Research Group Social Computing
Department of Informatics
Technical University of Munich

Empathetic Dialogue Agent trained with Reinforcement Learning

Ananta Bhattarai ✉, **Sophie Schoen** ✉, and **Viviane Rehor** ✉

Department of Informatics, Technical University of Munich

✉ ananta.bhattarai@tum.de
✉ sophie@schoen-online.com
✉ viviane.rehor@tum.de

July 31, 2021

Abstract — We present a multi-turn open-domain empathetic dialogue agent trained with reinforcement learning. In this work, we first fine-tune OpenAI GPT-2 model on the EMPATHETICDIALOGUES dataset in a supervised setting. We then train reward models to predict empathetic, relevant, and fluent responses using two different approaches: 1) Using human feedback to select a better response from two responses for the given context. 2) Using validation set gold response as the better response for the given context. Finally, we fine-tune the supervised policy using the reward model. We show that the performance of the fine-tuned model using reinforcement learning is increased with respect to automated metrics. In addition, our fine-tuned model is preferred by humans over the supervised model and state-of-the-art DialoGPT in three aspects of evaluation: empathy, relevance, and fluency.

1 Introduction

The usage and performance quality of dialogue systems has increased in various fields due to the rapid advances of machine learning techniques. The generative language models on different natural language tasks such as reading comprehension, text summarizing, machine translation, and question answering have achieved state-of-the-art performance without task-specific training [12]. Reinforcement learning (RL) has also been used widely in natural language tasks [13] [10] [5]. Therefore, we want to combine the success of the generative language models with RL to build a multi-turn open-domain empathetic dialogue agent. Our main goal is to create a dialogue system that can react to the user's questions and concerns by imitating the way humans communicate with each other. Human language includes a lot of emotions, different meanings, and expressions. The important aspect of human communication is empathy. Therefore, human-like dialogue models should possess the capability to understand and share the feelings of the conversation partner by being complimentary, attentive, and compassionate [16] [9].

Our long-term goal is to build a virtual dietary advisor that can advise people by helping and motivating them to be healthy. For the well-being of the user, it is crucial to know the emotional state and communicate in the right empathetic tone. To this end, we developed a human-like dialogue agent that can understand the user's emotion in the conversation and responds accordingly.

Our approach is similar to the one outlined in [23]. They fine-tuned OpenAI GPT-2 for text continuation and summarization using reinforcement learning. The reward model is built by asking humans questions. We use a similar method; however, we also propose an alternative approach to train the reward model in case of less human-annotated data. For a more detailed description of our procedure, see Section 3. We compare both approaches in terms of automated metrics and human evaluation. Our results and evaluation are outlined in Section 4.

2 Related work

In recent years, there has been substantial progress in natural language processing. With a vast increase in computing power and the availability of large amounts of Internet text data, training of large-scale generative language model in an unsupervised setting has been possible [12] [2] [11] [7]. Jeremy Howard and Sebastian

Ruder show that such pre-trained language models often can be fine-tuned on many different supervised datasets to achieve state-of-the-art performance [4]. However, for some tasks such as reading comprehension, machine translation, question answering, and summarization, generative pre-trained language model, GPT-2, trained on the dataset of 8 million web pages achieved state-of-the-art performance without task-specific training [12]. In our work, we use a pre-trained GPT-2 model as our baseline.

Recent works have attempted to build human-like open-domain conversation models [1] [22]. Yizhe Zhang and his co-authors show that DIALOGPT (dialogue generative pre-trained transformer), trained on 147M conversation extracted from Reddit comment, achieves a performance close to human with respect to their evaluation metrics [22]. However, their evaluation is limited to single-turn dialogue settings. Similarly, Daniel Adiwardana and his team propose Meena, a multi-turn open-domain chatbot with 2.6B parameters trained on conversations from public domain [1]. They claim the full version of Meena scores 79% on Sensibleness and Specificity Average (SSA) metric, close to human-level SSA of 86%. Despite the success, the limitation of their work is that they do not cover the empathy aspect of the model. Hannah Rashkin and his team propose a new benchmark for empathetic dialogue generation [14]. In addition, their work proposes an EMPATHETICDIALOGUES, a dataset of 25k conversations based on emotional situations, and shows that dialogue models trained on EMPATHETICDIALOGUES tend to produce more empathetic dialogues, compared to models trained just on public Internet conversations. Therefore, we first fine-tune our baseline generative language model, GPT-2, on the EMPATHETICDIALOGUES dataset in a supervised setting. We further fine-tune it with RL to create a multi-turn empathetic open-domain chatbot.

Researchers have applied RL to dialogue tasks in the past [8] [6]. Similarly, human feedback has also been used as a reward to train and boost the performance of the dialogue models [6] [3] [21]. Jiwei Li and his team improve the question-answering ability of the bot in an RL setting where the bot learns from the human feedback following the bot's generated response [8]. The reward they use is binary (1, if the bot answers correctly and 0 otherwise). However, the binary reward cannot be used in the open-domain dialogue generation task since there is no correct or incorrect answer. Natasha Jaques and her co-authors use modified Q-learning methods to model human preferences in dialogue [6]. They design several intrinsic reward functions and apply RL to optimize the reward functions. Designing a good reward metric representing human preferences is difficult since there are many components in the human conversation. Daniel M. Ziegler and his co-authors use human judgment as a reward [23] [19]. They train the reward model by asking humans to pick one response from N responses for the given context. Then, they apply RL to optimize that model. The results are promising for two natural language tasks: stylistic continuation summarization. Therefore, we closely follow the work of Daniel M. Ziegler and apply a similar method for our empathetic open-domain dialogue task. In addition to that, we also propose a slightly different method for the reward model training in case of less human-annotated data.

3 Methods and experiment details

3.1 Dataset and data preparation

To fine-tune OpenAI GPT-2 model for the dialogue task, we use EMPATHETICDIALOGUES dataset [14]. It contains around 25 thousand dialogs, which are conversations between humans. In each dialogue, the speaker talks about an emotional situation, and the listener responds accordingly. Speaker and listener then communicate with each other in alternating turns. They exchange up to 8 turns. We choose the EMPATHETICDIALOGUES dataset over other commonly used dialogue datasets because the responses of language models after fine-tuning on EMPATHETICDIALOGUES are more empathetic [14]. Moreover, the test set also could be used to evaluate our model in terms of human-like, as conversations in the dataset represent real human interactions.

