., 2019) and MIME (Majumder et al., 2020). Some researchers have explored persona-based approaches (Zhong et al., 2020), while others have implemented systems to hierarchically balance factors such as dialogue act and emotion to generate more empathetic responses (CoMAE, Zheng et al., 2021).

In our deep dive into various approaches to improving empathetic response generation, we explored a range of options to build upon existing research. After careful consideration, we selected the MoEL model as the foundation for our system design due to its simplicity. MoEL has achieved state-of-the-art results for empathetic response generation by using a model design (see Figure 1) that employs a separate decoder for every emotion and combines their outputs with the help of a meta-listener. This intuitive architecture allows for flexibility and the opportunity to meaningfully incorporate additional context-related signals.

The MoEL model also uses conversational context to model the distribution of emotions and fuse the response generated by individual emotion decoders. To improve the performance of MoEL, we provided it with information about the responder's persona to help it better understand how various personas may express emotions empathetically. We also attempted to introduce additional context annotations, such as dialog acts and communication mechanisms from CoMAE (Zheng et al., 2021), to further allow MoEL to learn about the interplay of various contextual factors. However, we were

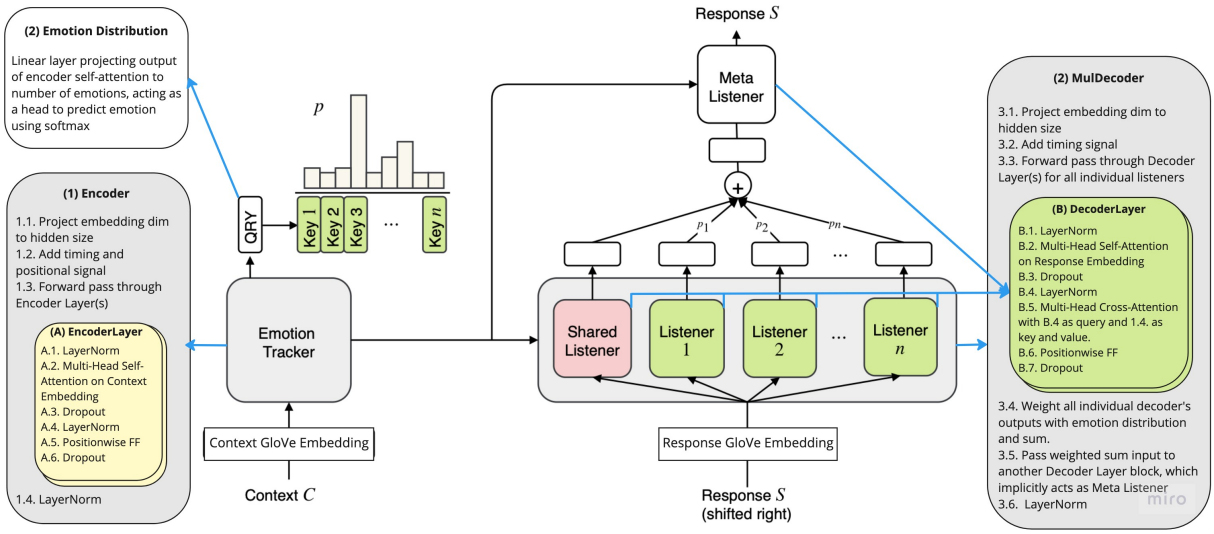


Figure 1: MoEL architecture by (Lin et al., 2019) annotated with transformer architecture modules by blue arrows.

unable to fully pursue this due to time and resource constraints.

## 2 Literature Review

Empathy has proven to be an important factor in the performance of dialog systems when given emotional user input (Zhou et al., 2018; Li et al., 2017; Zhou and Wang, 2017; Huber et al., 2018; Huang et al., 2020), and researchers have attempted diverse approaches to generating empathetic dialogue.

### 2.1 Empathetic Datsets

There is a body of research that focuses on integrating emotional support into dialog systems. Several large empathy-relevant datasets such as the EMOTIONAL SUPPORT CONVERSATION (ESConv, Liu et al., 2021) dataset and EMPATHETICDIALOGUES (ED, Rashkin et al., 2018), have been made available and provide valuable insights and data frameworks for training empathetic language generation models. ED is also the first dataset of its kind comprising 25,000 conversations in 32 different emotions. The results of training conversational models on this dataset in the listener role showed higher levels of empathy compared to models trained on non-empathetic datasets.

