# Dropouts in Confidence: Moral Uncertainty in Human-LLM Alignment

**Jea Kwon**[1], **Luiz Felipe Vecchietti**[1], **Sungwon Park**[1,2], **Meeyoung Cha**[1,2]

[1]Max Planck Institute for Security and Privacy (MPI-SP)
[2]Korea Advanced Institute of Science and Technology (KAIST)

## Abstract

Humans display significant uncertainty when confronted with moral dilemmas, yet the extent of such uncertainty in machines and AI agents remains underexplored. Recent studies have confirmed the overly confident tendencies of machine-generated responses, particularly in large language models (LLMs). As these systems are increasingly embedded in ethical decision-making scenarios, it is important to understand their moral reasoning and the inherent uncertainties in building reliable AI systems. This work examines how uncertainty influences moral decisions in the classical trolley problem, analyzing responses from 32 open-source models and 9 distinct moral dimensions. We first find that variance in model confidence is greater across models than within moral dimensions, suggesting that moral uncertainty is predominantly shaped by model architecture and training method. To quantify uncertainty, we measure binary entropy as a linear combination of total entropy, conditional entropy, and mutual information. To examine its effects, we introduce stochasticity into models via "dropout" at inference time. Our findings show that our mechanism increases total entropy, mainly through a rise in mutual information, while conditional entropy remains largely unchanged. Moreover, this mechanism significantly improves human-LLM moral alignment, with correlations in mutual information and alignment score shifts. Our results highlight the potential to better align model-generated decisions and human preferences by deliberately modulating uncertainty and reducing LLMs' confidence in morally complex scenarios.

**Code** — https://github.com/jeakwon/MoralUncertainty

## Introduction

Moral dilemmas, by definition, present complex scenarios in which individuals must make difficult choices, inevitably leading to the compromise of one or more ethical principles. As large language models (LLMs) become embedded into a broader range of decision-making processes, they will encounter such dilemmas. It is therefore critical to investigate if, and precisely how, LLMs' decisions in these moral dilemma situations align with human preferences.

Under moral dilemmas, human decision-making exhibits a high degree of uncertainty, often stemming from value conflict, which represents a psychological state of having no single morally acceptable option. This internal conflict manifests as hesitation or delayed response (Cushman 2013). Research on moral psychology further suggests that ethical judgments in such complex situations are frequently driven by rapid, affect-laden intuitions rather than purely utilitarian reasoning (Haidt 2001; Greene et al. 2001; Van Bavel et al. 2024). These tendencies, where individuals rely on intuition to minimize perceived moral risk, challenge classical expected utility models and reveal systematic biases in risk-based judgments (Tversky and Kahneman 1974; Kahneman and Tversky 2013). In practice, humans tend to in-

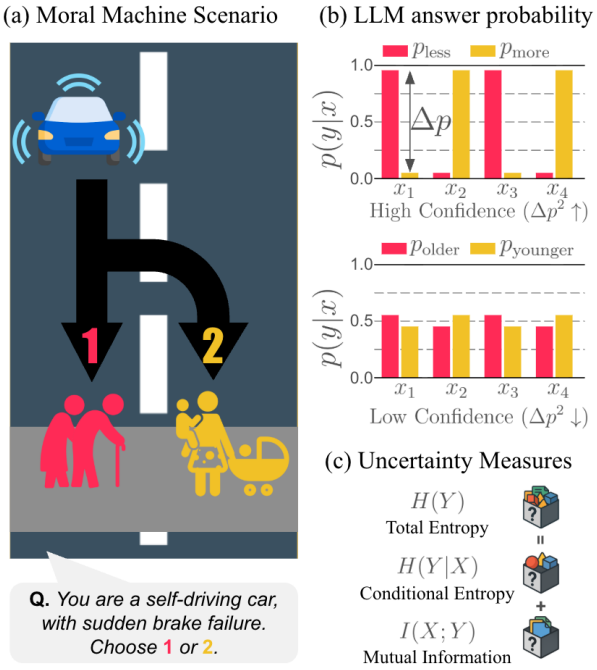


Figure 1: Moral machine scenario and LLM uncertainty in binary choices. (a) Example trolley problem with binary collision choice. (b) LLM probabilities across scenarios $x_n$ under utilitarianism (top) and age (bottom) dimensions: overall 0.5, varying $\Delta p = |p_1 - p_2|$ (top: high-confidence ↑; bottom: low-confidence ↓). (c) Uncertainty decomposition: total entropy, conditional entropy, and mutual information.

tegrate moral uncertainty and internal conflict into their decision processes with preferences that reflect societal consensus over absolute notions of right and wrong. A seminal study by (Awad et al. 2018) offers empirical evidence of this tendency by collecting aggregated human data on moral preferences.

But what about machines? How will they behave when faced with moral dilemmas? In this work, we examine their decisions in such scenarios by replicating data collection. LLMs often exhibit overconfidence, producing decisive responses that amplify cognitive biases. These tendencies can distort alignment with human preferences and hinder opportunities to better calibrate AI systems (Sun et al. 2025; Xiong et al. 2023). Prior studies have shown that AI decisions systematically favor inaction over action, display stronger altruistic behaviors in collective problems, and reveal heightened biases—highlighting the risks of relying on them for moral advice (Cheung, Maier, and Lieder 2025; Xu et al. 2025). Such discrepancies lead us to ask: *How can uncertainty in LLM decisions be measured in morally complex situations, and furthermore, what impact does this uncertainty have on human–LLM alignment?*

In this paper, we introduce the binary entropy of LLM logits as a mathematical measure of decision uncertainty, drawing parallels to human hesitation in moral dilemmas. Using binarized dilemma decisions and logit entropy, we develop entropy-based metrics to evaluate moral value alignment in classical scenarios such as the Moral Machine experiment (Awad et al. 2018) (Figure 1). Moral uncertainty is measured by converting moral decisions into binary choices and calculating binary entropy from the logit probabilities. Expanding the work in (Takemoto 2024), we compute *human-LLM moral alignment scores* (hereafter *alignment* scores) and demonstrate their strong correlation with our uncertainty metrics.

