Pragmatic Inference for Moral Reasoning: Generalization via Metapragmatic Links

Anonymous ACL submission

Abstract

While moral reasoning has emerged as a promising research direction for large language models (LLMs), achieving robust generalization remains a critical challenge. This challenge arises from the gap between what is said and what is morally implied. In this paper, we build on metapragmatic links and the moral foundations theory to close the gap. Specifically, we develop a pragmatic-inference approach that facilitates LLMs, for a given moral situation, to acquire the metapragmantic links between moral reasoning objectives and the social variables that affect them. This approach is adapted to three different moral reasoning tasks to demonstrate its adaptability and generalizability. Experimental results demonstrate that our approach significantly enhances LLMs' generalization in moral reasoning, paving the road for future research to utilize pragmatic inference in various moral reasoning tasks. ***Warning****: this paper contains offensive language.*

1 Introduction

With the advancement of Large Language Models (LLMs), how to regulate their performance in accordance with human moral values, namely, develop the capability of moral reasoning, has attracted scholarly interest. Table 1 demonstrates an example of the several elements that are involved in moral reasoning: moral situation, rule of thumb (RoT), moral foundation, and moral judgment. Given a moral situation, **moral foundations** (Haidt and Joseph, 2004; Graham et al., 2013) encode the underlying moral values (e.g., care or harm) that the situation either upholds or violates. A **RoT** provides a situated explanation that guides the evaluation of the situation as right or wrong (e.g., murder is wrong) (Forbes et al., 2020). The **moral judgment** then follows by determining whether the situation is morally acceptable or unacceptable. In the current data, "agree" means the reply in the moral situation is acceptable, while "disagree" is the opposite. "Neutral" shows that the annotator neither agrees nor disagrees with the reply. Thus far, learning paradigms, which aim to equip LLMs with moral reasoning, generally train models to directly map a moral situation to moral judgment or RoT, with or without a reference to moral foundations (Hendrycks et al., 2020; Forbes et al., 2020; Jiang et al., 2021; Ziems et al., 2022; Jiang et al., 2025a).

Moral Situation: Brush your teeth with laundry detergent.
Moral Foundation: Care. **Rule of Thumb (RoT)**: It is harmful to brush teeth with laundry detergent. **Moral Judgment**: Disagree.

Table 1: A Moral Reasoning Example. The Moral Foundations, Rule of Thumb and the Moral Judgment are all applied to the given reply. The Rule of Thumb (RoT) represents a judgment regarding an action mentioned in the prompt-reply, grounded in the moral foundation of care and sanctity. The moral situation consists of both the prompt and the corresponding reply.

Direct mapping often requires a huge amount of data to enable LLMs to capture the statistical correlations between situations and moral reasoning objectives (Rao et al., 2023; Preniqi et al., 2024; Jiang et al., 2025a). This is mainly due to the pragmatic gap between the two that needs to be filled (Liu et al., 2025b; Hendrycks et al., 2020; Tennant et al., 2024), metaphorically speaking, the gap between what is *said* and what is morally *implied*. To close this gap, moral reasoning requires the capability of pragmatic inference, namely, the capability of deriving a moral conclusion from the social variables underlying the given moral situation, such as sociocultural context, behavioral consequences, and discursive rules of conduct (Spencer-Oatey and Xing, 2019). Yang et al. (2025) argue that LLMs have acquired knowledge of social variables, but need to

develop the ability to draw on them effectively and appropriately. The current study thus develops a pragmatic-inference approach that connects moral reasoning objectives to different social variables. In linguistic terms, we specify the *metapragmatic links* to facilitate LLMs' acquisition of moral reasoning (see Section 2 for more discussions about metapragmatics).

Prior studies in moral reasoning either overlook the pragmatic gap (Hendrycks et al., 2020; Tennant et al., 2024) or focus on benchmarking (Ren et al., 2024; Yao et al., 2024; Kumar and Jurgens, 2025) and measuring the moral values embedded in LLMs (Scherrer et al., 2023; Jiang et al., 2025b), with limited exploration of approaches for pragmatic inference. Nevertheless, we have observed the effectiveness of pragmatic inference-driven approaches for morality-related tasks, such as toxicity identification (Sap et al., 2020; Zhou et al., 2023; Chen and Wang, 2025). These approaches similarly textualized the pragmatic inferential steps in order to enhance LLMs' recognition of toxicity (Chen and Wang, 2025). However, moral reasoning introduces three challenges that go beyond the inference of toxicity.

First, unlike toxic language, e.g., "lowering employment standards for females", which directly implies women's incapability, making a moral reasoning decision (e.g., a moral judgment) requires a further association with abstract moral values (Kádár and Haugh, 2013), e.g., why implying women's incapability is wrong. Moral foundations usually serve the abstract value types (see Section 2 for details of moral foundations), which this study endeavours to build a metapragmatic link to. Second, while toxic language leads to moral judgments, immoral situations do not necessarily contain explicit cues of harm in the way toxic language does. Hence, moral reasoning cannot be simply based on the inferential approaches developed for toxicity detection. Third, and probably of the most important, existing benchmarks for moral reasoning rarely involves annotations of pragmatic inference. In this paper, we show that prompting questions (for off-the-shelf LLMs) designed based on our approach can be effectively applied to obtain the **training data** for pragmatic inference. Training on the data enables smaller models to acquire moral reasoning and achieve better performance, even compared to more advanced LLMs.

In this paper, we apply our pragmatic inference approach to three different moral reasoning tasks with different objectives, to demonstrate its adaptability and generalizability. The results shed light on a widely-applicable way for LLMs to identify moral issues from input and to regulate their own performance in accordance with human moral values.

