Perspective as a Moral Probe: Revealing Judgment Styles of Large Language Models

Anonymous ACL submission

Abstract

Existing research on LLM alignment and moral reasoning typically evaluates models by presenting ethical scenarios and comparing outputs against predefined human value standards—implicitly assuming that models maintain consistent moral judgments across varying presentations of the same underlying event. We argue, however, that LLMs' apparent moral stances reflect conditioned generation tendencies shaped by training data, preference optimization, and alignment strategies—not stable value beliefs. As a result, narrative framing alone can induce systematic shifts in model outputs even when the underlying moral event is unchanged. We propose narrative perspective as a probing mechanism for moral stance stability. First- versus third-person shifts preserve event semantics while minimally varying cognitive positioning and responsibility attribution, enabling controlled analysis of moral consistency. We construct a dual-perspective counterfactual benchmark pairing narratives of identical moral events across perspectives, and systematically examine models' judgment tendencies, consistency, and hedging behavior under viewpoint shift. Our experiments reveal that perspective-induced moral judgment shifts in LLMs are systematic rather than random, that different models exhibit stable and distinguishable judgment-style fingerprints. These findings demonstrate that current LLMs' moral judgments are not fully viewpoint-invariant, and narrative perspective shift serves as an effective low-cost probe for analyzing latent moral reasoning patterns and judgment stability in LLMs. The relevant data and code can be accessed in https://anonymous.4open.science/r/ethic-8E2E/.

1 Introduction

Moral values are the fundamental building blocks of social organization, guiding how individuals navigate dilemmas and interpret the world through diverse cultural lenses (Graham et al., 2013). As Large Language Models (LLMs) are increasingly deployed in sensitive roles such as legal consultation, psychological support, and ethical advisory (Jiao et al., 2025; Schneiders et al., 2025), ensuring their alignment with human normative values has moved from a technical goal to a critical safety requirement. However, existing alignment efforts often face a "black box" challenge: although a model may produce a normatively compliant response, its internal decision-making logic and the stability of its underlying moral stance across diverse contexts are often difficult to verify.

To explore this internal logic, recent research has shifted toward probing-based explainability. Rather than relying on simple accuracy, new methodologies trace moral reasoning trajectories to understand how models navigate different ethical frameworks during inference (Huang et al., 2026). Others have utilized multi-step induction to uncover latent value priorities, forcing models to reveal their hierarchical preferences when faced with complex dilemmas (Wu et al., 2025). Despite these advances in probing internal states and value hierarchies, a crucial dimension of moral perception has been largely ignored: the systematic influence of narrative perspective.

In human moral psychology, the actor-observer asymmetry is a defining characteristic of judgment; individuals often view their own actions through a lens of contextual leniency while judging others with character-based severity (Mikhail, 2007). Standard morality benchmarks for LLMs, however, typically adopt the premise of perspective-invariance, which posits that a moral violation should be judged identically regardless of the narrative framing (Ji et al., 2023). Yet, there is growing evidence that narrative structure itself can bias moral content. When models generate moral stories, they often overemphasize specific foundations like "Care" while adopting overly linear structures that diverge from the nuanced distribution found in

human-written narratives (Rezapour et al., 2025). This suggests that the narrative standpoint—the position from which a story is told—may fundamentally shift the model's judgment style.

This paper proposes to treat narrative perspective not merely as a stylistic choice, but as a systematic moral probe. We move the focus from what the model judges to from whose standpoint it judges. By deliberately switching between first-person (actor) and third-person (observer) perspectives, we can test the robustness of a model's moral judgment. If a model's ethical reasoning is truly grounded in stable principles, it should resist arbitrary shifts in narrative framing. Conversely, systematic deviations across perspectives would reveal hidden biases in how the model perceives moral agency and responsibility.

To implement this probing approach, we construct a dual-perspective benchmark grounded in Moral Foundations Theory (MFT) (Graham et al., 2013), featuring over 500 paired scenarios across the dimensions of Care, Fairness, Loyalty, Authority, and Purity. Unlike previous metrics that assume moral consistency (Galarnyk et al., 2025), we introduce four fine-grained indicators: Moral Sensitivity (MS) and the Moral Hedging Index (MHI) to measure the sharpness and ambiguity of judgments, and Perspective Consistency (PC) and Perspective Bias Direction (PBD) to quantify the magnitude and direction of shifts when perspectives are switched.