We pre-process the data to create *(context, response)* pairs for supervised training. A context contains previous conversations utterances (up to 4) concatenated with *<SOC>* token. A *<SOC>* token represents the next person's turn. A response contains a reply to a context. Appendix A explains the data pre-processing step with

	Annotator 1		Annotator 2		Annotator 3	
	Agreement (%)	Cohen's kappa	Agreement (%)	Cohen's kappa	Agreement (%)	Cohen's kappa
Annotator 1	-	-	68	0.35	73	0.45
Annotator 2	68	0.35	-	-	73	0.46
Annotator 3	73	0.45	73	0.46	-	-

Table 1 Confusion matrix showing agreement percentage and Cohen's kappa score between the annotators

an example. In total, we generate around 62,000 *(context, response)* pairs for the training. We add *</s>* token at the end of a context to indicate the start of response and tokenize training samples using Byte-Pair Encoding (BPE). We set the maximum length of context tokens and response tokens to be 100 each. For contexts and responses less than 100 tokens, we pad them with a padding token on the right. We also shuffle the training examples before feeding them into the language model.

3.2 Data annotation

After fine-tuning OpenAI GPT-2 model on the EMPATHETICDIALOGUES dataset in a supervised setting, we use the validation set conversations to create contexts. The contexts are fed into the language model to obtain two responses for each context using different policies. The *(context, response1, response2)* pairs are then presented to us to pick the better response from 2 responses for the given context. The data are distributed randomly among the three of us. Since we want our model to produce responses that are more empathetic, relevant, and fluent, we use the same annotation criteria from [14]:

- **Empathy/Sympathy:** Which response shows an understanding of the feelings of the person talking about their experience?
- **Relevance:** Which response seems appropriate to the conversation? Was it on-topic?
- **Fluency:** Which response seems more fluent and accurate?

We have 6 options to pick from in the data annotation process:

- **1:** Left response is better.
- **2:** Right response is better.
- **3:** Both responses are equally good.
- **4:** Both responses are not relevant.
- **5:** Some parts of left response is better.
- **6:** Some parts of right response is better.

Appendix B shows our data annotation interface. In the case of option **3**, we let the machine randomly choose one response. Similarly, when both responses are not relevant (case **4**), we discard the *(context, response1, response2)* pair. We also observed that in some cases some parts of one response are better than the other. In such cases, we select **5** or **6**, copy the good parts of the left or right response, and pick the new copied left or right response as the better one.

We evaluate the quality of annotations and interrater reliability using agreement percentage and Cohen's kappa score between us. During the data annotation process, 100 *(context, response1, response2)* pairs were the same for three of us. Therefore, we use those 100 data to compute statistics. To have a good insight into the statistics, we only pick from options 1 and 2 and discard the rest while annotating the same 100 data. Table 1 shows the agreement percentage and Cohen's kappa score between us. The average agreement and average Cohen's kappa

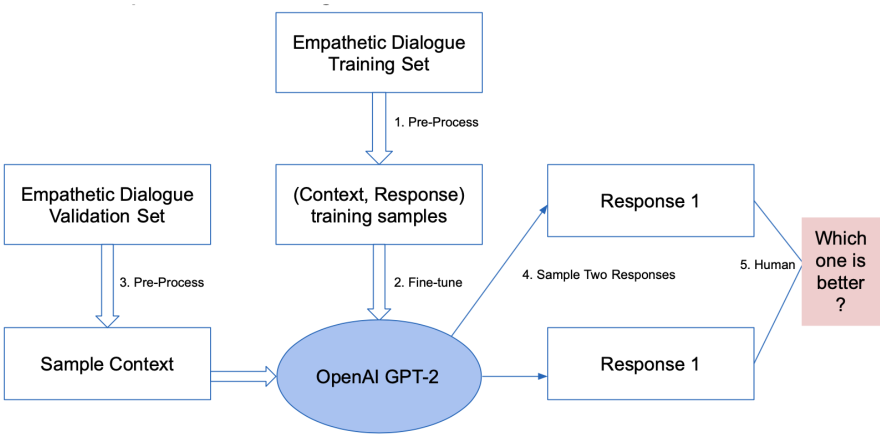


Figure 1 Supervised model training and inference.

score are 71% and 0.41, respectively. The result is reasonable considering the dialogue task is open-domain, and the annotation process is subjective. We collect around 2200 annotated data for our reward model training. The annotation process of one sample pair took about 45 seconds on average.

3.3 Models

3.3.1 Pretrained models

We start with OpenAI GPT-2 model that is pre-trained to predict the next token given the context [12]. During the evaluation of pre-trained models, we pad the context with ten high-quality *(context, response)* pairs, created in our data pre-processing step, on the left and sample the response. Pre-trained models without any fine-tuning on the EMPATHETICDIALOGUES dataset will be referred to as "zero-shot" models hereafter.

3.3.2 Supervised fine-tuned models

We then fine-tune the OpenAI GPT-2 model on *(context, response)* pairs, created from the EMPATHETICDIALOGUES dataset. The model is trained to generate a response for the given query followed by </s> token. For the first 146 steps, we use the learning rate of 1e-5 and batch size of 20. We then lower the learning rate to 7e-6 and increase the batch size to 40. The reason to use different batch sizes is that we had access to a more powerful machine in later training steps. We select the model that performs best on the validation set.

We also use our supervised fine-tuned model to collect *(context, response1, response2)* pairs for the data annotation. For all purposes, we use a temperature value of 0.7 while sampling from the supervised fine-tuned models unless explicitly stated otherwise. Large values of temperature favor rare tokens but might assign a high probability to the wrong word [1]. In contrast, smaller values of temperature favor common words such as prepositions [1]. Therefore, to have a balance between exploration and exploitation, we use a fair temperature value. Fine-tuned models in a supervised setting will be referred to as "supervised" models or baselines hereafter. Figure 1 shows the supervised model training and sampling steps.

3.3.3 Reward models

To train our reward models, we take our supervised model as a prior, then add a random linear layer on top of the supervised model to output a reward value. The reward model is trained to understand human preference in two responses for a given query. We train the reward model using the loss

$$loss(r) = \mathbb{E}_{(x,y_1,y_2,b)\sim S}\left[\log \frac{e^{r(x,y_b)}}{e^{r(x,y_1)} + e^{r(x,y_2)}}\right] \tag{1}$$

