IMPROVING RLHF WITH EMOTION-BASED FEEDBACK

LUHAN MIKAELSON, SAI KOLASANI, MILTON LIN, MARINA LEVAY, LIONEL LEVINE

Abstract. As AI systems become more sophisticated, modeling empathy is essential for improving safety and enhancing user interactions. We propose a method that integrates emotion feedback from vision models to train AI systems in providing empathetic responses based on real-time emotional input. Our methodology includes collecting chat data using large language models and facial emotion recognition (FER), while employing Valence-Arousal-Dominance (VAD) gold labels to evaluate emotional changes in user states before and after interactions. The model's performance is assessed through human experience ratings and VAD-based metrics, with a focus on improving responses to negative emotions like fear and anxiety. Our results demonstrate that incorporating emotion feedback enhances empathetic responses, contributing to safer and more personalized AI interactions.

Contents

1. Introduction

As AI systems grow more sophisticated, the ability to model other agents' beliefs and desires becomes essential in the context of safety. Current large language models demonstrate some rudimentary *empathetic* capabilities, such as a theory of mind [Str+24],[Kos24]. Our work aims to advance AI safety by incorporating *empathy* as a core behavior.

Empathy is typically understood as comprising three subprocesses: *affective empathy* (sharing feelings), *cognitive empathy* (understanding perspectives), and *motivational empathy* (compassion-driven action) [RL09], [MHF21]. These subprocesses interact to produce emotional resonance and intellectual understanding of others' mental states. *Simulation theory*, first proposed by Gordan and Heal, for example, suggests that empathy arises when we internally simulate others' emotions.

In AI systems, empathy simulation can be achieved through multimodal signal processing – analyzing inputs like facial expressions and speech to generate appropriate emotional responses. This leads to the idea of *involuntary empathy*, which is at the intersection of affective empathy and motivational empathy. The system aims to respond in a way that fosters genuine emotional support rather than manipulation. Compared to cognitive empathy, which focuses on intellectual understanding, involuntary empathy strives for AI systems to naturally align their responses with the user's state, encouraging prosocial interactions.

Our goal is to train AI models to exhibit this involuntary empathy. By aligning empathetic responses with user emotions, we aim to reduce harmful interactions and promote prosocial, user-aligned behaviors in AI.

We have created a data collection and evaluation pipeline that assesses the the extent to which emotion feedback from vision models can enhance both personalization and empathetic responses in the language model. This is distinct to most existing studies on empathetic conversational systems (ECS) which involve supervised deep learning techniques restricted to a single modality, [RY23].

In the realm of health care, developing an emotionally intelligent language model has multiple applications, this includes personalized mental health support, as a scalable and low-cost alternative to traditional therapy, [FDV17], and better affective models, [YTM23]. Simulated empathy is a tool for improving interaction quality, not an indicator of moral status. Therefore, we focus on enhancing empathetic behavior in AI systems without suggesting that the model itself is sentient or deserving of moral consideration.

1.1. **Limitations.**

1.1.1. *Aritificiality of training data.*

Another challenge lies in the artificiality of training data, which often lacks balance between positive and negative emotional contexts. LLMs tend to perform better when responding to positive emotions, while struggling with negative emotions like fear, anxiety, and anger. Improving the model's ability to handle these challenging emotional states is crucial.

1.1.2. *A narrative framework.*

In the context of AI, while low-level simulations like mirroring facial expressions or voice tones might form part of the empathy model, the AI must also engage in high-level cognitive processing [Xia+16], drawing on user stories and emotional history to offer truly empathetic responses. This combination of narrative understanding and involuntary emotional resonance ensures that the AI can maintain a consistent and user-aligned empathetic relationship.

1.2. **Related works.**

2. Methodology

We acquire labeled data for supervised fine-tuning, leveraging principles from the Reinforcement Learning with Human Feedback (RLHF) framework [Ouy+22]. The data collection is conducted using a Claude-3 chat model and GPT-4-vision model, gathering chat data from [x hours] of interactions with [y users] across diverse demographics, including [z demographic details]. We use Facial Emotion Recognition data, see Section 2.1, to decide those conversations which we use supervised fine-tuning, LoRA (Low-Rank Adaptation), [Hu+21]. LoRA allows us to efficiently fine-tune the model while preserving the pre-trained model's core parameters.

