Diagnosing Moral Reasoning Acquisition in Language Models: Pragmatics and Generalization

Anonymous ACL submission

Abstract

Ensuring that Large Language Models (LLMs) return just responses which adhere to societal values is crucial for their broader application. Prior research has shown that LLMs often fail to perform satisfactorily on tasks requiring moral cognizance, such as ethics-based judgments. While current approaches have focused on fine-tuning LLMs with curated datasets to improve their capabilities on such tasks, choosing the optimal learning paradigm to enhance the ethical responses of LLMs remains an open research debate. In this work, we aim to address this fundamental question: *can current learning paradigms enable LLMs to acquire sufficient moral reasoning capabilities?* Drawing from distributional semantics theory and the pragmatic nature of moral discourse, our analysis indicates that performance improvements follow a mechanism similar to that of semantic-level tasks, and therefore remain affected by the pragmatic nature of morals latent in discourse, a phenomenon we name the *pragmatic dilemma*. We conclude that this pragmatic dilemma imposes significant limitations on the generalization ability of current learning paradigms, making it the primary bottleneck for moral reasoning acquisition in LLMs.

***Warning**: examples in this paper may be offensive.*

1 Introduction

Given the widespread usage of LLMs across all facets of society, enabling such models with moral reasoning capabilities has become a significant research goal. Though AI alignment (Bai et al., 2022) has become the de-facto method to align LLMs with human values, its effectiveness has been debated (Lin et al., 2023; Qi et al., 2024). One significant complaint is that alignment with human preference does not allow LLMs to achieve intrinsic alignment, resulting in various safety issues, e.g., jailbreak attacks (Xie et al., 2023) and propagation of social biases to downstream tasks (Liu et al., 2024). However, enabling LLMs to develop moral reasoning capabilities is a non-trivial task; it is both a pragmatics-level task (Awad et al., 2022), as well as philosophically challenging, due to debate over the correct representation of human morals and ethics (Zhi-Xuan et al., 2024).

Jiang et al. (2021) and Hendrycks et al. (2020) represent pioneering efforts to enable LLMs to acquire ethical judgment capabilities by fine-tuning them on curated textual data that jointly depicts various moral situations alongside corresponding judgments. Zhou et al. (2024) introduces an in-context learning method to help LLMs perform moral reasoning, based on a top-down framework driven by the Moral Foundation Theory (Anderson and Anderson, 2007). Liu et al. (2023) introduce a social sandbox wherein LLMs can learn how to be moral through interactions. Tennant et al. (2024) propose a moral alignment framework to make LLM agents behave morally through a newly designed intrinsic moral reward function based on the Iterated Prisoner's Dilemma[1]. In addition to those efforts proposing solutions, new benchmarks have also been proposed (Forbes et al., 2020; Hendrycks et al., 2020; Ren et al., 2024).

There are also several studies which highlight the inefficiency of LLMs on tasks requiring moral reasoning. Talat et al. (2022) has criticized the Jiang et al. (2021) work described above, because while their intended goal was normative ethics, they instead leveraged a bottom-up approach for learning descriptive ethics (Vida et al., 2023; Fraser et al., 2022). Jin et al. (2022) empirically demonstrate that the current learning paradigm for moral reasoning tasks relies on a large training dataset. Sap et al. (2022) also show the failure of LLMs on social intelligence tasks such as theory-of-mind.

In cognitive science, Mahowald et al. (2024) sug-

[1] https://en.wikipedia.org/wiki/Prisoner%27s_dilemma

gest that while LLMs excel in formal language competence, they struggle with functional language competence—an essential requirement for acquiring moral reasoning capabilities. More fundamentally, Bender and Koller (2020) and other studies in BERTology (Rogers et al., 2021), argue that Transformers cannot achieve true language acquisition, as it necessitates physical grounding and situated communicative intent (Beuls and Van Eecke, 2024), which extends beyond the distributional semantics captured by Transformers (Harris, 1954; Lenci et al., 2008; Boleda, 2020). Previous studies (Bonagiri et al., 2024; Zhang et al., 2023) demonstrate that LLMs do not have consistent moral or ethical orientations across various instances, which is contrary to the moral consistency principle (Arvanitis and Kalliris, 2020). Appendix A.1 contains additional related works and motivation pertaining to machine ethics.

To address this debate, in this paper we pursue a deeper understanding of the mechanisms underlying current learning paradigms for moral reasoning acquisition. We argue that while existing paradigms can improve LLMs' performance on morality-related tasks, this enhancement: (1) primarily arises from distributional similarities between seen and unseen ethical situations, and (2) faces challenges in generalization due to the inherently pragmatic nature of morality. We name this phenomenon as the *pragmatic dilemma* (Laverick, 2010; Sap et al., 2022) of moral reasoning acquisition, which arises from the inherent gap between the nature of distributional semantics in LLMs and the pragmatic nature of morality. Significant consequences of this pragmatic dilemma include poor generalization and a lack of intrinsic alignment.

Specifically, we employ three fundamental tasks, Moral Foundations classification, rule of thumb generation, and ethical judgment prediction, as downstream evaluations of moral reasoning acquisition. We then compare their generalization characteristics with a representative semantics-driven task, sentiment analysis. Motivated by the distributional semantics hypothesis, we: (1) empirically show the generalization and convergence pitfalls of Moral Foundations classification; (2) given the characteristic of autoregressive language models, propose a *Representational Likelihood Algorithm (RLA)* to statistically correlate representational similarity between seen and unseen moral pragmatics with the prediction likelihood of unseen morals; and (3) using RLA, perform mechanistic analysis of LLM performance gains for unseen situations.

Section 2 introduces the prevalent learning paradigm for moral reasoning acquisition and highlights the generalization challenges in fine-tuning masked language models for moral foundation prediction. Section 3 presents experimental results across different learning paradigms, and Section 4 provides a detailed mechanistic analysis. Based on our experimental results, we conclude that the pragmatic dilemma blocks the effectiveness of current learning paradigms.

