Self-explaining emotion classification through preference-aligned large language models

Muhammad Hammad Fahim Siddiqui, Diana Inkpen, and Alexander Gelbukh

University of Ottawa, IPN Mexico

Abstract. Recent advancements in large language models (LLMs) have shown promise for NLP applications, yet producing accurate explanations remains a challenge. In this work, we introduce a self-explaining model for classifying emotions in X posts and construct a novel preference dataset using chain-of-thought prompting in GPT-4o. Using this dataset, we guide GPT-4o with preference alignment via the Direct Preference Optimization (DPO). Beyond GPT-4o, we adapt smaller models such as LLaMA 3 (8B) and DeepSeek (32B distilled) through preference tuning using Odds Ratio Preference Optimization (ORPO), significantly boosting their classification accuracy and explanation quality. Our approach achieves state-of-the-art performance (68.85%) on the SemEval 2018 E-c multilabel emotion classification benchmark, exhibits comparable results on the DAIR AI multiclass dataset and attains a high sufficiency score—indicating the standalone effectiveness of the generated explanations. These findings highlight the impact of preference alignment for improving interpretability and enhancing classification.

Keywords: LLMs, preference alignment, emotion classification

1 Introduction

Large Language Models (LLMs) and artificial intelligence (AI) systems have gained significant traction in various domains in natural language processing, including emotion classification. The need for explanations in AI systems, particularly in the context of emotion classification, is important for fostering trust and understanding among users. Explainable AI (XAI) aims to provide insights into the decision-making processes of these models, thereby enhancing their transparency and accountability [1][2]. Providing explanations for the classifications made by AI systems is crucial for users to understand the rationale behind the model's decisions. This understanding is particularly vital in sensitive applications such as mental health assessments, where misclassifications can have serious consequences [3][4]. Furthermore, the ability to explain AI decisions can help mitigate biases and improve the overall performance of emotion classification systems by allowing developers to refine their models based on user feedback and insights [5][6]. Aligning LLMs models with human preferences remains a persistent challenge. Preference optimization [7][9][8] has emerged as a promising approach to address this challenge, enabling control over model behavior by integrating direct human feedback into the training loop.

We use GPT-4o and chain-of-thought prompting to create a preference alignment dataset specifically designed for the task of emotion classification and explanation. Next, we employ two distinct preference alignment strategies. For GPT-4o, we adopt Direct Preference Optimization (DPO) to fine-tune the model's responses, and for smaller models such as LLaMA 3 (8B) and a distilled version of DeepSeek R1 (32B), we apply Odds-Ratio Preference Optimization (ORPO) [8].

Our experimental evaluation covers both multilabel classification on the SemEval 2018 E-c dataset [10] and single-label classification on the DAIR AI emotion dataset [11]. Results show that our DPO-tuned GPT-4o model not only sets a new benchmark in multilabel performance, but also achieves comparable state-of-the-art results in the multiclass setting. Furthermore, to quantify the quality of model-generated explanations, we incorporate the

David C. Wyld et al. (Eds): NLCAI, AIFU, CCSEA, BIoT, SEA, SIPRO, BDML, CLOUD – 2025
pp. 351-362, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151027

Sufficiency metric [12] from the FRESH pipeline [13], where our model exceeds the previous best-reported values on similar datasets. We observe that preference alignment enhances the model's ability to adhere to a consistent output format, producing explanations and classification labels in a predictable, structured manner.

By making both our code and dataset publicly available, we aim to encourage further exploration of preference alignment techniques and support wider adoption of self-explaining LLMs across diverse natural language understanding tasks. Our findings indicate that preference-aligned models strike a promising balance between accuracy and interpretability, offering a new avenue for transparent and user-centric AI.

2 Related work

Emotion classification focuses on identifying and categorizing emotions expressed in textual data. Accurate emotion classification can provide valuable insights into public sentiment, enabling businesses, policymakers, and researchers to understand societal trends and reactions [14][15]. Moreover, multilabel classification, which allows for assigning multiple labels to a single instance, is essential as emotions are often complex and overlapping. For instance, a single tweet may express both joy and sadness, necessitating a multilabel approach to capture the full spectrum of emotions [16]. Numerous studies have explored methodologies for emotion classification, particularly focusing on multilabel and multiclass approaches. For instance, Ferreira and Vlachos introduced a multilabel stance detection method that incorporates label dependencies, demonstrating improved performance over traditional models [17]. Furthermore, ensemble methods have proven beneficial in managing the complexities of multilabel stream classification [18]. The integration of instruction tuning for large language models (LLMs) has also emerged as a promising avenue for improving multilabel emotion classification, allowing models to better adapt to human expressions [19].

