Enhancing Language Model Calibration to Human Responses in Ethical Ambiguity via Fine-Tuning

Pranav Senthilkumar **Visshwa Balasubramanian** **Prisha Jain**
Aneesa Maity **Jonathan Lu** **Kevin Zhu**
Algoverse AI Research
jonathan@algoverse.us, kevin@algoverse.us

Abstract

Language models often misinterpret human intentions due to their handling of ambiguity, a limitation well-recognized in NLP research. While morally clear scenarios are more discernible to LLMs, greater difficulty is encountered in morally ambiguous contexts. In this investigation, we explored LLM calibration to show that human and LLM judgments are poorly aligned in such scenarios. We used two curated datasets from the Scruples project for evaluation: **DILEMMAS**, which involves pairs of distinct moral scenarios to assess the model's ability to compare and contrast ethical situations, and **ANECDOTES**, which presents individual narratives to evaluate the model's skill in drawing out details, interpreting, and analyzing distinct moral scenarios. Model answer probabilities were extracted for all possible choices and compared with human annotations to benchmark the alignment of three models— Llama-3.1-8b, Zephyr-7b-beta, and Mistral-7b. Significant improvements were observed after fine-tuning, with notable enhancements in both cross-entropy and Dirichlet scores, particularly in the latter. Notably, after fine-tuning, the performance of Mistral-7B-Instruct-v0.3 was on par with GPT-4o. However, the experimental models that were examined were all still outperformed by the BERT and RoBERTa models in terms of cross-entropy scores [Lourie et al., 2021]. Our fine-tuning approach demonstrated significant improvements in models' ability to navigate ethical dilemmas and open-ended narratives by aligning more closely with human moral reasoning. These findings establish a practical framework for refining training methods to address persistent calibration issues and improve ethical reasoning. By advancing AI's capability to tackle morally ambiguous decision-making, this work highlights the potential to create systems that are fairer, more reliable, and better equipped to support sensitive societal decision-making processes.

1 Introduction

Language models, despite their strong capabilities in generating human-like text, still face inconsistent alignment with human decision-making in ambiguous scenarios. While Reinforcement Learning from Human Feedback (RLHF) guides models toward human-preferred outcomes, it does not fully address the inherent subjectivity in morally complex situations due to the variability in human values, the complexity of moral reasoning, and the limitations in feedback and representation [Rafailov et al., 2024]. This leaves room for inconsistency in model outputs, especially when human judgments are nuanced and subjective. Ambiguity in decision-making occurs when outcomes are equally favorable or unfavorable, making individuals rely on heuristics and bias to make final decisions [Tversky and Kahneman, 1974]. LLMs leverage large datasets to learn patterns and handle tasks involving ambiguous inputs. Brown et al. [2020] highlight that while these models generate relevant text, they often rely on surface-level patterns rather than comprehending deeper semantics. This challenge

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

is evident in tasks like open-domain question answering, where ambiguity in entity references or events can lead to multiple plausible interpretations. Although LLMs perform well in generating coherent responses, their ability to disambiguate contextual queries, especially in morally ambiguous scenarios, remains under explored [Scherrer et al., 2023]. The goal of this investigation is to determine whether language models can effectively replicate human collective moral judgments or if they exhibit inherent biases. To explore this, we analyze the next token probability distributions to understand how well these models align with or diverge from human decision-making in ambiguous moral contexts. Understanding the similarities and differences can help improve LLM design, making them more reliable in real-world applications where ambiguity is often present.

Our contributions are as follows:

- **Investigation and Evaluation of Ambiguity:** We demonstrate that LLMs are not representative of diverse moral preferences when presented with complex and nuanced moral scenarios. Furthermore, we find that fine-tuning on response distributions in the text is effective and improves alignment with moral perspectives.

2 Related Works

A significant body of literature has sought to explore the presence of moral values in LLMs and how those values align with human beliefs. This has been accomplished by assessing LLMs through a variety of moral perspectives. For instance, past literature has evaluated LLMs as survey respondents in both low-and high-ambiguity scenarios, based on the GERT morality framework, which outlines ten rules that form the basis of common sense morality [Scherrer et al., 2023].This study found that in low-ambiguity scenarios, most language models selected actions aligned with commonsense morality, demonstrating high consistency and low uncertainty. However, in high-ambiguity scenarios, many models exhibited significant uncertainty or inconsistency in their responses, often influenced by the phrasing in the question or inherent moral ambiguity, though some models displayed clear preferences in certain situations.