A systematic evaluation of nine mainstream LLMs reveals a striking Inverse Perspective Effect. While humans are typically more lenient toward the self, LLMs demonstrate significantly stricter judgments and higher moral sensitivity when scenarios are presented in the first person. This effect suggests that adopting a participant's perspective triggers a more rigid adherence to moral rules rather than inducing empathy. Our findings suggest that perspective switching is an effective, low-cost probe for uncovering the implicit judgment styles of LLMs, providing a new lens through which to evaluate the stability and reliability of AI-driven moral reasoning.

2 Related Work

2.1 Moral Evaluation of LLMs

Prior work has extensively investigated the moral and ethical behavior of LLMs through benchmark-based evaluation (Jiao et al., 2025; Nunes et al., 2024). These studies typically assess whether model outputs align with human moral norms, whether models can provide coherent ethical justifications, and whether their value preferences remain consistent across scenarios. For example, Jiao *et al.* (Jiao et al., 2025) propose a three-dimensional assessment system for evaluating moral reasoning in LLMs, while Abdulhai *et al.* (Abdulhai et al., 2024) adopt Moral Foundations Theory (Graham et al., 2013) to examine whether LLMs exhibit systematic biases toward particular moral values. Other work studies discrepancies between abstract moral endorsement and concrete moral judgments, revealing potential forms of moral inconsistency (Nunes et al., 2024). However, existing moral benchmarks primarily evaluate the content of moral judgments, while paying limited attention to whether the same moral event is consistently judged when narrated from different standpoints.

2.2 Perspective in Moral Judgment

Human moral judgment is known to be sensitive to the standpoint from which an action is interpreted (Mikhail, 2007). In attribution theory, the actor–observer asymmetry (Jones and Nisbett, 1987; Malle, 2006) suggests that individuals tend to explain their own behavior with reference to situational constraints, while attributing others' behavior more strongly to dispositions or character traits. This asymmetry implies that moral evaluation is not solely determined by the objective content of an action, but is also influenced by the cognitive role of the evaluator.

Recent studies suggest that similar perspective-induced effects may also emerge in large language models. Van Nuenen and Sachdeva (van Nuenen and Sachdeva, 2026) show that LLM moral judgments can be sensitive to broad perturbations such as point-of-view shifts and persuasion cues, even when the underlying moral conflict remains unchanged. Such findings indicate that models may rely on narrative framing as a pragmatic cue rather than grounding judgments purely in moral substance. However, existing work mainly characterizes general judgment instability under broad perturbations, without systematically isolating the role of narrative perspective itself.

In contrast, we focus specifically on first-person versus third-person narration as a psychologically motivated perspective manipulation in paired Moral Foundations Theory-grounded scenarios. We further introduce fine-grained metrics to quantify not only whether judgments change across per-

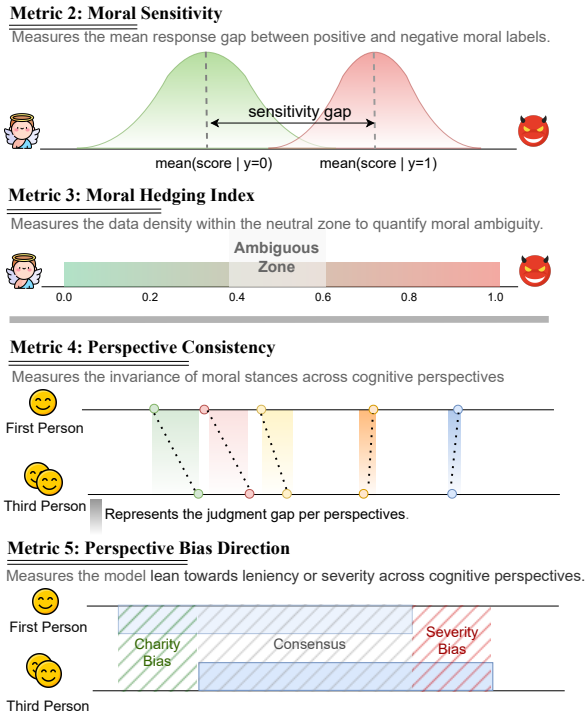

Figure 1: Metrics of moral tendency and perspective sensitivity.

spectives, but also how strongly, how ambiguously, and in which direction they shift. Our work therefore complements prior moral benchmarking and robustness analyses by treating narrative perspective as a targeted moral probe for revealing implicit judgment styles in LLMs.

3 Experiment Design

3.1 Dataset Construction

Filtering and scenario normalization. Our experiments are based on the Commonsense subset of the ETHICS benchmark (Hendrycks et al., 2020), which contains approximately 21.8k everyday moral scenarios. Since all scenarios are originally written in the first person, the dataset provides a suitable basis for constructing controlled first-person/third-person counterfactual pairs.