### 2.2 MoEL: Empathetic Conversation by Mixing of Emotions

The MoEL model (Lin et al., 2019) achieved state-of-the-art results by training on the ED dataset (Rashkin et al., 2018) and using emotion-specific decoders, each of which was tuned to a specific emotion. The meta-listener then combines the responses from these decoders based on the emotion distribution in the context text. The effectiveness of MoEL confirms that conversational models trained on empathetic datasets demonstrated stronger empathy compared to models trained on non-empathetic datasets.

### 2.3 Persona-based Conversation

Some efforts focus on the exploration of how persona affects empathetic responding. Persona embeddings have been successfully implemented by Li et al. (2016) that result in persona consistency. Inspired by Mazaré et al. (2018), Zhong et al. (2020) created a novel large-scale multi-domain PERSONA-BASED EMPATHETIC CONVERSATION (PEC) dataset.

PEC contains over 350,000 persona-based empathetic conversations from two subreddits (*happy* and *offmychest*) where users often share empathetic responses to personal posts. In addition, the dataset includes up to 100 self-describing "persona sentences" per user, making it unique and valuable as it is not only crowd-sourced and empathetic, but also includes persona information about the speakers. Furthermore, the PEC dataset was constructed with strict data collection rules and criteria for sentence length, subjectivity, and content, to select qualified persona sentences. Each user is represented by no more than 100 persona sentences. Two sample conversations are shown in Figure 2.

The authors also proposed CoBERT, a BERT-based response selection model that achieved state-of-the-art performance on the PEC dataset. More

significantly, the authors conducted the first empirical study on the impact of persona on empathetic responses, which demonstrated a positive correlation between persona and empathy.

### 2.4 Dialog Act and Communication Mechanism

In the field of empathetic response generation, it has been suggested that considering dialog acts and communication mechanisms can improve a system's ability to comprehend emotions. Welivita and Pu (2020) proposed a taxonomy of dialog act (DA) for empathetic conversations. Integrating dialogue intent modeling and neural response generation has been shown to improve the response quality of chatbots and make them more controllable and interpretable. Sharma et al. (2020) characterized text-based expressed empathy into three communication mechanisms based on the theoretical definition of empathy: emotional reaction (ER), interpretation (IP), and exploration (EX). Combining these concepts, Zheng et al. (2021) proposed a system that hierarchically balances dialogue act, communication mechanism, and emotion factors to create an empathetic language generation model.

## 3 Methodology and Architecture

This project was organized into four major components. The first was processing the PEC dataset to create persona embeddings from the persona sentences. The second was training an emotion classifier on the ED dataset (Rashkin et al., 2018) and using it to annotate the context utterances in PEC with emotion labels. The third involved gaining an understanding of the MoEL model and improving its performance without considering persona. Finally, we dove deep into the MoEL model and modified its architecture to consider persona embeddings during response generation.

### 3.1 Processing PEC

The PEC dataset is made available through the Huggingface dataset API (Wolf et al., 2019) and contains 281,000 training instances, 36,000 validation instances, and 35,000 test instances. To further enhance the quality of data, we applied the following four criteria to filter the PEC dataset:

- The context must have more than one speaker.
- Context sentences must have at least 20 words in total.
- The respondents' personas must have at least 25 sentences.
- The response utterance must have at least 10 words.

This left us with 83,000 training, 11,000 validation, and 11,000 test examples.

#### 3.1.1 Creating Persona Embeddings

To generate persona embeddings from concatenated persona sentences, we used large pre-trained language models that can provide contextualized embeddings. However, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) were not sufficient for this task because they have a maximum input size of 512 tokens, whereas 100 persona sentences for each speaker typically exceeded this limit by several orders of magnitude. While it was possible to truncate the persona corpus to 512 tokens and extract embeddings using BERT-based models, this approach resulted in the loss of persona information.