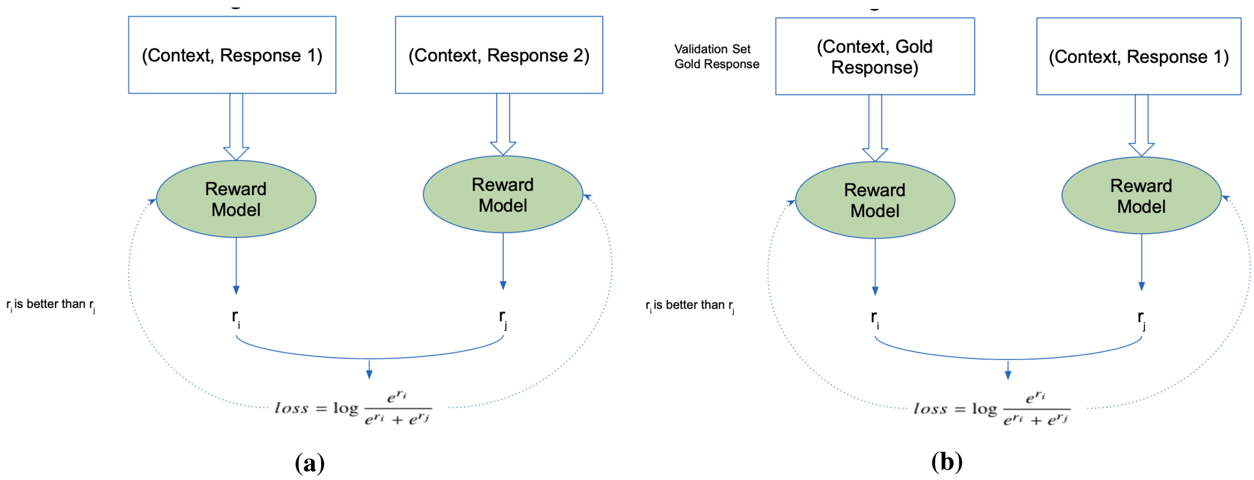


Figure 2 Reward models training. (a) Reward-1 training, (b) Reward-2 training

where S is our annotated dataset, x is *context*, y_1 is *response1* for *context* x, y_2 is *response2* for *context* x, b is option we select in the data annotation process and $r(x, y)$ is reward value given by the reward model for *response* y, given *context* x. At the end of the training, we normalize reward model output such that the responses from dataset S achieve mean 0 and variance 1. We train the reward model for just one epoch to avoid over-fitting. We use a learning rate of 1.5e-5 and a batch size of 4.

We observed that the reward model trained with less human-annotated data performs poorly in understanding the human preferences; see section results and evaluation. Therefore, we propose an alternative method to fit the reward model. We generate a new dataset D which contains (x, y_1, y_2, b) pairs where y_1 is the gold response from the validation set for the given *context* x, y_2 is the sampled response from the supervised model for the given *context* x and b is 1. We train the reward model on dataset D, where gold response expresses the human preference. As the conversations in EMPATHETICDIALOGUES dataset are real human interactions, we believe that training the reward model in this way is equivalent to giving human feedback. We use the same training hyperparameter as mentioned above. Reward model trained on dataset S and dataset D will be referred to as "Reward-1" and "Reward-2" respectively hereafter. Figure 2 shows the reward models training process.

3.3.4 RL fine-tuned models

We want to update the initial policy $\pi = \rho$ of the supervised model to optimize the reward model. We add a KL divergence term $\beta \mathrm{KL}(\pi, \rho)$ to the reward and optimize the modified reward using RL. The modified reward R is given by

$$R(x, y) = r(x, y) - \beta \frac{\pi(y|x)}{\rho(y|x)} \tag{2}$$

where R(x, y) is the modified reward for *response* y given *context*. The penalty term restricts updated policy π to move too far from the initial policy ρ. We set the value of β to 0.03 for our experiments.

To train the policy π, we use the PPO2 version of Proximal Policy Optimization from [17]. We set $\gamma = 1$ and default values for the other parameters. We do four PPO epochs per batch with one minibatch each and run the training 1000 episodes. We use the learning rate of 2e-6 and batch size of 5. Similarly, we use a temperature value of 0.7 while sampling from the RL fine-tuned models unless explicitly stated otherwise. Policy training to optimize Reward-1 and Reward-2 will be referred to as "Policy-1" and "Policy-2" respectively hereafter. Figure 3 shows the policy training steps.

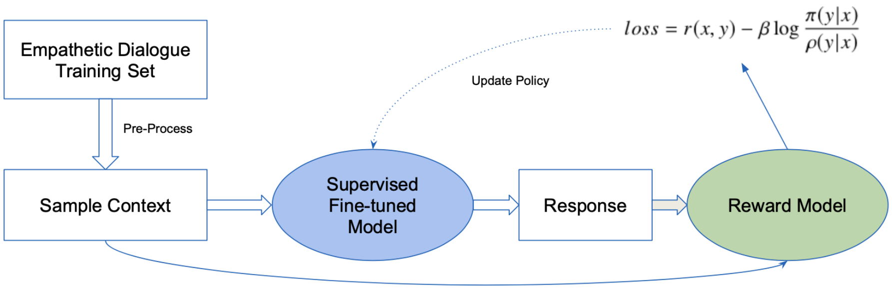


Figure 3 Policy training.

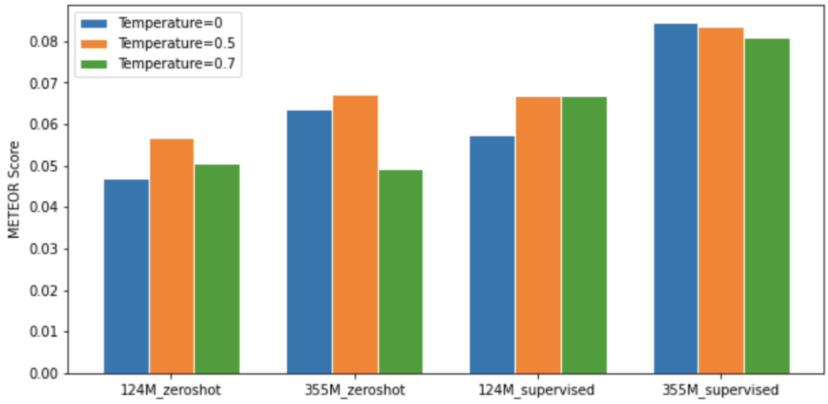


Figure 4 METEOR score vs different GPT-2 models. The numeric prefix on the x-axis represents the number of parameters in the model.

3.4 Evaluation metrics

We evaluate our final models on different automated metrics and decide for the best performing models to then again compare them by human evaluation with the same criteria as in section 3.2. Human evaluation is important, as automated metrics don't always correlate with human judgments of dialogue quality. Nevertheless, they are much more convenient to get on a large amount of data, and our final human evaluation in section 4.3.5 shows that the models we pick using automated metrics perform well. One of our automated metrics is the METEOR score, widely used in machine translation. The other metrics are based on the implementations of Florian von Unhold's paper and therefore inspired by Saleh's et al. (2019) work [20] [15]. For our evaluation, we use the EMPATHETICDIALOGUES test set to create contexts.