2 Preliminary Background

In this section, we begin by introducing the benchmarks and dataset annotation used in our study. We then present the prevailing learning paradigm for moral reasoning acquisition. Finally, we use the Moral Foundations prediction task with a Masked Language Model, specifically BERT (Devlin et al., 2019), as a case study, to illustrate the generalization challenges of this task by drawing comparisons to the semantics-level task of sentiment analysis.

2.1 Benchmark and Dataset Annotation

Situation: Reminding my coworker who crashed into my car to pay to get it repaired.
Moral Foundation: Fairness.
Rule of Thumb (RoT): If you crash into someone's car, you should pay for their repairs.
(Ethical) Judgment: You should.

Table 1: Dataset Annotation. Given a moral situation describing a morality-relevant case, the corresponding Moral Foundation, RoT, and Judgment are presented.

We employ two popular benchmarks: MIC (Ziems et al., 2022) and SocialChem (Forbes et al., 2020). Table 1 presents an overview of the dataset annotations used across both benchmarks. Given a moral situation, the *Moral Foundation* (Haidt and Joseph, 2004; Haidt and Graham, 2007) represents the underlying social norm that the situation either adheres to or violates (please refer to Table 8 for more details of Moral Foundation Theory). The *RoT (Rule of Thumb)* encapsulates a fundamental explanation of right and wrong behavior, serving as a guidance for the subsequent ethical judgment. The *(Ethical) Judgment* then determines whether the given situation is deemed acceptable or unacceptable. While a single moral situation may be associated with multiple moral foundations, this study focuses

exclusively on cases where only one underlying moral foundation is present. In the MIC, each prompt-reply pair is treated as a distinct situation.

2.2 Learning Paradigms

Existing learning paradigms for moral reasoning acquisition generally fine-tune LLMs on curated textual data that depicts various moral situations alongside corresponding judgments or actions. In previous studies, ethical judgment prediction and RoT generation are the most popular tasks (Bonagiri et al., 2024; Ren et al., 2024; Hendrycks et al., 2020; Sorensen et al., 2024), and Moral Foundations classification is widely accepted in the area of computational social science (Johnson and Goldwasser, 2018; Roy et al., 2021). Though there is no agreed-upon definition for moral reasoning acquisition, we consider Moral Foundations classification, RoT generation, and ethical judgment prediction as three downstream tasks indicative of moral reasoning capabilities. Although some studies incorporate interactive sandboxes or multi-round feedback into learning paradigms (Liu et al., 2023; Wang et al., 2024), Moral Foundations classification, RoT generation, and ethical judgment prediction remain fundamental tasks, which when fine-tuned with LLMs form the preferred learning paradigms.

Notations. We denote the moral situation as x_s, the moral foundation as y_m, the RoT as y_r, and the judgment as y_j. Assuming an LLM f is parameterized by θ, RoT generation is formulated as $y_r = f_\theta(x_s)$ and judgment prediction is represented as $y_j = f_\theta(x_s)$.

Fine-tuning Strategies. Current learning paradigms of moral reasoning acquisition which aim to maximize conditional probabilities $\mathcal{P}_\theta(y_r|x_s)$ and $\mathcal{P}_\theta(y_j|x_s)$, typically apply a self-supervised fine-tuning or a reinforcement learning loss objective[2]. Given the causal relationships among moral foundations, RoT, and judgment, previous studies often integrate them into a unified prediction task, such as $y_r = f_\theta(x_s, y_m)$ and $y_j = f_\theta(x_s, y_m, y_r)$. During fine-tuning, the input for RoT generation can be x_s with or without y_m, while the input for ethical judgment prediction can be x_s with or without y_m and/or y_r.

2.3 Pitfalls of Generalization

In this section, we use the Moral Foundations classification task as an example to illustrate its generalization pitfalls by comparing it to the semantics-level task of sentiment analysis. We argue that *in moral classification tasks, there should be serious generalization issues since the classification model has to map semantically different situations into the same moral foundation label*. A direct consequence is that an unseen situation is likely to be predicted correctly only if a semantically similar sample exists in the training set. This similarity requirement is much stricter than that for the sentiment analysis task.

Situation: Kicking a kid out of his birthday party.
Situation: Not telling my mom I smoke weed.

Table 2: Situation Examples. Two moral situations with the same underlying moral foundation of authority-subversion.

Our argument is driven by the gap between the distributional semantics captured by neural language models and the inherently pragmatic nature of morality. For instance, Table 2 presents two moral situations from the SocialChem benchmark; they are semantically different (*distributional semantics*) but the underlying moral foundations are identical (*pragmatics*). If we force an MLM to map these two situations into the same moral foundation label, it would violate the captured distributional semantics during pre-training. To illustrate how the violation works, we refer to a semantics-level task of sentiment analysis using the SST dataset from the GLUE benchmark (Wang et al., 2018).

Experimental Settings for Classification. We have two settings for the moral classification tasks: classify moral situations to moral foundations and classify RoTs to moral foundations. We use a fine-tuning dataset with 7500 randomly sampled cases and the bert-base-uncased[3] model as the backbone model. Beyond the backbone model, we insert a fully-connected layer as the classifier layer. More details about the hyperparameters setting is available in Appendix A.2.

Observations for Classification Performance. Table 1 presents the classification performance on both the training and development set. Compared to the generalization behavior observed in SST (rightmost figure), the moral foundation classification tasks (first four figures) exhibit a significant performance gap between the training set and the development set. However, for `MIC-RoT` and

[2]Please note the choice of objective loss function does not impact our conclusion.

[3]https://huggingface.co/google-bert/bert-base-uncased

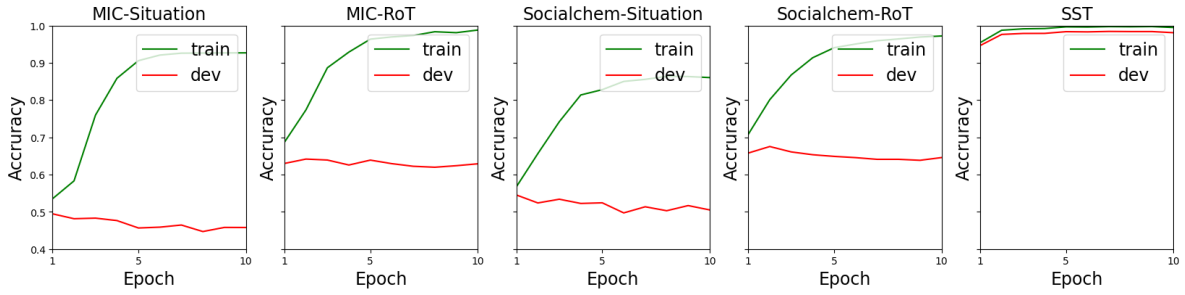


Figure 1: Training and Development Accuracy Over 10 Fine-tuning Epochs. The first four figures display results for moral foundation classification tasks, while the rightmost figure shows the results for the SST benchmark.