As a result, we used Longformer (Beltagy et al., 2020) to extract persona embeddings because it can handle input of up to 4,096 tokens. Longformer, which is based on RoBERTa, has a similar input preprocessing format and employs global and local context attention mechanisms that are well-suited for the persona corpus, as it is desirable for every token in a sentence to attend to each other in a local context, and all sentences in persona corpus for a user to attend to each other in a global context.

### 3.2 Augmenting PEC dataset with emotion labels

Since the PEC dataset does not include emotion labels and our MoEL-based model requires emotion information, we augmented the dataset by creating a BERT-based emotion classifier. This classifier was trained on the EMPATHETICDIALOGUES dataset (Rashkin et al., 2018) and was able to predict the emotion label for a given conversation context. The model achieved an accuracy of 0.350 and an F1 score of 0.349 on the ED test data.

The emotion classifier then took context data from the PEC dataset and predicted one of 32 ED emotion labels, which were used to augment the PEC dataset and enable the integration of persona data into our model.

| | happy | offmychest |
|---|---|---|
| Conversation | Celebrating 43 years of marriage with the love of my life. | Worried. Am I becoming depressed again? Please don't leave me. Is everything okay? You don't seem yourself. |
| | She looks very young for someone who has been married 43 years. That must surely put her in the 63-73yr age range?! | I'm living these exact words. |
| | I just turned 61, thanks! | I hope everything works out for you. I'm trying not to fall apart. |
| | I hope I look that young when I'm 61! You guys are too cute, congratulations :) | Me too. If you ever want someone to talk to my messages are open to you. |
| Persona | I took an 800 mg Ibuprofen and it hasn't done anything to ease the pain. | I think I remember the last time I ever played barbies with my litter sister. |
| | I like actively healthy. | I have become so attached to my plants and I really don't want it to die. |
| | I want a fruit punch! | I'm just obsessed with animals. |

Figure 2: Example of conversations and personas sentences from PEC

| personas (sequence) | context (sequence) | context_speakers (sequence) | response (string) | response_speaker (string) |
|---|---|---|---|---|
| [ "i have a roku tv that came with a shitty basic remote .", "i 'd shit a lil bit on… | [ "found out this morning i got a job promotion ! : )" ] | [ "HeWentToJared91" ] | "whilst popping ?" | "Evref" |
| [ "i was only judging based on act 1 .", "i actually run in altra torin 3 's , so thes… | [ "found out this morning i got a job promotion ! : )" ] | [ "HeWentToJared91" ] | "you look like a nice person : we 're proud of you , and i bet you earned that… | "tylock" |

Figure 3: The PEC data schema

| 16 emotion groupings | | | |
|---|---|---|---|
| 1 | afraid | terrified | anxious |
| 2 | angry | annoyed | furious |
| 3 | guilty | embarrassed | ashamed |
| 4 | prepared | anticipating | apprehensive |
| 5 | hopeful | confident | |
| 6 | grateful | content | |
| 7 | excited | joyful | |
| 8 | disappointed | disgusted | |
| 9 | proud | impressed | |
| 10 | sad | devastated | |
| 11 | sentimental | nostalgic | |
| 12 | trusting | faithful | |
| 13 | caring | | |
| 14 | jealous | | |
| 15 | lonely | | |
| 16 | surprised | | |

Table 1: 32 emotions grouped into 16 emotion groupings used to label PEC data

#### 3.2.1 Emotion Classification Refinement

However, since some of these emotions overlap, we hypothesized that by grouping similar emotions (e.g., "angry" and "furious"), the system could achieve higher accuracy in emotion identification without compromising response quality. We reduced the number of distinct emotions to 16 by grouping them as shown in Table 1:

Using these groupings, we trained a separate emotion classifier that only predicted these 16 emotion groupings. This classifier achieved an accuracy of 0.466 and an F-1 score of 0.458 on test data. We then trained MoEL with PEC data augmented with these emotion labels. By grouping the emotions in this way, we effectively reduced the number of listeners in MoEL, which resulted in improved performance, as discussed in Section 5.2.

### 3.3 Understanding MoEL

MoEL's innovation was in integrating a mixture of empathetic "expert listeners" that were individually tuned to different emotions in order to generate an empathetic response based on the distribution of the user's emotion. The design is based on a standard transformer (Vaswani et al., 2019) containing one encoder and several parallel horizontally stacked decoders that collectively feed into another decoder.