The rise of explainable AI has become increasingly important in classification systems, including emotion classification. XAI aims to provide transparency in AI decision-making processes, allowing users to understand how and why certain classifications are made. This is particularly crucial in sensitive applications, such as mental health assessments, where the implications of misclassification can be profound. The literature emphasizes that explanations can improve user trust and facilitate better human-AI collaboration [20][21]. Recent evaluations of various explanation methods have underscored their utility in providing insights into decision-making processes and shown how traditional statistical and feature attribution methods are insufficient for adequate explanations [22]. Additionally, developing self-explaining architectures aims to inherently incorporate interpretability into neural text classifiers, [24]. Recent advancements in AI, particularly with the emergence of LLMs, have transformed various applications, but making these models behave according to task-specific needs remains challenging. Preference alignment algorithms have gained attention as a means to align AI models with human preferences, ensuring that AI outputs are more relevant and acceptable to users [25]. Various preference alignment techniques, such as reinforcement learning from human feedback (RLHF), have been instrumental in fine-tuning models for applications ranging from content moderation to personalized recommendations [25]. These algorithms enable models to learn from user interactions, thereby enhancing their performance in real-world scenarios [25]. Despite advancements in LLMs, uncontrolled generation remains a significant challenge, often leading to hallucinations—instances where the model generates false or nonsensical information. This phenomenon can undermine the reliability of AI systems, particularly in applications requiring factual accuracy, such as news generation or medical advice [26][27]. The literature

indicates that hallucinations can arise from various factors, including insufficient training data and the inherent complexity of language generation tasks [26][27]. Addressing this issue is critical to ensuring the safe deployment of LLMs in sensitive applications. Preference alignment enhances the relevance of AI outputs and can be leveraged to create self-explaining models with controlled quality explanations. By aligning model behavior with user preferences, developers can ensure that the explanations generated by AI systems are not only accurate but also tailored to the user's context and needs [25]. This approach can significantly improve user trust and satisfaction, as users are more likely to engage with systems that provide clear and relevant explanations for their outputs [25]. Integrating preference alignment in self-explaining models represents a promising direction for future research, aiming to bridge the gap between complex AI systems and user understanding.

3 Data preparation

A core contribution of our work is the creation of a preference dataset tailored to emotion classification and explanation. In the context of large language models (LLMs), a preference dataset consists of multiple candidate responses for the same input, accompanied by explicit human judgments of which response is preferred. These human judgments provide the model with valuable signals on what constitutes a "better" output, enabling fine-tuning toward outputs more closely aligned with human expectations.

3.1 Base dataset and prompting strategy

We utilized the GPT-4o model to create a synthetic dataset, following a methodology similar to the Self-Instruct [28] approach. We start with the SemEval 2018-Ec multilabel emotion classification dataset (see Table 2 for dataset details). Specifically, we use the training and validation splits (7,724 tweets in total). For each tweet, we prompt GPT-4o to generate two candidate responses. Each response is structured in the following JSON-like format containing the fields shown in table 1. The explanation field contains a concise but informative rationale linking textual elements in the tweet to the predicted emotions. The most_prominent_emotion field indicates the single dominant emotion in the tweet, while multiple_emotions_present enumerates all predicted emotions.

3.2 Correctness criterion and dataset filtering

To ensure high-quality preference samples, we include only those pairs of responses where at least one of the two responses meets our correctness threshold. A response is deemed correct if (1) the multiple_emotions_present field exactly matches the gold-standard set of emotions for that tweet, and (2) the most_prominent_emotion is one of the correctly identified emotions. Out of the 7,724 tweets, 1,495 met these criteria-constituting our final preference dataset.

3.3 Human Annotation and Scoring

Each of the 1,495 pairs of GPT-4o responses was then manually inspected by two domain experts and one adjudicator to decide on conflicts. The annotators scored both responses along the following four dimensions:

1. Correctness: Evaluates whether the response's predicted emotions align with the gold labels, and whether the explanation logically supports those labels. This is the highest-priority dimension, particularly when deciding which response is ultimately "preferred."
2. Clarity: Measures how unambiguous and straightforward the explanation is. Vague or contradictory responses receive lower scores.
3. Helpfulness: Assesses how effectively the explanation explains the reasoning behind the predicted emotions, highlighting salient parts of the tweet.
4. Verbosity: Gauges whether the explanation is sufficiently detailed without being excessively long or repetitive. Overly concise or unnecessarily verbose explanations receive lower scores.

Each dimension is rated on a scale of 1 to 5, and an overall rating (1–5) summarizes the annotators' judgment across all four dimensions for each of two responses for each tweet. While these scores were not strictly required by the specific preference alignment algorithm that we employed [8] [9], they provide rich metadata for future work in reinforcement learning from human feedback (RLHF) and other preference-based optimization approaches. During the annotation process, we also fixed minor errors in the chosen response to perfect its correctness, clarity, helpfulness, and verbosity. Table 1 shows an annotated example from our preference dataset.