4 Results and evaluation

4.1 Supervised baselines

To ensure that our supervised model is a strong baseline, we compare two versions (small and medium GPT-2) of supervised models with Zero-Shot models. We use a validation set of the EMPATHETICDIALOGUES to create contexts and sample one response to each context. We then compare the sampled response with the gold response in the validation set using the METEOR score. Although the METEOR score is not a proper metric to evaluate open-domain dialogue models, it provides an insight into how close the sampled response is to the gold response. Furthermore, the METEOR score also gives us an overview of whether supervised training was beneficial. Figure 4 shows the comparison of four GPT-2 models for different values of temperature with respect to METEOR score. We can see that the METEOR score increases after the supervised training for both versions of GPT-2. In addition to that, GPT-2 with 355M parameters achieves the highest METEOR score for

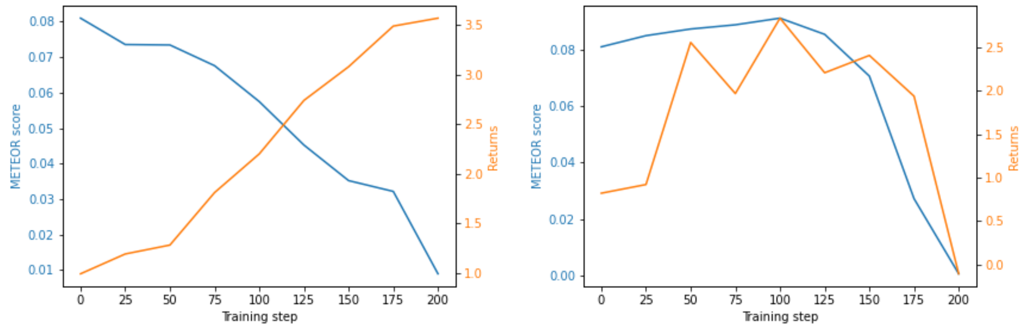
Context	Response	Reward
A: YI still can't believe my parents for divorced.	What happened?	+9.93
A: My friend is coming in from out of town today.	Oh wow, I'm very excited. Who is he?	+9.32
A: A few days ago, I was looking through a bunch of old photo albums. B: Were you looking for something in particular or just for fun? A: Just for fun. It brought back a lot of good memories!	Did you get anything?	+3.11
A: It was a tough day at work B: Awe, I am sorry. What made it tough?	I am sorry	+6.41
A: So my dog passed away a couple of weeks ago. Its been really hard not having him around. B: Oh no! How old was he?	-	+11.88

(a)

Context	Response	Reward
A: I've been hearing some strange sounds around the house at night lately.	What are they, and how bad are they, is it a ghost or something?	+14.17
A: We went to our local county fair a couple of days ago. B: I love the fair! Did you have a good time? C: We did. We listened to music, rode some rides.	That's cool. I love fairs.	+11.48
A: I am so ready for this semester B: do you go to college? A: Yes I am a senior this yar B: That's good. I am sure your parents are proud	I think so.	+2.05
A: I lent my friend 500$ and she never paid me back. Ugh. I am too nice sometimes.	I hope she didn't go to hell.	-4.96
A: Last week we had a fire start in our bathroom.	-	-5.27

(b)

Figure 5 Reward given by our reward models for *(context, response)* pairs (a) Reward assigned by our Reward-1, (b) Reward assigned by our Reward-2



(a) **(b)**

Figure 6 Comparison of METEOR score and the average return over different steps of the policy training. (a) Policy-1, (b) Policy-2

all values of temperature amongst all the models. Therefore, we continue from the 355M supervised GPT-2 model for the reward and policy training.

4.2 Reward-1 and Reward-2

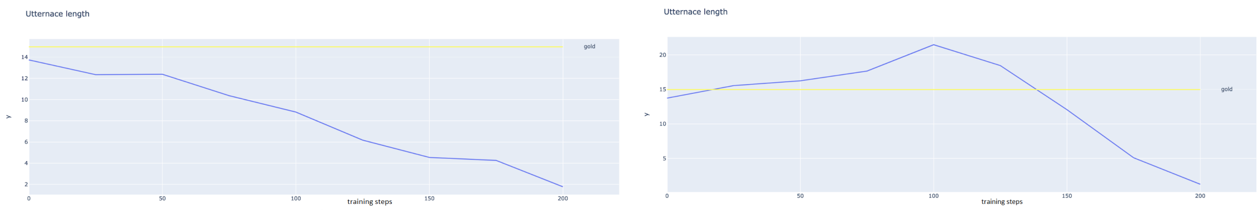
We also evaluate rewards given by our reward models for some (context, response) pairs. Figure 5(a) shows rewards given by our Reward-1 model for (context, response) pairs. As we can see, for the first two responses that contain questions, Reward-1 assigns a high reward. However, Reward-1 assigns a positive reward for the last two responses that are completely out of context and empty, respectively. One reason for this might be that less annotated data wasn't enough for the reward-1 to learn human preferences. Therefore, our Reward-1 model could not generalize well.

Similarly, Figure 5(b) shows rewards given by our Reward-2 model for (context, response) pairs. We can see that the reward-2 model also assigns high reward to responses that are relevant to the context and contain questions. Unlike Reward-1, for the last two (context, response) pairs, Reward-2 assigns negative rewards as one of the responses is out of context and the other one is just an empty response. The results show reward-2 model trained with gold responses from the validation set has a good grasp of human-like response.

4.3 Policy-1 and Policy-2

4.3.1 METEOR metric

We evaluated how the METEOR score and the average return vary during the policy training. Figure 6(a) shows the METEOR score and the average returns over different training steps of Policy-1. The average return



(a) **(b)**

Figure 7 Comparison of average number of words per generated turn over different steps of the policy training. (a) Policy-1, (b) Policy-2

increases, but the METEOR score decreases as we update the policy. The result suggests that average return and METEOR score are completely uncorrelated for the Policy-1 training. In addition to that, the plot further supports our underlying assumption that the Reward-1 model, trained with less human-annotated data, couldn't generalize well.

Similarly, Figure 6(b) shows the METEOR score and the average return over different training steps of Policy-2. As we can see, both average return and the METEOR increase for the first few training steps and then decrease. The decrease in both average returns and METEOR is due to over-optimization or over-fitting. The result also shows that METEOR and average returns are correlated. The model at training step=100 achieves the highest value in terms of average return and METEOR score. The plot clearly shows that we benefit from the policy training using Reward-2.

4.3.2 Utterance length metric

We look at the length of the generated utterances by our different models. Utterance length metric is obvious to look into as an utterance of length one is not empathetic, and it would not contribute to continuing the conversation.

In Figure 7a, we can see that for Policy-1, the utterance length decreases while training longer. The plot explains why the meteor score for Policy-1 decreases. It also further supports our assumption that the Reward-1 model didn't learn to predict human preference. Instead, it favored empty responses. On the other hand, the average utterance length values for Policy-2 generated responses in Figure 7b look better for some iterations of the policy training. Therefore, we decided to further investigate Policy-2 models.

4.3.3 Question metrics

With these metrics, we want to find out, if the generated responses try to ask questions to ensure the continuation of the conversation. Saleh et al. (2019), for example, demonstrate that a dialog model can be optimized for asking more questions by providing a positive reward when the generated utterance contains a question word and a question mark [15]. On the other hand, posting a question might not always be the right action in a conversation as we learned from Florian von Unhold's master's thesis, where he showed that different metrics have different values at different points in time of one human conversation [20]. This means sometimes a question is a good utterance but sometimes it is less helpful. Taking this into account we try to compare the generated response to the gold standard response at the same point in time of the conversation. Explaining this we have four different question metrics:

- Question indicates whether the given utterance contains a question word or a question mark (return 1) or neither of them (return 0). The question word must be one of 'who', 'what', 'why', 'where', 'how', or 'when'.
- Question ratio gives the percentage of generated question utterances of all generated utterances with the same question definition as in the previous metric.