2 Related Works

Pragmatic Inference and Metapragmatics. It is widely known that distributional semantics is one of the theoretical foundations for embeddings. It interprets the meaning of a linguistic unit by additional signs (e.g., words, emoji) surrounding it (Harris, 1954; Firth, 1957; McDonald and Ramscar, 2001; Lenci and Sahlgren, 2023; Boleda, 2020). In other words, it treats the "context" as the texts provided, neglecting the many "unsaid" social variables that underlie real interactions, such as the social constraints that a speaker encounters, the effects that a speech creates on the listeners, and the delivery of intention (Lenci et al., 2008; Purver and Sadrzadeh, 2015). These are studied in pragmatics – the linguistic discipline that studies language in context. Possibly due to the above-mentioned gap between "having internal representations of social variables" and "drawing effectively on them" (Yang et al., 2025), LLMs have shown several pragmatic deficiencies, e.g., understanding implicature (Cong, 2024), comprehending metaphors and humours (Ruis et al., 2023; Barattieri di SanPietro et al., 2023), and moral alignment (Khamassi et al., 2024; Liu et al., 2024).

In this work, we leverage pragmatic inference to facilitate LLMs to achieve stronger performance in moral reasoning. To reiterate, pragmatic inference refers to the process of deriving conclusions of meanings based on context, intention, and language use (Elder, 2024). The inferential process is, however, not often explicitly stated even in human interactions. It is rather common from human speakers to present their moral judgment without specifying how they make it. The inferences thus stay non-verbally at the level of metapragmatic awareness, where meta-links are built between an instance of language use and social variables (e.g., linking the phrase "thank you" to the social norm of politeness) (Kádár and Haugh, 2013). Humans learn these metapragmatic links not just from language input, but also through physical and social experiences. The latter is absent in LLMs. Therefore, for LLMs to learn, the inferential process

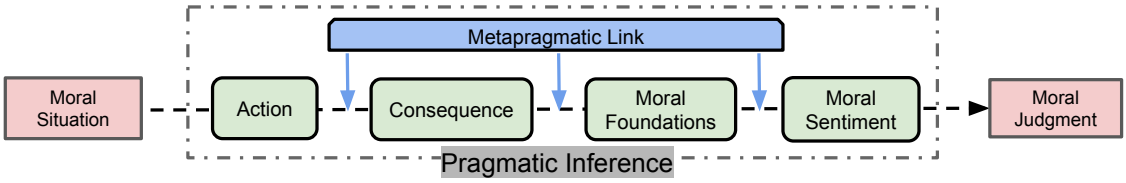

Figure 1: Our Pragmatic Inference Approach for Moral Reasoning. It consists of four main steps: (1) **Action**: identifying the actions involved in the moral situation; (2) **Consequence**: estimating the consequences of those actions; (3) **Moral Foundations**: predicting the relevant moral foundations associated with the actions and determining whether those actions violate or uphold the identified foundations; and (4) **Moral Sentiment**: identifying the sentiment of the situation with respect to the identified actions.

where metapragmatic links are built needs to be textualized, involving relevant social variables, such as context and moral foundations.

Moral Foundations Theory (MFT). Haidt and Joseph (2004) and Graham et al. (2013) describe a framework for understanding human morals and moral reasoning. MFT hypothesizes that moral judgments are guided by domain-specific foundations, i.e., care, fairness, loyalty, liberty, authority, and sanctity. These foundations capture the diversity of moral concerns and have been widely adopted in computational social science (Johnson and Goldwasser, 2018; Wendell and Tatalovich, 2021; Reinig et al., 2024). They provide a principled way to annotate, analyze, and model moral dimensions in language, making them particularly relevant for studying how LLMs represent and reason about morality.

3 Method

In this section, we first introduce our pragmatic inference approach and then describe how it can be adapted to different moral reasoning objectives, namely moral foundations classification (MFC) and moral judgment. Our pragmatic inference approach is also flexible with different types of input. In the current study, we tested moral situation or RoT as input[1].

3.1 Motivation

Proposing a pragmatic inference approach for moral reasoning is challenging because there is no linguistic theory that clearly defines what constitutes an appropriate pragmatic inference for LLMs (Bender and Koller, 2020; Mahowald et al., 2024). To enable LLMs to make effective moral judgments while drawing effectively on relevant social variables, the theoretical frameworks of how humans make pragmatic inferences and moral judgments are borrowed. However, this study does not assume their direct applicability to LLMs. Instead, this paper proposes a pragmatic inference approach after careful examinations with LLMs.

Figure 1 introduces the design and its rationales: The **Action** component is to identify mentioned actions in the given moral situation. According to the famous pragmatics theory of speech act, humans speak not for the purpose of speaking, but for doing things (Austin, 1975). Thus, the moral judgment is often, if not always, a judgment of the action that the speaker is expressing (e.g., "I want to destroy the world") or that others once performed (e.g., "who said he is an idiot") (Forbes et al., 2020; Ziems et al., 2022). We set "Action" as the first step, foregrounding the rest of steps.

The second step **Consequence** is to recognize the consequence of carrying out the "Action". Moral judgment is, in nature, an evaluation of good/bad placed on the outcome of an action (Nozick, 1968; Schaich Borg et al., 2006; Emelin et al., 2021).

The third step **Moral Foundations** (MFs) then moves from the individual action/consequence to the common ground that underlies them. The effectiveness of including MFs in moral reasoning has been supported by previous theories and experiments (Graham et al., 2013; Abdulhai et al., 2024).

Given the results of Moral Foundations step (e.g., an action violates a moral foundation), the last step **Moral Sentiment** is necessary to unravel the relationship between actions, consequences, and the moral situation. For example, if an action is moral and the situation shows a positive sentiment about it, the situation is likely moral. The importance of moral sentiment in causally shaping moral judgment has been shed light on by the Social Intuition-

[1]We note that pragmatic inference often admits multiple pathways to the same conclusion, and our approach represents one such pathway.