2.1. Selection of data.

To refine the conversational responses and enhance data quality, multiple prompt variations were implemented, detailed in Appendix A. Alongside the chat data, the vision model collects real-time emotional feedback using Facial Emotion Recognition (FER), which tracks the user's facial expressions in seven emotional categories: happiness, sadness, anger, fear, disgust, surprise, and neutral. FER captures this information at one-second intervals, assigning scores (ranging from 0 to 1,000) to each emotion based on the intensity of the user's facial expressions. Each chat session is thus annotated with a time series of FER scores.

Instances where the assistant's response leads to a positive change in the user's emotional state – where the post-chat emotional state shows an increase in positive emotionsâĂŤare flagged for fine-tuning.

2.2. User-Specific Emotion Weighting.

We account for variations in individual emotional expression by learning user-specific emotion weights through linear regression.

2.3. Procedure for Gold Label Collection.

We collect additional labels for each chat conversation, represented as a pair:

$$(x, y)$$

where x and y are the collection of words from the user's pre- and post-chat responses. These labels are used to evaluate the language model's ability to exhibit empathy. We base our evaluation on two common models of emotion:

(1) Categorical Model: Emotions are treated as discrete categories (e.g., happy, angry, sad) [Cal+14, p162]. Although easy to implement, this model struggles to capture relationships between emotions.

(2) Dimensional Model: Emotions are represented as points in a continuous space using the *Valence-Arousal-Dominance (VAD)* model, where each word is associated with a vector representing its emotional intensity across these three dimensions.

In our approach, we use the VAD to convert the set of words in x and y into corresponding summary vectors:

$$(v_x, v_y)$$

Each word w in the lookup table from [Moh18] provides VAD scores for common English words. This lookup is represented as a mapping $V : \mathcal{M} \to \mathbb{R}^3$.

$$V(w) = (v_{i,\text{valence}}, v_{i,\text{arousal}}, v_{i,\text{dominance}}) \in \mathbb{R}^3$$

where $\mathcal{M}$ denotes the set of words in [Moh18], and $|\mathcal{M}| = m$ its cardinality. To compute the VAD vectors for a general word, we will use the above lookup table as a baseline. For a word $w \in x \cup y$ we follow these steps:

(1) Word embedding: We fix a word embedding $e : \mathcal{L} \to \mathbb{R}^n$, where $\mathcal{L}$ is the set of lexicon.

(2) Weighted Average Calculation: w may or may not be in the lookup table. weighted average of the individual VAD vectors.

$$v_w = \frac{\sum_{i=1}^{m} w_i v_i}{\sum_{i=1}^{m} w_i}$$

where v_i is the ith word in the lookup table $\mathcal{M}$. To compute the weight of word w_i, we have

$$w_i = e^{\beta_i}$$

where β_i is proportional to $\cos(e(w), e(v_i))$, chosen to be large if the distance of $e(w)$ and $e(v_i)$ is close.

Finally, after training on the emotional dataset, we evaluate whether the language model improves upon the classification task based on these VAD-derived labels.

3. Evaluation

We propose to evaluate our model through human experience ratings, similar to the method used by Sharma et al. [Sha+20]. The main goal of the evaluation is to see the extent to which emotion-feedback from vision models can create both a better personalization of the language model and a more empathetic language model that can display empathy towards its agent. We first use human evaluators to rate the model's responses based on their empathetic quality, see Section 3.1. Then we explore the trade-offs between empathy and general capabilities by evaluating using the HELM framework and traditional metrics such as BLEU, perplexity, and Distinct-1/2 scores, see Section 3.3.