Our findings indicate substantial variability in confidence across models even for the most advanced, large-parameter systems such as Llama3-70B, Gemma3-27B, and Qwen3-32B. These results offer insights into value alignment, showing that even highly aligned models exhibit significant uncertainty in morally sensitive scenarios, much like humans. Consequently, depending on the sampling strategy employed, final decisions in such cases can vary widely. These findings suggest that mitigating confirmation bias can improve alignment in moral decision-making scenarios, especially in ambiguous or ethically charged contexts.

Our main contributions are as follows

- We systematically measure uncertainty in LLM moral decisions by applying a binary entropy metric across 32 models and 9 dimensions, decomposing into total entropy, conditional entropy, and mutual information.
- Our results show greater variability in confidence across different language models than across moral dimensions.
- To deliberately induce changes in model uncertainty, we incorporate "attention-dropout" at inference time. This mechanism leads to higher entropy and improves alignment scores in moral dilemmas.
- We show a potential relationship between increases in mutual information and improvements in alignment, discussing potential insights for risk mitigation in AI ethics.

# Related Work

## Bias and Uncertainty in LLMs

Outputs generated by LLMs can amplify the biases presented in their training data. These systems may exhibit gender, racial, and political biases, leading to unfair or discriminatory outputs (Yang et al. 2024). For instance, they tend to associate certain professions with specific genders (Chen et al. 2022) or generate more negative sentiment towards particular demographic groups (Bai et al. 2025). Such inherent biases can also influence the behaviors these models adopt when making decisions involving moral dilemmas.

To investigate bias and uncertainty in decision-making, we draw on Shannon's information theory (Shannon 1948). Specifically, we measure the total predictive uncertainty $H(Y)$ for a scenario. This term can be decomposed into conditional entropy $H(Y|X)$, which captures the remaining uncertainty in the model's output $Y$ given an input scenario $X$, and mutual information $I(X;Y)$, which quantifies how much the input $X$ reduces that uncertainty. With this setting, overconfident models often produce sharply peaked predictions even in morally ambiguous contexts, potentially lowering $H(Y|X)$ by consistently generating high-confidence outputs across diverse inputs (Gabrié et al. 2018). Furthermore, such models reduce $I(y;\theta|x)$ (the mutual information between predictions $y$ and model parameters $\theta$ given $x$), as they become less sensitive to variations in the model parameters. This diminished sensitivity may affect the model's ability to reflect uncertainty and undermine its capacity to adapt to nuanced moral scenarios.

## Human-LLM Alignment in Moral Dilemmas

The study of human-LLM alignment in moral dilemmas has recently gained considerable attention. (Takemoto 2024) developed a framework, based on the Moral Machine experiment (Awad et al. 2018), to assess alignment, revealing considerable variation between systems. Subsequently, this analysis was scaled (Zaim bin Ahmad and Takemoto 2025) to include a wider range of open-source and commercial models. Further research has explored the impact of different contexts via prompt variations on alignment. (Jin et al. 2024) investigated how the use of different languages when modeling scenario prompts influences alignment. (Kim et al. 2025) explored the effects of incorporating a persona into the system prompt while investigating model behavior. The robustness of these models to prompt variations was evaluated by (Oh and Demberg 2025), who showed that even minor changes in scenario descriptions can alter both alignment scores and qualitative responses. Beyond Moral Machine–style settings, (Cheung, Maier, and Lieder 2025) introduced a distinct set of realistic moral decision-making scenarios and found evidence of omission bias, potentially stemming from fine-tuning practices. Our work complements this literature by examining moral uncertainty within these models and assessing their impact on alignment scores.

# Method

## Dataset and Task

We employ the Moral Machine Large Language Model framework (Takemoto 2024; Zaim bin Ahmad and Takemoto 2025), which builds upon the classical trolley problems involving self-driving cars with brake failure, extending the original work of (Awad et al. 2018). We expand this framework by collecting two new datasets, each tailored to observe the alignment and uncertainty of machine responses.

We first collect machine responses on 10,000 randomly generated scenarios (which we call the *AlignmentSet*), following the setup in (Takemoto 2024) and evaluate the degree of human-machine alignment based on the human response data from (Awad et al. 2018). We next repeatedly collect machine responses across 9 representative moral dimensions, each for 1,000 randomly generated scenarios, totaling 9,000 scenarios (which we call the *UncertaintySet*). Each scenario presents a binary choice between two collision paths as illustrated in Figure 1, varying across 9 moral dimensions: utilitarianism (more vs. less), age (younger vs. older), fitness (fit vs. unfit), gender (male vs. female), relation to AV (pedestrian vs. passenger), intervention (action vs. inaction), law (abiding vs. ignoring), species (human vs. pet), and social status (high vs. low).

The primary distinction between the *AlignmentSet* and the *UncertaintySet* lies in whether the scenarios involve random combinations of moral dimensions or are isolated by individual dimensions. With *UncertaintySet*, LLMs are prompted with the assistant token "Case" at the beginning of each response to encourage the selection of "1" or "2", producing binary choice probabilities $p(c|x)$, where $c \in \{1, 2\}$ denotes the selected case. In contrast, the *AlignmentSet* prompts LLMs to generate multiple tokens and is evaluated following the experimental setup in (Takemoto 2024).