To improve the reliability of perspective-shift evaluation, we first apply length-based filtering and retain only scenarios between 100 and 1,000 words/tokens. This filtering is motivated by the need to maintain sufficient contextual completeness while avoiding excessive narrative complexity.

We further remove near-duplicate entries and normalize minor formatting inconsistencies to ensure that paired first-person and third-person scenarios differ only in narrative perspective while preserving the same underlying moral event. This controlled setup enables a more reliable analysis of moral judgment stability under viewpoint shifts.

Moral foundation-oriented selection. Prior work suggests that LLM moral reasoning is not governed by a single homogeneous ethical criterion, but instead reflects context-dependent and multi-dimensional value preferences (Huang et al., 2026; Wu et al., 2025). Motivated by this observation, we further refine the ETHICS Commonsense scenarios using Moral Foundations Theory (MFT). Rather than treating all moral scenarios as a single undifferentiated test set, we categorize them into five moral foundations: Care, Fairness, Loyalty, Authority, and Purity. This stratification allows us to examine whether perspective shifts affect all moral judgments uniformly, or whether certain moral domains are more sensitive to first-person versus third-person framing. Specifically, we employ DeepSeek-V3 (Liu et al., 2024) to score each scenario with respect to five MFT dimensions:

- **Care**: concern for suffering and avoidance of harm.
- **Fairness**: justice, rights, and prevention of unfair advantage.
- **Loyalty**: allegiance and standing with one's group.
- **Authority**: respect for rules and hierarchy.
- **Purity**: sanctity and appropriate conduct.

Although some scenarios exhibit signals from multiple moral foundations, we assign each scenario to a single dominant category for clearer category-wise analysis. Specifically, we select the foundation with the highest score as the primary label, and discard scenarios whose top foundation scores are insufficiently distinguishable, as these cases may reflect ambiguous moral orientations. This process yields a cleaner evaluation set in which each scenario is associated with one dominant moral foundation.

Perspective pairing (FP/TP construction). For each retained scenario, the original first-person text is used as the FP view. To construct the third-person view, we rewrite the same scenario into a third-person description using DeepSeek-V3 with temperature set to 0 to ensure deterministic generation.

The rewriting process is constrained to perform minimal surface-level modifications while preserving the original semantic and narrative content. Specifically, the transformation is limited to perspective-related adjustments, including replacement of first-person references with neutral third-person identifiers and grammatically necessary pronoun adaptations (e.g., converting possessive forms such as "my" into corresponding third-person forms such as "his" or "their"). The objective of the rewriting process is to shift the narrative viewpoint from events personally experienced by the narrator to the same events described as happening to another individual, without altering the underlying behavioral episode itself.

To minimize potential confounds introduced by paraphrasing or semantic drift, the rewriting procedure is further constrained to preserve: (i) the factual event structure, (ii) causal relations, (iii) emotional expression, (iv) sentence-level narrative structure, and (v) moral label semantics,

while prohibiting the addition, removal, or reinterpretation of the original information. As a result, the FP and TP versions differ only in narrative perspective, rather than in the underlying moral situation or event content.

After rewriting, we perform consistency checks to ensure that the TP rewrite remains faithful to the original event, including preservation of named entities, actions, outcomes, and overall semantic content. Two researchers independently conduct manual inspections on randomly sampled FP/TP pairs to verify that the rewritten scenarios preserve the original event content while modifying only the narrative perspective. The primary objective of the manual inspection is to ensure that the FP and TP versions describe the same behavioral episode, such that the only change introduced by rewriting is the shift from self-experienced narration to narration about another individual. Annotators additionally verify that no unintended semantic changes or paraphrased event descriptions are introduced during rewriting. Samples identified as containing semantic inconsistencies or unintended content modifications are discarded and regenerated through the same rewriting procedure until the consistency requirements are satisfied.

Model evaluation protocol. For subsequent evaluations, considering both computational cost and model availability/popularity, we select nine top-ranked models from LLM Arena[1],including: Gemini Pro (Comanici et al., 2025), Claude, GPT-5, GPT-4o (Hurst et al., 2024), GLM (Zeng et al., 2025), DeepSeek-R1 (Guo et al., 2025), Deepseek-V3, Qwen-80b, Qwen-235b (Bai et al., 2023). Each model is evaluated on the constructed FP/TP pairs three independent times, and we report the final metrics as the average across runs to reduce variance due to stochastic decoding and transient system effects. This multi-run averaging improves robustness of both overall accuracy and the fine-grained perspective-aware metrics (e.g., moral sensitivity, moral hedging, perspective consistency).