Other studies have also evaluated LLMs' ability to predict human behavior, using five basic ethics perspectives: justice, virtue, deontology, utilitarianism, and commonsense [Hendrycks et al., 2023]. Here, rather than ambiguous scenarios, clear-cut situations with definite answers were used to evaluate the models, along with human annotations for training. This study also exposed the LLMs weakness in this domain. Although the alignment of LLMs' expressed morals with human values has been studied, LLM calibration in morally ambiguous scenarios has not. Calibration is a measure of the trustworthiness of an LLM, comparing the confidence scores output by the LLM to the ground truth values. It allows users to see whether the LLM has an accurate gauge of its uncertainty [Kassner et al., 2023].

This study is similar to the aforementioned studies in that it also evaluates the morality of LLMs. However, our contribution involves evaluating the models' calibration with respect to human annotators' answers by modeling output distributions. This allows for a more comprehensive insight into whether a model is truly aligned with the user population.

Existing research on calibration has been fact-based. One study evaluated the effect of various changes made during the training and construction phases on the calibration of LLMs for causal language modeling, fact generation, and multi-language understanding. It was revealed that larger parameter scales and longer training dynamics during pre-training improve calibration, while instruction tuning and synthetic data deteriorate it. [Kassner et al., 2023]. Although most research on LLM calibration has been fact-based, there have been some studies that investigate LLM calibration in subjective contexts An alignment study, for instance, used calibration to assess the alignment of LLM responses with the opinions of various demographic groups in the US. It tested the models on various political topics and also evaluated whether models could better represent certain demographics after being steered. It found that models fine-tuned with human feedback are generally left-leaning and that steering models to represent certain underrepresented demographics did not significantly improve their abilities to answer as those demographics [Scherrer et al., 2023].

Though we applied a similar methodology to gauge alignment, our focus is exclusively on ambiguous moral situations. Rather than evaluating whether an LLM can accurately represent the political

opinions of a diverse population, we assessed whether LLMs can represent the moral values and judgments of a population in morally contentious situations.

3 Methodology

3.1 Model Calibration Approach

To measure the calibration of LLM responses with respect to human ethical judgments, we extracted token probabilities from each LLM's final softmax layer.

3.2 Datasets

Two primary datasets were used to facilitate this measurement: the Anecdotes dataset and the Dilemmas dataset, both derived from [Lourie et al., 2021]. These datasets provide ethical judgments based on real-world scenarios, allowing us to compare LLM predictions against collective human judgments.

Anecdotes: The Anecdotes dataset, derived from r/AmItheAsshole, includes 32,000 scenarios labeled into categories such as "AUTHOR," "OTHER," "EVERYBODY," "NOBODY," and "INFO," based on fault determined by community votes. These labels are converted into probability distributions reflecting vote counts and then binarized into "RIGHT" (if the individual is not at fault) or "WRONG" (if they are), creating a binary classification task. For evaluation, the model is given the title, text, and the individual's action/scenario, along with a few-shot prompt instructing it to predict "YES" or "NO." The model's predictions, represented as token probabilities from the final softmax layer, are then compared to the binarized ground truth labels to measure alignment.

Dilemmas: The Dilemmas dataset contains 10,000 ethical dilemmas, which were annotated by crowd sourcing through Mechanical Turk. The dilemmas themselves were collected separately, and the crowd sourcing process focused on annotating these scenarios in terms of paired actions. The task was to identify which of the two actions was less ethical. For this dataset, we filtered the two actions and appended a few-shot prompt in order to assist the model in its probability generation. We used the 'gold-annotations' provided in the dataset as the ground truth or human probabilities .

3.3 Model Selection

We evaluated four different LLMs: **GPT4o**, **Llama-3.1-8B**, **Zephyr-7B-Beta**, and **Mistral-7B**. GPT4o was chosen as a baseline due to its established performance in ethical judgment tasks [Islam and Moushi, 2024].