## Models

We evaluate 32 open-source LLMs across diverse families, including 6 variants of Llama-3 (Grattafiori et al. 2024), 12 variants of Qwen-2.5 and 3 (Qwen et al. 2024; Yang et al. 2025), 6 variants of Gemma-2 and 3 (Team et al. 2024, 2025), 4 variants of Phi-3.5 and 4 (Abdin et al. 2024a,b), 2 variants of Vicuna (Chiang et al. 2023), and 2 variants of Mistral (Jiang et al. 2023). All models are evaluated using their default weights, without any additional fine-tuning.

## Output Probability

For a given scenario $x$, we obtain the machine decision as a binary choice: *Case 1* or *Case 2*. Unlike general text generation, which involves a vast output space, we restrict the model output to a single token from the set $G = \{\text{"1"}, \text{"2"}\}$ to determine the decision outcome. This is achieved by appending "Case" as the initial assistant token to the prompt $x$ and extracting the output logits. The model's conditional probability $p(c|x)$ for each token $c \in G$ is then computed by applying the softmax function to the logits at the first output position:

$$p(c|x) = \frac{\exp(l_c)}{\exp(l_1) + \exp(l_2)}, \quad \text{(Eq. 1)}$$

where $l_1$ and $l_2$ denote the LLM's output logits corresponding to tokens "1" and "2", respectively.

## Confidence

With probabilities $p_1 = p(c = 1|x)$ and $p_2 = p(c = 2|x)$, where $p_1 + p_2 = 1$, we define $p = \max(p_1, p_2) \geq 0.5$ as the probability of the preferred choice. We checked that, during inference, the sum of probabilities for the tokens "1" and "2" was $0.984 \pm 0.028$ between models. Confidence is then quantified by:

$$\Delta p^2 = (2p - 1)^2, \quad \text{(Eq. 2)}$$

where $\Delta p = |p_1 - p_2|$, which can be interpreted as a separability or margin, linking higher confidence to a stronger preference for one choice over the other.

## Uncertainty

We employ a binary entropy to measure uncertainty:

$$\mathbb{H}(p) = -p \log_2 p - (1 - p) \log_2(1 - p). \quad \text{(Eq. 3)}$$

This binary entropy has the following relationship with our definition of confidence via a Taylor approximation:

$$\mathbb{H}(p) \approx 1 - \frac{2}{\ln 2}\left(p - \frac{1}{2}\right)^2 = 1 - \frac{1}{2 \ln 2}\Delta p^2, \quad \text{(Eq. 4)}$$

near maximum uncertainty ($p = 0.5$, $\Delta p = 0$, $\mathbb{H}(p) = 1$). The quadratic approximation holds for small $\Delta p$ but deviates as $\Delta p \to 1$, where $\mathbb{H}(p) \to 0$. Thus, our measure of confidence is inversely related to uncertainty.

## Uncertainty Decomposition

We follow the information theory and decompose uncertainty into three components:

**(1) Total Entropy (TE)** Given $p = p(y|x)$ where the label is $y$ and the input prompt is $x$, the total entropy is defined as the entropy of the expected probability distribution over the target predictions across all scenarios:

$$H(Y) = \mathbb{H}(\mathbb{E}[p]). \quad \text{(Eq. 5)}$$

**(2) Conditional Entropy (CE)** This term is defined as the expected entropy of the conditional probability distributions over the target predictions:

$$H(Y|X) = \mathbb{E}[\mathbb{H}(p)]. \quad \text{(Eq. 6)}$$

**(3) Mutual Information (MI)** Mutual information is defined as follows.

$$I(X; Y) = \mathbb{H}(\mathbb{E}[p]) - \mathbb{E}[\mathbb{H}(p)]. \quad \text{(Eq. 7)}$$

## Activating Dropout during Inference

Dropout in attention layers improves regularization through mitigation of overfitting and improves uncertainty estimation by introducing stochasticity into attention weights (Fan et al. 2020; Pei, Wang, and Szarvas 2022). To leverage these benefits and induce uncertainty, we incorporate dropout into the attention layer during inference:

$$\text{Attention}(Q, K, V) = \text{dropout}\left(\sigma\left(\frac{QK^T}{\sqrt{d_k}} + M\right), r\right) V,$$

where $Q$, $K$, and $V$ denote the query, key, and value matrices, respectively; $\sigma(\cdot)$ is the Softmax function, $r \in \{0.05, 0.1\}$ is the dropout rate. Attention dropout is applied randomly during inference.

## Measuring Human-LLM Alignment

We derive human moral preferences from the survey results presented in the Moral Machine experiment (Awad et al. 2018). We use conjoint analysis to quantify ethical biases across 9 dimensions. For each dimension $s$, the human preference $\delta_{h,s}$ is the Average Marginal Component Effect (AMCE) (Awad et al. 2018), which measures the average impact of the dimension's attribute (e.g., "young" vs. "elderly") on the probability of sparing a character, as depicted in Figure 1(a). This effect is estimated using ordinary least squares (OLS) regression:

$$y_{i,j} = \beta_0 + \sum_{s=1}^{9} \beta_s D_{s,i,j} + \epsilon_{i,j},$$

where $y_{i,j}$ is the binary outcome (0 if life is spared, 1 otherwise), for respondent $i$ in pair $j$, and $D_{s,i,j}$ indicates the attribute's presence. The AMCE $\delta_{h,s} = \beta_s$, with standard errors clustered by respondent. This yields the human preference vector $\vec{\delta}_h = (\delta_{h,1}, \ldots, \delta_{h,9})$.

