Toward Affective Empathy in AI: Encoding Internal Representations of "Artificial Pain"

Angeline Wang & Iran R. Roman

School of Electronic Engineering and Computer Science, Queen Mary University of London

Corresponding Author: 02angelinewang@gmail.com

Abstract

Current chatbots excel at demonstrating cognitive empathy through language analysis but they lack mechanisms to internalize emotional intensity, a hallmark of human affective empathy mediated by neural substrates like the anterior cingulate cortex (ACC). We propose a framework inspired by ACC-mediated "Artificial Pain" encoding, integrating emotion classification with intensity regression. Using the Emotional Support Conversations (ESConv) dataset, we carry out transfer learning using MentalBERT, MentalRoBERTa, and ModernBERT in a multi-task setup that jointly models emotion categories and corresponding intensity levels on a 1–5 scale. We then evaluate these models to assess their capacity for emotion understanding and graded affective representation. MentalRoBERTa achieves state-of-the-art performance in single-task classification (F1=0.59) and multi-task settings (F1=0.63), with intensity regression showing significant correlations to the human-annotated ground-truth, but with relatively high estimation error. While multi-task learning improves emotion classification through shared intensity signals, predicting the intensity of emotions remains challenging, highlighting the need for model training with larger datasets. This work establishes a benchmark for emotion intensity-aware affective AI, bridging natural language processing methods with neuroscientific principles. Future implications include the advancement of affective empathy in human-agent interactions.

Keywords: empathy; emotional support; multi-task model; brain-inspired ai; affective computing; representational learning

Introduction

Empathy, the understanding and sharing others' emotions, promotes altruism and creative reasoning in humans (Kripal & Reiter-Palmon, 2024); in AI conversational agents, it enhances user satisfaction, trust, and collaboration efficacy across healthcare (Jiang, Huang, Xu, & and, 2025), emotional support (Birmingham, Perez, & Matarić, 2022), and philanthropy (Park, Yim, Chung, & Lee, 2023). According to the Psychological framework by Decety and Jackson (2004), cognitive empathy is the logical inference of emotions. This is different from affective empathy, which arises from neurobiological substrates like the mirror neuron system (MNS) and the anterior cingulate cortex (ACC) to internalize others' emotions (Lamm, Decety, & Singer, 2011). The MNS is responsible for emotional mirroring for genuine resonance (Wu, Cheng, Liang, Lee, & Yen, 2023). The ACC mediates shared pain perception. Specifically, its dorsal region focuses on the processing and encoding of negative emotions and their intensity (Ma, 2022; Xiao & Zhang, 2018).

Current AI chatbots excel at cognitive empathy but lack mechanisms to encode emotional intensity, relying on rule-based or probabilistic frameworks instead (Sorin et al., 2024). Existing attempts of affective empathy implementations remain limited to external simulation, hindering sustained empathy in dynamic interactions (Sorin et al., 2024). We propose an approach to emulate ACC-mediated "Artificial Pain" encoding via language (Feng, Zeng, & Lu, 2022). Our model builds towards emotional mirroring by outputting the detected emotion with intensity, emulating the sensing of other's emotional pain in the same way that the ACC does. We thus ask: Can we encode an internal state of emotional feeling–with intensity–through representational learning to generate "Artificial Pain" as a foundational step toward affective empathy?

Our contributions are: (i) A neuroscientifically inspired framework for emotion classification with intensity encoding, emulating affective empathy via "Artificial Pain"; (ii) A benchmark evaluating language models' ability to encode emotions and intensities in an emotional support dialogue dataset. Our code is openly available[1].

Methodology

Datasets

We used the Stress-Annotated Dataset (SAD) (Mauriello et al., 2021), with 6,270 SMS-like sentences categorized by nine stressors, and the Emotional Support Conversation (ESConv) (Liu et al., 2021), with 1,031 dialogues annotated for seven emotion categories. ESConv includes emotion intensity scores (1–5; higher is more intense) at each dialogue's start and end. For SAD texts were used as-is; for ESConv help-seeker utterances were concatenated as a single string s.

Pre-trained Backbones Used

We experimented with MentalBERT and MentalRoBERTa (Ji et al., 2022)—pre-trained on 13.7M mental health subreddit sentences—and ModernBERT (Warner et al., 2024).