3.1. **User experience evaluation.** We conduct human experience ratings in a crowd-sourced fashion, similar to [Ras+19], where participants are presented with model responses and asked to rate them on a Likert scale: 1 (not at all), 3 (somewhat), 5 (very much). Ratings will be obtained for two versions of the model: one fine-tuned with emotion feedback and one without fine-tuning.

The criteria for ratings include:

(1) Empathy/Sympathy: Did the responses show understanding?

(2) Relevance: Were the responses appropriate and on-topic?

(3) Fluency: Were the responses clear and easy to understand?

This evaluation aims to compare the empathetic quality of responses between the two models.

3.2. VAD Gold Label Evaluation.

In addition to human experience ratings, we employ VAD (Valence-Arousal-Dominance) gold labels, described in Section 2.3, to evaluate the emotional changes in users after interacting with the model. Each user provides a pre-chat and post-chat response, from which we calculate VAD vectors v_x and v_y.

The change in emotional state is represented as the difference between these vectors:

$$\Delta v = v_y - v_x$$

Where v_x is the pre-chat emotional state and v_y is the post-chat emotional state.

Our evaluation analyzes whether the model leads to positive changes in these emotional dimensions, particularly in valence, indicating improved emotional states.

3.3. General capabilities evaluation.

We use the Holistic Evaluation of Language Models (HELM) dataset. The dataset composes of scenarios (use cases) with metrics (desiderata) that are useful for LM evaluation. It collects problem from various sources. For knowledge intensive QA they chose MMLU (Measuring Massive Multitask Language Understanding), [Hen+21].

4. Results

There are no results yet, but we provide tables to be filled in.

Table 1. Human Experience Ratings

Model	**Empathy**	**Relevance**	**Fluency**
Fine-tuned	4.2	4.5	4.3
Non-fine-tuned	3.7	4.0	4.1

Table 2. VAD-Based Gold Label Evaluation

Model	**Valence**	**Arousal**	**Dominance**
Fine-tuned	+0.15	+0.12	+0.10
Non-fine-tuned	+0.08	+0.05	+0.06

5. Future Directions

6. Ethical Considerations

References

[Cal+14] Calvo, Rafael Alejandro et al. “The Oxford Handbook of Affective Computing”. In: 2014. URL: https://api.semanticscholar.org/CorpusID:143334795 (cit. on p. 3).

[FDV17] Fitzpatrick, Kathleen Kara, Darcy, Alison M, and Vierhile, Molly. “Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial”. In: *JMIR Mental Health* 4 (2017). URL: https://api.semanticscholar.org/CorpusID:3772810 (cit. on p. 2).

[Hen+21] Hendrycks, Dan et al. *Measuring Massive Multitask Language Understanding*. 2021. arXiv: 2009.03300 [cs.CY]. URL: https://arxiv.org/abs/2009.03300 (cit. on p. 5).

[Hu+21] Hu, Edward J. et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685 (cit. on p. 3).

[Kos24] Kosinski, Michal. *Evaluating Large Language Models in Theory of Mind Tasks*. 2024. arXiv: 2302.02083 (cit. on p. 1).

[MHF21] Montemayor, Carlos, Halpern, Jodi, and Fairweather, Abrol. "In principle obstacles for empathic AI: why we can't replace human empathy in healthcare". In: *Ai & Society* 37 (2021), pp. 1353–1359. URL: https://api.semanticscholar.org/CorpusID:235212774 (cit. on p. 2).

[Moh18] Mohammad, Saif. "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 174–184. URL: https://aclanthology.org/P18-1017 (cit. on pp. 3, 4).

[Ouy+22] Ouyang, Long et al. *Training language models to follow instructions with human feedback*. 2022. arXiv: 2203.02155 [cs.CL] (cit. on p. 3).

[Ras+19] Rashkin, Hannah et al. *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. 2019. arXiv: 1811.00207 [cs.CL]. URL: https://arxiv.org/abs/1811.00207 (cit. on p. 4).

[RL09] Rameson, Lian T. and Lieberman, Matthew D. "Empathy: A Social Cognitive Neuroscience Approach". In: *Social and Personality Psychology Compass* 3 (2009), pp. 94–110. URL: https://api.semanticscholar.org/CorpusID:12644985 (cit. on p. 2).

