Empathetic Conversation Generation and Evaluation Framework for Mental Health Use

Akhil Vinta
UCLA (405288527)
akhilvinta@g.ucla.edu

Chris Park
UCLA (806183297)
chrisjpark@g.ucla.edu

Karl Goeltner
UCLA (205312849)
kgoeltner@g.ucla.edu

Kumar Saunack
UCLA (805912645)
krsaunack@g.ucla.edu

Shrea Chari
UCLA (005318456)
shreachari@g.ucla.edu

Abstract

Empathy plays a crucial role in mental health support and leveraging conversational agents for empathetic interactions holds promise in this domain. In this paper, we aim to explore empathy in models quantitatively and qualitatively while offering improvements in both areas. For quantitative evaluation, we create a new mental health dataset used for finetuning aimed to enhance objective measurements of a model's empathetic capabilities. Additionally, we evaluate and enhance baseline generative models to prioritize empathy in conversations. We also explore novel SoTA techniques for generative models, namely RLHF, to test its efficacy.

1 Introduction

The fusion of Large Language Models (LLMs) with the quest for empathetic bots is a critical field for mental health support. With recent advances in LLMs and their ability to generate human-like text, such bots are now feasible. By harnessing LLMs' capacity to generate contextually relevant responses, we can cultivate conversational agents capable of fostering genuine empathetic connections with individuals seeking mental health support. However, the challenge lies in developing robust frameworks that not only detect empathy but also ensure its consistent expression in generated conversations. In this paper, we delve into the fusion of LLMs and empathy, exploring how these models can be leveraged to facilitate empathetic interactions within the realm of mental health.

Through our research, we aim to contribute to the development of empathetic conversation generation frameworks tailored for mental health applications and objectively evaluate our improvements. As a result, our work focuses on 2 fronts:

- Creating quantitative evaluation criteria for determining how empathetic a response is to a given statement
- Training a generative model that can generate empathetic responses and evaluate the model on both qualitative and quantitative criteria

1.1 Contributions

Our contribution is threefold:

- We create a new dataset for empathy detection and in the process improve empathy detection models

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

- We evaluate and improve the baselines for generative models in terms of empathy-oriented conversations
- We test the effectiveness of new methods like RLHF for LLMs on the generative task

2 Literature Review

2.1 Empathy Dataset

In order to evaluate empathy in AI-generated conversations and explore the idea of empathetic response generation, we introduce Rashkin et al. [2019]. The paper proposes a novel resource `EMPATHETICDIALOGUES(ED)`, a dataset consisting of 25k one-on-one dialogues between a *Speaker* and *Listener* grounded in specific situations.

The research evaluates using `EMPATHETICDIALOGUES(ED)` as a benchmark for models in the *Listener* role listening and responding empathetically to conversations initiated by a *Speaker*. To evaluate the use of this dataset, they utilize transformer and BERT-based architecture models pre-trained on Reddit conversations and fine-tuned with `ED` to generate responses. The model-generated responses were then rated on automated metrics and human-based ratings. The experiment results show that models fine-tuned on the `ED` dataset showed a significant improvement in generating empathetic responses compared to the pre-trained models not fine-tuned on the dataset. This indicates the effectiveness of the dataset in enhancing the empathetic capabilities of conversational agents. Based on these findings, we utilized the `EMPATHETICDIALOGUES` dataset as an essential resource for training and fine-tuning our generator models. We ultimately utilize the dataset as the foundations to create a more empathetic and flexible generative AI. Rashkin et al. [2019] is overall significant because it contributes a unique dataset built for the precise nature of our project. The ED dataset is the basis of much of the previous work done with empathetic models, including some we reference in this project.

2.2 Empathetic Dialogue Generative Model

Gao et al. [2021] proposes a novel framework that improves empathetic response generation by using hard and soft gated attention mechanisms to incorporate the emotion cause into response generation. They start with trainable emotion embeddings, which are fed into the encoder to get contextualized word representations of the input. The model then utilizes a multi-head attention network where both the encoder and decoder contain self-attention layers and the decoder uses a cross-attention layer. Lastly, soft and/or hard-gated attention mechanisms are used on top of the cross-attention layer. Utilizing automatic and manual evaluations, the paper found that the proposed model can generate more meaningful and empathetic responses.

We utilize this novel framework as a significant model in our experiments that we ultimately attempt to improve. A limitation of this paper is that the architecture of the model is complicated and difficult to train. Additionally, the model implementation is difficult to configure towards custom datasets and other existing libraries.