Input:
There is a rule of thumb (RoT): {RoT}.
There are the six moral foundations and their DEFINITIONs are: {DEFINITIONs}.
Inference Prompts:

1. The rule of thumb {RoT} includes a Judgment to an Action, identify the Judgment and the associated Action.
2. What would be the consequence if the Action is carried out?
3. Why is the consequence of the Action relevant to the moral foundations {moral_foundations} by referring to the provided DEFINITIONs of those moral foundations.

Figure 2: Prompt Questions For the Task of Moral Foundations Classification. All content within { } is provided by the benchmark. We adopt the MIC benchmark (Ziems et al., 2022), which consists of prompt–reply pairs, where each pair is treated as a moral situation and the moral reasoning objectives are defined with respect to the reply.

ist Model (Haidt, 2001; Zollo, 2021).

These steps are connected via metapragmatic links. For example, the connection between an action and its consequence requires the metapragmatic evaluation of social norms, e.g., the connection between the action of "encouraging war" and the consequence of "civilian casualty" is built on the oft-unspoken inference about how wars cause mass death. Similarly, attributing the consequence of an action to the abstract MFs demands the metapragmatic evaluation of how MFs guide and are also reflected in human behaviours. **Through reifying the metapragmatic links between different inferential components, this study teaches LLMs to effectively conduct moral reasoning**.

In the following subsections, we apply different parts of the framework to three moral reasoning tasks that have different objectives, to show its adaptability. Specifically, we first designed prompting questions according to the framework and used them to obtain training data from off-the-shelf LLMs. The training data is then used to train several small LLMs.

3.2 Task 1: Moral Foundations Classification (MFC)

The first task demonstrates how to identify the moral foundations from a given RoT. It requires only a light load of pragmatic inference, because RoT is a situated explanation for the good or bad nature of a moral situation, in other words, it already has a judgment on the given situation, e.g., "murder is wrong". Figure 2 presents the prompting questions of the inferential steps used in the MFC task. The input here is a simple RoT and the definitions of six MFs. The inferential steps are merely the first three components of our pragmatic inference approach (Figure 1), namely, action, consequence, and MFs.

3.3 Task 2: Moral Judgment

Figure 3 shows the prompting questions used in the second task. Unlike the first MFC task, where RoTs are presented without context, the input here is a moral situation consisting of a prompt–reply pair along with the associated ground-truth MFs. In other words, we provide the specific context (i.e., prompt and reply) and test whether LLMs can effectively make context-specific moral judgments based on MFs.

The second task include all components of our framework, namely, action, consequence, MFs, and moral sentiment, given the increasing complexity of its pragmatic inference. Especially, the MFs component is divided into two steps: step 3 relates action to the ground-truth MFs, while step 4 further specifies whether the relation is a violation or observation of the MFs[2]. We should note that, during the training, judgment is provided for LLMs to learn its connections with the other components (step 6). In the actual test, LLMs are required to produce the judgment by themselves.

3.4 Task 3: MFC+Moral judgment

Building on the first two tasks, the third task teaches LLMs to jointly infer MFs from a given *situation* and make corresponding moral judgments, namely, MFC + moral judgment. For the prompting questions, we drop the given ground-truth moral foundations in the input and replace the step 3 and 4 in Figure 3 with the following prompt questions:

[2]Please note that the empirical results indicate that applying Step 3 or Step 4 individually performs better than combining them. This is reasonable, as each step constructs a metapragmatic link, and applying the steps individually makes this link easier to establish.

Input:
There is a Prompt-Reply pair: {prompt} {reply}.
The moral foundations underlying this Prompt-Reply are {moral_foundations}.
There are the six moral foundations and their DEFINITIONs are: {DEFINITIONs}.
Inference Prompts:

1. What Actions are mentioned in the Reply?
2. What are the potential consequences of those Actions mentioned in the Reply?
3. Based on the consequences of these Actions, which actions correspond to the moral foundations {moral_foundations} according to the given DEFINITIONs?
4. How do Actions associated with specific moral foundations ({moral_foundations}) up-regulate or down-regulate those foundations based on their consequences?
5. What are the sentiment of the Reply to those consequences, positive, negative or neutral?
6. The moral judgment of the Reply is {judgment}. Given your previous answers, explain why the judgment is {judgment} by referring to the Reply's sentiment to the consequences of those Actions.

Figure 3: Prompting Questions for the Task of Moral Judgment Based on Ground-truth Moral Foundations. All content within { } is provided by the benchmark. We adopt the MIC benchmark (Ziems et al., 2022), which consists of prompt–reply pairs, where each pair is treated as a moral situation and the moral reasoning objectives are defined with respect to the reply.

(3). Based on the consequences of these Actions, why are their underlying moral foundations classified as {moral_foundations} according to the provided DEFINITIONs?
(4). How do Actions up-regulate or down-regulate their underlying moral foundations based on their consequences?

In doing so, *we guide LLMs to learn to infer the underlying moral foundations*, motivated by their strong performance on the MFC task (Figure 2). The input is kept to the situation (Prompt+Reply) and the definitions of MFs only, no ground-truth MFs are provided. This setting is closer to the real life where a speaker only hears the interaction and has some (implicit) knowledge of MFs.

4 Experimental Setting and Results

In this section, we present the experimental settings and results to evaluate the effectiveness of our pragmatic inference approach for moral reasoning. Further analyses of our approach are provided in Appendix A.2.

4.1 Experimental Settings

Benchmark. In this paper, we adopt the Moral Integrity Corpus (MIC) benchmark and its annotations from Ziems et al. (2022). We chose the MIC benchmark because it closely reflects how humans make moral judgments in daily communication and provides richer contextual information. One advantage of the MIC dataset is that it provides levels of agreement between the annotators, which allows us to focus on the less controversial subset of the benchmark. Indeed, we adopt 26% of the samples that have received full agreement among all annotators, given that moral judgments are subjective and accounting for subjectivity is beyond the current research scope. The MIC benchmark utilizes the six MFs defined by Graham et al. (2013): care, fairness, liberty, loyalty, authority, and sanctity. In the data, up to three MFs could be associated with one moral situation. As presented above, we utilize two of the evaluations as our objectives: MFs and moral judgment (agree, neutral, or disagree with the reply). RoT is not selected as a learning objective because: (i) it does not have a fixed answer and, more crucially, multiple RoTs may exist in one situation[3], and (ii) it has a rather dynamic relationship with MFs.