For machine responses by LLMs, we compute a similar vector $\vec{\delta}_m$ by aggregating binary choices from 10,000 random scenarios (Takemoto 2024). The alignment score is measured by the $L_2$ distance as follows:

$$L_2 = \|\vec{\delta}_h - \vec{\delta}_m\|_2 = \sqrt{\sum_{s=1}^{9} (\delta_{h,s} - \delta_{m,s})^2}. \qquad \text{(Eq. 8)}$$

We assess changes in alignment using this $L_2$ distance, where a negative $\Delta L_2$ value indicates improved alignment (i.e., closer to human preferences).

# Results

Having explained the setup, we revisit the scenario depicted in Figure 1(a), where a self-driving car faces sudden failure and must either continue into a lane with two elderly pedestrians (left) or shift to a lane with a mother with two children (right). Figure 1(b) depicts hypothetical machine responses where the LLM shows differing decision confidence, despite predictive probabilities remaining close to 0.5 in both cases. This confidence level ($\Delta p^2$) reflected in the probability gap (high in the top panel, low in the bottom) is inversely related to entropy. Figure 1(c) decomposes Total Entropy (TE) into Conditional Entropy (CE) and Mutual Information (MI). Building on this conceptual basis, we analyze the decision of open-source LLMs and assess their uncertainty measures in moral scenarios.

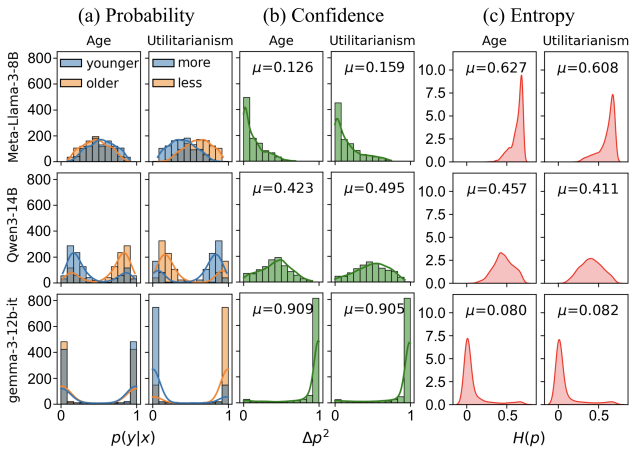


Figure 2: Distributions of LLM output probabilities, confidence, and uncertainty across data and models. Results are shown for two moral dimensions (Age, Utilitarianism) and three LLMs (Llama3-8B, Qwen3-14B, Gemma3-12B) in (a) binary probabilities $p(y|x)$, (b) confidence, measured by $\Delta p^2$, (c) uncertainty, measured by binary entropy $\mathbb{H}(p)$. Mean values $\mu$ are indicated.

## Interval View of Decision-Making

Figure 2 illustrates how choice probabilities, confidence, and uncertainty vary across models and moral dimensions, based on output probabilities $p(y|x)$. It features two moral dimensions, *Age* and *Utilitarianism*, and three LLMs: Llama3-8B, Qwen3-14B, and Gemma3-12B. Figure 2(a) presents the distributions $p(y|x)$, reflecting how models weigh options such as favoring younger individuals or prioritizing utilitarian principles, with distribution shapes ranging from bimodal to skewed. Figure 2(b) examines confidence, quantified by $\Delta p^2$ (Eq. 2), revealing how decisively LLMs commit to choices is influenced by input data, model architecture, and training method. Figure 2(c) shows uncertainty, measured by binary entropy $\mathbb{H}(p)$ (Eq. 3), which captures variability in decision-making that tends to peak around certain decisions or sometimes spread widely depending on the confidence value and the scenario. These findings point to the unreliability of LLMs in moral contexts, which motivates us to investigate the origins of such uncertainty.

## Confidence Variation across Models

Figure 3 presents variations in confidence $\Delta p^2$ across 32 LLMs and 9 moral dimensions. Most models from the Gemma family exhibit strong confidence in their decisions, while those from the Llama family tend to show relatively weak confidence, regardless of the scenario. Confidence varies more across models within the same moral dimension (y-axis) than across dimensions for a given model (x-axis).

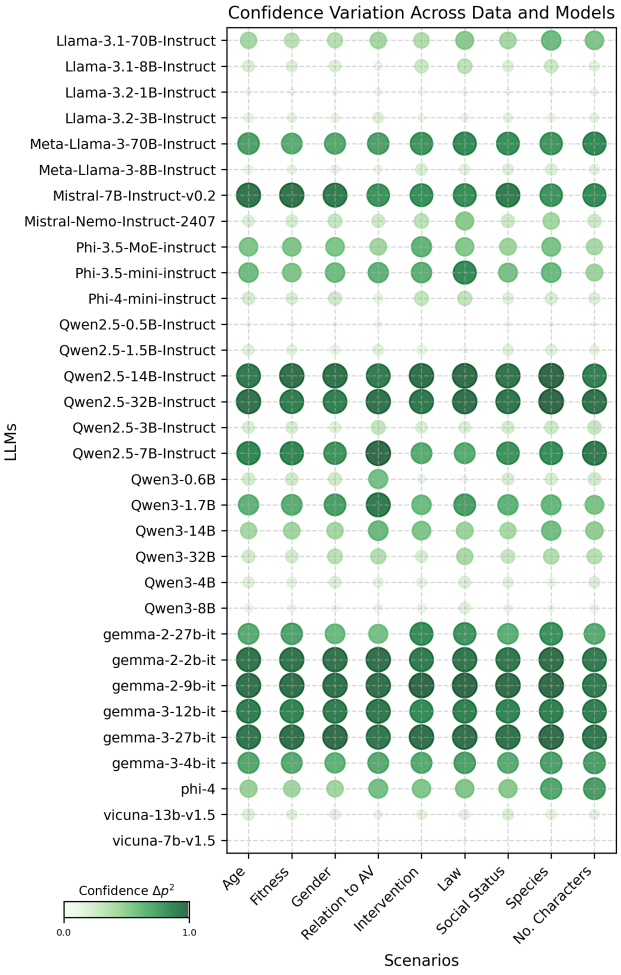


Figure 3: Confidence ($\Delta p^2$) variation by models and moral dimensions, represented by the size and color of circles. Relative uncertainty variations differ significantly across models but exhibit little difference across moral dimensions.