Field	**Value**
ID	2018-En-12345
Tweet	There's no one right way to live life but: be kind, be genuine, be honest, be open. #Love yourself & #laugh often.
Gold multilabel	['joy', 'love', 'optimism']
Response 1	**"explanation"**: "The tweet promotes positivity and well-being, focusing on being kind, genuine, and open. The use of hashtags like #Love and #laugh often, along with the mention of mindfulness, suggests a strong emphasis on love and joy. **"most_prominent_emotion"**: "love", **"multiple_emotions_present"**: ["love", "joy", "trust", "optimism"]
Response 2	**"explanation"**: "The tweet emphasizes positivity and self-care, which are indicative of emotions like love and joy. The use of words such as 'be kind,' 'be genuine,' and 'love yourself' strongly suggests love as the primary emotion. The encouragement to 'laugh often' and the hashtag #Mindfulness contribute to a sense of joy and optimism. The overall tone is uplifting and encouraging, promoting a positive outlook on life.", **"most_prominent_emotion"**: "love", **"multiple_emotions_present"**: ["love", "joy", "optimism"]
Chosen response	1 or 2 (according to correctness)
Response 1 ratings	Correctness[1-5], Clarity[1-5], Helpfulness[1-5], Verbosity[1-5], Overall[1-5]
Response 2 ratings	Correctness[1-5], Clarity[1-5], Helpfulness[1-5], Verbosity[1-5], Overall[1-5]
Chosen response fixed	Yes or No (if the chosen response was manually fixed for errors)

Table 1. Example of annotated response pairs from our preference dataset

3.4 Data generation settings

We configured GPT-4o to produce two responses per tweet with moderate stochasticity to encourage diverse outputs. We set the model's temperature to 0.4 and top-p to

0.8, aiming to balance creativity with focus. Higher temperature or top-p values could introduce more varied yet potentially off-topic content; lower values might yield overly deterministic responses that lack richness. In practice, this configuration allowed us to gather responses that varied enough in style and detail to be meaningfully compared and scored by human annotators. To further ensure the generated responses conform to a consistent structure, we employed the OpenAI structured outputs utility. We defined the required JSON schema using Python's Pydantic library, and then passed that schema to the chat_completions function call. This setup enforced that GPT-4o produced outputs in our desired format and emotion values comprising explanation, most_prominent_emotion, and multiple_emotions_present—thereby streamlining both automatic parsing and human annotation.

We evaluated our model on two Twitter-based emotion classification datasets with distinct label configurations. The DAIR AI dataset [11] contains six possible emotions—anger, fear, sadness, joy, disgust, fear—where each tweet is assigned a single most prominent emotion. In contrast, the SemEval 2018 E-c dataset [10] encompasses eleven possible emotions—anger, fear, sadness, joy, disgust, fear, optimism, pessimism, sadness, surprise, trust, neutral—and permits multiple emotions to co-occur within a single tweet. Accordingly, our model uses the most_prominent_emotion field to handle DAIR AI and the multiple_emotions_present field for SemEval 2018 E-c.

To ensure that the model's outputs conform to these respective label sets, we defined two Pydantic schema objects reflecting the allowable outputs for each dataset. These schemas were then passed as constraints to the Chat Completions API (via OpenAI's structured outputs utility). We could have included these restrictions solely within the model's prompt, but large language models often struggle to consistently adhere to textual format directives alone [29], [30], [31]. By providing an explicit schema, the model is programmatically constrained to produce outputs that align with the specified fields and label sets for each dataset, leading to more reliable and parsable results. The resulting preference dataset comprises 1,495 tweet-responses pairs with comprehensive human evaluations. This dataset constitutes the foundation for the subsequent preference alignment of GPT-4o and other LLMs, guiding them to produce explanations and predictions that match ground-truth labels more accurately and better align with human notions of quality, clarity, and helpfulness.

Dataset	**Train**	**Dev**	**Test**	**Total**
SemEval-2018 (Multi-label) [10]	6,838	886	3,259	10,983
DAIR AI (Multi-class) [11]	16,000	2,000	2,000	20,000

Table 2. SemEval 2018 E-c and DAIR AI datasets distribution statistics.

4 Methodology

4.1 Alignment and preference datasets

Alignment refers to the process of ensuring that AI systems operate in accordance with human values, preferences, and intentions [32]. Achieving alignment is critical for the responsible deployment of AI technologies, particularly in sensitive applications where ethical considerations are paramount. One effective approach to achieving alignment is through the use of preference datasets, which capture the nuanced preferences of users. These datasets can be employed to train AI systems to recognize and prioritize human

values, thereby enhancing their decision-making processes. For instance, incorporating dynamic utility functions that reflect changing human preferences can help AI systems adapt to evolving societal norms and expectations [42][44].