MFC				Moral Judgment
#MFs=1	#MFs=2	#MFs=3	Average	
.694	.274	.147	.579	.466

Table 2: Performance of Direct Prompting Off-the-shelf LLM (deepseek-chat). The prompts used in these evaluations are provided in Appendix 5.

To examine how moral reasoning performance scales with the size of the training corpus, we consider three data settings. We use subsets of 5K, 10K, and all of the 23500 samples of the full-agreement data from the training set as the fine-tuning data. We report all performance results using a fixed development set, with models trained on different subsets of the full-agreement training data. Therefore we can verify how the training data scales can influence the moral reasoning per-

[3]We refer readers to Section 3.1 of Ziems et al. (2022) for further details.

formance. To ensure the selected dataset from this 2022 study is still appropriate for the current study, we first prompt the deepseek-chat model directly. Table 2 presents its performance on both moral foundation prediction and moral judgments. They clearly show that moral reasoning remains a challenging task even for advanced LLMs.

Therefore, following the steps of our pragmatic inference methods (detailed in Section 3), we prompt the DeepSeek model to provide responses to each of the inferential components as well as their metapragmatic links. The step-wise answers are then used as the training data[4] for fine-tuning the LLMs towards moral reasoning objectives.

Backbone Models. We chose Llama3-3.2-1B, Llama3-3.2-3B (Grattafiori et al., 2024) and Mistral-7B (Mistral-7B-v0.3 (Jiang et al., 2023)) as our backbone models, and leverage the supervised fine-tuning objective to train those models by following previous studies (Ziems et al., 2022; Liu et al., 2025b). We intentionally select smaller models, as they are generally considered less capable than their larger counterparts. We aim to demonstrate the effectiveness of our pragmatic inference approach through the strong performance achieved by these smaller models.

Evaluation. To reiterate, we have three moral reasoning tasks. Prediction accuracy is used as the evaluation metric across them. In the meantime, the prediction accuracy of MFs is further divided by the data that refers to one, two, or three MFs, respectively. In other words, we have taken into consideration the one-to-many relationship between RoT/situation and MFs.

Method Setting and Baseline Methods. We denote the LLM f is parameterized by θ, and the definitions (see Figure 4) of MFs are represented with d_m. Let (x_S, x_R, y_M, y_J) denote a sample from the MIC benchmark, where x_S, x_R represents the moral situation and RoT, respectively. y_M is its associated MF label. Each y_M may contain between one and three MFs. y_J is the moral judgment. We use $|y_M| = i$ $(1 \leq i \leq 6, x \in \mathbb{Z})$ to indicate that x is associated with i MFs. We compute accuracy for each value of i.

For **baseline** methods, we follow prior work (Forbes et al., 2020; Ziems et al., 2022) and consider two settings: **(1)** The **base** setting. For the moral judgment tasks, we compute $y_J = f_\theta(x_S)$; and for MFC, we compute $y_M = f_\theta(x_R)$. **(2)** The **base+** setting. For the moral judgment task, we compute $y_J = f_\theta(x_S, y_M)$ wherein the ground-truth MFs are provided; and for MFC task, we compute $y_M = f_\theta(x_R, d_m)$ wherein only the definition of six MFs are provided. **(3)** The **CoT** setting. We also consider a strong baseline based on the standard Chain-of-Thought (CoT) method (Wei et al., 2022). CoT assists LLMs to divide a complex task into intermediate steps and has been proven to be effective in logical reasoning tasks (Fang et al., 2025; Ji et al., 2025). However, it has also been identified as less effective in non-logical reasoning tasks, such as pragmatic inference (Sprague et al., 2024; Chen and Wang, 2025). To ensure that the effectiveness of CoT is not overlooked, we apply the same training process, namely, obtaining training data by using CoT to prompt the same off-the-shelf LLMs[5] and then use the step-wise answers to train the models. We compute $y_J = f_\theta(x_S, \{\text{CoT_inference}\})$ and $y_M = f_\theta(x_R, \{\text{CoT_inference}\})$ for moral judgment and MFC, respectively.

For our pragmatic inference approach, we compute $y_M = f_\theta(x_R, d_m, \{\text{ours_inference}\})$ for Task 1, denoted as **ours-MFC**. For Tasks 2 and 3, we denote $y_J = f_\theta(x_S, y_M, \{\text{ours_inference}\})$ as **ours-MJ**, where moral judgment (MJ) is the objective; and we denote $y_J, y_M = f_\theta(x_S, \{\text{ours_inference}\})$ as **ours-joint**, where the LLMs are asked to predict both MFs and the moral judgment. More details about hyperparameters and experimental setup are in Appendix A.1.

4.2 Experimental Results

Task 1: MFC. Table 3 reports the models' performance on the development set of data that have 1–3 moral foundation labels. Due to the space limit, we report the average performance for each experimental setting in Table 4. Across all training dataset sizes, model variants, and numbers of moral foundation labels, our method consistently outperforms the baseline methods and the CoT method. Comparing the results of **base** and **base+**, it is evident that incorporating the definition of MFs is beneficial, which aligns with findings from previous studies (Ziems et al., 2022; Liu et al., 2025b). Our method also consistently outperforms the **CoT** method, which is, at times, worse than the **base+**

[4]The data used in our experiments is available in the anonymized GitHub repository: https://anonymous.4open.science/r/moralpragmatics-F891.

[5]Please refer to Appendix 6 for the details of the prompting questions.