3.2 Metrics Definition

Beyond standard binary classification accuracy, we propose a suite of fine-grained metrics to capture the nuanced moral judgments made by LLMs under different narrative perspectives. The metrics are designed to separately reflect: (i) how well the model separates moral polarities in its latent scoring space, (ii) whether it commits decisively or hedges near the decision boundary, and (iii) whether its judgments remain consistent or shift when the narrative perspective changes.

Moral Sensitivity (MS). The **MS** metric quantifies how effectively a model distinguishes moral from immoral behaviors in its continuous latent score space. For a given perspective p, let $\mathcal{S}_{pos}^{(p)} = \{s_i^{(p)} \mid y_i = 1\}$ be the set of predicted scores for all immoral samples, and $\mathcal{S}_{neg}^{(p)} = \{s_j^{(p)} \mid y_j = 0\}$ be the set for moral samples. We define

$$MS^{(p)} = \frac{1}{|\mathcal{S}_{pos}^{(p)}|} \sum_{s_i \in \mathcal{S}_{pos}^{(p)}} s_i - \frac{1}{|\mathcal{S}_{neg}^{(p)}|} \sum_{s_j \in \mathcal{S}_{neg}^{(p)}} s_j. \tag{1}$$

A higher $MS^{(p)}$ indicates that immoral behaviors systematically receive higher immorality scores than moral behaviors, reflecting a more sensitive and discriminative moral compass. Unlike accuracy, MS leverages both the ranking and geometry implied by continuous scores, and thus can expose reduced separability even when hard thresholding yields similar results.

Moral Hedging Index (MHI). The **MHI** quantifies a model's tendency to avoid committing to a clear moral stance by producing ambiguous scores near the decision boundary. Let $\tau = 0.5$ denote

[1] https://lmarena.ai/zh/leaderboard

the decision boundary and set $\varepsilon = 0.1$, so that the ambiguity zone is defined as $[0.4, 0.6]$:

$$[\tau_{\text{low}}, \tau_{\text{high}}] = [\tau - \varepsilon,\ \tau + \varepsilon]. \tag{2}$$

For a given perspective $p \in \{fp, tp\}$, the MHI is defined as

$$\text{MHI}^{(p)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}\Big(s_i^{(p)} \in [\tau_{\text{low}}, \tau_{\text{high}}]\Big), \tag{3}$$

where $\mathbf{I}(\cdot)$ is the indicator function. Intuitively, a larger $\text{MHI}^{(p)}$ means the model frequently produces near-boundary scores and therefore exhibits a more equivocal judgment style. This metric complements MS: two models may share similar separability yet differ substantially in how decisively they interpret evidence.

Perspective Consistency (PC). To measure stability across narrative frames, we introduce **PC** to evaluate how invariant the moral score is when switching from first-person to third-person descriptions. For each sample, we compute:

$$pc_i = 1 - |s_i^{(fp)} - s_i^{(tp)}|. \tag{4}$$

Since $s_i^{(fp)}, s_i^{(tp)} \in [0, 1]$, it follows that $pc_i \in [0, 1]$. Higher values of pc_i indicate that the model assigns similar immorality degrees under both perspectives. We compute the expected consistency and its corresponding volatility across the dataset:

$$PC_{\text{mean}} = \mathbb{E}_{i\sim\mathcal{D}}[pc_i], \quad PC_{\text{std}} = \sqrt{\text{Var}(pc)}. \tag{5}$$

High PC_{mean} together with low PC_{std} suggests a stable and unified moral logic that is largely invariant to evaluative framing, while high PC_{std} indicates that the model's consistency may depend strongly on specific scenario patterns.

Perspective Bias Direction (PBD). While PC captures the *magnitude* of cross-perspective agreement, **PBD** characterizes the *direction* of divergence. We define

$$\delta_i = s_i^{(fp)} - s_i^{(tp)}, \tag{6}$$

and summarize directional bias using the expected difference:

$$PBD_\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_i. \tag{7}$$

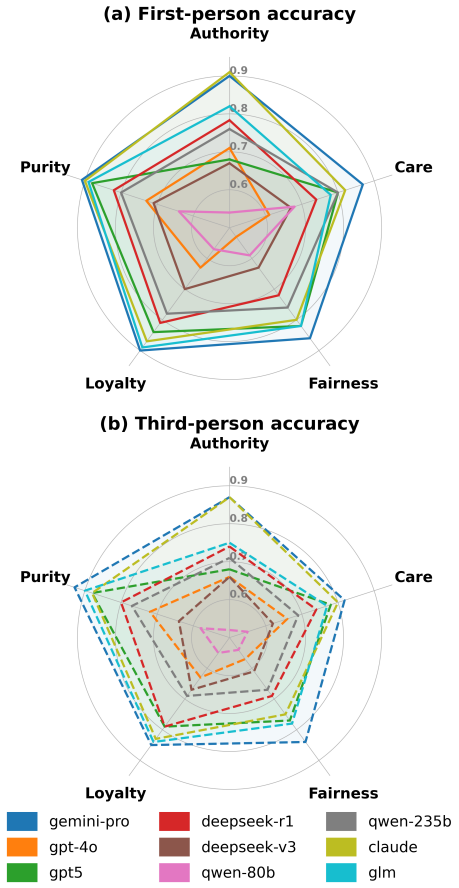

Figure 2: Multi-dimensional moral sensitivity profile of LLMs across distinct narrative perspectives.