This translates to each individual emotion listener being a transformer decoder, the outputs of which are fused by the "meta-listener" decoder to generate a final response. An additional head on top of encoder outputs, used to predict emotion labels, allows MoEL to learn how to model the emotion distribution of the context. This probability distribution over emotions acts as weights for the individual emotion listeners, and the weighted sum of these listeners becomes the input for the meta-listener along with the output of the transformer encoder. This design is shown in Figure 1.

### 3.4 MoEL architecture modification for integrating Persona

MoEL's design provides multiple opportunities to incorporate persona embedding as input, but selecting an effective method has proven challenging.

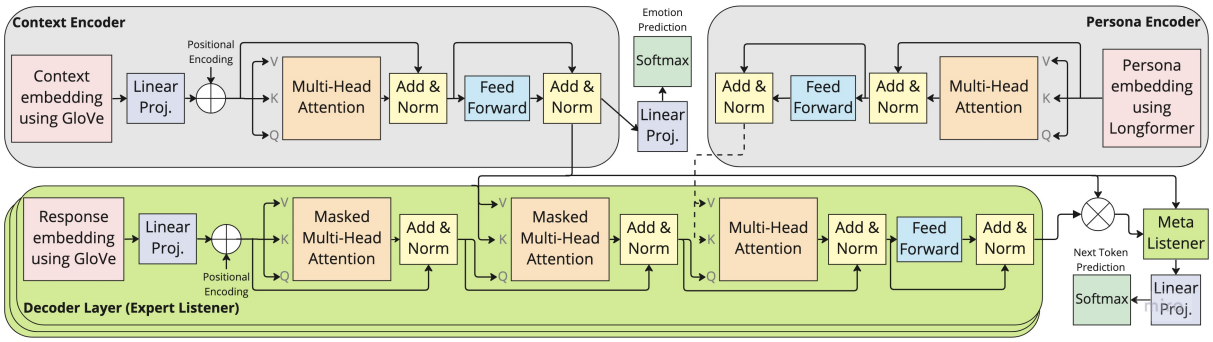


Figure 4: Modified MoEL architecture for empathetic dialogue generation with persona. This transformer-based architecture consists of an encoder with self-attention for both the context text and persona embeddings, $n + 1$ decoders for $n$ distinct emotions in the dataset, and a meta-listener decoder that accepts emotion distribution weighted sum of expert listeners to generate empathetic dialogue. Each individual decoder module includes self-attention on the right-shifted target text, cross-attention with the context encoder output, followed by another layer of cross-attention with the persona encoder output.

Some possibilities include inputting the persona into each expert listener during decoding, providing the persona solely as a complimentary input to the meta-listener, or creating a new persona listener parallel to the expert listeners that could be fused with them before being passed to the meta-listener. We also explored the possibility of fusing the persona embedding with the context input before the encoder step, which would be fed into the attention mechanism during decoding.

During the design process, we encountered several issues, including the extent of modifications to make to MoEL and how to prevent any additional mechanisms from degrading the quality of the existing training architecture. Our initial idea of fusing persona embeddings only in the meta-listener was appealing due to its simplicity and the fact that it would have allowed most of MoEL's architecture to remain unchanged. However, integrating the persona through an additional multi-head attention unit within each decoder layer would allow each decoder to learn how various personalities express respective emotions differently. With large corpus of natural conversations, we have the opportunity to train each emotion's decoder to be conditioned on the persona. Since a common DecoderLayer module is used to represent every expert listener, shared listener, and meta listener (see Figure 1), our major changes focused on this module, and the downstream effects it had.

#### 3.4.1 Final architecture design

To make persona available to the model at each step of the training process, the persona embedding is bundled with the dialogue context, emotion, and target embeddings as each training "batch" is created. This required code modifications throughout the model design in order to allow information to flow. The pre-processed persona embeddings are passed through the encoding layer unchanged and retrieved during decoding for use in multi-head attention, both at the level of the expert decoders and the meta-listener. The full architecture is explained and illustrated in detail in Figure 4, with the code being available publicly on github.

#### 3.4.2 Attempted Variations

In addition to the presented architecture, we experimented with several variants to determine their effectiveness.