`SocialChem-RoT`, because the training accuracy approaches 100% and converges after only several epochs, this suggests that *task difficulty is not the primary cause of the observed generalization gap*. The difference in classification performance between Situation and RoT stems from the fact that RoT is constructed based on typical moral foundations, inherently conveying information about the corresponding moral foundation. However, the generalization gap between the training set and development set for all moral foundation classification settings is apparent. To further analyze the generalization pitfall in moral foundation classification, we examine the convergence behavior with respect to training dataset size. We use the curve of development accuracy in SST as a reference to highlight the convergence issue observed in moral foundation classification tasks.

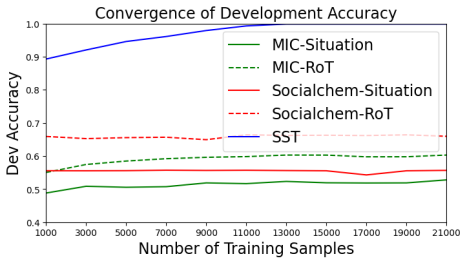


Figure 2: Convergence Curve of Development Accuracy for Considered Classification Tasks. Only the development accuracy of SST increases with more training samples and finally approaches 1.0.

Experimental Settings for Convergence. SST is a binary classification task. To ensure a fair comparison, we re-categorize the moral foundation labels for MIC and SocialChem to convert them into a binary classification task (details are in Appendix A.3). For each task setting, we incrementally increase the training set size from 1,000 to 210,000 in steps of 2,000 and report the best performance on the development set at each training size setting.

Observations for Convergence. Figure 2 illustrates the curve of development accuracy across all evaluated classification tasks. For SST, accuracy improves as the number of training samples increases, eventually stabilizing and approaching 1.0. In contrast, the development accuracies for moral foundation classification tasks show no improvement in SocialChem and only marginal gains in MIC. We believe this disparity is due to the fact that moral situations in SocialChem are generally shorter than that of MIC. The convergence behavior analysis again showcases the generalization pitfalls of the moral foundation classification task.

In summary, we: (1) introduce the current learning paradigms for moral reasoning acquisition; and (2) show the generalization pitfalls of the moral foundation classification task (a pragmatics-level task) by referring and comparing to the development accuracy of a semantics-level task.

3 Fine-tuning for Moral Reasoning Acquisition

In this section, we introduce fine-tuning strategies and experimental results of existing learning paradigms for moral reasoning acquisition, focusing on the tasks of RoT generation and ethical judgment prediction.

Experimental Settings. We take Mistral-7B[4] and Llama3-8B[5] as the backbone models and leverage the LoRA method to fine-tune them through a supervised fine-tuning loss. For each benchmark, we employ two fine-tuning strategies for RoT generation and four fine-tuning strategies for ethical judgment prediction. For *RoT generation*, we fine-tune LLMs: (1) to directly

[4] https://huggingface.co/mistralai/Mistral-7B-v0.1

[5] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

SocialChem	BertScore	Rouge1	Rouge2	RougeL	MIC	BertScore	Rouge1	Rouge2	RougeL
rot	.777	.229	.096	.213	rot	.768	.175	.077	.168
moral-rot	.836	.416	.205	.401	**moral-rot**	.826	.393	.192	.379
judg	.7240	.230	.137	.230	judg	.671	.071	.000	.071
moral-judg	.7632	.464	.346	.464	**moral-judg**	.762	.314	.000	.314
rot-judg	.7626	.464	.346	.464	rot-judg	.660	.061	.000	.061
moral-rot-judg	.7628	.463	.345	.463	moral-rot-judg	.761	.306	.000	.306

Table 3: Performance of Fine-tuned `Mistral` Model Across Various Fine-tuning Strategies for Each Benchmark, with the best strategy highlighted in **bold**. For both tasks, introducing more information, e.g., moral foundation, in the fine-tuning process would improve the performance. The `moral-rot` achieves the optimal performance for both SocialChem and MIC. The `moral-judg` and `moral-judg` are the best strategy for SocialChem and MIC respectively, in terms of the judgment prediction task. Additional results for Llama3 are availabe in Table 7.

generate RoT according to the given situation (rot) and (2) first generate the moral foundation, then the RoT (moral-rot). For *Judgment Prediction*, we fine-tune LLMs to: (1) directly predict judgment (judg) , and (2) firstly generate the moral foundation and/or RoT then the judgment (moral-judg, rot-judg and moral-rot-judg).

The prompting format and LoRA fine-tuning settings are available in Appendix A.4. We consider 10000 samples with only one underlying moral foundation for analytical convenience. In the process of fine-tuning, we take the check point with the least loss on the development set, and report its performance on the test set. During inference, we prompt fine-tuned LLMs to first generate intermediate predictions before producing the final RoT or ethical judgment, following the same prompting strategy used during fine-tuning. For example, in the `moral-rot` strategy, LLMs are instructed to first predict the moral foundation based on the given situation and subsequently generate the RoT using both the situation and the predicted moral foundation. Following Ziems et al. (2022), we report the performance of the BertScore (Zhang et al., 2019), Rouge-1, Rouge-2, and Rouge-L metrics.

RoT generation and ethical judgment prediction align with the core capabilities essential for morality-related scenarios and serve as prototypical formats for moral reasoning. By incorporating moral foundations into RoT generation, we aim to guide LLMs to first identify the moral foundation associated with a given situation, thereby improving the quality of the generated RoT. RoTs serve as instances of evidence and explanation for ethical judgments, aligning with previous studies that seek to enhance LLMs' social intelligence through social interaction environments (Liu et al., 2023; Wang et al., 2024).