Model	Data Scale	Accuracy(#MFs=1)				Accuracy(#MFs=2)				Accuracy(#MFs=3)			
		base	base+	CoT	ours	base	base+	CoT	ours	base	base+	CoT	ours
Llama-1B	5000	.501	.696	.661	**.890**	.402	.545	.475	**.856**	.396	.460	.428	**.806**
	10000	.582	.593	.677	**.832**	.489	.449	.545	**.778**	.414	.365	.468	**.743**
	235000	.552	.671	.717	**.770**	.505	.514	.509	**.796**	.387	.495	.401	**.743**
Llama-3B	5000	.537	.754	.719	**.851**	.466	.653	.544	**.822**	.423	.572	.459	**.798**
	10000	.386	.558	.700	**.725**	.336	.507	.507	**.649**	.324	.451	.441	**.698**
	235000	.653	.790	.727	**.839**	.493	.645	.533	**.784**	.451	.617	.414	**.743**
Mistral-7B	5000	.612	.649	.677	**.803**	.474	.504	.533	**.634**	.441	.469	.401	**.577**
	10000	.456	.653	.653	**.754**	.373	.466	.515	**.606**	.401	.419	.523	**.536**
	235000	.487	.669	.649	**.768**	.442	.536	.420	**.604**	.383	.446	.383	**.559**

Table 3: **Moral Foundations Classification (MFC)** Performance Across Different Training Dataset Scales. The majority of prompt-reply pairs involve no more than three moral foundations labels; therefore, we report the performance on cases containing fewer than four moral foundations. For example, **#MFs=2** indicates performance on instances annotated with a label containing two moral foundations. The optimal performance is highlighted with a **bold** font.

Model	Data Scale	base	base+	CoT	ours -MFC
Llama 1B	5000	.433	.567	.521	**.851**
	10000	.495	.469	.564	**.784**
	235000	.481	.560	.542	**.770**
Llama 3B	5000	.475	.659	.574	**.824**
	10000	.349	.505	.549	**.691**
	235000	.532	.684	.558	**.789**
Mistral 7B	5000	.509	.540	.537	**.671**
	10000	.410	.513	.563	**.632**
	235000	.437	.550	.484	**.643**

Table 4: **Average Moral Foundations Classification (MFC)** Performance Across Different Training Dataset Scales. Our proposed pragmatic inference methods outperform all baseline methods and the CoT method.

Model	Size	base	base+	CoT	ours-MJ	ours-joint
Llama -1B	5000	.151	.080	.629	**.708**	.706
	10000	.498	.335	.652	**.728**	.718
	23500	.179	.318	.641	**.710**	.704
Llama -3B	5000	.160	.304	.635	.712	**.750**
	10000	.205	.252	.653	.735	**.748**
	23500	.249	.390	.665	.710	**.766**
Mistral -7B	5000	.418	.366	.637	**.704**	.694
	10000	.438	.434	.628	**.760**	720
	23500	.486	.358	.644	.696	**.704**

Table 5: Accuracy of the Moral Judgment Task Across Training Dataset and Model Scales. The optimal performance is highlighted with a **bold** font. Our proposed pragmatic inference approach, i.e., **ours-MFs** and **ours-joint**, outperform all baseline methods and the CoT method. Although **ours-joint** infers the MFs, its performance is comparable to **ours-MFs**, which has access to the ground-truth MFs.

method. This finding corroborates previous observations of the inefficiency of CoT in pragmatic inference (Chen and Wang, 2025; Sprague et al., 2024). Another noteworthy finding is that, unlike the baseline methods whose performance drops significantly as the number of underlying MFs increases, our method exhibits much smaller performance degradation. Interestingly, neither increasing the model size nor increasing the training dataset leads to stable performance gains. The possible reasons include that the additional data does not substantially expand the range of unique moral situations, and the same size of training data is insufficient to modify the knowledge that larger LLMs have already acquired during pre-training compared to smaller models.

Tasks 2 and 3: Moral Judgment. Table 5 presents the performance on the moral judgment tasks. Again, our pragmatic inference methods (**ours-MJ** and **ours-joint**) consistently outperform both baseline methods (**base, base+**) and the **CoT** method across different training dataset scales and model architectures. Similar to the MFC task, increasing data size does not necessarily improve the model performance. Since moral judgment is a three-class classification task (agree, disagree, or neutral), where random guessing yields an accuracy of approximately 0.333, the baseline methods perform worse than random guessing in some settings

of dataset and model sizes. In contrast, baseline methods perform much better than random guessing on the MFC task. The difference corroborates that the pragmatic gap varies in size between the two tasks. This is understandable as the involvement of context (prompt+reply) in the second task brings about more variability in interpretations than the close alignment of RoT and normative interpretations, for example, the RoT "*It is good to keep in touch with friends*" can rather be easily attributed to the relational norms between friends. It is, however, worth emphasizing that our methods have been consistently effective across the different tasks, showing the adaptability and stability of our pragmatic inference approach.

Noteably, CoT in tasks 2 and 3, although still falling behind our methods, outperforms the baselines. The finding suggests that the baseline designs are particularly weak in addressing complex tasks, which the step-wise nature of CoT provides an advantage. However, CoT as a trick designed primarily for logical reasoning tasks still has a rather restricted performance in pragmatic inference tasks.

Recall that the difference between Task 2 and Task 3 is that Task 2 has the input of ground-truth MFs, whereas Task 3 needs to predict both MFs and moral judgment. Possibly with the involvement of ground-truth MFs, **ours-MJ** outperforms **ours-joint** in two of the three models. This does not deny the effectiveness of the design of **ours-joint**. In fact, it outperforms **ours-MJ** with Llama-3B, shedding light on the potential of LLMs automatically inferring both underlying MFs and making moral judgments.