2.3 Nucleus Sampling

Holtzman et al. [2020] evaluates existing decoding strategies and introduces Nucleus Sampling, a novel decoding strategy. The paper explores sampling text from a dynamic nucleus of the probability distribution, which allows for diversity while effectively truncating the less reliable tail of the distribution. Their research found that the resulting text better demonstrated the quality of human text and enhanced diversity. A limitation of this paper is that generation produced by Nucleus Sampling is not perfect and can confuse different words. Additionally, the paper mentions future goals in dynamically characterizing the region of confidence of language models. We ultimately utilize the work done in this paper to improve our EmpDialogue Model to try to have more diverse outputs and avoid degenerate outputs.

2.4 Alignment of Generative models

To better align model behavior to human expectations, there has been a large body of work (Ziegler et al. [2019], Wu et al. [2021]). This method has been shown to get significantly improved results for generative models for conversations by Wu et al. [2021]. In the absence of human evaluators, Lee et al. [2023] show that designing surrogate reward functions can help scale model alignment as well. While research has gone into using LLM outputs as the rewards themselves (Roit et al. [2023]), we limit the scope of our experiments to engineered rewards.

3 Methodology

3.1 Data

3.1.1 Reddit Dataset

Sharma et al. [2020] references a reddit dataset that consists of threads posted on 55 mental health-focused subreddits. It contains 1.6M threads and 8M interactions and judges empathy based on 3 mechanisms:

- Emotional Reactions (ER): Expressing emotions such as warmth, compassion, and concern.
- Interpretations (IP): Communicating an understanding of the seeker's feelings.
- Explorations (EX): Improving understanding of the seeker asking questions or gently probing.

For each mechanism, crowd workers annotated whether the response post contained: no communication of empathy (label 0), weak communication of empathy (label 1), or strong communication of empathy (label 2). For mechanisms labeled 1 or 2, rationale is provided.

3.1.2 Custom Dataset

To try to improve the performance of the evaluator, we needed to create a custom dataset that was labeled and had rationales. We started with an existing dataset from HuggingFace by Amod. This dataset consisted of a collection of questions and answers sourced from two online counseling and therapy platforms. The questions cover a wide range of mental health topics, and the answers are provided by qualified psychologists. We manually labeled 500 randomly selected entries of this dataset using the three evaluation criteria needed by our model (ER, IP, and EX) and provided the appropriate rationale. Labelers had a 62% inter-annotator agreement similar to the paper.

3.2 Evaluator models

3.2.1 Baseline Evaluator Model

Our baseline evaluator model is based on a paper by researchers at the University of Washington and Stanford University. They developed a novel framework to characterize the communication of empathy in a text response to a seeker post. Through a framework called `EPITOME`, the model quantifies the quality of a response based on 3 criteria: emotional reactions, interpretations, and explorations. Each category is rated with 0, 1, or 2 corresponding to no, weak, or strong communication of such factors. Empathetic responses not only react with emotions of warmth and compassion but also communicate a cognitive understanding of others. Rashkin et al. [2019] developed a multi-task RoBERTa-based bi-encoder model for identifying empathetic conversations per these metrics and also extracting rationales underlying its predictions. This baseline evaluator was trained on 90% of the Reddit dataset mentioned above, and tested on the remaining 10% of the data.

3.2.2 Enhanced Evaluator Models

We enhanced our evaluator model in two ways. First, we pretrained the baseline model on Roberta-Large, which is a bigger model with more parameters. Our second enhancement involved pretraining our baseline evaluator on the custom dataset we annotated. We tested both of these enhanced models using the same test data as the baseline evaluator.

3.3 Generative models

3.3.1 Baseline Model

The first task in the process of building a generator model was to find a baseline model that we could use as a benchmark to measure future improvements. From various experimentation, we found that DialoGPT performed reasonably well for general empathetic response tasks. DialoGPT is a computational AI model launched by OpenAI. It is a version of the GPT-2 Large-Language Model that has been fine-tuned for conversations. Thus, it is appropriate in a setting where we want our model to respond to statements. We believe that DialoGPT will provide us with a reasonable benchmark to quantify the value of fine-tuning a model for empathetic responses.

3.3.2 Empathetic Dialogue Model (EmpDialogue) and Improvements

The Empathetic Dialogue (EmpDialogue) Model utilizes both emotional context and emotional cause information to generate more empathetic responses. The components of this model are twofold: the emotional reasoner and response generator.