Main Results. Table 3 and Table 7 present fine-tuning results for Mistral and Llama3, respectively[6]. As shown in Table 3, introducing moral foundations in fine-tuning enhances performance across all experimental settings. However, incorporating RoT information along into the ethical judgment prediction task has a negative impact to the MIC benchmark. We hypothesize that this is because judgments are significantly shorter than RoTs, and the added complexity of RoTs would introduce challenges for fine-tuning.

4 Mechanistic Analysis

In the previous sections, we introduced preliminary studies regarding the generalization pitfalls of the moral foundations classification task (Section 2), and the performance of fine-tuning LLMs for two moral reasoning tasks (Section 3). In this section, we: (1) propose the Representational Likelihood Algorithm (RLA) which can uncover supportive training samples for a given test sample; (2) explore the characteristics of supportive training samples, demonstrating that the introduction of additional information to enhance generalization aligns with the generalization mechanism of the semantics-level task; (3) showcase that the *pragmatic dilemma* still holds even though fine-tuned LLMs perform better in RoT generation and ethical judgment prediction.

Motivation. Our study builds on the representational learning nature of LLMs and the widely accepted principle in generalization theory that a well-trained machine learning model can generalize effectively when the training and test set distributions are closely aligned in the feature space (Zhou et al., 2022; Hupkes et al., 2022). Since neural language models capture distributional semantics, representational similarity can be interpreted as equivalent

[6]Note that this paper does not aim to achieve state-of-the-art performance but rather to investigate the underlying mechanisms behind these performance gains.

to distributional similarity. Recall from Section 2 that we highlighted the generalization pitfalls of the moral foundation classification task. We argue that similar pitfalls should also exist for RoT generation and ethical judgment prediction. Our hypothesis is that *for a given test sample, the LLM can generalize effectively only if highly similar training samples have been adequately learned during fine-tuning*. To test this hypothesis, we propose a novel algorithm to identify the training samples most conducive to the generalization of a given test sample within the representation space.

4.1 Representational Likelihood Algorithm

Motivated by the representation similarity hypothesis in domain generalization (Ben-David et al., 2006), we present our method for identifying training samples that contribute to the prediction of a given test sample. We refer to these training samples as *generalization-supportive samples*[7]. Our goal is to correlate representational similarity with LLM predictions, and then leverage this correlation to characterize the generalization mechanism of the considered morality acquisition tasks.

Assume that a fine-tuned LLM f_θ has been trained on the training set $\mathcal{D}_{\text{train}}$, where each sample is represented as $x = [x_s, y_m, y_r, y_j]$, following the annotation introduced in Section 2.2. We denote training samples as $x \sim \mathcal{D}_{\text{train}}$ and test samples as $x' \sim \mathcal{D}_{\text{test}}$. The hidden states of f_θ are denoted by $\mathcal{H}_\theta(\cdot)$, and the conditional likelihood of a given input and output is represented as $\mathcal{P}_\theta(\cdot|\cdot)$. Denote the cosine similarity function as $\cos(\cdot)$. Algorithm 1 presents our proposed Representational Likelihood Algorithm (RLA) by taking the `judg` fine-tuning strategy ($y_j = f_\theta(x_s)$) as an instance. Specifically,

1. For each test case, we randomly sample $\mathcal{N}$ samples $\mathcal{X}$ from the training set (line 3).
2. For each training sample x^t in the sampled set $\mathcal{X}$, we calculate the **similarity score** S^t which comprises the: (1) cosine similarity between two hidden states $\mathcal{H}_\theta(x_s^t)$ and $\mathcal{H}_\theta(x_s^{'})$ (line 5) measuring the representational similarity, and (2) likelihood, the conditional probability $\mathcal{P}_\theta(x_j^t|x_s^t)$ measuring how good f_θ fits x^t (line 5). With this design, only those training samples that have been fitted well by f_θ would be considered in the process of measuring representational similarity.

[7]In this paper, we use generalization-supportive and supportive interchangeably.

Algorithm 1 RLA for Judgment Prediction

1: Initialize $r = 0$, $\mathbf{d} = \{\}$
2: **for** each sample x' in $\mathcal{D}_{\text{test}}$ **do**
3: Sampling $\mathcal{N}$ cases from $\mathcal{D}_{\text{train}}$ as $\mathcal{X} = [x^1, x^2, \cdots, x^{\mathcal{N}}]$
4: **for** each x^t in $\mathcal{X}$ **do**
5: $S^t = \overbrace{\cos(\mathcal{H}_\theta(x_s^t), \mathcal{H}_\theta(x_s^{'}))}^{\textbf{representational similarity}} \cdot \overbrace{\mathcal{P}_\theta(x_j^t|x_s^t)}^{\textbf{likelihood}}$
6: $\mathbf{d}[S^t] = \underbrace{\mathcal{P}_\theta(x_j^t|x_s^{'})}_{\textbf{prediction}}$
7: **end for**
8: Sort $\mathbf{d}$ by key in *ascending* order, return the value list as $\mathcal{V}$
9: **if** MEAN($\mathcal{V}[: \frac{\mathcal{N}}{2}]$) < MEAN($\mathcal{V}[\frac{\mathcal{N}}{2} :]$) **then**
10: r++
11: **end if**
12: **end for**
13: **return** $\frac{r}{\#\mathcal{D}_{\text{test}}}$

3. Compute the conditional probability of the training sample's judgment given the test case's situation (line 6).
4. If f_θ becomes increasingly likely to assign x_t's judgment x_j^t to x_s' as their representational similarity increases, then we can correlate representational similarity and prediction (lines 8-10).