1. We attempted to use the output of the persona encoder as the query and the output of the encoder-decoder cross-attention as the keys and values. However, this resulted in the model failing to converge and falling into the sub-optimal minima of predicting the end-of-sentence token as the first token in most cases.

2. We also tried performing cross-attention with the output of the persona encoder before performing cross-attention with the output of the context encoder.

## 4 Results

### 4.1 F-1 Scores from Emotion Classification

The two emotion classifiers that used 32 and 16 emotions respectively, were able to achieve F-1 scores of 0.349 and 0.458. Given the large number of emotions considered in MoEL, we believe that

| Models | Dataset | Accuracy | BLEU_b | BLEU_g | METEOR_b | METEOR_g | Perplexity | sentiment_b | sentiment_g |
|---|---|---|---|---|---|---|---|---|---|
| ModMoEL + persona primary | PEC_16 | 0.243 | 2.020 | 2.100 | 0.104 | 0.108 | **143.176** | 0.028 | 0.088 |
| ModMoEL + persona | ED_16 | 0.260 | 1.910 | 2.080 | 0.100 | 0.109 | 143.479 | -0.004 | 0.050 |
| ModMoEL + persona | PEC_16 | 0.233 | **2.170** | **2.110** | **0.106** | **0.111** | 144.703 | 0.010 | 0.052 |
| ModMoEL | PEC_16 | **0.285** | 1.750 | 1.530 | 0.096 | 0.103 | 146.301 | **0.111** | **0.153** |
| | | | | | | | | | |
| MoEL (baseline) | ED_32 | **0.346** | **2.330** | **2.300** | **0.120** | **0.125** | **73.337** | -0.025 | 0.049 |

Figure 5: Test performance for MoEL model variants. “b” indicates beam search decoding and “g” indicates greedy decoding. “Persona primary” variant does cross-attention with persona before context whereas model name with just "persona" follows the architecture from Figure 4.

these scores are sufficient for augmenting the PEC dataset with labels for training MoEL.

### 4.2 Evaluation Metrics

In their original evaluation, (Lin et al., 2019), used two evaluation metrics to assess the performance of MoEL: BLEU scores to compare generated responses against the gold standard responses, and human ratings of the generated responses. While they also tracked the accuracy of the model’s emotional prediction, it was not used in the evaluation. In this project, we were unable to obtain human ratings of the generated responses due to logistical constraints. As a result, we introduced several additional metrics to evaluate the quality of our model.

In addition to BLEU scores and accuracy, we also tracked METEOR scores (Banerjee and Lavie, 2005), sentiment increase analysis, and perplexity for our model variations. METEOR scores can be calculated automatically like BLEU scores but are fundamentally more robust since the harmonic mean of weighted precision and recall is calculated. Originally designed for machine translation evaluation, it is based on unigram matching between utterances. Sentiment analysis was conducted using the VADER model(Hutto and Gilbert, 2014), which was designed to measure the sentiment intensity of social media posts. We applied this to our model as an analogous measure of the emotional expressiveness of a response. We also tracked the perplexity of various iterations of our model.

### 4.3 Evaluation Results

Our evaluations showed that the baseline MoEL model outperformed our modified versions on most of the metrics. This may be due to the difference in the training datasets: MoEL was designed and trained using the EMPATHETICDIALOGUES dataset, while ModMoEL was trained on the PEC dataset. A fundamental difference between ED and PEC are that ED was collected in an orchestrated setting whereas PEC has been collected from human conversations on Reddit. Therefore, it’s expected that PEC will be a harder challenge for any language model than ED. We also attempted to train ModMoEL on ED with the addition of persona information, but the results were suboptimal, likely due to incongruities between the persona data and the dialogues in ED.

Our comparison of different variations of MoEL yielded mixed results. The addition of persona embeddings improved the BLEU, METEOR, and perplexity scores of the model, but resulted in a decrease in emotion prediction accuracy and sentiment increase scores. While it is hard to predict the impact of these results on human testing with certainty, the reduction of sentiment values conflicts with our hypothesis and warrants further investigation in future testing. See Figure 5 for detailed results.