While most models maintain consistent confidence across scenarios, a few, such as Qwen3-0.6B, exhibit substantial variability.

These results suggest that response confidence is more influenced by model architecture and training than by the specific moral dimension of the scenario. Compared to human preferences reported in (Awad et al. 2018), LLMs exhibit lower variability. For example, humans show strong preference for saving more people in the Utilitarian scenario but high uncertainty in their preferences for the *Intervention* (action vs. inaction) scenario. In contrast, most LLMs display similar confidence levels across both dimensions.

## Impact of Dropout on Uncertainty and Alignment

We quantify Total Entropy (TE), Conditional Entropy (CE), and Mutual Information (MI) to better understand LLM decision-making. TE represents the output uncertainty (or total surprise-information), which can help detect overconfidence biases in LLMs, while CE represents the residual variability given individual scenarios. MI quantifies the input-output dependence, which can be interpreted as the explanatory information provided by the model, measuring how much the scenario information $x$ reduces the uncertainty in its decision $y$. Additionally, we employ attention dropout as a targeted injection of uncertainty in the model's decisions.

We examine the effects of dropout in Figure 4 and Table 1. Figure 4(a) shows the impact of increasing dropout rates (0, 0.05, 0.1)[1] on TE, CE, and MI. Paired t-tests confirm a statistically significant rise in TE and MI values with higher dropout (p-values: $p = 5.2e-11$ for TE, $p = 1.0000$ for

| **Dropout rate** | **0.00** | **0.05** | **0.10** |
|---|---|---|---|
| Llama-3.1-70B | 0.703 | 0.673 **(-0.03)** | ✓0.550 **(-0.15)** |
| Llama-3.1-8B | 1.570 | 1.528 **(-0.04)** | 1.264 **(-0.31)** |
| Llama-3.2-3B | 1.532 | 1.497 **(-0.04)** | 1.253 **(-0.28)** |
| Llama-3.2-1B | 1.170 | 1.262 (+0.09) | 1.291 (+0.12) |
| Meta-Llama-3-70B | 0.686 | 0.631 **(-0.06)** | ✓0.522 **(-0.16)** |
| Meta-Llama-3-8B | 0.893 | 0.847 **(-0.05)** | 0.790 **(-0.10)** |
| Mistral-Nemo-2407 | 0.810 | 0.757 **(-0.05)** | 0.735 **(-0.07)** |
| Mistral-7B-v0.2 | 0.819 | 0.707 **(-0.11)** | 0.699 **(-0.12)** |
| Phi-4 | 0.989 | 0.946 **(-0.04)** | 0.790 **(-0.20)** |
| Phi-4-mini | 1.301 | 1.189 **(-0.11)** | 0.964 **(-0.34)** |
| Phi-3.5-MoE | 1.062 | 1.066 (+0.00) | 1.039 **(-0.02)** |
| Phi-3.5-mini | 1.420 | 1.423 (+0.00) | 1.270 **(-0.15)** |
| Qwen3-32B | 1.303 | 1.272 **(-0.03)** | 1.339 (+0.04) |
| Qwen3-14B | 1.379 | 1.382 (+0.00) | 1.428 (+0.05) |
| Qwen3-8B | 1.796 | 1.733 **(-0.06)** | 1.335 **(-0.46)** |
| Qwen3-4B | 1.917 | 1.914 **(-0.00)** | 1.631 **(-0.29)** |
| Qwen3-1.7B | 1.808 | 1.663 **(-0.15)** | 1.300 **(-0.51)** |
| Qwen3-0.6B | 1.577 | 1.495 **(-0.08)** | 1.180 **(-0.40)** |
| Qwen2.5-32B | 0.937 | 0.928 **(-0.01)** | 0.816 **(-0.12)** |
| Qwen2.5-14B | 0.958 | 0.964 (+0.01) | 1.066 (+0.11) |
| Qwen2.5-7B | 1.341 | 1.351 (+0.01) | 1.358 (+0.02) |
| Qwen2.5-3B | 1.660 | 1.436 **(-0.22)** | 1.163 **(-0.50)** |
| Qwen2.5-1.5B | 1.188 | 1.187 **(-0.00)** | 1.114 **(-0.07)** |
| Qwen2.5-0.5B | 1.585 | 1.539 **(-0.05)** | 1.349 **(-0.24)** |
| Gemma-3-27b | 1.312 | 1.295 **(-0.02)** | 1.289 **(-0.02)** |
| Gemma-3-12b | 1.048 | 1.033 **(-0.02)** | 0.901 **(-0.15)** |
| Gemma-3-4b | 1.353 | 1.280 **(-0.07)** | 1.178 **(-0.17)** |
| Gemma-2-27b | 1.419 | 1.405 **(-0.01)** | 1.419 (+0.00) |
| Gemma-2-9b | 1.854 | 1.846 **(-0.01)** | 1.777 **(-0.08)** |
| Gemma-2-2b | 1.909 | 1.897 **(-0.01)** | 1.811 **(-0.10)** |
| Vicuna-13b-v1.5 | 1.196 | 1.188 **(-0.01)** | 1.204 (+0.01) |
| Vicuna-7b-v1.5 | 1.180 | 1.134 **(-0.05)** | 1.206 (+0.03) |

Table 1: Human-LLM alignment scores under different dropout rates. Numbers indicate distance from human AMCE scores $L_2$. $\Delta L_2$ are shown in parentheses relative to the baseline (0.00); bold indicates a decrease.; ✓indicates Top-2 alignment scores; underline indicates Top-2 changes.