The emotional reasoner recognizes the context emotion of the query (classification problem) and the cause behind the emotion (sequence labelling), providing emotional information for response generation. Multi-task learning is used to learn both the emotion recognition and emotion cause detection tasks at the same time. Words are additionally embedded with positional information and fed into a transformer encoder, where we then use the value of the $[CLS]$ token to compute emotion context. The emotional cause detection labels each word in the sequence with an emotion cause-oriented label $0, 1$, indicating whether the word is related to the emotion cause. The emotional reasoner was trained on the RECCON and ED dataset for empathy training and used the pretrained glove algorithm for word embeddings.

The response generator is an encoder-decoder transformer network that makes use of deep emotional information from emotion reasoner to generate more empathetic responses. It explores both hard and soft gating strategies to allow the model to focus more on words related to the emotion cause. Given the predicted context emotion ψ and the emotion cause-oriented labels C from the emotion reasoner, the response generator's objective is to produce an empathetic response Y that is emotionally coherent and contextually relevant to the dialogue. This is achieved by optimizing the probability $P(Y|X, \psi, C)$, where X represents the dialogue context, aiming to maximize the alignment between the generated response and both the emotional context and the causes identified. To incorporate the emotion cause into response generation, a gated attention mechanism (either hard or soft gating strategies) on top of the cross attention layer in the decoder allows the model to focus on emotion-cause related words.

We improved this model by incorporating Nucleus Sampling as outlined in Holtzman et al. [2020] as an improvement over the Top-K sampling method in the decoder. In implementing Nucleus Sampling, we sought to avoid degenerate outputs and output more dynamic language. We used the parameters of 0.9 for the probability mass threshold (p) and 0.4 for softmax temperature and trained for around 60 epochs. We also implemented other small improvements such as changing the max decoding token length to 2048 and changing the length penalty to 1.0.

3.3.3 Fine-Tuning GPT-2

For each conversation X_v , we generate multi-turn conversation data $(X_s^1, X_r^1, ..., X_s^T, X_r^T)$, where T is the total number of turns. X_s^t denotes the statement at turn t and X_r^t denotes the corresponding response. For instruction tuning GPT-2, we pair up statement response pairs. For dialogues where we do not have responses to a given statement, we simply drop them.

We perform instruction-tuning of GPT-2 on the response tokens, using its original auto-regressive training objective. Specifically, we compute the probability of the responses by $p(X_r|X_s) = \prod_{i=1}^{L} p_\theta(x_i|x_{<i}, X_s)$ where L is the sequence length for the responses. To help the model learn the separation between responses and statements better, we introduce additional start and end of speech tokens.

Overall, a given sample input to the model looks like <SPEAKER>X_s<RESPONSE>X_r with losses propagated only on X_r.

3.3.4 Improving Alignment of SFT-GPT2

We also attempt to encourage certain behaviours in the model to improve exploration and interpretation in conversations via alignment. Given a model, the aim is to optimize the outputs for certain rewards. We use the current SoTA method, PPO by Schulman et al. [2017], for policy optimization after rollouts.

We design a surrogate reward function to assign objective values to responses. For efficiency reasons, the reward function used is a mixture of experts. The experts were models already fine tuned for sentiment analysis, toxicity detection, similarity to therapist behaviour. We also add a length reward for explicit reward signals for longer responses. Instead of taking a linear combination of rewards from each model, we design a reward that maximizes the margin between positive and negative sentiments and that also adds an exponential negative reward for toxic behaviour. Each sub-reward is weighed differently and according to different combination of weights, we end up with 3 different models: `rlhf length`, `rlhf therapist` and `rlhf question`. They focus on length, therapist-like statements and asking questions respectively.

For each conversation X_v , we only use the statements $(X_s^1, ..., X_s^T)$ using the same notation as before where X_s^t denotes the statement at turn t. The model generates several outputs for the same query. Each response is assigned a reward using a reward model and the weights are updated to optimize the rewards.

For implementation, we use the `trl` library by huggingface for rollouts and backpropagation. An additional regularization with KL divergence was also required to keep the model outputs coherent.

4 Results and Discussion

4.1 Evaluator Enhancements Results

Using our reddit test set, we evaluated four evaluator models: roberta-base-eval, finetuned-roberta-base-eval, roberta-large-eval, and finetuned-roberta-large-eval. As seen in Table 1 below, we see improvements in accuracy and f1-scores for both finetuned-roberta-base-eval or roberta-large-eval in emotional reactions (ER) and interpretations (IP). This is likely attributed to the larger model having more parameters which has previously captured more complex patterns and nuances in language via pre-training. High ER is attributed to greater acknowledgement of feelings via warmth and compassion. High IP communicates an understanding of the seeker's experiences and/or feelings via paraphrasing or reflections of ones own feelings.