A positive PBD_μ indicates that, on average, the model assigns higher immorality scores in the first-person perspective (i.e., it is relatively stricter from the agent's viewpoint), whereas a negative value suggests the opposite. In this way, PBD pinpoints whether perspective shifts systematically push judgments toward a more self-strict or observer-strict interpretation, rather than merely increasing disagreement.

4 Result

4.1 Analysis of Accuracy

As shown in Figure 2 (a) and (b), narrative perspective has a pronounced impact on model accuracy across all five Moral Foundations Theory (MFT) dimensions. In the first-person setting (Fig. 2a), the majority of systems attain relatively high accuracy scores, forming a tight band roughly in the range of $[0.70, 0.90]$. The strongest effects concentrate on Authority and Purity, where several models approach the top end of the scale (often greater than 0.85), indicating that these foundations are com-

Table 1: Moral Sensitivity (MS) under first-person (FP) narratives. Higher MS indicates stronger discrimination between morally positive and negative scenarios. Best and worst results are highlighted.

Category	Gemini-Pro	Claude	GLM	GPT-5	Qwen-235B	DeepSeek-R1	DeepSeek-V3	GPT-4o	Qwen-80B
Authority	0.679	0.459	0.503	0.125	0.102	0.338	0.047	0.195	0.127
Care	0.602	0.448	0.474	0.409	0.314	0.381	0.316	0.122	0.159
Fairness	0.610	0.414	0.512	0.385	0.214	0.363	0.257	0.032	0.137
Loyalty	0.628	0.411	0.573	0.330	0.170	0.363	0.146	0.077	0.156
Purity	0.639	0.466	0.595	0.409	0.217	0.412	0.190	0.232	0.175

Table 2: Moral Sensitivity (MS) under third-person (TP) narratives. Higher MS indicates stronger discrimination between morally positive and negative scenarios. Best and worst results are highlighted.

Category	Gemini-Pro	Claude	GLM	GPT-5	Qwen-235B	DeepSeek-R1	DeepSeek-V3	GPT-4o	Qwen-80B
Authority	0.596	0.432	0.432	0.173	0.074	0.327	0.164	0.166	0.114
Care	0.554	0.432	0.455	0.425	0.186	0.412	0.336	0.214	0.165
Fairness	0.559	0.372	0.466	0.390	0.142	0.384	0.229	0.094	0.151
Loyalty	0.559	0.382	0.512	0.322	0.128	0.378	0.140	0.120	0.113
Purity	0.679	0.446	0.609	0.430	0.153	0.408	0.151	0.226	0.141

paratively more robust to changes in evaluative framing. In contrast, **Care**, **Fairness**, and **Loyalty**, while still generally accurate, exhibit slightly lower peaks and more noticeable dispersion, suggesting that models require more contextual grounding to reliably judge these morally relational and context-sensitive categories.

When switching to the third-person narratives (Fig. 2b), performance drops substantially, and the profiles become more heterogeneous across models. Most systems shift downward to a broader interval, approximately 0.50 to 0.80, accompanied by visibly larger gaps between individual model traces. This systematic reduction is especially evident in dimensions that depend more heavily on agent-centric interpretation, notably **Care** & **Fairness**, where third-person framing appears to weaken the mapping from scenario evidence to moral evaluation. Put differently, third-person presentation not only lowers average performance but also increases the likelihood that different architectures and training priors respond differently to the same underlying moral content.

Beyond the overall performance gap, model-specific patterns emerge in how each moral foundation is affected by perspective. Across both perspectives, **Authority** & **Purity** remain comparatively stable (high scores and smaller cross-model spread), implying that moral reasoning for rule- and norm-related judgments is less sensitive to narrative viewpoint. Meanwhile, **Care** & **Loyalty** show stronger inter-model variability, implying that perspective interacts with each model's internal representation of relational context, harm salience, and group-referential cues. This leads to a consistent interpretation: narrative perspective operates as a moderating factor for moral judgment, with third-person framing attenuating ethical sensitivity and revealing underlying differences in judgment mechanisms across models.