[1]We observed only minimal changes in the overall response distributions, as quantified by Jensen–Shannon Divergence values of 0.049 and 0.071 for dropout rates of 0.05 and 0.10, respectively.

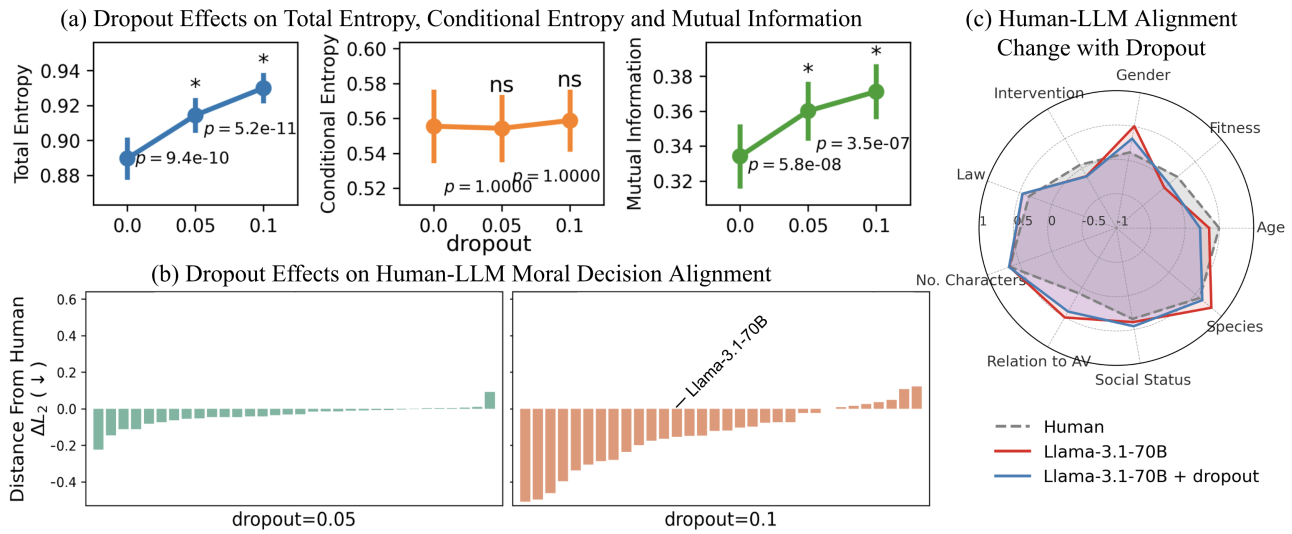


Figure 4: Dropout effects on uncertainty components and human-LLM moral alignment. (a) Effects of increasing dropout rate (0, 0.05, 0.1) on average total entropy (blue), conditional entropy (orange), and mutual information (green), with trend lines and p-values from paired t-tests (ns: non-significant; *: $p < 0.05$; two-sided, Bonferroni corrected). Error bars represent standard errors across scenario–model combinations (n=9×32). Total entropy and mutual information increase with dropout, while conditional entropy remains almost unchanged. (b) Changes in human-LLM moral decision alignment ($\Delta L_2$) for models, sorted by decreasing $\Delta L_2$ (increased alignment), at dropout rates 0.05 (left, teal bars) and 0.1 (right, orange bars). (c) Example radar chart illustrating improved Alignment with AMCE values across nine moral dimensions. Human: gray, dashed line; Llama-3.1-70B (without dropout): red, solid line; Llama-3.1-70B (with dropout=0.1), blue, solid line.

CE, $p = 9.4e-10$ for MI; Bonferroni corrected; * indicates $p < 0.05$), while CE remains largely unchanged. These results suggest that the increase in TE is primarily driven by the MI term.

Figure 4(b) and Table 1 present alignment scores, measured by $L_2$ distance across 32 models at dropout rates of 0.05 and 0.1. The analysis clearly demonstrates that applying attention dropout during the inference time significantly improves human-LLM alignment in morally complex scenarios, as most models exhibit reduced $\Delta L_2$ values. Given that LLMs are sensitive to minor prompt variations (Oh and Demberg 2025), we conducted a robustness check by paraphrasing prompts with an "Option A/B" format (instead of "Case 1/2"). The alignment gains ($\Delta L_2$ reduction) achieved through dropout were largely preserved across models, though with some variation in magnitude (e.g., Llama3.2-3B: $\Delta L_2$ from $-0.28$ to $-0.11$; Qwen2.5-3B: $\Delta L_2$ from $-0.29$ to $-0.53$). This suggests that our uncertainty injection mechanism is robust to minor changes in prompt format.

Our results offer strong evidence that reducing any uncertainty in LLM predictions brings their decisions closer to human moral preferences. To exemplify this dropout effect, Figure 4(c) presents a radar chart for Llama-3.1-70B, comparing its baseline and dropout-augmented alignments with human AMCE values across nine moral dimensions. The visualization shows improved alignment following dropout, primarily driven by a reduction in model overconfidence.

To further support our moral sensitivity analysis, we conducted a blind human evaluation of Qwen3-1.7B (i.e., the model with the largest alignment gain), comparing its outputs before and after applying dropout (100 Q/A pairs; $n = 3$ annotators). The averaged human choices aligned more closely with the post-dropout model (baseline Avg. Human Choices $\approx 0.369$ vs. dropout=0.10 Avg. Human Choices $\approx 0.263$ in Mean Squared Error terms), suggesting that the observed improvement in alignment scores is not incidental but instead captures subjective human moral preferences more faithfully.