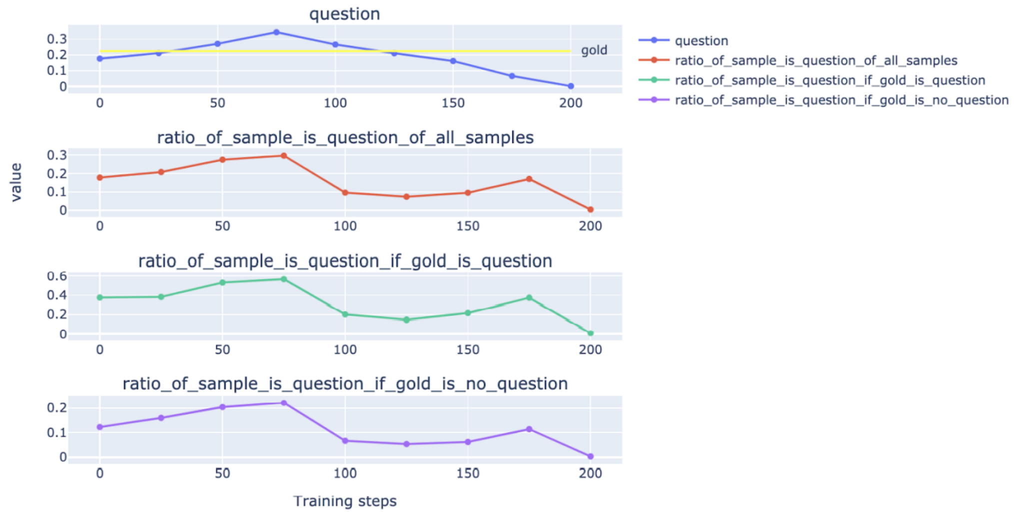


Figure 8 Average question metric values for generated turns from Policy-2 on training steps 0 to 200

- Question ratio of gold questions gives the percentage of generated question utterances of all generated utterances only looking at the utterances where the corresponding gold standard was a question with the same question definition as in the previous metric.
- Question ratio of gold no questions gives the percentage of generated question utterances of all generated utterances only looking at the utterances where the corresponding gold standard was no question with the same question definition as in the previous metric.

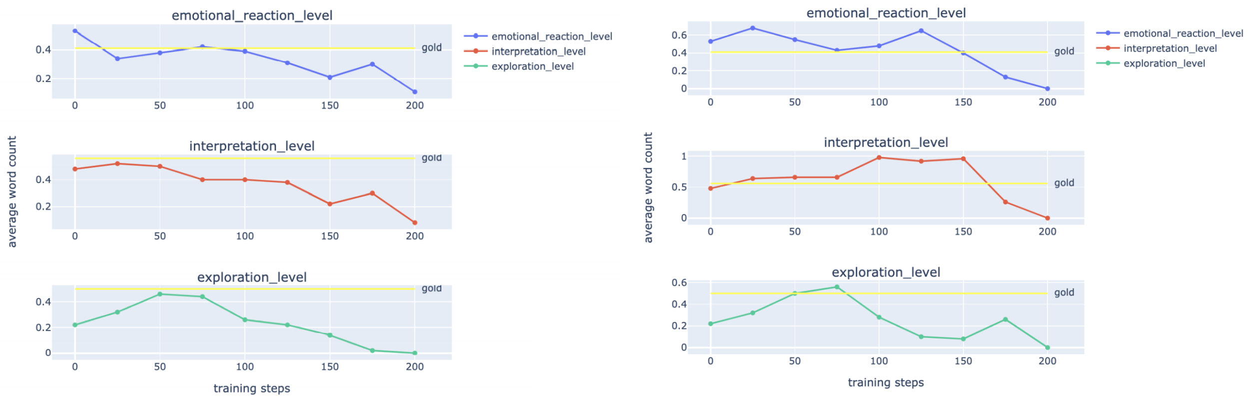
In Figure 8 we can see again after the meteor score in Section 4.3.1 a clue for the model at training step 100 being a good model for human evaluation in the end. The question metric values, as we can see in the first graph, are very similar to the average metric value for the gold conversation turns. That means the model produces nearly as many questions as the gold responses from the EMPATHETICDIALOGUES dataset.

Additionally, there is another model that looks promising. At training step 75 all average question metric values are very high. For the first two graphs, this means just that there is a lot of generated conversation turns with questions in them(around 30 % as you can see on the y-axes). For the last two graphs, this means that in cases where the corresponding gold response was a question a high (30%) number of generated responses were also questions and vice versa for the cases where the corresponding gold response was not a question in the last graph. As a result, the question metrics give evidence to further investigate Policy-2 models at training steps 75 and 100.

4.3.4 Empathy metrics

Arguably most important for our purpose is a model that generates empathetic utterances with high emotional, interpretation, and exploration level values. Again we used the metric implementation and trained models that are used for getting the metric values of Florian von Unhold's master's thesis [20].

He trained the models himself based on the code, ideas, and descriptions of instances of the RoBERTa-based bi-encoder model for identifying empathy by Sharma et al. (2020) where each instance predicts one of the three metrics [18]. Sharma's EPITOME framework of expressed empathy defines three empathy communication mechanisms: emotional reaction, interpretation, and exploration. It was trained on two datasets one from TalkLife.com, which is the largest global peer-to-peer mental health support network. The other one being 55 selected subreddits about depression and mental health. All empathy metric values are given on a scale of [0,1,2] where 2 is the best value.



(a) **(b)**

Figure 9 Average empathy metric values per generated turn over different steps of the policy training. (a) Policy-1, (b) Policy-2

- Emotional reaction level measures if an utterance shows warmth and compassion in reacting to the previous utterance.
- Interpretation level shows if an utterance shows an understanding of feelings and experiences of the previous utterance.
- Exploration level shows if an utterance brings new topics and emotions into the conversation regarding feelings and experiences of the previous utterance.

On Figure 9b, we see that for at least emotional reaction level and interpretation level the average values over the generated responses from Policy-2 model at training step 100 are again promising because they are either very high (interpretation level) or similar to the average value from the golf responses (emotional reaction level).

A second good number of training steps seems to be 75. Here both emotional reaction level and interpretation level are similar to the values of the average gold responses and exploration level is at its highest point. This means that the generated responses show even more ability to bring new topics into the conversation and continue it than the gold responses. Thus these metrics also point into the direction of investigating the models from Policy-2 at training steps 75 and 100.

Another thing we found out is that the empathy metrics on the generated responses from the models of Policy-1 can be seen in Figure 9a. Here nearly all values are falling during training which again shows that the training policy with Reward-1 seems somehow not to work with the current amount of annotated data by us.