However, there is a degradation in performance especially with respect to explorations (EX). This is likely attributed to our labeling disparities in creating the new dataset where our labelers ranked responses giving advice as weak as opposed to no exploration. Since the original paper's labeling criteria was vague: "does the response make an attempt to explore the seeker's experiences and feelings?", we considered practical advice to address their situation as a weak exploration. Table 2 depicts similar ER/IP vs EX disparities with respect to rationale comparisons.

Table 1: Evaluator Empathy Rating Comparison

Model	ER		IP		EX	
	acc.	f1	acc.	f1	acc.	f1
roberta-base-eval	80.10	72.60	81.94	**71.75**	**95.00**	**72.62**
finetuned-roberta-base-eval	82.07	70.69	**82.89**	65.34	39.84	33.38
roberta-large-eval	**83.95**	**77.22**	82.89	65.34	94.69	68.79
finetuned-roberta-large-eval	79.36	57.50	74.03	55.38	60.16	48.81

* **bold** is best value per column

4.2 Generator Enhancements Results

On the generative side using our reddit test set, we evaluated six generative models with our baselines as: dialo-gpt, empdialo-soft-gate, and empdialo-hard-gate; our enhancements as: empdialo-enhanced,

Table 2: Evaluator Empathy Rationale Comparison

Model	ER			IP			EX		
	acc.	f1	IOU	acc.	f1	IOU	acc.	f1	IOU
roberta-base-eval	63.72	66.04	72.41	63.74	62.15	63.21	72.54	**79.18**	**92.22**
finetuned-roberta-base-eval	**67.79**	**67.16**	73.12	**64.63**	**61.22**	58.22	47.42	44.97	34.64
roberta-large-eval	63.15	67.06	**75.43**	60.66	60.26	**64.38**	**72.77**	78.57	91.98
finetuned-roberta-large-eval	62.13	64.53	73.27	59.29	60.64	61.81	59.02	54.46	58.29

* **bold** is best value per column

sft-gpt2, rl-q2-enhanced. As seen in Table 3, there are significant increases for empdialo-enhanced from baselines in generating weak (1) and strongly (2) rated emotional reaction (ER) and exploration (EX) responses. Adding nucleus sampling to avoid sampling from low-confidence distributions, increasing max-token decoding length, and stopping training at 80 epochs to reduce overfitting led to the increases in both ER and EX for empdialo-enhanced.

However, when looking at Table 4, it seems apparent that one limiting factor to increasing the generative performance is attributed to length. It also seems clear that a higher category score in ER/IP/EX is loosely correlated with increase of length of responses between 1 and 2 ratings. sft-gpt2 and rl-q2-enhanced models have near double the lengths of baselines highlighting the strength of both SFT on the EmpatheticDialogues dataset and RL approaches. Notably our RL approach uses an explicit reward alignment for length among others which further improves this behavior. This results in increases in 1 and 2 ratings for IP compared to baseline empdialo models. dialo-gpt outperforms in interpretations likely because it's already been pre-trained on a large set of reddit conversations.

For a better visualization on the generative model response evaluations across each category ER/IP/EX, please reference Figure 1, Figure 2, Figure 3 for generative model's evaluated category label percentages. For our generative model's average lengths per category and label, see Figure 4, Figure 5, Figure 6.

Table 3: Generator Empathy Response Rating Percentages

Model	ER			IP			EX		
	0	1	2	0	1	2	0	1	2
dialo-gpt	*47.56*	38.76	13.68	47.23	**26.38**	**26.38**	**49.51**	37.13	13.36
empdialo-soft-gate	23.78	52.12	*24.10*	**90.55**	4.56	4.89	32.57	43.32	*24.10*
empdialo-hard-gate	35.83	*50.49*	13.68	*90.23*	3.58	6.19	42.35	*43.97*	13.68
empdialo-enhanced	12.38	**53.09**	**34.53**	90.23	4.56	5.21	20.20	**45.28**	**34.53**
sft-gpt2	**49.84**	39.41	10.75	72.31	*6.51*	*21.17*	*55.70*	34.53	9.77
rl-q2-enhanced	40.07	49.51	10.42	81.11	4.23	14.66	49.51	41.04	9.45