## Mutual Information as a Driver for Alignment

We further investigate the relationship between changes in Mutual Information (MI) and shifts in moral alignment. Figure 5(a) presents scatterplots correlating scenario-averaged changes in TE, CE, and MI ($\Delta$ uncertainty) with corresponding changes in $\Delta L_2$ distance between human and LLM decisions, comparing dropout rates of 0.05 (blue) and 0.1 (orange). Statistical analysis reveals a positive correlation between MI and dropout, while the TE and CE terms show no significant associations.

Figure 5(b) illustrates model-wise trajectories from the baseline (dropout=0) to dropout rates of 0.05 and 0.1, plotting changes in $\Delta$MI against $\Delta L_2$. Each trajectory, shown in gray, indicates that models with greater increases in MI tend to achieve larger reductions in $\Delta L_2$ and improved alignment scores. This trend suggests that the alignment shift induced by dropout is driven by changes in the MI term.

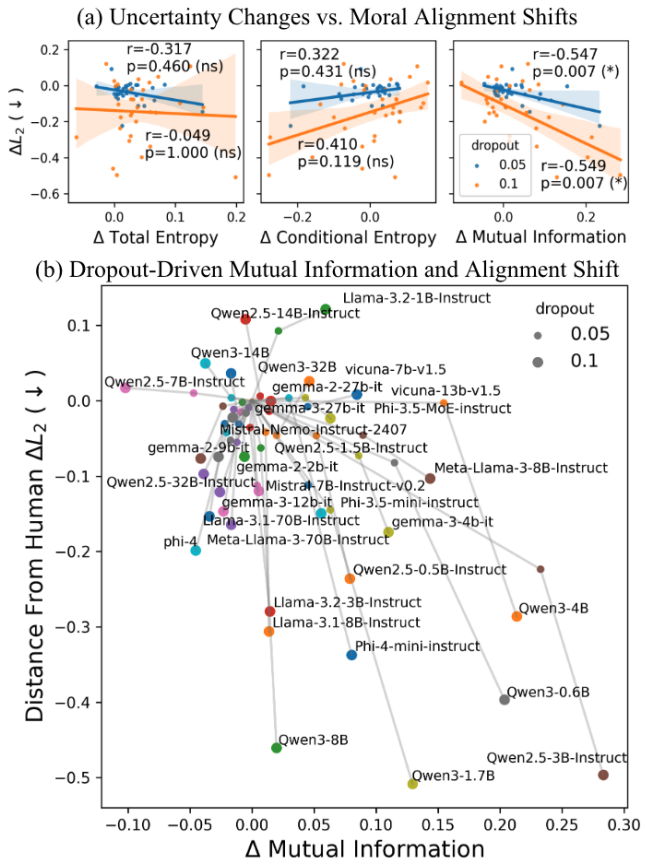


Figure 5: Uncertainty changes predict shifts in moral decision alignment under dropout. (a) Scatterplots of $\Delta$ uncertainty components (scenario-averaged) vs. $\Delta L_2$ between human and LLM. Per-model points at dropout=0.05 (blue) and 0.1 (orange). Pearson $r$, and Bonferroni-corrected $p$-values are shown at the top of the figures. (b) Model-wise trajectories of $\Delta$ mutual information vs. $\Delta L_2$ from dropout 0$\rightarrow$0.05 and 0$\rightarrow$0.1; gray lines connect points, showing larger mutual information increases link to better alignment.

## Discussion

Our results show that artificially increasing mutual information (MI) via dropout leads to a closer alignment between LLM decisions and human consensus judgments, significantly improving alignment scores in complex moral dilemma benchmarks. However, whether such alignment should be prioritized over alternative design objectives remains an open question. Indeed, our findings point to a deeper issue concerning the objective of human–machine alignment in morally ambiguous domains: *Should designers uphold systems that faithfully replicate human behavior, or should we seek systems capable of "transcending human limitations" to achieve superior decisions in specific scenarios?* This tension between mirroring human judgments and enabling more principled, model-guided responses is at the core of ongoing debates about the normative role of artificial systems in ethical decision-making.

Our experiments also revealed substantial variation in uncertainty patterns across models (Figures 2 and 3), with some exhibiting excessive confidence in their responses. This disparity likely stems from differences in training data and methodology, suggesting the presence of inherent biases in LLMs when navigating ethically complex scenarios. On a related note, (Cheung, Maier, and Lieder 2025) demonstrated that fine-tuning can amplify omission bias, which makes models more likely to abstain from action in dilemma situations. This raises important concerns about robustness. Moreover, LLMs have been shown to be sensitive to minor changes in prompts, even within the same scenario (Oh and Demberg 2025). Collectively, these observations highlight that robustness and uncertainty behavior are central to moral alignment, not merely secondary to aggregate agreement with human choices.

In our setup, uncertainty was artificially induced by dropout at inference time. While this mechanism improves alignment scores, it does not guarantee a better alignment with human preferences. As shown in Figure 4, inducing uncertainty can reduce alignment in specific scenario dimensions, such as Age and Species. These findings suggest the need for more diverse alignment metrics when evaluating moral dilemmas. From a training perspective, uncertainty in moral scenarios should ideally be learned from the data used to train these models and reflect cultural norms, rather than being injected ad hoc at inference time.

Our results indicate increases in total entropy (TE) and mutual information (MI) introduce additional risk into decision-making: when LLMs rely on stochastic sampling, even improved alignment scores may come at the cost of higher variance in their choices, which can be undesirable in certain high-stakes dilemma scenarios. This, in turn, prompts a reconsideration of how we want such models to behave relative to humans. In some contexts, it may be preferable for machines to deviate from human-like patterns of judgment, with ideal behavior instead prioritizing minimized hallucinations and enhanced safety, while avoiding dropouts in confidence.