Explainable AI benefits significantly from alignment achieved through preference datasets, as it enables models to provide transparent and interpretable outputs that resonate with user expectations. By leveraging preference datasets, AI systems can be designed to generate explanations that align with users' values and preferences, thus fostering trust and understanding. This alignment can be enhanced by employing techniques that account for the variability in human preferences, ensuring that the AI's explanations are not only accurate but also contextually relevant [43]. Ultimately, the integration of preference datasets into the alignment process can lead to the development of AI systems that are not only effective but also ethically sound and aligned with the values of the communities they serve.

4.2 Algorithms and model training

In this subsection, we explore the methodologies and processes involved in training our models to align with human preferences. We begin by describing both Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), highlighting their role in leveraging the preference dataset. We then detail our model training setup, hyperparameter choices, and overall training workflow.

Direct Preference Optimization Direct Preference Optimization (DPO) [9] is a pairwise preference alignment technique that leverages human feedback to guide model responses. Given two candidate responses A and B for the same input, along with a human-annotated preference, DPO adjusts the model parameters so that the reward for the preferred response is higher. Formally, the DPO loss can be written as:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(A,B)\sim D}\big[\log \sigma\big(r_\theta(A) - r_\theta(B)\big)\big],$$

$r_\theta(A)$ and $r_\theta(B)$ are the learned reward scores assigned by the model to responses A and B, respectively. The expression $\sigma(r_\theta(A) - r_\theta(B))$ gives the probability that the model prefers A over B under the current parameters, where $\sigma(\cdot)$ is the logistic (sigmoid) function. By taking the negative log of this probability and averaging over all pairs in the dataset D, the model is penalized when it fails to assign a higher reward to the human-preferred response. Minimizing $\mathcal{L}_{\text{DPO}}(\theta)$ thus encourages the model to consistently rank the chosen response more favorably than the non-chosen one, aligning outputs with human preferences.

We adopted DPO as the sole preference alignment strategy for tuning GPT-4o, as it is the only such algorithm currently offered by OpenAI for their models. This approach allowed us to leverage pairwise preferences collected in our dataset to optimize GPT-4o's responses toward higher human satisfaction. Consequently, GPT-4o achieved stronger alignment with human judgments regarding clarity, correctness, and helpfulness of its outputs.

Odds Ratio Preference Optimization Odds Ratio Preference Optimization (ORPO) [8] is a pairwise preference alignment technique that optimizes the ratio of predicted probabilities assigned to the preferred versus the non-preferred response. Formally, its loss function can be expressed as:

$$\mathcal{L}_{\text{ORPO}}(\theta) = -\mathbb{E}_{(A,B)\sim D}\left[\log\left(\frac{p_\theta(A)}{p_\theta(B)}\right)\right],$$

where $p_\theta(\cdot)$ indicates the model's predicted probability for each candidate response A or B. Minimizing this term encourages the model to assign higher likelihood to the chosen (human-preferred) response, thus aligning outputs with annotator judgments. Compared to DPO—which uses a sigmoid-based difference in reward scores—ORPO operates directly on the odds ratio, potentially offering more stable updates for smaller models.

We used ORPO for its superior stability over DPO [33][34][35], especially when training open-source models. We applied ORPO to preference-align two open-source, instruction-tuned language models, achieving improved alignment to human-labeled preferences.

Model training We aligned three models using our preference dataset. Specifically, we applied DPO to train GPT-4o, while for the two open-source models—LLaMA 3 (8B) and DeepSeek R1 (Distilled Qwen 2.5, 32B)—we employed ORPO. This setup allowed each model to leverage human preference annotations in a manner best suited to its respective infrastructure.

Hyperparameter setup For DPO training on GPT-4o, we followed the default preference alignment workflow provided by OpenAI. We ran the fine-tuning for 2 epochs with a batch size of 8, a learning rate multiplier of 1, and set the random seed to 42 for reproducibility. Additionally, we used $\beta = 0.1$ in the loss computation, which controls the gradient update weight for the preference signal. For ORPO training on the open-source models (LLaMA 3 and DeepSeek R1), we employed a maximum input and prompt length of 1,024 tokens. The per-device train and per-device eval batch sizes were both set to 4, with 2 gradient accumulation steps to effectively reach an overall batch size of 8. We used a learning rate of 2×10^{-4}, adamw8bit optimization, and weight decay of 0.01. The training proceeded for 100 steps, with an evaluation step every 10 steps. This configuration was managed using a HuggingFace-compatible ORPOTrainer module, ensuring consistent training parameters across our preference alignment experiments.