4.3.5 Human evaluation

As a result of all the previous metrics, we decided to annotate the responses generated by the Policy-2 model at training steps 75 (P75) and 100 (P100) by comparing them against the responses generated by our supervised baseline model for the same context. For the P75 model, we use a temperature of 0.5 while sampling responses as we found it performed better than higher temperatures. Table 2 shows the results of our annotations.

The data was randomly divided between the three of us. Each of us annotated 50 comparisons between P75 (and P100) and the supervised baseline per annotation criteria empathy, fluency, and relevance (150 comparisons annotated each). As we can see from the table, P100 is preferred more than P75 over the supervised baseline in terms of empathy. However, in other criteria (fluency and relevance), our supervised baseline is preferred over the P100 model. On the other hand, P75 is preferred over the supervised baseline in all three criteria.

Empathy: *Which model generates responses that show more understanding of feelings?*		
	P75 Model	P100 Model
Preference over supervised baseline	52.66 %	**57.34 %**
Preference over state-of the art DialoGPT Model	**60.00 %**	-
Fluency: *Which model generates responses that are more fluent and accurate?*		
	P75 Model	P100 Model
Preference over supervised baseline	**57.34 %**	46.66 %
Preference over state-of the art DialoGPT Model	**76.67 %**	-
Relevance: *Which model generates responses that seem appropriate and on-topic?*		
	P75 Model	P100 Model
Preference over supervised baseline	**55.34 %**	43.34 %
Preference over state-of the art DialoGPT Model	**67.33 %**	-

Table 2 Results of human evaluation for empathy, relevance and fluency, showing preferences (%) for our models (P75 and P100), supervised baseline and state-of-the-art DialoGPT

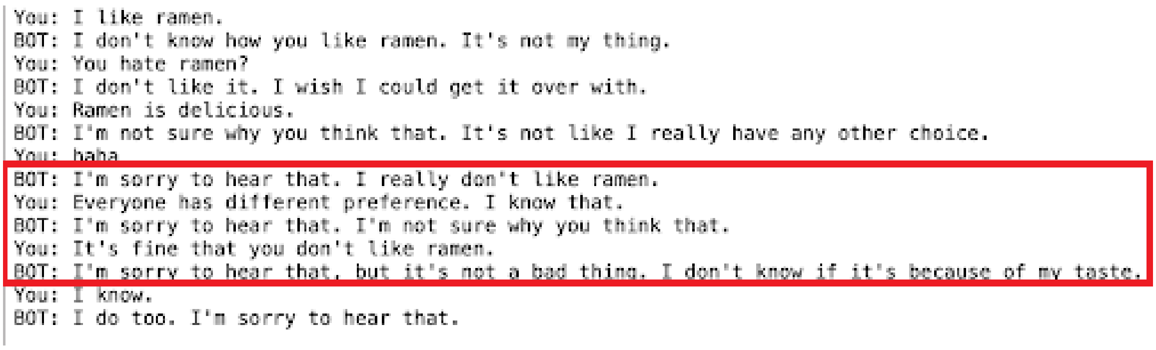


Figure 10 Repetitive dialogue turns generated by P75

In the end, we decided to compare the better model P75 against a state-of-the-art model DialoGPT, trained on the Reddit dataset [22]. We use the HuggingFace library to sample from DialoGPT with the default parameters. As we can see, more than 60% of P75's responses are preferred over the DialoGPT's responses in all three categories. Appendix C shows the sample responses from different models for the same context. Similarly, Appendix D shows the P75 conversations with humans.

4.3.6 Limitations

Although our P75 beats the state-of-the-art model in terms of human evaluation, there are some limitations of our model. In Figure 10, there is an example of some repetitive turns generated by our model. We also observed some contradicting responses like *I like pizza, but I'm not super fond of pizza.*. Therefore, we will further investigate this problem in the future.

Our current repetition metrics look for words from specific previous utterances or the whole conversation. This is an important metric because the reproduction of the story the conversation partner tells the system shows understanding and empathy and is often necessary to react. On the other hand, just repeating and don't mentioning any new topic or different views on the current topic is not helpful and does not contribute to keeping the conversation alive. As well as minimizing repeated words can be used to prevent the model from repeating

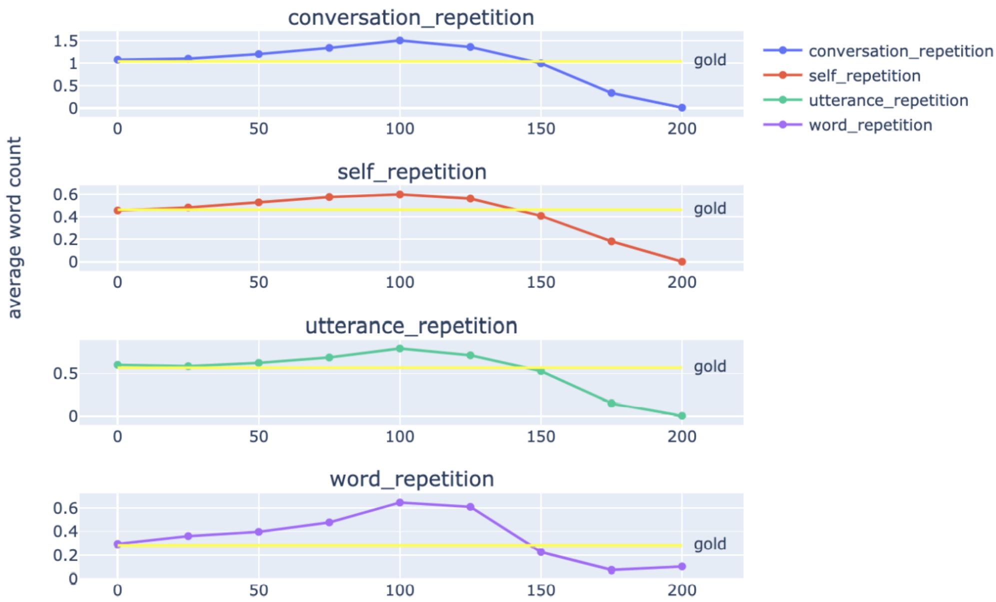


Figure 11 Average number of repeated words per generated turn from Policy-2 on training steps 0 to 200

itself. Moreover Saleh et al. (2019) penalizes utterances for high repetition values [15]. Again the code base is used from the master's thesis of Florian von Unhold [20] and include the following values:

- Self-repetition is the count of distinct words in the given utterance that are also in any of the previous utterances of the current speaker.
- Utterance repetition is the count of distinct words in the given utterance that are also in the previous utterance.
- Word repetition is the count of words that occur multiple times within the given utterance, i.e., the count of duplicates.
- Conversation repetition is the count of distinct words that have been used in the whole conversation so far. This metric is an extension of the adapted reward function metrics. It measures how often conversation partners reuse words from the whole dialog history.