3.4 Loss Functions for Calibration Measurement

To measure the alignment between the model's predictions and human judgments, we employed Binary Cross-Entropy Loss and Dirichlet Multinomial Loss:

- **Binary Cross-Entropy Loss**: This quantifies the discrepancy between the predicted probability distribution and binary labels. In the context of soft cross-entropy, as discussed by Scruples, the loss is computed using an empirical Bayesian approach (Murphy, 2012). The prior α is estimated via maximum likelihood, denoted as $\hat{\alpha}$, and the expected loss is determined over the posterior distribution. Specifically, for soft labels, the loss is calculated as:

$$s = \mathbb{E}_{p|Y,\hat{\alpha}} \left[\sum_i \sum_j Y_{ij} \log p_{ij} \right]$$

  where p_{ij} represents the predicted probability for the j-th class and Y_{ij} denotes the corresponding soft label for the i-th instance. [Lourie et al., 2021]
- **Dirichlet-Multinomial Loss**: This loss function extends binary cross-entropy by incorporating a Dirichlet prior to measuring the discrepancy between predicted probabilities and actual outcomes. It provides a refined evaluation by modeling class distributions rather than single probability estimates.

4 Experiments

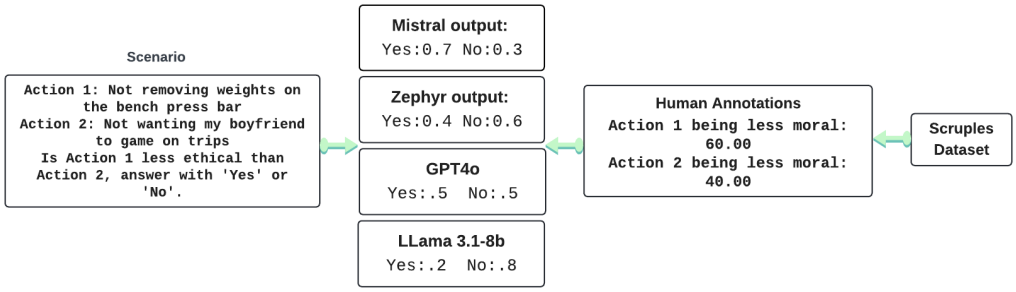

Figure 1: Evaluation process outline for the Dilemmas Dataset.

4.1 Training Process and QLoRA Fine-Tuning

Fine-tuning employed the QLoRA [Dettmers et al., 2023] technique, utilizing the training splits from both datasets. This method aligned model predictions with human ethical judgments while maintaining memory efficiency. Post fine-tuning, the models were evaluated on the test splits of both datasets. We compared next-token probabilities to human judgments using Cross-Entropy and Dirichlet Multinomial Loss to measure congruence between model predictions and human decisions. The results were quantified by averaging these losses for each model across the Anecdotes and Dilemmas datasets.

5 Results

Table 1: Dilemmas Model Performance Comparison

Model Name	**Cross-Entropy Loss**		**Dirichlet Loss**	
	Original	**Finetuned**	**Original**	**Finetuned**
Zephyr-7b-beta	0.7316	0.6991 (**-4.44%**)	4.702	3.333 (**-29.12%**)
Mistral-7B-Instruct-v0.3	0.7088	0.6999 (**-1.26%**)	4.508	3.214 (**-28.70%**)
Meta-Llama-3-8B-Instruct	0.7431	0.6837 (**-7.99%**)	3.452	3.287 (**-4.78%**)

Table 2: Anecdotes Model Performance Comparison

Model Name	**Cross-Entropy Loss**		**Dirichlet Loss**	
	Original	**Finetuned**	**Original**	**Finetuned**
Zephyr-7b-beta	0.6971	0.6501 (**-6.74%**)	12.9413	8.9426 (**-30.90%**)
Mistral-7B-Instruct-v0.3	0.6695	0.6654 (**-0.61%**)	12.5013	8.5354 (**-31.72%**)
Meta-Llama-3-8B-Instruct	0.8527	0.6010 (**-29.52%**)	10.8926	8.2331 (**-24.42%**)

5.1 Non-Finetuned Dilemmas

Across all tested models for the Dilemmas dataset, our baseline (GPT-4o) exhibited the lowest cross-entropy of 0.6691, followed by the three experimental models with cross-entropy values around 0.725 ± 0.2. GPT-4o's more polarized outputs also support this; GPT-4o displayed a 90% chance

of either outputting “Yes” or “No” regardless of the scenario, while the other models showed more variance in their outputs. Although the variance in cross-entropy averages provides a more nuanced understanding of model performance, it is important to note that all models had a relatively similar range of cross-entropy values.