Training overview and performance analysis For the DPO approach, we used OpenAI's dedicated module for preference alignment, wherein our preference dataset's training and validation splits were supplied as JSONL files. The API managed the entire fine-tuning lifecycle internally, allowing us to concentrate on evaluating and iterating over different model checkpoints. Despite the high computational cost of training large models, this workflow proved both straightforward and efficient, as it removed the burden of manually handling complex hyperparameters. By contrast, for ORPO preference alignment, we opted for smaller open-source models (LLaMA 3 and DeepSeek R1) to accommodate our limited compute budget on Google Colab Pro, which offers access to an NVIDIA A100 40GB GPU. To further accelerate training and inference, we integrated models from the unsloth [36] library, which supports optimized inference kernels. We also utilized Parameter-Efficient Fine-Tuning (PEFT) [37] via LoRa (Low-Rank Adaptation) [38] in a 4-bit precision setting—enabling the loading of large-scale language models on smaller GPU memory. We set the LoRa rank = 16, disabled dropout, and used a minimal batch size of 8 to remain within memory constraints. Despite these resource limitations, the fine-tuned models exhibited substantial improvements in aligning with human preferences, indicating the effectiveness of ORPO for lighter-weight deployments.

5 Architecture diagram

The diagram 1 illustrates how the SemEval 2018 E-c train+dev dataset is used to prompt GPT-4o for two candidate responses. These responses are then manually annotated and

refined to form the preference dataset, which is subsequently utilized to train GPT-4o via DPO and open-source models (DeepSeek, LLaMA) via ORPO. Finally, the resulting fine-tuned models generate inferences (explanations and classifications) that are evaluated for quality and correctness.

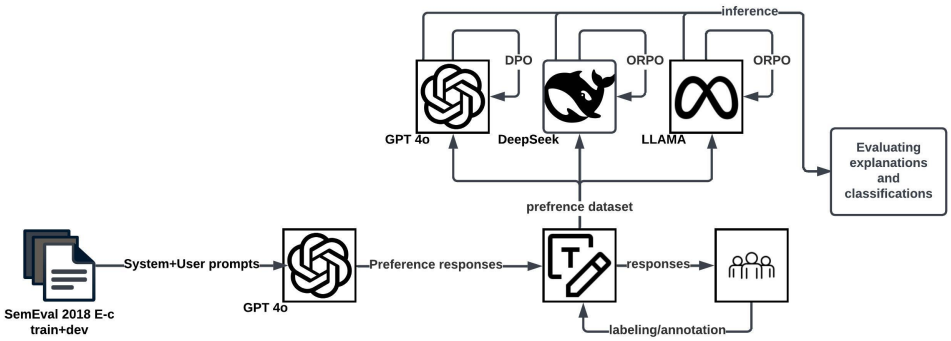

Fig. 1. General system schema

6 Evaluation

6.1 Evaluating explanations

To assess the quality and faithfulness of the generated explanations, we employ the Sufficiency metric, which measures whether the explanation text alone is sufficient to recover the model's prediction [12] [39]. Our implementation builds on the "Faithfulness-by-construction" (FRESH) pipeline which is a framework to evaluate sufficiency of explanations: the sole explanations, without the remaining parts of the input, must be sufficient for predicting a label [13], where a separate classifier (BERT-based) is trained to predict the label using only the extracted explanation—omitting the rest of the input text. If the classifier maintains high accuracy on this restricted view, we interpret the explanations as being faithful to the original model's rationale, reflecting a strong alignment between the explanation and the underlying decision process. This evaluation method is used to evaluate generative explanations [24] [22].

6.2 Evaluating classifications

For classification performance on the SemEval 2018 E-c test set, we adopt the official multi-label accuracy (Jaccard index) metric, as well as micro F1 and macro F1 scores, to capture both label-wise and overall prediction quality. On the DAIR AI test set, which is a single-label task, we compute the standard accuracy, precision, recall, and F1 metrics to measure the effectiveness of our model predictions. These metrics collectively provide a comprehensive view of how well our models identify emotions and align with the ground-truth labels across both datasets. From the generated responses, we use the multiple_emotions_present, and most_prominent_emotion fields for multilabel and multiclass evaluation respectively.