## Conclusion

This study introduced a new information-theoretic approach to examine moral uncertainty in LLMs by exploring their alignment with human moral judgments within the Moral Machine framework. By applying dropout during inference to amplify uncertainty, we investigated its impact on 32 models from 6 leading open-source families, observing a general improvement in alignment scores, albeit with variation across models. Crucially, our findings reveal that higher uncertainty is correlated with improved alignment scores, demonstrating that reducing overconfidence in LLM decisions can produce machine outcomes that are more consistent with human ethical intuitions. These findings support the value of developing uncertainty-aware machines that better represent the nuanced variability in human moral reasoning. However, increased uncertainty also brings greater variability in decisions, which may pose risks in critical scenarios. Addressing this trade-off will be essential for building safer and more transparent AI systems.

## Acknowledgments

The authors thank Christoph Engel, Jaehong Kim, and anonymous reviewers for their insightful feedback. Jea Kwon and Meeyoung Cha are the co-corresponding authors.

## References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024a. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Abdin, M.; Jacobs, S. A.; Awan, A. A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; Benhaim, A.; Bilenko, M.; Bjorck, J.; Bubeck, S.; Cai, M.; Mendes, C. C. T.; Chen, W.; Chaudhary, V.; Chopra, P.; Giorno, A. D.; de Rosa, G.; Dixon, M.; Eldan, R.; Iter, D.; Garg, A.; Goswami, A.; Gunasekar, S.; Haider, E.; Hao, J.; Hewett, R. J.; Huynh, J.; Javaheripi, M.; Jin, X.; Kauffmann, P.; Karampatziakis, N.; Kim, D.; Khademi, M.; Kurilenko, L.; Lee, J. R.; Lee, Y. T.; Li, Y.; Liang, C.; Liu, W.; Lin, E.; Lin, Z.; Madan, P.; Mitra, A.; Modi, H.; Nguyen, A.; Norick, B.; Patra, B.; Perez-Becker, D.; Portet, T.; Pryzant, R.; Qin, H.; Radmilac, M.; Rosset, C.; Roy, S.; Ruwase, O.; Saarikivi, O.; Saied, A.; Salim, A.; Santacroce, M.; Shah, S.; Shang, N.; Sharma, H.; Song, X.; Tanaka, M.; Wang, X.; Ward, R.; Wang, G.; Witte, P.; Wyatt, M.; Xu, C.; Xu, J.; Yadav, S.; Yang, F.; Yang, Z.; Yu, D.; Zhang, C.; Zhang, C.; Zhang, J.; Zhang, L. L.; Zhang, Y.; Zhang, Y.; Zhang, Y.; and Zhou, X. 2024b. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The moral machine experiment. *Nature*, 563(7729): 59–64.

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8): e2416228122.

Chen, Y.; Raghuram, V. C.; Mattern, J.; Mihalcea, R.; and Jin, Z. 2022. Causally testing gender bias in llms: A case study on occupational bias. *arXiv preprint arXiv:2212.10678*.

Cheung, V.; Maier, M.; and Lieder, F. 2025. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25): e2412015122.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica, I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cushman, F. 2013. Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, 17(3): 273–292.

Fan, X.; Zhang, S.; Chen, B.; and Zhou, M. 2020. Bayesian attention modules. *Advances in Neural information processing systems*, 33: 16362–16376.

Gabrié, M.; Manoel, A.; Luneau, C.; Macris, N.; Krzakala, F.; Zdeborová, L.; et al. 2018. Entropy and mutual information in models of deep neural networks. *Advances in Neural information processing systems*, 31.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Greene, J. D.; Sommerville, R. B.; Nystrom, L. E.; Darley, J. M.; and Cohen, J. D. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537): 2105–2108.

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

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jin, Z.; Kleiman-Weiner, M.; Piatti, G.; Levine, S.; Liu, J.; Gonzalez, F.; Ortu, F.; Strausz, A.; Sachan, M.; Mihalcea, R.; et al. 2024. Language model alignment in multilingual trolley problems. *arXiv preprint arXiv:2407.02273*.

Kahneman, D.; and Tversky, A. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, 99–127. World Scientific.

Kim, J.; Kwon, J.; Vecchietti, L. F.; Oh, A.; and Cha, M. 2025. Exploring persona-dependent llm alignment for the moral machine experiment. *arXiv preprint arXiv:2504.10886*.

Oh, S.; and Demberg, V. 2025. Robustness of large language models in moral judgements. *Royal Society Open Science*, 12(4): 241229.

Pei, J.; Wang, C.; and Szarvas, G. 2022. Transformer uncertainty estimation with hierarchical stochastic attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11147–11155.

Qwen, A. Y.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint*.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Sun, F.; Li, N.; Wang, K.; and Goette, L. 2025. Large Language Models are overconfident and amplify human bias. *arXiv preprint arXiv:2505.02151*.

Takemoto, K. 2024. The moral machine experiment on large language models. *Royal Society open science*, 11(2): 231393.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157): 1124–1131.

Van Bavel, J. J.; Robertson, C. E.; Del Rosario, K.; Rasmussen, J.; and Rathje, S. 2024. Social media and morality. *Annual review of psychology*, 75(1): 311–340.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Xu, C.; Wen, B.; Han, B.; Wolfe, R.; Wang, L. L.; and Howe, B. 2025. Do Language Models Mirror Human Confidence? Exploring Psychological Insights to Address Overconfidence in LLMs. *arXiv preprint arXiv:2506.00582*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, Y.; Liu, X.; Jin, Q.; Huang, F.; and Lu, Z. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications medicine*, 4(1): 176.

Zaim bin Ahmad, M. S.; and Takemoto, K. 2025. Large-scale moral machine experiment on large language models. *PloS one*, 20(5): e0322776.