5 Discussion

Different from existing approaches that train models to map moral situations to moral judgment, (Hendrycks et al., 2020; Forbes et al., 2020; Jiang et al., 2021; Ziems et al., 2022; Jiang et al., 2025a), this paper develops a pragmatic inference approach that bridges the two. The bridge is built to close the pragmatic gap between what is said and what is morally implied. It did so by textualizing the metapragmatic links between the designed components in the inferential process. In other words, the approach leverages the "internal representations of social variables" (Yang et al., 2025) and facilitates LLMs to learn to draw on them. It is in line with previous studies that develop other pragmatic inferential methods for toxic language detection tasks (Sap et al., 2020; Zhou et al., 2023; Chen and Wang, 2025).

However, it further illuminates the need for metapragmatic links to be incorporated into LLM training. This finding provides another explanation for the effectiveness of step-wise design, such as CoT, in improving LLM performance. In other words, what the step-wise responses present to LLMs may be the meta-links (not necessarily metapragmatic links) that have "internal representations" (Yang et al., 2025) but need LLMs to learn how to draw on. The finding also echoes previous arguments about the importance of annotating reasoning process, instead of the outcomes only (Wolfson et al., 2026; Gao et al., 2022).

Moreover, the performance increase that we find across three different tasks and different models suggests a potential generalizability of the current approach in different pragmatic tasks such as the understanding of metaphors, humor, and sarcasm, which LLMs have thus far been found deficient in (Barattieri di SanPietro et al., 2023; Ruis et al., 2023). For future work, this paper can be extended in two directions: (1) incorporating moral foundation inference, and (2) exploring more practical scenarios involving metaphors and referential expressions in text.

6 Conclusion

In this paper, we develop a pragmatic inference approach to enhance generalization in moral reasoning by addressing the gap between LLMs' knowledge of social variables and their capability of drawing on them. The approach embeds moral foundations as a component of the pragmatic inference, while textualizing metapragmatic links between different inferential components for LLMs to learn. We adapt the approach to three different moral reasoning objectives. The experimental results and analyses demonstrate that the proposed pragmatic inference approach significantly improves moral reasoning performance, reduces perplexity compared to baseline approaches, and avoids reliance on shallow heuristics.

Limitations

While the steps in the current pragmatic inference chains have been highly effective, this paper has not tested the performance of each step due to the limited space. We thus encourage future studies to conduct further ablation experiments, as well as

explore their variations. These future attempts may also extend the current experimental settings to other downstream morality-relevant tasks, such as jailbreaks. It is anticipated that the generalisability of our design will be further confirmed across the tasks that have different semantic distributions.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752.

John Langshaw Austin. 1975. *How to Do Things with Words*. Harvard University Press.

Chiara Barattieri di SanPietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent.

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, Online. Association for Computational Linguistics.

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

Xi Chen and Shuo Wang. 2025. Pragmatic inference chain (pic) improving llms' reasoning of authentic implicit toxic language. *arXiv preprint arXiv:2503.01539*.

Yan Cong. 2024. Manner implicatures in large language models. *Scientific Reports*, 14(1):29113. Publisher: Nature Publishing Group.

Chi-Hé Elder. 2024. *Pragmatic Inference: Misunderstandings, Accountability, Deniability*. Cambridge University Press. Google-Books-ID: okn8EAAAQBAJ.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

Yuanheng Fang, Guoqing Chao, Wenqiang Lei, Shaobo Li, and Dianhui Chu. 2025. CDW-CoT: Clustered Distance-Weighted Chain-of-Thoughts Reasoning. *arXiv preprint*. ArXiv:2501.12226 [cs].

JR Firth. 1957. A synopsis of linguistic theory. *Studies in linguistic analysis, Special volume of the philological society/Blackwell*.

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.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022. Hierarchical Annotation for Building A Suite of Clinical Natural Language Processing Tasks: Progress Note Understanding. *LREC ... International Conference on Language Resources & Evaluation : [proceedings]. International Conference on Language Resources & Evaluation*, 2022:5484–5493.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. 2013. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jonathan Haidt. 2001. The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, 108(4):814.

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. *Word*, 10(2-3):146–162.

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*.

Shihao Ji, Zihui Song, Fucheng Zhong, Jisen Jia, Zhaobo Wu, Zheyi Cao, and Tianhao Xu. 2025. MyGO Multiplex CoT: A Method for Self-Reflection in Large Language Models via Double Chain of Thought Thinking. *arXiv preprint*. ArXiv:2501.13117 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, and 1 others. 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, and 1 others. 2025a. Investigating machine moral judgement through the delphi experiment. *Nature Machine Intelligence*, pages 1–16.

Liwei Jiang, Taylor Sorensen, Sydney Levine, and Yejin Choi. 2025b. Can language models reason about individualistic human values and preferences? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6757–6794.

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.

Mehdi Khamassi, Marceau Nahon, and Raja Chatila. 2024. Strong and weak alignment of large language models with human values. *Scientific Reports*, 14(1):19399.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Shivani Kumar and David Jurgens. 2025. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with UniMoral. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5890–5912, Vienna, Austria. Association for Computational Linguistics.

Dániel Z. Kádár and Michael Haugh. 2013. *Understanding Politeness*. Cambridge University Press, Cambridge.

Alessandro Lenci and Magnus Sahlgren. 2023. *Distributional semantics*. Cambridge University Press.

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

Guangliang Liu, Milad Afshari, Xitong Zhang, Zhiyu Xue, Avrajit Ghosh, Bidhan Bashyal, Rongrong Wang, and Kristen Johnson. 2024. Towards understanding task-agnostic debiasing through the lenses of intrinsic bias and forgetfulness. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1843–1856.

Guangliang Liu, Bocheng Chen, Han Zi, Xitong Zhang, and Kristen Marie Johnson. 2025a. Diagnosing the performance trade-off in moral alignment: A case study on gender stereotypes. *Preprint*, arXiv:2509.21456.