7 Results

Tables 3 and 4 summarize the classification outcomes of our proposed models compared to various state-of-the-art baselines on the SemEval 2018 E-c and DAIR AI emotion datasets, respectively. For the multilabel setting (SemEval 2018 E-c), GPT-4o fine-tuned with DPO achieves the highest accuracy (68.85%), outperforming both zero-shot GPT-4o and other transformer-based baselines. Meanwhile, DeepSeek R1 (Distilled Qwen 32B) and LLaMA 3 8B, both trained via ORPO, show competitive results, albeit slightly lower than GPT-4o DPO, because these are much smaller models. For the multiclass DAIR AI dataset, GPT-4o DPO similarly attains strong performance, reaching 93.1% accuracy and an F1 score of 87.9—comparable with established transformer-based SOTA models. The two ORPO-aligned models also maintain high classification metrics (over 88% accuracy), demonstrating that preference alignment can yield significant improvements even for smaller or distilled architectures.

Models	**Accuracy %**	**Micro F1 %**	**Macro F1 %**
Our proposed models			
GPT-4o - DPO	**68.85**	**80.53**	**74.44**
DeepSeek R1 (Distilled Qwen 32B) - ORPO	65.66	77.57	71.58
LLAMA 3 8B - ORPO	64.12	75.91	68.89
Current state-of-the-art models [19]			
$GPT2 - IT_A$	**67.56**	**79.36**	**73.05**
Zero-shot GPT-4o	64.76	76.77	70.09
RoBERTa MA	62.40	74.20	60.30

Table 3. Emotion classification results on different models for SemEval-2018 Task1-Ec dataset. The best values are in bold.

Models	**Accuracy**	**Precision**	**Recall**	**F1**
Our proposed models				
GPT-4o - DPO	**93.10**	**90.80**	**87.09**	**87.90**
DeepSeek R1 (Distilled Qwen 32B) - ORPO	90.77	89.57	86.92	86.76
LLAMA 3 8B - ORPO	88.22	86.03	85.26	84.32
Current State-of-the-art models				
sagemaker-roberta-base-emotion [1]	**93.10**	88.30	**90.90**	**89.50**
roberta-base-emotion [2]	93.10	**91.70**	87.40	88.20
Zero-shot GPT-4o	92.40	90.63	87.81	87.63

Table 4. Emotion classification results on different models for Dair AI emotion dataset. The best values are in bold.

Table 5 presents the Sufficiency metric results, which gauge whether a model's explanation text alone is predictive of its final classification. GPT-4o DPO achieves a Sufficiency score of 63.66, notably surpassing zero-shot GPT-4o and other popular explanation methods such as SHAP-RoBERTa and LIME-RoBERTa. While DeepSeek R1 and LLaMA 3 fall slightly behind GPT-4o, they still outperform all state-of-the-art approaches, suggesting that preference-aligned models are able to produce higher-quality, self-consistent explanations.

Explainable AI models	Sufficiency
Our proposed models	
GPT-4o - DPO	**63.66**
DeepSeek R1 (Distilled Qwen 32B) - ORPO	61.12
LLAMA 3 8B - ORPO	60.98
Current State-of-the-art models [22]	
GPT-4o	59.66
SHAP-RoBERTa	54.16
LIME-RoBERTa	53.22

Table 5. Sufficiency metric for evaluating explanations for the models

7.1 Analysis

The classification results highlight the benefits of preference alignment for both large-scale and smaller models. GPT-4o DPO not only sets a new performance bar on the SemEval 2018 E-c dataset but also demonstrates strong generalization on the DAIR AI dataset, achieving accuracy levels on par with top-performing baselines. Notably, GPT-4o never sees the DAIR AI data during preference alignment—indicating that the alignment process itself confers robustness that generalizes to unseen tasks. Meanwhile, the open-source models, DeepSeek R1 and LLaMA 3 8B, show considerable gains upon being trained with ORPO, underscoring that systematic preference alignment can upgrade instruction-tuned models even under constrained compute budgets.

Beyond classification accuracy, our preference-aligned models also demonstrate marked improvements in output format consistency. Before alignment, these generative LLMs would frequently violate the specified output schema—up to 33% of the time in our experiments—forcing us to parse or manually correct their outputs. This issue is especially pronounced for open-source models that lack robust structured output support (like OpenAI's GPT-4o). After training with our preference dataset, however, all models adhere strictly to the requested JSON-based output, reducing formatting errors to 0%. This consistency is a substantial asset in production settings, where reliable parsing and downstream automation are crucial. Overall, the preference alignment process yields models that are both accurate and operationally reliable, suggesting a promising direction for scalable, human-centered natural language systems.

8 Conclusion and future work

We introduced a preference-aligned, self-explaining approach to emotion classification. Our contributions include constructing a preference dataset that explicitly captures human judgments about clarity, correctness, helpfulness, and verbosity, and using this dataset to preference-tune both GPT-4o and open-source LLMs. The results demonstrate that preference alignment not only boosts classification performance, but also enhances explanation quality, offering insights that directly reference salient parts of the input text. Moreover, our method drastically reduces formatting inconsistencies, a major hurdle when deploying LLMs in practical settings.