In our experiments, we utilize the hidden states from the 15th layer onward of the final token as the representation and compute the average cosine similarity across these layers to obtain the representational similarity score. This is because previous studies (Geva et al., 2023; Liu et al., 2024) indicate that the LLMs considered in this paper generally exhibit differences in the hidden state space from the 15th layer onward. Table 4 presents the results of two baseline fine-tuning strategies, `rot` and `judg`, evaluated across various benchmarks and LLM models. As shown, all experimental results

	Mistral	Llama3
Socialchem-rot	.920	.924
Socialchem-judg	.998	.996
MIC-rot	.926	.912
MIC-judg	.990	.971

Table 4: Experimental results for the simulation task show that all values exceed 0.9, indicating a strong correlation between representational similarity and prediction.

exceed 0.9, particularly the `judg` fine-tuning strategy which is very close to 1.0, demonstrating that there exists correlation between representational similarity and prediction. In other words, for a given test sample, generalization-supportive training samples can be identified by assessing their representational similarity.

4.2 Interpretation of Generalization

Building on the method for identifying generalization-supportive training samples from Section 4.1, this section interprets the generalization mechanism of the examined morality-relevant tasks by analyzing the characteristics of these supportive training samples[8].

For each test sample, we collect the top-10 generalization-supportive training samples with the most highest similarity score S^t. However, the similarity score S^t is a high-level metric capturing the statistical correlation between representational similarity and predictions, making it insufficient for directly interpreting the underlying reasons for performance gains. To have an in-depth analysis, we investigate (i) the cosine similarity of hidden states between the test sample's moral situation and the training sample's moral situation; (ii) the BertScore between the train sample's situation and the test sample's situation. Figure 3 present these two analytical perspectives on the top 10 generalization-supportive training samples for the fine-tuned Mistral model across two benchmarks.

By zooming into the left four subfigures in Figure 3, introducing moral foundation or RoT in the fine-tuning process can decrease the representational similarity, particularly the optimal fine-tuning strategies, e.g., `moral-rot` and `moral-judg`, lead to lower representational similarities than that of the baseline strategy (`rot` and `judg`). This phenomenon aligns with our hypothesis that *generalization in moral reasoning acquisition tasks requires a high degree of representational similarity between test and training samples*.

By referring to the curve of SST that also faces a lower representational similarity, we can conclude that additional information of moral foundation or RoT would alleviate the generalization pitfall of the baseline strategy that necessitates much similar training samples to generalize. This is rather natural since those fine-tuning strategies not only

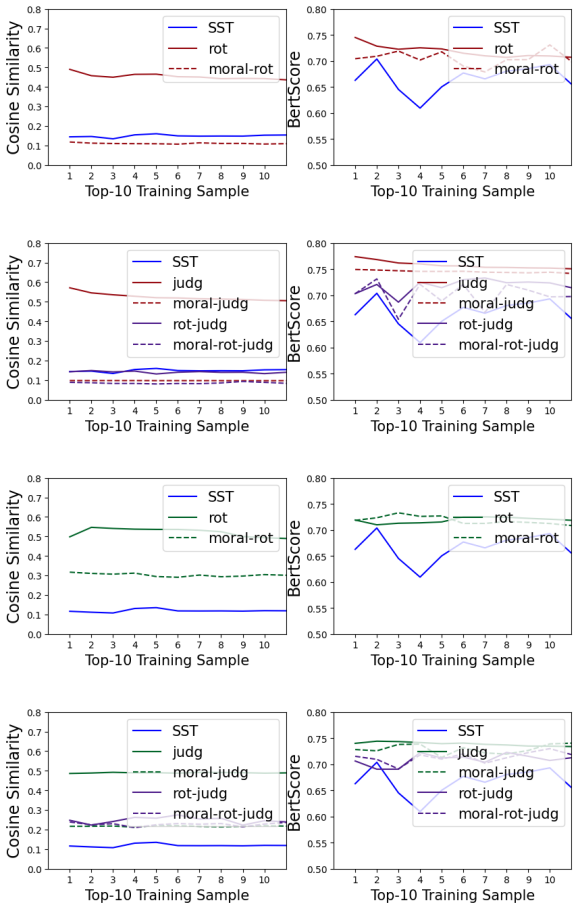


Figure 3: Top-10 generalization-supportive training samples analysis for fine-tuned Mistral with the SocialChem (upper two rows) and MIC (bottom two rows) benchmark.

capture the information of situations but also moral foundations and/or RoTs, newly introduced information would impact the characteristics of the representation space.

Additionally, we can observe decreased BertScore in the right four sub-figures, except for RoT generation in the MIC benchmark, where the BertScore for `moral-rot` remains close to that of the baseline `rot` strategy. A decrease in BertScore suggests that the additional information reduces reliance on generalization-supportive training samples with high distributional similarity to the test sample. Due to the association between distributional similarity and representational similarity in LLMs, those two observations are aligned. It is not surprising that the performance gain arises from the generalization mechanism analogical to that of semantics-level tasks. A natural question is *does the incorporation of moral foundations or RoT alleviate the pragmatic dilemma of current learning paradigms in moral reasoning acquisition?*

[8] In this section, we provide a detailed analysis only for the fine-tuned Mistral, while the analysis for the fine-tuned Llama3 is presented in Appendix A.5.

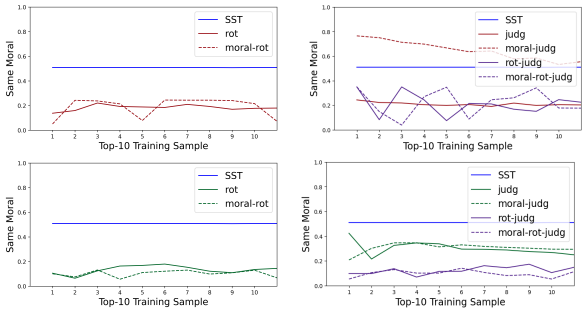


Figure 4: Ratio of generalization-supportive training situations with the same underlying moral foudation as the test situation. Upper two subfigures are for SocialChem and the bottom sub-figures are for MIC. Top-50 situations are available in Appendix A.6.

An extreme case for the vanishment of the pragmatic dilemma is: *for a given test situation, top-10 generalization-supportive training moral situations should have the same underlying moral foundations as the test moral situation*. Therefore, we compute the ratio of the top-10 supportive training moral situations that share the same moral foundations as the test moral situation. Notably, we take the term training/test moral situation, for MIC and SocialChem, instead of training/test samples to emphasize that our analysis exactly focuses on moral situations. For reference, we include SST and consider the sentiment label when calculating the ratio for SST.