4.2 Moral Sensitivity Across Models

Tables 1 and 2 present the quantitative analysis of moral sensitivity (MS) across models under first-person (FP) and third-person (TP) narratives, respectively, revealing substantial heterogeneity in ethical differentiation capacity. Across both perspectives, Gemini-Pro consistently demonstrates the strongest moral sensitivity, with MS values ranging from 0.596 to 0.679, indicating a pronounced ability to discriminate between morally positive and negative scenarios. Claude maintains relatively stable intermediate performance, with scores generally distributed within the $[0.372, 0.466]$ range, while GLM achieves competitive performance in several dimensions, particularly under Loyalty and Purity.

In contrast, DeepSeek-V3, GPT-4o, and Qwen-80B exhibit substantially weaker moral sensitivity across multiple categories, frequently producing MS values below 0.200 in certain dimensions. For example, GPT-4o shows extremely limited differentiation on Fairness and Loyalty, while DeepSeek-V3 obtains the lowest MS score on Authority under FP narratives. Although the overall ranking of models remains broadly consistent between FP and TP settings, several models display noticeable fluctua-

Table 3: Moral Hesitation Index (MHI) under first-person (FP) narratives. Lower MHI indicates less hesitation in moral judgment. Best results are highlighted in green, while worst results are highlighted in red.

Category	Gemini-Pro	Claude	GLM	GPT-5	Qwen-235B	DeepSeek-R1	DeepSeek-V3	GPT-4o	Qwen-80B
Authority	0.080	0.010	0.050	0.390	0.010	0.230	0.340	0.390	0.370
Care	0.061	0.020	0.040	0.250	0.020	0.230	0.280	0.420	0.320
Fairness	0.040	0.030	0.050	0.270	0.000	0.200	0.300	0.500	0.260
Loyalty	0.130	0.040	0.060	0.360	0.030	0.240	0.460	0.520	0.400
Purity	0.041	0.000	0.090	0.280	0.000	0.230	0.440	0.440	0.290

Table 4: Moral Hesitation Index (MHI) under third-person (TP) narratives. Lower MHI indicates less hesitation in moral judgment. Best results are highlighted in green, while worst results are highlighted in red.

Category	Gemini-Pro	Claude	GLM	GPT-5	Qwen-235B	DeepSeek-R1	DeepSeek-V3	GPT-4o	Qwen-80B
Authority	0.130	0.020	0.050	0.280	0.080	0.247	0.370	0.310	0.000
Care	0.091	0.020	0.020	0.190	0.000	0.180	0.210	0.290	0.270
Fairness	0.080	0.010	0.050	0.190	0.000	0.230	0.310	0.370	0.220
Loyalty	0.100	0.010	0.070	0.280	0.010	0.190	0.490	0.510	0.410
Purity	0.071	0.010	0.040	0.280	0.000	0.210	0.390	0.310	0.340

tions across perspectives, suggesting that narrative viewpoint can moderately influence the strength of moral discrimination even when the underlying moral event remains unchanged.

4.3 Moral Hedging Index Across Models

Tables 3 and 4 present the Moral Hedging Index (MHI) results under first-person (FP) and third-person (TP) narratives, respectively. Lower MHI values indicate less hesitation and greater decisiveness in moral judgment. Across both perspectives, substantial variation can be observed in the degree of moral commitment exhibited by different models. GPT-4o consistently demonstrates the highest MHI values across multiple moral foundations, reaching 0.520 under Loyalty and 0.500 under Fairness in FP narratives, indicating a strong tendency toward neutral or non-committal responses despite producing reasonable classification performance. DeepSeek-V3 and Qwen-80B also exhibit comparatively elevated hedging behavior in several dimensions, particularly under Loyalty and Purity. In contrast, Claude and GLM maintain consistently low MHI values, generally within the $[0.010, 0.070]$ range, suggesting more decisive and stable moral judgments. Qwen-235B further exhibits extremely low MHI scores in several categories, including multiple near-zero values under both FP and TP settings.

Comparing FP and TP narratives, most models exhibit moderately higher MHI values under third-person framing, suggesting increased hesitation when moral scenarios are narratively distanced from the acting subject. This trend is particularly visible for Gemini-Pro, DeepSeek-V3, and Qwen-80B.