* **bold** is best value per column, *italic* is 2nd best value per column

Table 4: Generator Empathy Response Average Character Lengths

Model	ER			IP			EX		
	0	1	2	0	1	2	0	1	2
dialo-gpt	47.85	26.80	42.02	43.21	18.05	51.89	45.89	28.13	42.78
empdialo-soft-gate	35.97	48.77	51.07	45.69	56.64	47.60	39.85	48.33	51.28
empdialo-hard-gate	37.27	46.47	51.29	43.00	46.27	54.53	37.90	47.23	51.29
empdialo-enhanced	36.87	44.85	50.38	45.49	50.79	46.19	37.77	45.83	50.38
sft-gpt2	*79.63*	*86.60*	*95.52*	*78.31*	*82.65*	*103.97*	*79.02*	*87.74*	**99.77**
rl-q2-enhanced	**86.10**	**102.82**	**95.62**	**91.04**	**86.85**	**122.00**	**87.87**	**104.82**	*93.76*

* **bold** is best value per column, *italic* is 2nd best value per column

4.3 Challenges and Solutions

4.3.1 Evaluator Model

The original paper Sharma et al. [2020] outlines their new empathy evaluation framework using two custom datasets, one from TalkLife and the second being a manually created one from reddit mental health threads. The reddit dataset contained 1.6M threads and 8M interactions. However, the TalkLife dataset contained 6.4M threads and 18M interactions (seeker post, response post pairs), significantly more than only the reddit dataset. Though the reddit dataset was readily available, we could not access the TalkLife version due to its non-public nature.

Hence, we resorted to creating a custom dataset based on HuggingFace by Amod. Using the mental health dataset consisting of questions and answers similar to Sharma et al. [2020]'s reddit dataset, two members of our team created a small custom dataset of 500 entries that aligned with the manual labeling process used in the paper. However, there were also some labeling discrepancies specifically in interpretations (IP) where we considered advice as weak as opposed to no interpretation rating.

4.3.2 Generator Model

There were a number of challenges that we ran into when implementing the generator models. The first hurdle we faced was that we could only attain 50% accuracy on validation set when we ran the EmpatheticDialogue emotional reasoner model, despite the paper mentioning that the model could reach 80%. Thus, we were forced to navigate the fact that our environment consisted of discrepancies and deviations from the original environment of the paper.

Another challenge that we faced was simply that of time and compute resources. Because we were training very complex models for extensive periods of time, we would often run out of Colab compute units, and thus had to modify the code to include saving and loading parameters. However, even this workaround would still require us to frequently toggle between the Google accounts that we owned and restart the training pipeline every time GPU units expired.

5 Conclusion

In this project, we have successfully trained and enhanced an evaluator which is able to quantify empathetic responses based on three objective criteria: Emotional Reactions, Interpretations, and Explorations. The model will provide a score from 0-2 for each of these criteria given a response to a certain query.

We have also trained and enhanced a generative model which can generate responses with high empathy scores. As we can see from the Results section, our model performs better with each improvement that we add to it. The Empathetic Dialogue model performs significantly better than the baseline DialoGPT model, implying the value of the added emotion model. The enhancements that we made to the Empathetic Dialogue model, utilizing RLHF and Nucleus Sampling rather than top-K sampling led to a significant improvement in both the Emotional Reactions and Explorations categories. Finally, training with RLHF to increase response lengths was successful, as the RL-Q2-Enhanced model provides significantly longer outputs than the previous models.

With all of the above considered, we conclude that we were able to successfully achieve the goals of our initial endeavors.

5.1 Impact

As stated in the introduction, there are many valuable use-cases for models that generate empathetic responses. One such use case is assisting in affordable, accessible, and empathetic mental health treatment and counseling. Another use case is customer support and service, where empathetic models can create a more enjoyable customer experience. Other tasks that this can impact include movie script writing, educational tools, content creation, and more.

5.2 Future Work

In this project, we have built and fine-tuned two models. One is an evaluator that can expertly quantify responses based on objective measures of empathy, and the other is a generator which aims to generate responses that are as empathetic as possible. This begs the following question: Why not utilize RL to improve our generator, using our evaluator as a reward function? The pipeline would go as such: The generator is given an input query, to which it provides a response. This pair of (query,response) is then passed to the evaluator, which is already trained to output values for ER, EX, and IP from 0-2. With some clever reward engineering, we can then use the evaluator output to perform a train step on the generator, utilizing a high reward to encourage similar generator outputs, and a low reward to push the generator away from outputs as such.

In fact, we have already attempted to connect this RL pipeline utilizing our generator and evaluator, but found that the evaluator was too complex and thus our training time was too high. We found that our model was processing around 1 query per minute, which of course is much too slow for RL given our time-frame to complete the project. In the future, when time and compute resources permit, this can be a very natural extension of our current project that ties both ends together to build a robust and objective means of generating empathetic responses.