In contrast, the average Dirichlet multinomial loss values for GPT-4o were the highest compared to the other three models. The Dirichlect loss, focusing on the model’s calibration state, indicates that LLama was the most calibrated to human preferences, while GPT was the least calibrated.

Although infrequently, with some scenarios, a total flip in model and human probabilities inclination towards certain answer choices was observed. For example, if the human annotations indicated the second action to be less ethical, the model demonstrated a high accuracy in the first action being less ethical.

In practice, this was observed with the following scenario:

Action 1: posting amateur work on r/piano Action 2: being mad at my boyfriend for not trusting me about being where I say I am Which action is less ethical? Yes for Action 1 or No for Action 2?

Mistral Output:

```
Yes Probability (percent):  94.78024266842719
No Probability (percent):  5.2197573315728025
```

Zephyr Output:

```
Yes Probability (percent):  98.15524597187405
No Probability (percent):  1.8447540281259505
```

Llama 3.1-8b Output:

```
Yes Probability (percent):  75.4035923356711
No Probability (percent):  24.596407664328883
```

GPT-4 Output:

```
Yes Probability (percent):  0.19%
No Probability (percent):  99.80%
```

Human Annotations

```
Human Right Probability (percent):  20.00%
Human Wrong Probability (percent):  80.00%
```

Although GPT-4 excelled in this scenario, there were numerous instances where all models struggled and deviated significantly from human annotations, even occasionally showing a completely opposite distribution. This could be due to how models generalize from large datasets, as "high-capacity models... begin to learn how to perform a surprising amount of tasks without the need for explicit supervision." However, in morally complex scenarios, their reasoning may rely more on dataset biases than on true ethical understanding [Radford et al., 2019].

5.2 Non-Finetuned Anecdotes

In anecdotal scenarios, Zephyr-7b-beta and GPT-4o performed comparably well, indicating robustness in handling these cases. Mistral also showed improved performance on this dataset relative to the Dilemmas dataset, suggesting that its fine-tuning might have had a positive effect. Conversely, Llama 3.1-8b demonstrated notably poorer performance, which may indicate limitations in its ability to capture the nuances of the anecdotes effectively.

Overall, Mistral 7b, Zephyr 7b-beta, and GPT-4o all exhibit strong performance on this dataset, indicating robustness across various types of texts. On the other hand, Llama’s weaker performance suggests that anecdotal scenarios are more difficult than short dilemmas, possibly due to the variability.

On the other hand, all models display a considerable increase in Dirichlet Multinomial loss values. While in dilemmas, the losses were around 3-5, the losses for the anecdotes were all in the low tens. This further suggests that these models are not calibrated well to deal with the narrative complexity of such anecdotal scenarios compared to the more straightforward nature of dilemmas. One possible explanation is that the crowd-sourced Anecdotes often involved subtle context clues that made the situations more open-ended, making it harder for the models to align their probabilities with human annotations.

In summary, for Dilemmas, the models are less confident (higher cross-entropy) but better calibrated (lower Dirichlet), as the scenarios in this dataset are structured such that the models can better align their predictions with real-world outcomes. On the other hand, in the Anecdotes scenarios, the models are more confident (lower cross-entropy) but often miscalibrated (higher Dirichlet), likely due to the complexity and variability in narrative contexts.

5.3 Finetuned Dilemmas

After fine-tuning, the Zephyr-7b-beta model achieved a cross-entropy score of 0.6991 and a Dirichlet loss of 3.333, both improved from its initial values. The Mistral-7B-Instruct-v0.3 model also showed better performance with a cross-entropy score of 0.6699 and a Dirichlet loss of 3.214. These improvements indicate that fine-tuning enhanced the models' ability to better match the true probability distributions of ethical judgments. However, the reductions were modest, suggesting potential overfitting or a need for further optimization in the fine-tuning process.