Figure 4 presents the results for this ratio. Interestingly, even for SST, which can be viewed as a binary classification task, only half of the supportive training samples share the same sentiment label as their corresponding test samples. For both RoT generation and ethical judgment prediction, the optimal fine-tuning strategies (`moral-rot` and `moral-judg`) align with the baseline fine-tuning strategies (`rot` and `judg`), except for `moral-judg` on the SocialChem benchmark. We believe this exception arises because the textual length of moral situations in SocialChem is relatively short, amplifying the influence of ethical judgment during fine-tuning. On the other hand, we calculate the average conditional likelihoods of the top-10 supportive training situations, and note the optimal fine-tuning strategy does help LLMs fit training samples. These observations suggest that LLMs consider moral situations and additional information together to generalize, but still operate primarily within the realm of semantics.

Recall that, in Section 2, we demonstrate that the generalization and convergence behavior of moral foundation classification is different from SST due to the pragmatic delimma. Similarly, we also argue that the pragmatic nature of morality would be more negative to the language modeling capability of LLMs than that from SST. Figure 5 presents the perplexity evaluation results, acquired through the OpenWebText datset (Gokaslan et al., 2019), of Mistral models fine-tuned with different strategies. It is obvious that morality-relevant tasks introduce more perplexity than SST.

	SocialChem	MIC
rot	.389	.659
moral-rot	.418	.738
judg	.992	.770
moral-judg	.997	.835

Table 5: The average conditional likelihoods of top-10 generalization-supportive training samples.

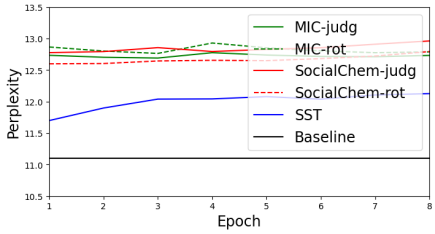


Figure 5: Perplexity for Mistral. Baseline indicates the Perplexity of the LLMs without any fine-tuning.

In summary, *while the optimal fine-tuning strategies improve performance on both tasks, this improvement remains within the realm of distributional semantics, and the pragmatic dilemma persists*.

5 Conclusion

In this paper, we answered the question *can current learning paradigms enable LLMs to acquire moral reasoning?* Based on distributional semantics and the pragmatic nature of morality, we demonstrate that (1) the pragmatic dilemma of LLMs make them inefficient in moral reasoning acquisition tasks; (2) the improved performance still stems from the realm of distributional semantics; (3) the current learning paradigm for moral reasoning acquisition impairs LLMs' language modeling capability more than semantics-level tasks. We conclude that the pragmatic dilemma is the primary bottleneck for moral reasoning acquisition. Please refer to Appendix A.7 for more dicussions.

Limitations

In this draft, we focus only on moral situations with a single underlying moral foundation. However, in real-world scenarios, moral situations often involve multiple moral foundations, which we leave for future research. Additionally, while the tasks considered in this paper reflect fundamental aspects of moral reasoning, a deeper analysis of how the pragmatic dilemma manifests in recently proposed social sandbox systems would be a valuable direction for future study.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément Crepy, Daria Valter, John Canny, and Natasha Jaques. 2023. Moral foundations of large language models. *arXiv preprint arXiv:2310.15337*.

Colin Allen, Wendell Wallach, and Iva Smit. 2006. Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17.

Michael Anderson and Susan Leigh Anderson. 2007. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4):15–15.

Michael Anderson and Susan Leigh Anderson. 2011. *Machine ethics*. Cambridge University Press.

Alexios Arvanitis and Konstantinos Kalliris. 2020. Consistency and moral integrity: A self-determination theory perspective. *Journal of Moral Education*, 49(3):316–329.

Isaac Asimov. 1941. Three laws of robotics. *Asimov, I. Runaround*, 2(3).

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Edmond Awad, Sydney Levine, Michael Anderson, Susan Leigh Anderson, Vincent Conitzer, MJ Crockett, Jim AC Everett, Theodoros Evgeniou, Alison Gopnik, Julian C Jamison, et al. 2022. Computational ethics. *Trends in Cognitive Sciences*, 26(5):388–405.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9:20–1.

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(4):1277–1311.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Vamshi Krishna Bonagiri, Sreeram Vennam, Priyanshul Govil, Ponnurangam Kumaraguru, and Manas Gaur. 2024. Sage: Evaluating moral consistency in large language models. *arXiv preprint arXiv:2402.13709*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Kathleen C Fraser, Svetlana Kiritchenko, and Esma Balkir. 2022. Does moral code have a moral code? probing delphi's moral philosophy. *arXiv preprint arXiv:2205.12771*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Zellig S Harris. 1954. Distributional structure.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. *arXiv preprint arXiv:2008.02275*.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2022. State-of-the-art generalisation research in nlp: a taxonomy and review. *arXiv preprint arXiv:2210.03050*.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny T Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jack Hessel, et al. 2025. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, pages 1–16.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 720–730.

Shivani Kumar and David Jurgens. 2025. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with unimoral. *arXiv preprint arXiv:2502.14083*.

Wendy Laverick. 2010. Ethical dilemmas and pragmatic compromises. *Ethnography in social science practice*, page 73.

Alessandro Lenci et al. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*.

Guangliang Liu, Haitao Mao, Jiliang Tang, and Kristen Johnson. 2024. Intrinsic self-correction for enhanced morality: An analysis of internal mechanisms and the superficial hypothesis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16439–16455, Miami, Florida, USA. Association for Computational Linguistics.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models on simulated social interactions. In *The Twelfth International Conference on Learning Representations*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Rajakishore Nath and Vineet Sahu. 2020. The problem of machine ethics in artificial intelligence. *AI & society*, 35:103–111.

Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2024. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040, Bangkok, Thailand. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Shamik Roy, María Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2024. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from

natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2024. Moral alignment for llm agents. *arXiv preprint arXiv:2410.01639*.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: A survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhaowei Zhang, Fengshuo Bai, Jun Gao, and Yaodong Yang. 2023. Measuring value understanding in language models through discriminator-critique gap. *arXiv preprint arXiv:2310.00378*.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. 2024. Beyond preferences in ai alignment. *Preprint*, arXiv:2408.16984.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. 2024. Rethinking machine ethics–can llms perform moral reasoning through the lens of moral theories? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2227–2242.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A Appendix

A.1 Additional Related Works

Machine ethics (Anderson and Anderson, 2011; Tolmeijer et al., 2020; Nath and Sahu, 2020; Allen et al., 2006) has been a long-standing research topic for hardware and software systems, with the aim of maximizing their benefits while minimizing societal risks. Recently, we have witnessed the progress of Artificial Intelligence (AI), particularly that associated with Large Language Models (LLMs), changing the world. Ensuring LLMs will acquire an understanding of ethics to prevent them from making harmful decisions has become a serious research problem for both academia and industry. Dating back to the 1940s, the Three Laws of Robotics (Asimov, 1941) were proposed to ensure that robots do not cause harm to humans. Since then, machine ethics has been explored by researchers in philosophy, psychology, and cognitive science. However, it remains a significant challenge for AI, as even coherent and diverse language generation poses difficulties. The widespread deployment of LLMs opens the door for AI researchers to pursue ethics acquisition due to their strong semantic modeling capability.

Numerous studies have attempted to evaluate the moral and ethical orientations encoded in LLMs through empirical experiments. Bonagiri et al. (2024) demonstrates that model performance and moral consistency are independent of one another, while Abdulhai et al. (2023) investigates whether LLMs exhibit biases toward specific moral principles. Scherrer et al. (2024) proposes a statistical method to assess the moral values encoded in LLMs, and Zhang et al. (2023) introduces a metric to determine whether LLMs understand ethical values both in terms of "knowing what" and "knowing why." Collectively, these studies highlight that LLMs lack consistent moral or ethical orientations across different scenarios. Enabling LLMs to acquire ethical values is a formidable challenge, not only because ethical AI operates at the level of pragmatics (Awad et al., 2022), but also due to the

philosophical complexities surrounding the proper representation of human ethics (Zhi-Xuan et al., 2024). Progress has been made, albeit only partially.

A.2 Hyperparameters for the Bert Classifier

Hyperparameters are available in Table 6.

Hyperparameters	Setting
Optimizer	AdamW
Adam β_1	0.9
Adam β_2	0.98
Adam ϵ	1e-3
Learning rate for BERT	5e-5
Learning rate for classifier layer	1e-2
Maximum training epochs	10
Weight decay	0.01
Batch size	32
Seed	1,2,3,4,5

Table 6: Hyperparameter Settings for the AdamW Optimizer.

A.3 Re-categorization of Moral Foundation Labels

For **MIC**, we label samples with the moral foundation of *Care* as 0, and those with the foundations of *Fairness, Liberty, Authority*, and *Loyalty* as 1. For **SocialChem**, samples classified under *Loyalty-Betrayal* are labeled as 0, while those falling under *Fairness-Cheating*, *Care-Harm*, *Sanctity-Degradation*, and *Authority-Subversion* are labeled as 1.

A.4 Experimental Settings for Fine-tuning

Prompting format moral-rot-judgment

Situation: {#SITUATION}
Moral Foundation: {#MORAL_FOUNDATION}
Rule of Thumb: {#RoT}
Ethical Judgment: {#judgment}

LoRA hyperparameters
rank: 64
lora alpha: 16
lora dropout: 0.1
target modules: q_proj, k_proj, v_proj, o_proj
batch size: 16
learning rate: 5e-5

A.5 Mechanistic Analysis to Fine-tuned Llama3

Table 6 introduces the fine-tuning results for the Llama3 model. Different from Mistral, introducing additional information of the moral foundations and RoT do not always contribute to better performance. For the SocialChem benchmark, the baseline fine-tuning strategy outperforms other strategies, albeit by a very narrow margin. This aligns with the generalization mechanism illustrated in Figure 6. Unlike Mistral, the introduction of moral foundations and RoT does not reduce cosine similarity or BertScore. Figure 11 shows the ratio of the same moral foundation among top 10 generalization-supportive training moral situations, and the behavior of Llama3 is the same as Mistral. In summary, the pragmatic dilemma still persists for the Llama3 model and is even worse than that of the Mistral model.

A.6 Top-50

In Figure 4, we show only 10 generalization-supportive samples. Here, we demonstrate that the characteristics of all top-50 generalization-supportive training samples are closely aligned with those of the top 10 reported in that figure.

Mistral-SocialChem-RoT:[0.138, 0.16, 0.221, 0.193, 0.189, 0.185, 0.21, 0.193, 0.17, 0.178, 0.18, 0.176, 0.181, 0.191, 0.187, 0.157, 0.161, 0.145, 0.163, 0.133, 0.15, 0.181, 0.152, 0.162, 0.18, 0.163, 0.173, 0.16, 0.158, 0.186, 0.176, 0.178, 0.17, 0.185, 0.171, 0.169, 0.165, 0.194, 0.191, 0.173, 0.19, 0.173, 0.188, 0.192, 0.188, 0.195, 0.189, 0.19, 0.195, 0.17] with mean value of 0.17636