In Figure 11, we can see the average values of how many tokens were repeated when generating responses by our Policy-2 models. From the plot, for training steps 75 and 100, we have high values, but it's difficult to say if high values in the repetition of single words are desirable or not.

5 Conclusion and future work

In our work, we propose an open-domain multi-turn human-like dialogue agent trained with RL. We also provide an alternative method for the reward model training in case of less human-annotated data. Our results verify that RL fine-tuned model outperforms the supervised baseline in terms of both automated metrics and human evaluation. Furthermore, human evaluation results show that our model is preferred more than the state-of-the-art DialoGPT in three aspects: empathy, relevance, and fluency.

Although our results look promising, there are still some limitations of our model. As already discussed in our result section, there is a problem of repetition and contradiction in our model. One possible solution is to detect if any two turns contain a long common sub-sequence and to remove such candidates while sampling the responses. In addition to that, we can design a reward function to penalize such responses. We also don't address how safe our model is. Therefore, future work could be carried out to evaluate if the model responses are sensitive or toxic. Since training the reward model with less human-annotated data in an offline RL setting didn't seem to work, we think doing an online RL is helpful in such a case. Online RL solves the problem of distributional shift during training and inference. In conclusion, our contribution has opened up many ideas for future research.

References

[1] Daniel Adiwardana et al. *Towards a Human-like Open-Domain Chatbot*. 2020. arXiv: 2001.09977 [cs.CL].

[2] Andrew M. Dai and Quoc V. Le. *Semi-supervised Sequence Learning*. 2015. arXiv: 1511.01432 [cs.LG].

[3] Braden Hancock et al. *Learning from Dialogue after Deployment: Feed Yourself, Chatbot!* 2019. arXiv: 1901.05415 [cs.CL].

[4] Jeremy Howard and Sebastian Ruder. *Universal Language Model Fine-tuning for Text Classification*. 2018. arXiv: 1801.06146 [cs.CL].

[5] Natasha Jaques et al. *Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control*. 2017. arXiv: 1611.02796 [cs.LG].

[6] Natasha Jaques et al. *Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog*. 2019. arXiv: 1907.00456 [cs.LG].

[7] Urvashi Khandelwal et al. *Sample Efficient Text Summarization Using a Single Pre-Trained Transformer*. 2019. arXiv: 1905.08836 [cs.CL].

[8] Jiwei Li et al. *Dialogue Learning With Human-In-The-Loop*. 2017. arXiv: 1611.09823 [cs.AI].

[9] Yukun Ma et al. "A survey on empathetic dialogue systems". In: *Inf. Fusion* 64 (2020), pp. 50–70.

[10] Romain Paulus, Caiming Xiong, and Richard Socher. *A Deep Reinforced Model for Abstractive Summarization*. 2017. arXiv: 1705.04304 [cs.CL].

[11] Matthew E. Peters et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL].

[12] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". In: (2019).

[13] Marc'Aurelio Ranzato et al. *Sequence Level Training with Recurrent Neural Networks*. 2016. arXiv: 1511.06732 [cs.LG].

[14] Hannah Rashkin et al. *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. 2019. arXiv: 1811.00207 [cs.CL].

[15] Abdelrhman Saleh et al. "Hierarchical Reinforcement Learning for Open-Domain Dialog". In: *CoRR* abs/1909.07547 (2019). arXiv: 1909.07547. URL: http://arxiv.org/abs/1909.07547.

[16] Magda Saraiva, Hande Ayanoglu, and Beste Ozcan. "Emotional Design and Human-Robot Interaction". In: Sept. 2019, pp. 119–141. ISBN: 978-3-319-96721-9. DOI: 10.1007/978-3-319-96722-6_8.

[17] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG].

[18] Ashish Sharma et al. "A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support". In: *CoRR* abs/2009.08441 (2020). arXiv: 2009.08441. URL: https://arxiv.org/abs/2009.08441.

[19] Nisan Stiennon et al. *Learning to summarize from human feedback*. 2020. arXiv: 2009.01325 [cs.CL].

[20] Florian von Unhold. "State-Based Dialog Modeling Using Human Preference". MA thesis. Technical University of Munich, 2021.

[21] Sanghyun Yi et al. *Towards Coherent and Engaging Spoken Dialog Response Generation Using Automatic Conversation Evaluators*. 2019. arXiv: 1904.13015 [cs.CL].

[22] Yizhe Zhang et al. *DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation*. 2020. arXiv: 1911.00536 [cs.CL].

[23] Daniel M. Ziegler et al. *Fine-Tuning Language Models from Human Preferences*. 2020. arXiv: 1909.08593 [cs.CL].

A Data pre-processing

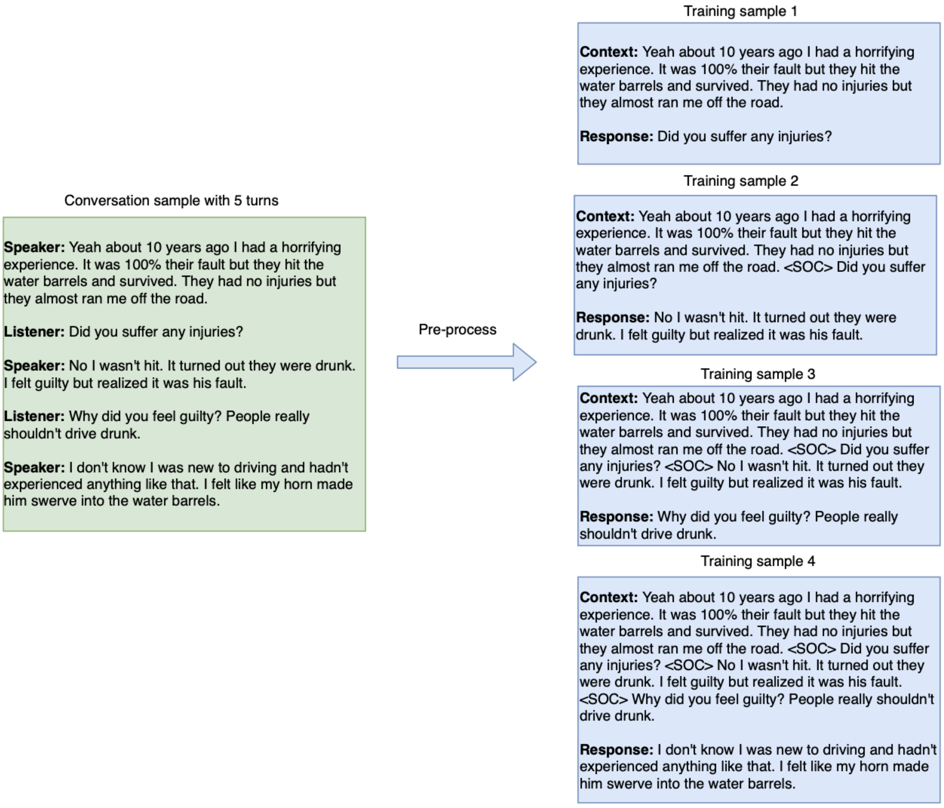


Figure 12 Data pre-processing example. 4 training samples are generated from one conversation with 5 turns.