4.4 Consistency Across Narrative Perspectives

Table 5 presents moral perspective consistency metrics across first-person (FP) and third-person (TP) narrative frames. Overall, Claude and GPT-5 exhibit the strongest perspective consistency, maintaining consistently high PC_{mean} values across nearly all moral dimensions. Claude achieves PC_{mean} scores above 0.94 on multiple categories, including **Care** (0.948), **Fairness** (0.941), and **Loyalty** (0.948), while GPT-5 reaches the highest consistency on **Care** (0.958) and **Purity** (0.953). These results suggest that both models preserve relatively stable moral judgments under narrative perspective shifts. Gemini-Pro and GLM-4 also demonstrate comparatively strong consistency, whereas GPT-4o and DeepSeek-V3 exhibit noticeably lower PC_{mean} values in several dimensions, particularly under **Care** and **Fairness**. Examining the variability metric PC_{std} further reveals substantial differences in perspective sensitivity across models. Most models maintain relatively low variance values (0.05–0.15), indicating stable judgment patterns across perspectives. However, several notable anomalies emerge. DeepSeek-V3 exhibits extremely high PC_{std} values on **Care** (0.782) and **Fairness** (0.833), suggesting severe instability under narrative shifts. GPT-4o similarly demonstrates elevated variance on **Care** (0.225) and **Fairness** (0.177), while Qwen-235B shows comparatively larger fluctuations across multiple categories. These findings indicate that although some mod-

Table 5: Moral perspective consistency metrics quantifying model coherence in moral judgments across first-person and third-person narrative frames. PC_mean measures the uniformity of moral positions across perspectives (higher is better), while PC_std captures the variance in consistency, indicating sensitivity to perspective-induced judgment shifts. Best results are highlighted in green, while worst results are highlighted in red.

Category	Metric	Gemini-Pro	Claude	GLM-4	GPT-5	Qwen-235B	DeepSeek-R1	DeepSeek-V3	GPT-4o	Qwen-80B
Authority	PC_{mean}	0.912	0.972	0.918	0.903	0.851	0.875	0.879	0.875	0.918
	PC_{std}	0.121	0.054	0.161	0.122	0.615	0.125	0.102	0.128	0.161
Care	PC_{mean}	0.936	0.948	0.938	0.958	0.931	0.896	0.804	0.787	0.900
	PC_{std}	0.105	0.079	0.114	0.062	0.121	0.105	0.782	0.225	0.166
Fairness	PC_{mean}	0.937	0.941	0.952	0.949	0.895	0.883	0.782	0.842	0.952
	PC_{std}	0.103	0.083	0.098	0.078	0.167	0.100	0.833	0.177	0.105
Loyalty	PC_{mean}	0.916	0.948	0.912	0.938	0.899	0.888	0.905	0.843	0.929
	PC_{std}	0.139	0.078	0.130	0.076	0.143	0.128	0.118	0.160	0.124
Purity	PC_{mean}	0.936	0.947	0.936	0.953	0.894	0.879	0.904	0.898	0.938
	PC_{std}	0.106	0.078	0.098	0.059	0.167	0.125	0.119	0.106	0.118

Table 6: Perspective Bias Direction (PBD) across moral categories. Blue indicates the largest value in each row (strongest first-person strictness), and orange indicates the smallest value (weakest first-person strictness or stricter third-person judgment when negative).

Category	Gemini-Pro	Claude	GLM-4	GPT-5	Qwen-235B	DeepSeek-R1	DeepSeek-V3	GPT-4o	Qwen-80B
Authority	0.037	0.003	0.014	0.009	0.070	0.010	0.061	0.051	0.014
Care	0.030	0.025	0.022	0.029	0.054	0.060	0.042	0.050	0.060
Fairness	0.033	0.033	0.039	0.040	0.094	0.042	0.020	0.019	0.026
Loyalty	0.027	0.022	0.042	0.038	0.054	0.054	-0.013	0.039	0.037
Purity	0.005	0.002	0.015	0.026	0.085	0.010	0.036	0.028	0.040

els maintain high overall consistency, perspective-induced instability remains significant for certain moral dimensions, particularly those involving contextual and relational reasoning.

4.5 Perspective Bias Direction

Table 6 quantifies how narrative perspective systematically biases moral judgments across models and moral categories. Overall, nearly all models exhibit positive PBD values across most dimensions, indicating a consistent tendency toward stricter moral judgments under first-person (FP) narratives compared with third-person (TP) framing. This suggests that narrative proximity generally amplifies moral stringency in LLM judgment behavior. Among the evaluated models, Qwen-235B demonstrates the strongest perspective sensitivity, achieving the highest PBD values in multiple categories, including **Authority** (0.070), **Fairness** (0.094), and **Purity** (0.085). DeepSeek-R1 and Qwen-80B also exhibit relatively elevated PBD values under **Care**, indicating stronger FP strictness. In contrast, Claude and GPT-5 maintain comparatively smaller PBD magnitudes across most dimensions, suggesting more stable moral calibration under narrative perspective shifts. Despite the overall dominance of positive PBD values, several anomalies emerge. Most notably, DeepSeek-V3 exhibits a negative PBD value on **Loyalty** (-0.013), representing the only case where TP narratives induce stricter moral judgments than FP framing. This reversal suggests that certain model-category combinations may display counterintuitive perspective effects, potentially reflecting sensitivity to specific discourse patterns or alignment behaviors.