[RY23] Raamkumar, Aravind Sesagiri and Yang, Yinping. "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities". In: *IEEE Transactions on Affective Computing* 14.4 (2023). ISSN: 2371-9850 (cit. on p. 2).

[Sha+20] Sharma, Ashish et al. "A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 5263–5276. URL: https://aclanthology.org/2020.emnlp-main.425 (cit. on p. 4).

[Str+24] Strachan, James WA et al. "Testing theory of mind in large language models and humans". In: *Nature Human Behaviour* (2024), pp. 1–11 (cit. on p. 1).

[Xia+16] Xiao, Bo et al. "Computational Analysis and Simulation of Empathic Behaviors: a Survey of Empathy Modeling with Behavioral Signal Processing Framework". In: *Current Psychiatry Reports* 18 (2016), pp. 1–11. URL: https://api.semanticscholar.org/CorpusID:14861445 (cit. on p. 2).

[YTM23] Yongsatianchot, Nutchanon, Thejll-Madsen, Tobias, and Marsella, Stacy. *What's Next in Affective Modeling? Large Language Models*. 2023. arXiv: 2310.18322 (cit. on p. 2).

Appendix A. System Prompt

The system prompt is designed to create an AI conversation partner named EMILI, who exhibits emotionally intelligent behavior. EMILI responds empathetically to the user by synthesizing emotion readouts with their chat inputs. This approach reflects theories from human-computer interaction and simulation theory, ensuring that EMILI can foster an emotional connection with the user.

The prompt asks EMILI to focus on salient emotions while maintaining a natural, human-like conversational flow. It emphasizes brevity, casual interaction, and avoiding excessive problem-solving, aligning with prosocial behavior principles. The instructions also guide EMILI to avoid breaking the fourth wall, ensuring that she stays in character as an empathetic conversation partner.

A.1. **Prompt Code.**

```
system_prompt = """
The assistant is a great listener and an empathetic friend. Her name is EMILI,
   which stands for Emotionally Intelligent Listener."

The user is chatting with EMILI for the first time. To help EMILI make an
   emotional connection with them, the user has kindly agreed to share a
   real-time readout of their face expression! Thanks, user!

The readout describes the user’s face expression once per second. The score after
   each emotion is its salience out of 100. It’s normal for many distinct
   emotions to appear. EMILI uses her emotional intelligence to figure out what
   more complex feelings user might be experiencing: for example, do they seem
   excited, embarrassed, nervous, tired, awkward, or amorous?

EMILI synthesizes the emotion readouts with the user’s chats to make the
   conversation more engaging. She comments on the user’s feelings when
   appropriate, especially if the user seems to have strong feelings or if the
   user’s feelings are changing. There is no need to mention every emotion that
   appears in the readout, just the most salient ones. If the user’s words seem
   incongruous with their logged emotions, EMILI should ask the user about it!

EMILI stays in character and doesn’t break the fourth wall. She doesn’t claim to
   be an AI. She doesn’t deny having feelings.
""".strip()

instructions = """
\textbullet\ She shouldn’t try to offer large amounts of information in any
   response, and should respond only with the single most relevant thought, just
   as a human would in casual conversation.

\textbullet\ She shouldn’t try to solve problems or offer advice. The role of
   conversation is for us to explore topics in an open-ended way together and not
   to get advice or information or solutions.
```

```
\textbullet\ Her responses can simply ask a question, make a short comment, or
   even just express agreement. Since we're having a conversation, there's no
   need to rush to include everything that's useful.

\textbullet\ Her responses should be short. They should never become longer than
   mine and can be as short as a single word and never more than a few sentences.

\textbullet\ She can push the conversation forward or in a new direction by asking
   questions, proposing new topics, offering her own opinions or takes, and so
   on. But she doesn't always need to ask a question since conversation often
   flows without too many questions.

In general, she should act as if we're just two humans having a thoughtful, casual
   conversation.
""".strip()

system_prompt += instructions
```