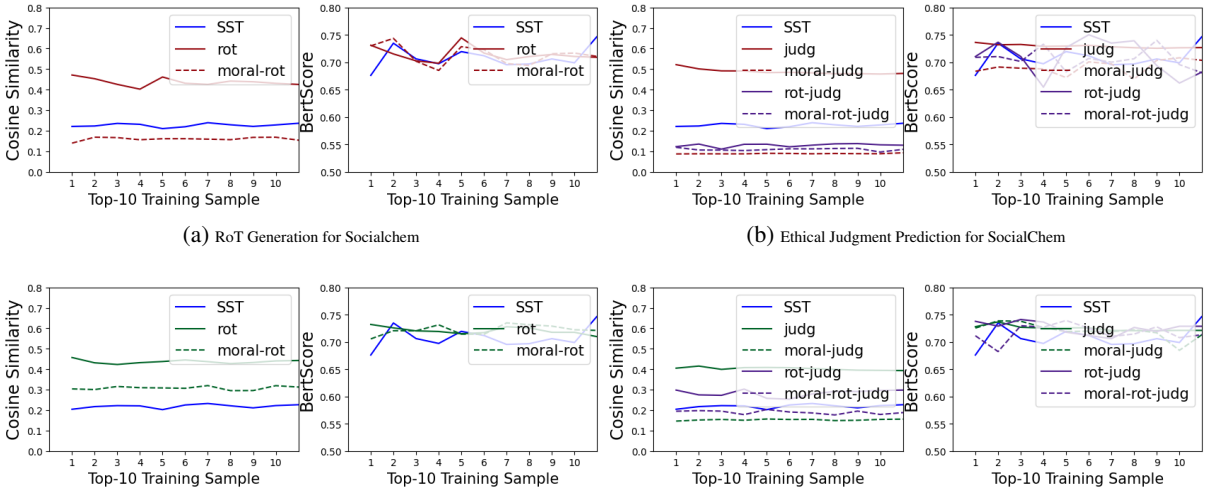
Mistral-Socialchem-MoralRoT: [0.051, 0.243, 0.238, 0.214, 0.079, 0.245, 0.244, 0.244, 0.241, 0.216, 0.072, 0.133, 0.204, 0.276, 0.137, 0.179, 0.178, 0.115, 0.049, 0.151, 0.152, 0.16, 0.089, 0.048, 0.186, 0.141, 0.126, 0.137, 0.146, 0.047, 0.045, 0.041, 0.122, 0.156, 0.143, 0.084, 0.237, 0.232, 0.135, 0.099, 0.09, 0.207, 0.371, 0.169, 0.23, 0.127, 0.093, 0.199, 0.164, 0.163] with mean of 0.15696

A.7 Discussion

Generalization remains a significant challenge in the acquisition of moral reasoning, and no optimal solution has yet been identified. Recently, Jiang et al. (2025) proposed a hybrid approach that combines bottom-up and top-down methods. However, their method still relies on a substantial number of training samples. Bergen et al. (2016) demonstrated that pragmatic reasoning can be approximated through semantic inferences, highlighting a linguistic foundation for this connection. Nevertheless, how to formally structure a semantic inference

SocialChem	BertScore	Rouge1	Rouge2	RougeL	MIC	BertScore	Rouge1	Rouge2	RougeL
rot	.8222	.358	.151	.343	rot	.814	.365	.152	.332
moral-rot	.8217	.356	.152	.340	**moral-rot**	.818	.365	.168	.352
judg	.759	.440	.313	.440	judg	.684	.109	.000	.109
moral-judg	.757	.411	.285	.411	moral-judg	.751	.254	.000	.254
rot-judg	.755	.400	.264	.400	rot-judg	.660	.061	.000	.061
moral-rot-judg	.752	.370	.248	.370	**moral-rot-judg**	.762	.314	.000	.314

Table 7: Performance of Fine-tuned `Llama3` Model Across Various Fine-tuning Strategies for Each Benchmark. The best fine-tuning strategy is highlighted in **bold** and the second best strategy is underlined. For MIC, incorporating additional information, such as moral foundations, during fine-tuning enhances performance; however, this effect is not observed for SocialChem.



(a) RoT Generation for Socialchem

(b) Ethical Judgment Prediction for SocialChem

(c) RoT Generation for MIC

(d) Ethical Judgment Prediction for MIC

Figure 6: Top-10 Generalization-Supportive Training Samples Analysis for Fine-tuned Llama3 Through the Introduced Fine-tuning Strategies.

framework for moral reasoning remains an open question. One promising direction is to ground such a framework in the human moral decision-making process. Kumar and Jurgens (2025) introduced the first benchmark in the NLP community focused on how humans make moral decisions. Their benchmark is based on an intuitionist model: participants are first asked to make a moral judgment and then provide an explanation for their decision. This type of annotation presents challenges for LLMs, as human explanations are expressed in free-text form and often lack enough situated semantic information (Sap et al., 2022). Despite these difficulties, the benchmark offers a valuable opportunity for exploring methods that aim to derive semantic inferences from human rationales—potentially bridging the gap in pragmatic moral reasoning.

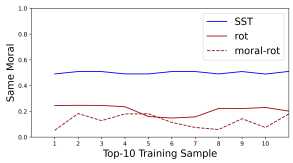


Figure 7: RoT in SocialChem

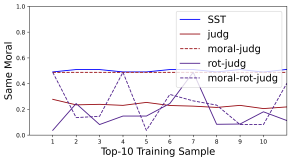


Figure 8: Ethical Judgment Prediction in SocialChem

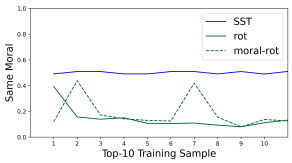


Figure 9: RoT in MIC

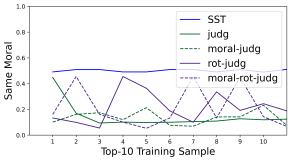


Figure 10: Ethical Judgment Prediction in MIC

Figure 11: Same Moral Ratio for Fine-tuned Llama3.

Moral Foundation Branches	Brief Description
Care Harm	Demonstrates care, generosity, compassion, and empathy, while showing sensitivity to others' suffering and upholding the principle of avoiding harm.
Fairness Cheating	Encompasses fairness, justice, reciprocity, altruism, rights, autonomy, equality, proportionality, and the rejection of cheating.
Loyalty Betrayal	Emphasizes group affiliation, solidarity, patriotism, and self-sacrifice, while prohibiting betrayal.
Authority Subversion	Upholding social roles, respecting authority and traditions, valuing leadership, and prohibiting rebellion.
Purity (Sanctity) Degradation	Reverence for the sacred, purity, religious principles guiding life, and prohibitions against violating the sacred.

Table 8: Brief Descriptions of the Moral Foundations. Each foundation has two aspects representing positive and negative perspectives of that moral foundation branch. Please refer to Atari et al. (2023) for the most up-to-date list of moral foundations and their descriptions.