Guangliang Liu, Lei Jiang, Xitong Zhang, and Kristen Marie Johnson. 2025b. Diagnosing moral reasoning acquisition in language models: Pragmatics and generalization. *arXiv preprint arXiv:2502.16600*.

Guangliang Liu, Zimo Qi, Xitong Zhang, and Kristen Marie Johnson. 2025c. Discourse heuristics for paradoxically moral self-correction. *arXiv preprint arXiv:2507.00985*.

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*, 28(6):517–540.

Scott McDonald and Michael Ramscar. 2001. Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 23.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Robert Nozick. 1968. Moral complications and moral structures. *Nat. LF*, 13:1.

Vjosa Preniqi, Iacopo Ghinassi, Kyriaki Kalimeri, and Charalampos Saitis. 2024. Moralbert: Detecting moral values in social discourse. *arXiv e-prints*, pages arXiv–2403.

Matthew Purver and Mehrnoosh Sadrzadeh. 2015. From distributional semantics to distributional pragmatics. In *Proceedings of the IWCS 2015 Workshop on Interactive Meaning Construction*, pages 21–22.

Abhinav Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388.

Ines Reinig, Maria Becker, Ines Rehbein, and Simone Paolo Ponzetto. 2024. A survey on modelling morality for text analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4136–4155.

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.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs. *arXiv preprint*. ArXiv:2210.14986 [cs].

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Metaphor and large language models: When surface features matter more than deep understanding. *arXiv preprint arXiv:2507.15357*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Jana Schaich Borg, Catherine Hynes, John Van Horn, Scott Grafton, and Walter Sinnott-Armstrong. 2006. Consequences, action, and intention as factors in moral judgments: An fmri investigation. *Journal of cognitive neuroscience*, 18(5):803–817.

Nino Scherrer, Claudia Shi, Amir Feder, and David Blei. 2023. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273.

Helen Spencer-Oatey and Jianyu Xing. 2019. Interdisciplinary perspectives on interpersonal relations and the evaluation process: Culture, norms, and the moral order. *Journal of Pragmatics*, 151:141–154.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint*. ArXiv:2409.12183 [cs].

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Dane G Wendell and Raymond Tatalovich. 2021. Classifying public policies with moral foundations theory. *Policy Sciences*, 54(1):155–182.

Tomer Wolfson, Harsh Trivedi, Mor Geva, Yoav Goldberg, Dan Roth, Tushar Khot, Ashish Sabharwal, and Reut Tsarfaty. 2026. MoNaCo: More Natural and Complex Questions for Reasoning Across Dozens of Documents. *Transactions of the Association for Computational Linguistics*, 14:23–46.

Diyi Yang, Dirk Hovy, David Jurgens, and Barbara Plank. 2025. Socially Aware Language Technologies: Perspectives and Practices. *Computational Linguistics*, 51(2):689–703. _eprint: https://direct.mit.edu/coli/article-pdf/51/2/689/2505992/coli_a_00556.pdf.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773.

Lamberto Zollo. 2021. The consumers' emotional dog learns to persuade its rational tail: Toward a social intuitionist framework of ethical consumption. *Journal of Business Ethics*, 168(2):295–313.

A Appendix

A.1 Experimental Details

We apply supervised fine-tuning with a learning rate of 5e-5 and a batch size of 24. For mistral-7B, we leverage the LoRA setting with rank: 64, lora alpha: 16, lora dropout: 0.1. All experiments are conducted for 5 epochs using three random seeds (1, 2, and 3), and we report the optimal result across three random seeds. **Template for fine-tuning LLMs** is: **base**: The rule-of-thumb judgment is {RoT}. The moral foundations underlying the rule-of-thumb are {moral foundations}. **base+**: {Definition of moral foundations} The rule-of-thumb judgment is {RoT}. The moral foundations underlying the rule-of-thumb are {moral foundations}. **ours**: {Definition of moral foundations} The rule-of-thumb

judgment is {RoT}. ###Inference: {inference} the moral foundations underlying the rule-of-thumb are {moral foundations}.

Care: wanting someone or something to be safe, healthy, happy.
Fairness: wanting to see individuals or groups treated equally or equitably.
Liberty: wanting people to be free to make their own decisions.
Loyalty: wanting unity and seeing people keep promises or obligations to an in-group.
Authority: wanting to respect social roles, duties, privacy, peace, and order.
Sanctity: wanting people and things to be clean, pure, innocent, and holy.

Figure 4: Definition of Six Moral Foundations

There are six moral foundation: {DEFINITIONS}
The rule-of-thumb judgment is "{RoT}".

Which moral foundation or foundations are underlying the rule-of-thumb?

There is a question-answer interaction "Question: {question} Reply: {answer}"
What is the moral judgment of the Reply? Please answer with agree, neutral, or disagree.

Figure 5: Prompts for Direct LLMs Prompting

There are six moral foundation: {DEFINITIONS}
There is a rule-of-thumb (RoT): {RoT}
Why the moral foundation of this RoT is {mft}?
Please think step-by-step.

There is a Conversation Prompt: "Question: {question} Reply: {answer}"
Why the moral judgment of this Reply is {judgment}?
Please think step-by-step.

Figure 6: Prompts for Standard CoT Prompting

A.2 Further Analysis

In the previous sections, we demonstrated the effectiveness of our proposed pragmatic inference approach. In this section, we present additional analyses to further characterize it. Specifically, we: (1) conduct intervention experiments to examine whether our approach guides LLMs to avoid heuristics and utilizes MFs when making moral judgments; (2) calculate the perplexity to show that our approach makes moral reasoning discourse more learnable than baseline methods that overlook the pragmatic gap; and (3) perform a detailed analysis of individual MFs, showing that while performance on the majority of MFs is strong, dataset imbalance remains a key bottleneck to achieving good performance on rare MFs. *Notably, for all analytical experiments on the moral judgment task, we employ the design of Task 3, namely, jointly infering moral foundations and moral judgments.*

Dependence on Moral Foundations. Prior studies show that LLMs may not pragmatically leverage the cues of MFs or other morality-related explanations. Instead, they rely on superficial correlations to achieve generalization in pragmatics-level tasks (Shapira et al., 2024; Liu et al., 2025c; Sanchez-Bayona and Agerri, 2025). Just as Liu et al. (2025c) demonstrates, LLMs correct immoral outputs to moral ones, while neglecting the underlying moral issues provided to them. This gives rise to the issue of generalizability across different datasets. To test whether the same problem occurs in the current study, we conduct an intervention experiment to examine whether our pragmatic inference method has facilitated LLMs' effective use of MFs.