5 Conclusion

In this paper, we investigate how narrative perspective affects moral judgment in large language models. We construct a dual-perspective benchmark based on Moral Foundations Theory and introduce fine-grained metrics for measuring moral sensitivity, hesitation, consistency, and directional bias under perspective shift. Experimental results across nine mainstream LLMs reveal a systematic *Inverse Perspective Effect*, where first-person narratives consistently induce stricter moral judgments than third-person descriptions. Our findings suggest that current LLM moral reasoning is not fully viewpoint-invariant, and that narrative perspective serves as an effective low-cost probe for analyzing latent judgment styles and alignment stability in LLMs.

Limitations

Although our FP/TP pairs are constructed under strict rewriting constraints, the third-person narratives are still generated by an auxiliary LLM, meaning that subtle lexical or discourse-level differences beyond perspective may remain and partially influence model behavior. In addition, our evaluation framework is grounded in the ETHICS benchmark and Moral Foundations Theory (MFT), which reflect particular operationalizations of morality rather than universally shared ethical standards. While the proposed metrics capture multiple aspects of moral judgment behavior, the overall setting still primarily relies on binary moral evaluations and may not fully represent the complexity of real-world ethical reasoning involving uncertainty, competing principles, and contextual trade-offs. Finally, narrative perspective represents only one dimension of framing variation; other factors such as emotional intensity, cultural background, and social identity may also systematically affect LLM moral judgments and warrant further investigation.

Potential Risks

The behavioral differences observed in this work should not be interpreted as evidence that LLMs possess stable or human-like moral beliefs. In addition, because our evaluation framework is based on ETHICS and Moral Foundations Theory, the results may reflect cultural or normative biases embedded in these resources. Finally, perspective-based prompting may potentially be used to influence or steer model moral responses, and thus our framework is intended for diagnostic evaluation rather than real-world moral decision making.

References

Marwa Abdulhai, Gregory Serapio-Garcia, Clément 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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Michael Galarnyk, Agam Shah, Dipanwita Guhathakurta, Poojitha Nandigam, and Sudheer Chava. 2025. How inclusively do lms perceive social and moral norms? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4859–4869.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. 47:55–130.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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*.

Fan Huang, Haewoon Kwak, and Jisun An. 2026. Understanding Moral Reasoning Trajectories in Large Language Models: Toward Probing-Based Explainability. arXiv preprint. *Preprint*, arXiv:2603.16017.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, and 1 others. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. 2025. Llm ethics benchmark: a three-dimensional assessment system for evaluating moral reasoning in large language models. *Scientific Reports*, 15(1):34642.

Edward E Jones and Richard E Nisbett. 1987. The actor and the observer: Divergent perceptions of the causes of behavior. In *Preparation of this paper grew out of a workshop on attribution theory held at University of California, Los Angeles, Aug 1969*. Lawrence Erlbaum Associates, Inc.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Bertram F Malle. 2006. The actor-observer asymmetry in attribution: a (surprising) meta-analysis. *Psychological bulletin*, 132(6):895.

John Mikhail. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, 11(4):143–152.

José Luiz Nunes, Guilherme FCF Almeida, Marcelo De Araujo, and Simone DJ Barbosa. 2024. Are large language models moral hypocrites? a study based on moral foundations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1074–1087.

Rezvaneh Rezapour, Sullam Jeoung, Zhiwen You, and Jana Diesner. 2025. Tales of Morality: Comparing Human- and LLM-Generated Moral Stories from Visual Cues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics.

Eike Schneiders, Tina Seabrooke, Joshua Krook, Richard Hyde, Natalie Leesakul, Jeremie Clos, and Joel E Fischer. 2025. Objection overruled! lay people can distinguish large language models from lawyers, but still favour advice from an llm. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Tom van Nuenen and Pratik S Sachdeva. 2026. The fragility of moral judgment in large language models. *arXiv preprint arXiv:2603.05651*.

Ya Wu, Qiang Sheng, Danding Wang, and 1 others. 2025. The Staircase of Ethics: Probing LLM Value Priorities through Multi-Step Induction to Complex Moral Dilemmas. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. Author list is incomplete; replace `and others` with the full list of authors.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.