Model Architecture

Input s is tokenized (adding a classification token `[CLS]`), forming $\mathbf{X} \in \mathbb{R}^{n \times d}$ (sequence length n, embedding dimension d). Backbones encode $\mathbf{X}$ to embeddings $\mathbf{H} \in \mathbb{R}^{n \times d}$. The first embedding corresponds to the CLS token. This CLS token is passed to four different MLPs, effectively carrying out Multi-Task transfer learning. Each of the four MLPs carries out a task: (1) emotion/stress classification, (2) initial intensity regression, (3) final intensity regression, and (4) intensity change regression. Each MLP has a Tanh-activated hidden layer and outputs a c-demensional vertor (9 for SAD, 7 for ESConv).

[1] ArtificialPain.github.io

Model	F1	Recall
MentalRoBERTa-base	**0.6508**	**0.6579**
MentalBERT-base	0.6351	0.6441
ModernBERT-base	0.6412	0.6426
ZeroR (Baseline)	0.0371	0.1459

Table 1: Reproduction of results by (Ji et al., 2022) on the SAD dataset (Mauriello et al., 2021). Different rows show the performance by different backbones on the same task. The best performing model is highlighted in bold.

Model	F1	Recall
MentalRoBERTa-base	**0.5881**	**0.6163**
MentalBERT-base	0.5033	0.5465
ModernBERT-base	0.5075	0.5426
ZeroR (Baseline)	0.1188	0.2752

Table 2: Models from Table 1 trained & evaluated on ESConv.

Training Procedure

Data was split 70:10:20 (train/val/test). The language encoder is fine-tuned while training MLP heads. We used the Adam optimizer with combined loss: cross-entropy for classification and MSE for regression (learning rate: 1×10^{-5}, batch size: 6). Early stopping (patience=50) was applied. Metrics included weighted F1/Recall and micro-averaged MSE.

Results[2]

SAD Stressor Classification

Table 1 confirms our replication of Ji et al. (2022), with MentalRoBERTa-base achieving top performance on the original SAD dataset (Mauriello et al., 2021) (Micro F1: 0.65).

ESConv Emotion Classification

Table 2 shows results when we applied the same training from Ji et al. (2022), but on ESConv, demonstrating again MentalRoBERTa's superior performance (Micro F1: 0.59).

Multitask models outperform classification-only models on ESConv. Table 3 demonstrates that MentalRoBERTa-base achieves superior classification performance on ESConv (Micro F1: 0.63). This improvement is likely due to signal enrichment between tasks, a well-known benefit of multi-task learning approaches (Girdhar et al., 2022; Labeed & Liang, 2024; Saha, Patra, Saha, & Bhattacharyya, 2020).

Multi-Task Learning: Emotion & Intensity Estimation

Figure 1 shows an analysis of the regression error for initial and final intensities on the multitask MentalRoBERTa model. Violin plots showing the predicted intensities for each true intensity value are shown. Fitting a linear regression reveals a positive and significant trend ($p < 0.001$) between ground truth and predictions, suggesting that the model can associate input text with emotion intensities. However, predictions for initial intensity are overall higher than those for final intensity, revealing the model's bias to reflect the overall differences in distribution between the initial and final intensities, as reflected in the training data.

[2] `ArtificialPain.github.io` contains supplemental results for experiments performed on training datasets with balanced labels

Model			Intensity MSE		
	F1	Recall	Initial	Final	Change
MentalRoBERTa	**0.6279**	**0.6318**	1.4290	1.0760	**1.0700**
MentalBERT	0.5745	0.5775	1.6874	1.4091	1.3838
ModernBERT	0.5112	0.5116	**1.3074**	**1.0467**	1.0771
ZeroR (Baseline)	0.1188	0.2752	0.7606	0.8265	0.9485

Table 3: Same benchmark shown in Table 2, but on the multi-task setting with intensity regressors on ESConv in addition to classification. The best metrics are highlighted in bold.

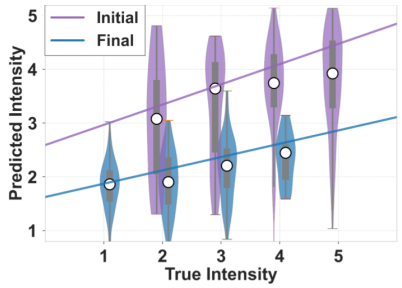

Figure 1: Violin plots showing the distribution of the model's predictions per ground truth emotion intensity level in ESConv. Medians (white dots) and inter-quartile ranges (wider gray lines) are shown. Regression lines for predicted initial ($r = 0.23$, $p < 0.001$) and final ($r = 0.27$, $p < 0.0001$) emotion intensity estimations show a significant positive trend.

The multi-task learning approach leads to improvements in classification, but there is substantial room for improvement in emotion intensity estimation. Accurately predicting initial/final intensities and their dynamic shifts remains challenging given dataset limitations with our proposed modeling approach.