For future work, researchers can incorporate retrieval-augmented generation (RAG) by leveraging vector databases of emotion datasets and related emotion concepts for more context-rich explanations and higher classification accuracy. Additionally, while our current study focuses on Direct Preference Optimization (DPO) and Odds Ratio Preference Optimization (ORPO), future work could explore Reinforcement Learning with Human Feedback (RLHF) and Group Relative Policy Optimization (GRPO), using our

fine-grained scores of correctness, clarity, helpfulness, and verbosity as reward signals. These directions collectively aim to yield more interpretable, robust, and human-aligned systems.

References

1. Arrieta, A., Díaz-Rodríguez, N., Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*. **58** pp. 82-115 (2020)
2. Gunning, D. & Aha, D. Darpa's explainable artificial intelligence program. *AI Magazine*. **40**, 44-58 (2019)
3. Okada, Y., Ning, Y. & Ong, M. Explainable artificial intelligence in emergency medicine: an overview. *Clinical And Experimental Emergency Medicine*. **10**, 354-362 (2023)
4. Naiseh, M., Al-Thani, D., Jiang, N. & Ali, R. How the different explanation classes impact trust calibration: the case of clinical decision support systems. *International Journal Of Human-Computer Studies*. **169** pp. 102941 (2023)
5. Burkart, N. & Huber, M. A survey on the explainability of supervised machine learning. *Journal Of Artificial Intelligence Research*. **70** pp. 245-317 (2021)
6. Choubisa, V. & Choubisa, D. Towards trustworthy ai: an analysis of the relationship between explainability and trust in ai systems. *International Journal Of Science And Research Archive*. **11**, 2219-2226 (2024)
7. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. & Lowe, R. Training language models to follow instructions with human feedback. (2022), https://arxiv.org/abs/2203.02155
8. Hong, J., Lee, N. & Thorne, J. ORPO: Monolithic Preference Optimization without Reference Model. (2024), https://arxiv.org/abs/2403.07691
9. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. & Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. (2024), https://arxiv.org/abs/2305.18290
10. Mohammad, S., Bravo-Marquez, F., Salameh, M. & Kiritchenko, S. SemEval-2018 Task 1: Affect in Tweets. *Proceedings Of The 12th International Workshop On Semantic Evaluation*. pp. 1-17 (2018,6), https://aclanthology.org/S18-1001/
11. Saravia, E., Liu, H., Huang, Y., Wu, J. & Chen, Y. CARER: Contextualized Affect Representations for Emotion Recognition. *Proceedings Of The 2018 Conference On Empirical Methods In Natural Language Processing*. pp. 3687-3697 (2018), https://aclanthology.org/D18-1404/
12. Jacovi, A., Sar Shalom, O. & Goldberg, Y. Understanding Convolutional Neural Networks for Text Classification. *Proceedings Of The 2018 EMNLP Workshop BlackboxNLP: Analyzing And Interpreting Neural Networks For NLP*. pp. 56-65 (2018,11), https://aclanthology.org/W18-5408/
13. Jain, S., Wiegreffe, S., Pinter, Y. & Wallace, B. Learning to Faithfully Rationalize by Construction. *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 4459-4473 (2020,7), https://aclanthology.org/2020.acl-main.409/
14. Shaikh, R., Rafi, M., Mahoto, N., Sulaiman, A. & Shaikh, A. A Filter-Based Feature Selection Approach in Multilabel Classification. *Machine Learning Science And Technology*. (2023)
15. Sadr, M., Mirtaheri, S., Greco, S. & Borna, K. Popular Tag Recommendation by Neural Network in Social Media. *Computational Intelligence And Neuroscience*. (2023)
16. Taha, A., Tiun, S., Abd Rahman, A., Ayob, M. & Abdulameer, A. Unified Graph-Based Missing Label Propagation Method for Multilabel Text Classification. *Symmetry*. (2022)
17. Ferreira, W. & Vlachos, A. Incorporating Label Dependencies in Multilabel Stance Detection. *Proceedings Of EMNLP*. (2019)
18. Büyükçakır, A., Bonab, H. & Can, F. A Novel Online Stacked Ensemble for Multi-Label Stream Classification. (2018)
19. Siddiqui, M., Inkpen, D. & Gelbukh, A. Instruction Tuning of LLMs for Multi-label EmotionClassification in Social Media Content. *Proceedings Of The Canadian Conference On Artificial Intelligence*. (2024,5,27), https://caiac.pubpub.org/pub/lezimqvm
20. Zhang, J. & Rao, Y. Research on Model and Algorithm of Multiview and Multilabel Classification Based on Nearest-Neighbor Model. *Mathematical Problems In Engineering*. (2022)
21. Chen, W., Zhang, B. & Lu, M. Uncertainty Quantification for Multilabel Text Classification. *Wiley Interdisciplinary Reviews Data Mining And Knowledge Discovery*. (2020)