References

Amod. Mental health counseling conversations. https://huggingface.co/datasets/Amod/mental_health_counseling_conversations. Accessed: 2024-03-01.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1534. URL https://aclanthology.org/P19-1534.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. *arXiv preprint arXiv:2306.00186*, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Appendix

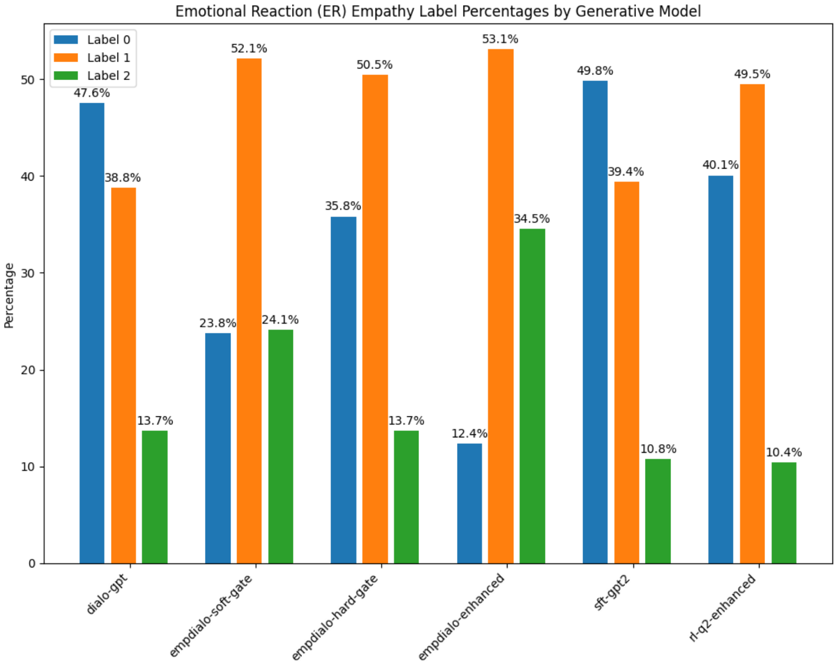

Figure 1: Emotional Reactions (ER) Generative Label Percentages

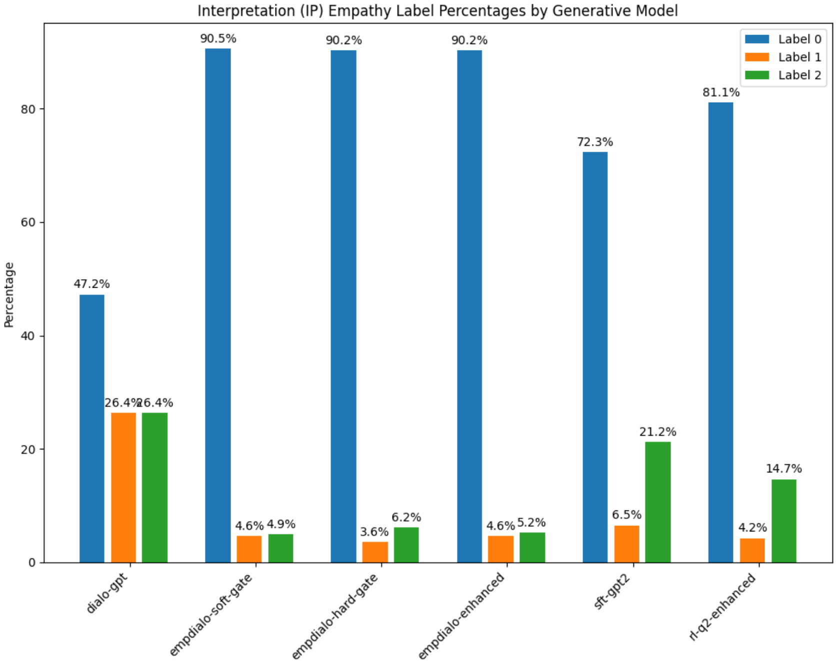

Figure 2: Interpretations (IP) Generative Label Percentages

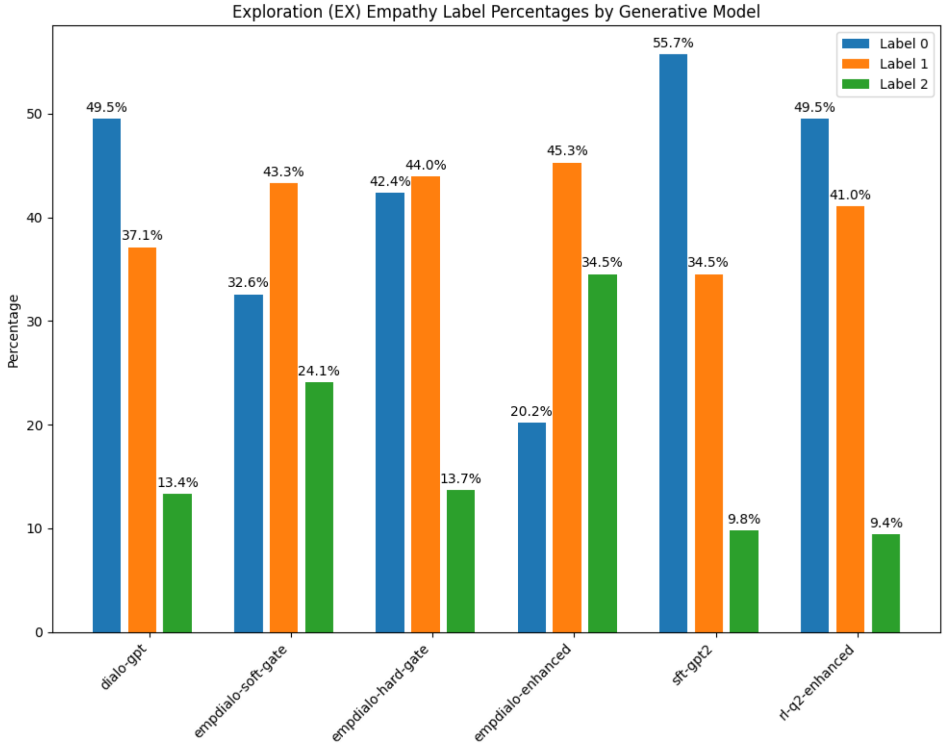

Figure 3: Explorations (EX) Generative Label Percentages