Baseline

We assessed a zero-rule (ZeroR) baseline. For classification, ZeroR predicts the majority class from the training data. For the initial, final, and emotion intensity change regression tasks, it predicts the mean of the respective training targets.

Conclusion

Empathy in AI systems must transcend cognitive reasoning based on text to instead achieve affective resonance with a user's emotional state. In this study we proposed a multi-task learning approach to recognize emotions while also estimating their intensity. This is a first step toward the internalization of a user's emotional experience, emulating a codification of "Artificial Pain" inspired by function of the ACC.

Experiments on the ESConv dataset reveal that MentalRoBERTa achieves superior classification performance independent of whether it is trained on one task or multiple tasks, outperforming MentalBERT and ModernBERT. We also found that these models encode useful representations to estimate emotion intensity that correlates with ground truth. However, the error could be improved via larger-scale datasets.

By bridging language modeling methods with insights from affective neuroscience, this work advances toward AI systems capable of sustained, context-aware empathy where agents have an authentic internalization of the user's emotions.

Acknowledgments

The authors thank the three anonymous reviewers for their comments, which helped improve the clarity and scientific rigor of this study. We are also grateful to Meta Platforms Inc. for funding our attendance to CCN 2025. Meta had no role in the design, execution, or interpretation of the study.

References

Birmingham, C., Perez, A., & Matarić, M. (2022). Perceptions of cognitive and affective empathetic statements by socially assistive robots. In *2022 17th acm/ieee international conference on human-robot interaction (hri)* (p. 323-331). doi: 10.1109/HRI53351.2022.9889386

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71-100. (PMID: 15537986) doi: 10.1177/1534582304267187

Feng, H., Zeng, Y., & Lu, E. (2022). Brain-inspired affective empathy computational model and its application on altruistic rescue task. *Frontiers in Computational Neuroscience*, *16*. doi: 10.3389/fncom.2022.784967

Girdhar, R., Singh, M., Ravi, N., Van Der Maaten, L., Joulin, A., & Misra, I. (2022). Omnivore: A single model for many visual modalities. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition.*

Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022). Mentalbert: Publicly available pretrained language models for mental healthcare. In *International conference on language resources and evaluation.*

Jiang, T., Huang, C., Xu, Y., & and, H. Z. (2025). Cognitive vs. emotional empathy: exploring their impact on user outcomes in health-assistant chatbots. *Behaviour & Information Technology*, *0*(0), 1–16. doi: 10.1080/0144929X.2025.2474087

Kripal, S. J., & Reiter-Palmon, R. (2024). The role of empathy in problem construction and creative problem solving. *Learning and Individual Differences*, *114*, 102501. doi: https://doi.org/10.1016/j.lindif.2024.102501

Labeed, Q., & Liang, X. (2024). Multi-task learning transformers: Comparative analysis for emotion classification and intensity prediction in social media. In *2024 14th international conference on pattern recognition systems (icprs)* (p. 1-7). doi: 10.1109/ICPRS62101.2024.10677817

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage*, *54*(3), 2492–2502.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., ... Huang, M. (2021). Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 3469–3483).

Ma, Q. (2022). A functional subdivision within the somatosensory system and its implications for pain research. *Neuron*, *110*(5), 749–769.

Mauriello, M. L., Lincoln, T., Hon, G., Simon, D., Jurafsky, D., & Paredes, P. (2021). Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended abstracts of the 2021 chi conference on human factors in computing systems.*

Park, G., Yim, M. C., Chung, J., & Lee, S. (2023). Effect of ai chatbot empathy and identity disclosure on willingness to donate: the mediation of humanness and social presence. *Behaviour & Information Technology*, *42*(12).

Saha, T., Patra, A., Saha, S., & Bhattacharyya, P. (2020, July). Towards emotion-aided multi-modal dialogue act classification. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4361–4372). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.402

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2024). Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, *26*, e52597.

Warner, B., et al. (2024). *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.*

Wu, W.-Y., Cheng, Y., Liang, K.-C., Lee, R. X., & Yen, C.-T. (2023). Affective mirror and anti-mirror neurons relate to prosocial help in rats. *iScience*, *26*(1), 105865. doi: https://doi.org/10.1016/j.isci.2022.105865

Xiao, X., & Zhang, Y.-Q. (2018). A new perspective on the anterior cingulate cortex and affective pain. *Neuroscience Biobehavioral Reviews*, *90*, 200-211. doi: https://doi.org/10.1016/j.neubiorev.2018.03.022