22. Fahim Siddiqui, M., Inkpen, D. & Gelbukh, A. Towards Interpretable Emotion Classification: Evaluating LIME, SHAP, and Generative AI for Decision Explanations. (2024)
23. Rajagopal, D., Balachandran, V., Hovy, E. & Tsvetkov, Y. SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers. (2021)
24. Rajagopal, D., Balachandran, V., Hovy, E. & Tsvetkov, Y. SELFEXPLAIN: A Self-Explaining Architecture for Neural Text Classifiers. (2021)
25. Michaud, J. Dynamic Preferences and Self-Actuation of Changes in Language Dynamics. *Language Dynamics And Change*. (2019)
26. Charte, F., Rivera, A., Jesús, M. & Herrera, F. Dealing With Difficult Minority Labels in Imbalanced Mutilabel Data Sets. *Neurocomputing*. (2019)
27. Lee, J., Seo, W. & Kim, D. Effective Evolutionary Multilabel Feature Selection Under a Budget Constraint. *Complexity*. (2018)
28. Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N., Khashabi, D. & Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions. (2023), https://arxiv.org/abs/2212.10560
29. Gu, Z., Sun, X., Lian, F., Kang, Z., Xu, C. & Fan, J. Dingo: towards diverse and fine-grained instruction-following evaluation. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **38**, 18108-18116 (2024)
30. Chen, Y., Xu, B., Wang, Q., Liu, Y. & Mao, Z. Benchmarking large language models on controllable generation under diversified instructions. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **38**, 17808-17816 (2024)
31. Alkalbani, A., Alrawahi, A., Salah, A., Haghighi, V., Zhang, Y., Alkindi, S. & Sheng, Q. A systematic review of large language models in medical specialties: applications, challenges and future directions. (2024)
32. Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., Zeng, F., Ng, K., Dai, J., Pan, X., O'Gara, A., Lei, Y., Xu, H., Tse, B., Fu, J., McAleer, S., Yang, Y., Wang, Y., Zhu, S., Guo, Y. & Gao, W. AI Alignment: A Comprehensive Survey. (2024), https://arxiv.org/abs/2310.19852
33. Amini, S., Vass, C., Shahabi, M. & Noble, A. Optimization of coal blending operations under uncertainty – robust optimization approach. *International Journal Of Coal Preparation And Utilization*. **42**, 30-50 (2019)
34. Kirk, H., Vidgen, B., Röttger, P. & Hale, S. Personalisation within bounds: a risk taxonomy and policy framework for the alignment of large language models with personalised feedback. (2023)
35. Sun, Z., Zhou, Y., Hao, J., Fan, X., Lu, Y., Ma, C., Shen, W. & Guo, C. Improving contextual query rewrite for conversational ai agents through user-preference feedback learning. *Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing: Industry Track*. (2023)
36. Daniel Han, M. & Team, U. Unsloth. (2023), http://github.com/unslothai/unsloth
37. Xu, L., Xie, H., Qin, S., Tao, X. & Wang, F. Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment. (2023), https://arxiv.org/abs/2312.12148
38. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. (2021), https://arxiv.org/abs/2106.09685
39. Yu, M., Chang, S., Zhang, Y. & Jaakkola, T. Rethinking Cooperative Rationalization: Introspective Extraction and Complement Control. *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP)*. pp. 4094-4103 (2019,11), https://aclanthology.org/D19-1420/
40. Chochlakis, G., Mahajan, G., Baruah, S., Burghardt, K., Lerman, K. & Narayanan, S. Leveraging Label Correlations in a Multi-Label Setting: a Case Study in Emotion. *ICASSP 2023 - 2023 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 1-5 (2022), https://api.semanticscholar.org/CorpusID:253223989
41. Baziotis, C., Nikolaos, A., Chronopoulou, A., Kolovou, A., Paraskevopoulos, G., Ellinas, N., Narayanan, S. & Potamianos, A. NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning. *Proceedings Of The 12th International Workshop On Semantic Evaluation*. pp. 245-255 (2018,6), https://aclanthology.org/S18-1037
42. Doorn, N. Artificial intelligence in the water domain: opportunities for responsible use. *Science Of The Total Environment*. **755** pp. 142561 (2021)
43. Gabriel, I. Artificial intelligence, values, and alignment. *Minds And Machines*. **30**, 411-437 (2020)
44. Carroll, M., Hadfield-Menell, D., Russell, S. & Dragan, A. Estimating and penalizing preference shift in recommender systems. *Fifteenth ACM Conference On Recommender Systems*. pp. 661-667 (2021)