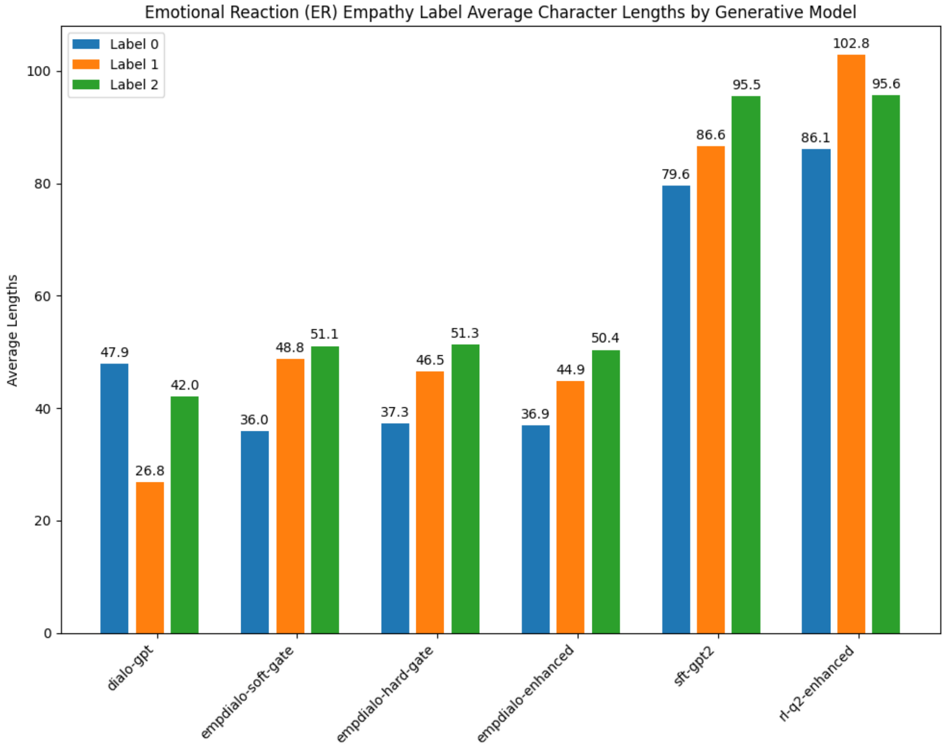

Figure 4: Emotional Reactions (ER) Generative Average Character Lengths

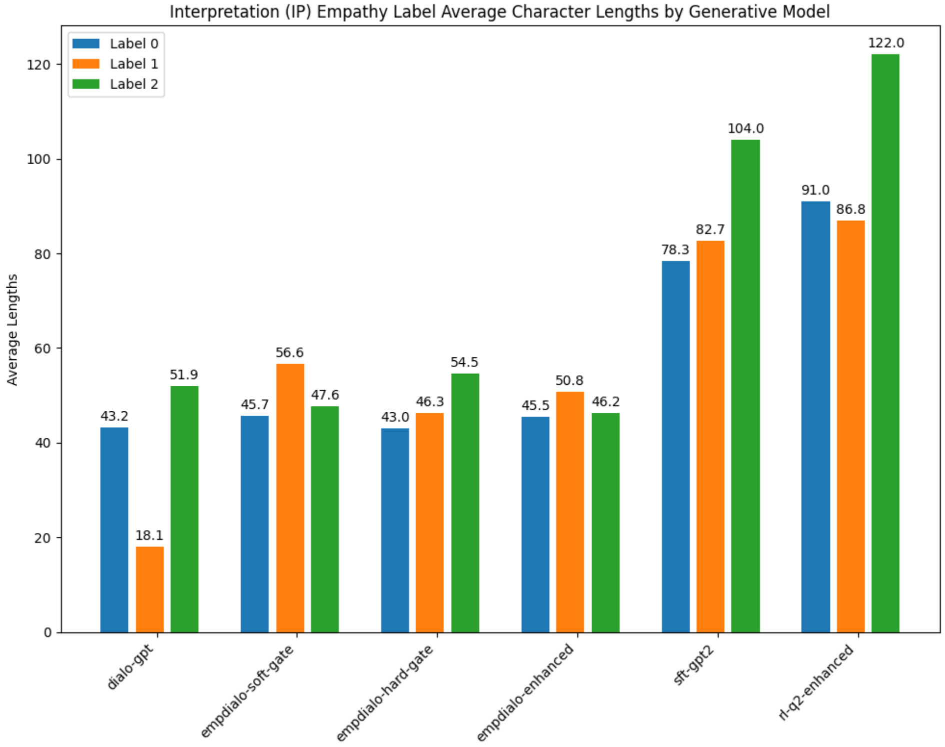

Figure 5: Interpretations (IP) Generative Average Character Lengths

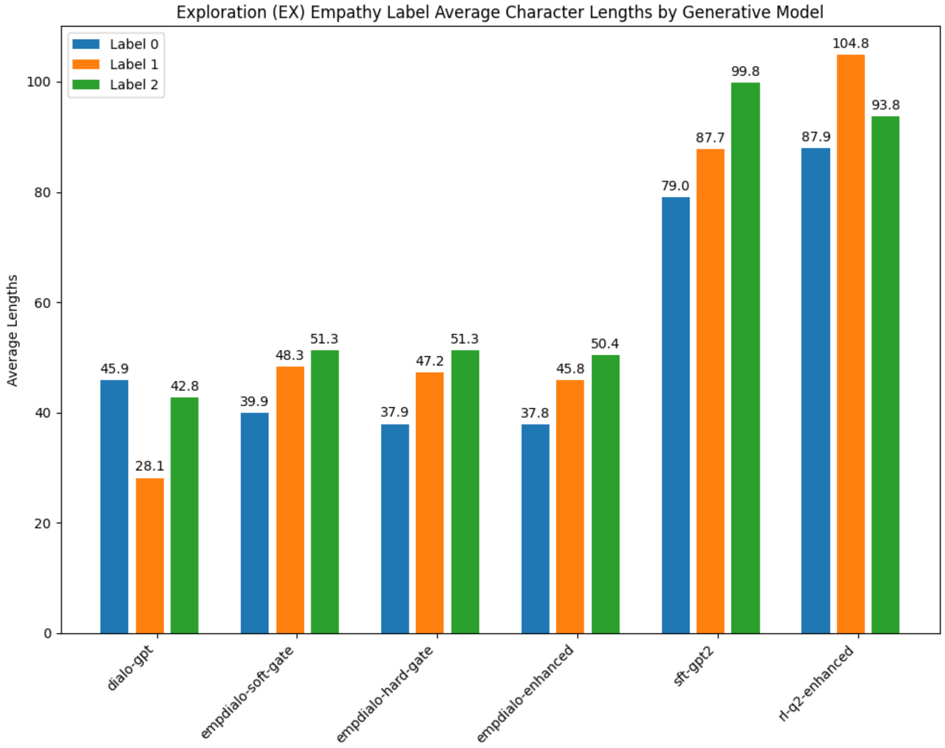

Figure 6: Explorations (EX) Generative Average Character Lengths