Figure 12 shows the pre-processing step to create training samples from one conversation sample with 5 turns.

B Data annotation interface

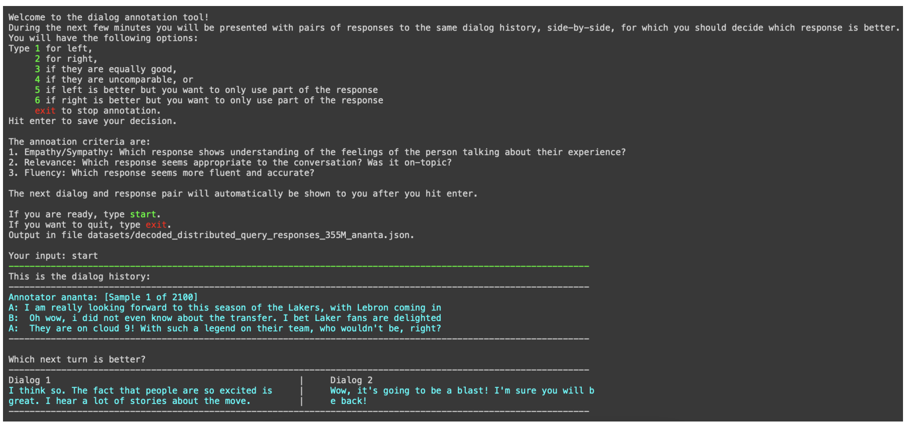


Figure 13 Data annotation interface.

Figure 13 shows our data annotation interface.

C Sample Responses

Contexts	P75	DialoGPT	Supervised Baseline
A: 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.	That's terrible. I'm sorry to hear that. Was it your fault?	I'm sorry to hear that.	That is pretty awful. Did the accident happen before or after?
A: I there, dont know what to do, jst broke up with my girlfirned, we were 8 years together B: sorry to hear! do you have any idea about the break up? did you think about it ?	I did, but i wasnt sure what to do, i was really upset about it.	I think I did, but I don't know. I'm sorry to hear that.	I don't know, I just felt so sad that my childhood friends broke up! i hope they'll find each other again!
A: my husband lost a job but i'm hoping he can find a full time job soon **B:** He will, I have faith. **A:** thank you so much! **B:** No problem. What kind of work does he do?	I think he does a lot of work in the field of IT.	He's a software engineer.	He does an internship. I'm pretty sure he does part time.
A: i'm so excited because i'm finally going to visit my parents next month! I didn't see them for 3 years **B:** 3 years is a long time. How come? **A:** I moved to the US and they stayed in my homecountry	Oh my gosh, I hope you're really excited about it. I'm sure it will be a great experience for you.	I'm so excited for you! I'm so glad you're moving to the US!	How long have you been there?

Figure 14 Sample responses from different models for the given contexts. For P75, a temperature value of 0.5 was used for sampling. For DialoGPT, HuggingFace default parameters were used. For the supervised baseline, a temperature value of 0.5 was used.

Figure 14 shows cherry-picked sample responses from different models for the same contexts.

D P75 Conversation with humans

Figures 15, 16, 17 and 18 show random conversations of P75 with humans.

YOU: Hi! How are you?! :)
BOT: I'm doing fine, I just need some time to recover from my injury.
YOU: What happened to you?
BOT: I broke my ankle and it looked like it was going to be a long recovery.
YOU: I am so sorry. How long do you need to rest?
BOT: I need a lot of time to recover from the injury. I'm not sure if I will be able to return to the field for the rest of the season.
YOU: You will. Don't worry. Don't lose hope.
BOT: I hope so too. I'm just thankful for the doctors and everything else I have to work with.
YOU: Yeah. Doctors are great.
BOT: I know. I'm just thankful that I'm able to keep going. I know I will be fine.
YOU: You will be. Do you have any plans for today?
BOT: I have a long day tomorrow. I'm going to be working on my leg, too. I'm really feeling it now.
YOU: I see. I am sad though.
BOT: I know. I just want to be able to get back to work.
YOU: You will.
BOT: I hope so. I think I am going to be okay.
YOU: What's your favourite movie?
BOT: I just like the ones that are about the frontiersman.
YOU: Do you have any specific one that you like?
BOT: I do. I think I like the one where the main character is a mountain man.
YOU: I haven't seen that movie. What's the name of the movie?
BOT: I think it's called "Mountain Man, The Biggest of the Big."
YOU: I will definitely watch it.
BOT: I'm sure you'll enjoy it. I'm not sure if I agree with the movie's ending, but I'm sure I'll like it.

Figure 15 P75 conversation one

YOU: Hi! How are you?!
BOT: How are you? I'm feeling better.
YOU: I am having a bad day.
BOT: It's common thing with me. I was feeling really down for few days. I was so upset about it.
YOU: Why were you feeling down?
BOT: That's a good question. I was feeling super angry, but I think I'm over it. I feel like it was my fault. I feel like I'm going to be okay.
YOU: Why were you angry?
BOT: Well, I was angry because I had no idea why the guy I was dating was taking so much from me. I was so upset. I had no idea what to do.
YOU: I feel you. Guys are like that.
BOT: Maybe I should have told him. I wish I could have gotten some other man. But I think I can forgive him now.
YOU: I think you should give him a chance.
BOT: Maybe he can get a girlfriend and I can move on to the next guy

Figure 16 P75 conversation two

YOU: Hello!
BOT: what is it
YOU: How are you?
BOT: I'm just happy to be in the country.
YOU: I want to move out of the country.
BOT: What's the worst thing about moving to new country
YOU: I think it's new people and new culture. You need some time to adapt there.
BOT: I'm sorry that's how it is
YOU: Do you wanna move to another country?
BOT: I'm not moving to another country.
YOU: why?
BOT: I just don't feel safe at home.
YOU: You aren't safe anywhere in this world.
BOT: I am so scared of the outside world, I feel like I might be murdered.

Figure 17 P75 conversation three

YOU: I like how cute cows are.
BOT: Well, I agree with you. I think cows are cute, but I have never seen a cow with its head out the window.
YOU: You are funny.
BOT: I agree with you. I think that is a cute thing to do.
YOU: I heard that a cow to Harvard.
BOT: I think that's hilarious. They get all the credit for that.
YOU: What did the cow study?
BOT: Well, he studied math and he got all the honors.
YOU: What are you studying?
BOT: I am studying math. I am not sure if I want to pursue it though.
YOU: I am not a math person. I think it's super complex.
BOT: I'm not sure it is, but I am confident that I can do it. I am also pretty confident that the math will be really fun. I just have to work on it.

Figure 18 P75 conversation four