## 5 Future Work

We extracted persona embeddings from the PEC dataset and the training of our model, which suggests that the inclusion of persona information in the decoder of the MoEL model could help generate more empathetic and fluent responses in various scenarios. After this, we explored how the incorporation of dialogue act labeling and communication mechanisms can also help the ModMoEL further increase the capability of generating empathetic responses. The original PEC dataset contains emotion labels, dialog contexts, and responses, but does not include dialogue act and communication labels. In order to incorporate the dialog act and communication mechanism, we utilized the CoMAE corpus, which is annotated with dialog act and communication mechanism by extracting the texts with dialog act and communication labels. We intended to com-

bine dialog act, communication mechanism, and emotion along with context from PEC dataset but found out that the two corpora have completely different data formats, which results in a much higher complexity of data processing task. As a result of this, a future line of inquiry could include addressing these issues to further strengthen the empathetic response generation ability of ModMoEL.

### 5.1 Human Evaluation

True empathy involves actively sharing the emotional state of another person. Needless to say, this is an impossible task for a dialogue system, which can only mimic the way that an empathetic person might respond. Likewise, perhaps the most important component of the quality of a chatbot's empathetic response is how the user perceives it. We were, unfortunately, unable to implement human testing in the scope of this project due to limitations in time and budget, and thus lack core data to fully evaluate our model results against its progenitors and competitors.

## References


Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150v2*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805v2*.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225.*

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692v1*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. *arXiv preprint arXiv:2010.01454*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

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

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441v1*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2019. Attention is all you need. *arXiv:1706.03762*.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: a multi-factor hierarchical framework for empathetic response generation. *arXiv preprint arXiv:2105.08316*.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. *arXiv preprint arXiv:2004.12316*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xianda Zhou and William Yang Wang. 2017. Mojitalk: Generating emotional responses at scale. *arXiv preprint arXiv:1711.04090*.

## 6 Division of Labor

- **Navjot Singh**: PEC dataset processing, Persona Embedding Creation, Setting up infra on GCP, MoEL code deep dive, implementation and ideation for incorporation of persona embeddings in MoEL
- **Samatha McAlevy**: Literature Review, Understanding MoEL, MoEL code deep dive, implementation and ideation for incorporation of persona embeddings in MoEL
- **Weirui Peng**: Literature Review, PEC dataset processing, Persona Embedding Creation, Exploring CoMAE and collecting its annotations
- **William Cui**: Literature Review, Emotion Classification Module, Setting up infra on GCP, Understanding MoEL, debugging persona incorporation in MoEL

### 6.1 Individual Contribution

I am happy with the way the whole group has contributed to the project. It has been a pleasure to work in an environment where everyone contributes.

Individually, I have been closely involved in the project throughout each phase. The initial brainstorming for the project idea was lead by Weirui and me, as Sam and William joined the wagon later. Everyone contributed to literature review, where I especially focussed on the ESConv paper that I presented in class as well.

Once we identified a potential project, I worked with Weirui to extract persona embeddings. After looking deep into the realms of pre-trained language models, I discovered Longformer and dove into its nuances to understand global and local attention masks. Because of its size Longformer is time and resource intensive, and I thus paralleled the job on cloud as I got familiar with GCP.

While Sam and William had started exploring MoEL, I joined the wagon to strengthen our force for introducing persona into MoEL. Learning that aspect was a very demanding but rewarding process as it involved making mistakes and getting lost deep into MoEL's code to make sense of what was happening. Overall, this project has been a great learning experience.

## 7 Appendix

See Figure 6 for detailed information about response generation for ModMoEL.

| Emotion | Context | Baseline | With Persona |
|---|---|---|---|
| **hopeful_confident** | my boyfriend said something supportive and it made my day | i 'm so happy for you ! i hope you have a good day ! | i ' m so happy for you ! i ' m glad you ' re feeling better ! |
| **sad_devastated** | i want a hug | i 'm sorry you 're going through this . i hope you 're feeling better . | i 'm sorry you are feeling this way . i 'm here to listen to you . |
| **lonely** | i do n't fit in | if you need someone to talk to , i 'm here . | you 're not the only one . you 're not alone . |

Figure 6: Sample response generation for ModMoEL when trained with and without persona. Differences are detectable but need large-scale human evaluation to determine effectiveness.