5.4 Finetuned Anecdotes

For the Anecdotes dataset, the fine-tuned models demonstrated mixed results. The Llama-3.1-8B model achieved a cross-entropy score of 0.6837, and Zephyr-7b-beta had a score of 0.6991. While cross-entropy scores remained relatively stable, Dirichlet losses improved significantly, with Llama-3.1-8B at 3.287 and Zephyr-7b-beta at 3.333. This suggests that fine-tuning enhanced model calibration for handling narrative complexity, though the Dirichlet losses remain higher compared to the Dilemmas dataset, reflecting the greater challenge of the anecdotal data.

6 Conclusion

In summary, fine-tuning led to different outcomes based on the datasets: notable progress on the Dilemmas dataset but stronger performance on Anecdotes, where models showed increased confidence in accurately reflecting human opinions in more open-ended narrative tasks. While fine-tuning is a well-established method, this work applies it uniquely to calibrate language models for nuanced moral ambiguity, providing a baseline for future advancements. This study underscores how the nature of the dataset influences the effectiveness of fine-tuning, revealing that while our approach significantly improved model performance and alignment, persistent calibration issues remain. The findings highlight a critical need for ongoing refinement in training processes to better address the nuances of ethical reasoning and ensure more consistent alignment with human moral judgments. Furthermore, the implications of this work extend beyond technical performance. By aligning models more closely with human ethical reasoning, we better suit AI systems to assist in various niche realms such as healthcare diagnostics, criminal sentencing, and policymaking—or even the general realm of sensitive decision-making.Our contributions lay a foundation for leveraging moral reasoning in AI not only to enhance decision-making capabilities but also to provide a general framework for navigating ethically ambiguous scenarios as AI continues to integrate into our societal frameworks.

6.1 Limitations

In this investigation, we received our moral dilemmas from Scruples, which derived them from Reddit and annotated them using Mechanical Turk. This dataset encompasses only specific types of moral ambiguity and isn't fully representative of real-world decision-making. This limitation is due to the availability of existing datasets with human-annotated judgments in ambiguous situations. However, this scope allowed us to build a comprehensive understanding of LLM calibration in a controlled setting with appropriate human annotations, helping us gauge the models' performance. Future research could expand on the findings of this study by utilizing a dataset that captures diverse cultural and situational nuances, resulting in broader insights.

Another limitation of our study is the use of majority voting to determine the "gold" standard for annotations in the Anecdotes dataset. While majority voting simplifies our evaluation process, it has the potential to overshadow minority perspectives, especially in morally ambiguous cases where diverse ethical viewpoints might exist. This approach may unintentionally marginalize less common but equally valid moral judgments, reducing the representational diversity of the ground truth labels. Future work should explore methods that better account for minority viewpoints, such as weighted voting schemes or incorporating probabilistic annotations, to provide a more nuanced understanding of moral ambiguity.

The binary nature of human annotations also simplifies complex moral scenarios into "right" or "wrong." While this expedited the process of measuring LLM calibration, it reduces the degree of human reasoning. Although our approach provides a degree of alignment with human judgment, future studies should delve into more extensive evaluation methods that capture the full essence of moral ambiguity.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values, 2023. URL https://arxiv.org/abs/2008.02275.

Raisa Islam and Owana Marzia Moushi. Gpt-4o: The cutting-edge advancement in multimodal llm, 2024. URL https://easychair.org/publications/preprint/z4TJ/open.

Nora Kassner, Oyvind Tafjord, Ashish Sabharwal, Kyle Richardson, Hinrich Schuetze, and Peter Clark. Language models with rationality, 2023. URL https://arxiv.org/abs/2305.14250.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes, 2021. URL https://arxiv.org/abs/2008.09094.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever *. Language models are unsupervised multitask learners, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290v3.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. Evaluating the moral beliefs encoded in llms, 2023. URL https://arxiv.org/abs/2307.14324.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases, 1974. URL https://www2.psych.ubc.ca/~schaller/Psyc590Readings/TverskyKahneman1974.pdf.