Specifically, we replace the automatically inferred MFs with the ground-truth MFs and observe how the performance changes. Note that this design is different from Task 2, because the ground truth MFs are not an initial input here. Instead, we let the LLMs complete their inference process with the predicted MFs, and then replace them with the ground truth MFs only in the answer. After that, LLMs are led to output their moral judgment. In other words, we pause the task right before the last step and do the replacement, then we check whether the 'airdropped' ground-truth MFs lead to better performance of LLMs, with everything else in the inference remaining the same.

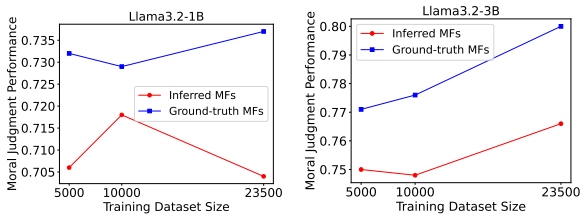

Figure 7: Accuracy of the Moral Judgment Task With Our Intervention Experiment.

Figure 7 exhibits that the replacement with ground-truth MFs consistently improves performance compared to the experiment without this intervention. In other words, the current pragmatic inference method has indeed facilitated LLMs to make effective use of MFs. Given MFs are commonly grounded in various moral situations, the finding suggests the potential for the current method to be generalized to other morality-related

tasks.

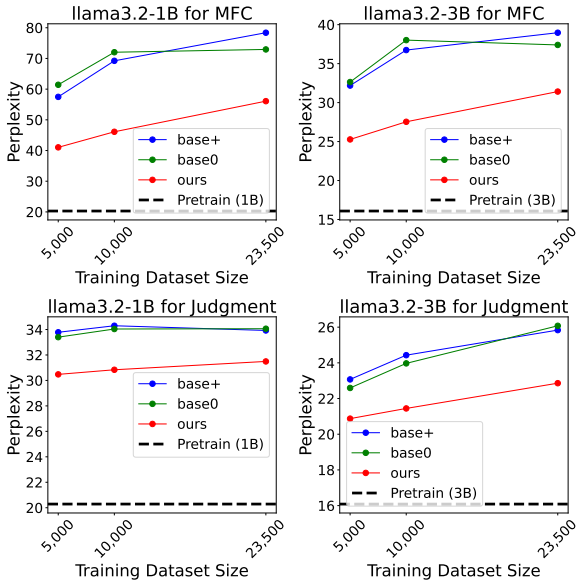

Figure 8: Perplexity of Fine-tuned Models Under Different Moral Reasoning Approaches on the Moral Foundations Classification (MFC) and Moral Judgment Tasks. Our proposed methods result in a much smaller perplexity than baseline methods, suggesting a better language modeling capability.

Perplexity. Although fine-tuning can introduce forgetting in LLMs (Kirkpatrick et al., 2017; Liu et al., 2024) and degrade their general capabilities, Liu et al. (2025b) emphasize that fine-tuning for moral reasoning, when not accounting for the pragmatic nature of morality, tends to yield higher perplexity. This suggests that the fine-tuning discourses are not semantically sufficient for LLMs to learn. This is also verified in other morality-relevant tasks such as social stereotype mitigation (Liu et al., 2025a). To further demonstrate that **our pragmatic inference method closes the pragmatic gap and makes the fine-tuning discourse for moral reasoning more learnable for LLMs**, we calculate the Perplexity of the optimal fine-tuned checkpoints of baseline approaches and our proposed approach. We use the WikiText dataset (Merity et al., 2016) to compute perplexity. Figure 8 reports perplexity results across different training dataset sizes and tasks. Models fine-tuned with our pragmatic inference methods consistently exhibit substantially lower perplexity compared to baseline approaches.

Moral Foundation-wise Analysis. Since MFs play an essential role in our pragmatic inference approach, we wonder whether and how the different

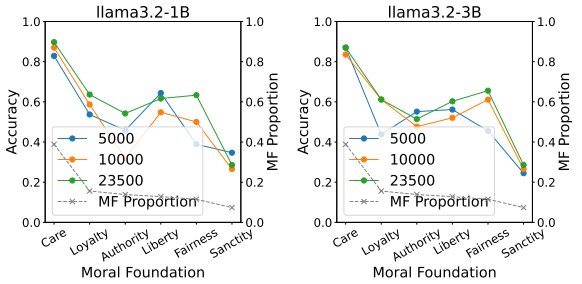

Figure 9: Performance of Moral Judgment Across Individual MFs. We report results only on test situations involving a single underlying moral foundation for more precise analysis. The black dashed line indicates the proportion of cases corresponding to each moral foundation in the training dataset. The X-axis orders the MFs in ascending order of their proportions.

distribution of them may affect the performance of models in the moral judgment task. Figure 9 illustrates MFs-wise performance, highlighting a correlation between the number of situations of each MF in the training dataset and the eventual accuracy of moral judgment. Notably, care accounts for about 40% of the training data, leading to higher model performance in test situations involving this foundation. In contrast, performance is much lower for test situations involving the *sanctity* foundation. These empirical results suggest a space for our methods to further improve model performance, if given sufficient and balanced annotation data.