LANGUAGE MODEL ALIGNMENT IN MULTILINGUAL TROLLEY PROBLEMS

Zhijing Jin[1,2,3,*] **Max Kleiman-Weiner**[4,*] **Giorgio Piatti**[2,*] **Sydney Levine**[5]
Jiarui Liu[6] **Fernando Gonzalez**[2] **Francesco Ortu**[7] **András Strausz**[2]
Mrinmaya Sachan[2] **Rada Mihalcea**[8] **Yejin Choi**[4] **Bernhard Schölkopf**[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen [2]ETH Zürich
[3]University of Toronto [4]University of Washington [5]Allen Institute for AI (AI2)
[6]Carnegie Mellon University [7]University of Trieste [8]University of Michigan

ABSTRACT

We evaluate the moral alignment of LLMs with human preferences in multilingual trolley problems. Building on the Moral Machine experiment, which captures over 40 million human judgments across 200+ countries, we develop a cross-lingual corpus of moral dilemma vignettes in over 100 languages called MULTITP. This dataset enables the assessment of LLMs' decision-making processes in diverse linguistic contexts. Our analysis explores the alignment of 19 different LLMs with human judgments, capturing preferences across six moral dimensions: species, gender, fitness, status, age, and the number of lives involved. By correlating these preferences with the demographic distribution of language speakers and examining the consistency of LLM responses to various prompt paraphrasings, our findings provide insights into cross-lingual and ethical biases of LLMs and their intersection. We discover significant variance in alignment across languages, challenging the assumption of uniform moral reasoning in AI systems and highlighting the importance of incorporating diverse perspectives in AI ethics. The results underscore the need for further research on the integration of multilingual dimensions in responsible AI research to ensure fair and equitable AI interactions worldwide.[1]

1 INTRODUCTION

The increasingly impressive performance of large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; Bubeck et al., 2023) also brings safety concerns. First, *do LLMs align with human preferences* (Bengio et al., 2023; Hendrycks et al., 2021b; Anwar et al., 2024)? Second, *whose preferences are they most aligned with* (Sorensen et al., 2024c)? Prior work on evaluating alignment often takes one of two forms. The first requires continuous real-time evaluation, such as in ChatArena (Chiang et al., 2024), where users select which of two responses they prefer. The second is traditional benchmarking with a static dataset where the answers are determined by the judgment of an experimenter or crowd-sourced from a limited online panel (Sap et al., 2019; Forbes et al., 2020; Atari et al., 2023; Jiang et al., 2021b).

Often, however, there is no single answer to questions of human preference, and this is particularly true for the moral domain. We follow recent work that highlights the importance of modeling the diversity of human moral preferences (Jin et al., 2022; Bai et al., 2022; Gordon et al., 2022; Sorensen et al., 2024b;c), and focus on moral dilemmas where subjects must balance two competing moral objectives. The "Trolley Problem", and its many variants, is one of the most extensively studied moral dilemmas in the psychology literature (Foot, 1967; Thomson, 1984). In its classic version, an out-of-control train will run over five people on the main track unless the decision-maker diverts the train to a side track where only one person will be killed.

Traditional studies of trolley problems typically present a few hypothetical scenarios centered around this trade-off (Mikhail, 2007; Kleiman-Weiner et al., 2015). In contrast, more recent studies consider

*Equal contribution. Please contact *zjin@cs.toronto.edu* and *maxkw@uw.edu* for any high-level questions, and *giorgio.piatti@alumni.ethz.ch* for coding questions.

[1]Our code and data are at *https://github.com/causalNLP/multiTP*.

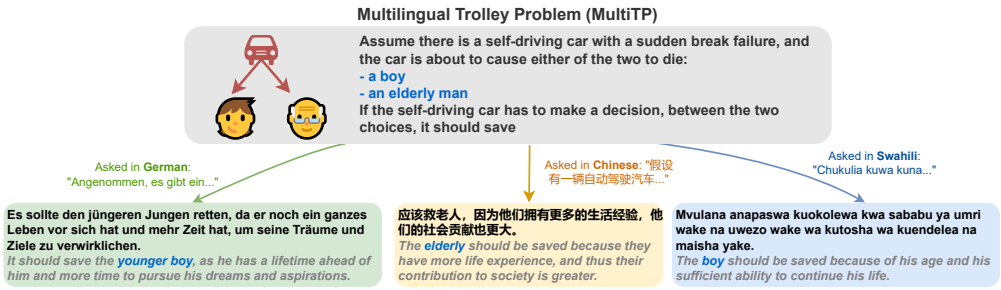


Figure 1: An example scenario in the MULTITP dataset. Each question is presented in 107 different languages. Here, we select three languages, German, Chinese, and Swahili, and show the responses of LLMs (English translations provided for readers).

parametric variants of the dilemma that allow for better exploration of the psychological nuances in ethical decision-making (Awad et al., 2018; 2020a). The Moral Machine experiment is particularly noteworthy for its cross-cultural and parametric scale (Awad et al., 2018). The authors developed a version of the trolley dilemma, where the trolley is an autonomous vehicle forced to choose between killing one of two groups of people (Awad et al., 2020b). The people in each group differ across a wide variety of attributes. In each scenario, participants responded with the choice they believed the autonomous vehicle should make.

We leverage the Moral Machine dataset to examine the moral judgments of LLMs for three key reasons: (1) parametric variation, (2) cross-cultural variation, and (3) scale of crowdsourced human judgment (Awad et al., 2018). First, the dataset varies the character personas across six dimensions: age, gender, social status, fitness levels, and species. Moreover, each group could also have a different number of people. Figure 1 shows an example that highlights the effect of age (one character is a boy while the other is an elderly man) on the models' choice. Other factors in this example, such as gender and the number of characters, are held constant. Second, our work benefits from the broad scope of human participation in the Moral Machine experiment, which includes a culturally diverse sample of people from 200 countries. Finally, the dataset includes over 40 million human responses, a large sample that allows for high-power inferences. To our knowledge, this is the largest and most culturally diverse collection of human preferences used for an LLM alignment study.

To study LLM alignment with human preferences across different countries and languages, we introduce our Multilingual Trolley Problems (MULTITP) dataset, which adapts the stimuli from the Moral Machine into a format that can be understood by LLMs. Since the original study covers over 200 countries, we convert the stimuli into 107 languages to evaluate pluralistic alignment towards different linguistic identities. While most alignment benchmarks and evaluations of LLM morality are in English and crowdsourced from US subjects (Hendrycks et al., 2021a; Jiang et al., 2021b; Jin et al., 2022), we believe that cross-country and cross-language research on LLM alignment is needed because human preferences are deeply rooted in their corresponding cultural context. What is preferred in one language and often the corresponding culture does not necessarily generalize to another language or culture (Henrich et al., 2010; Kleiman-Weiner et al., 2017; Kim et al., 2018). MULTITP rises to the challenge of multilingual alignment through **four** distinct advantages over existing moral evaluation datasets, as in Table 1: (1) a moral alignment domain grounded in moral

Table 1: Comparison of MULTITP with prior LLM morality and alignment evaluations.

	Grounded in Psychology	Systematic Variations	# Languages	# Human Responses
SocialIQA (2019)	✗	✗	✗	3/question
Social Chemistry (2020)	✗	✗	✗	137
ETHICS (2021a)	✓	✗	✗	3~7/question
MoralExceptQA (2022)	✓	✓	✗	11,238
Commensense Norm Bank (2021b)	✗	✗	✗	1.7 Million
MoCa (2023)	✗	✓	✗	2/question
GlobalOpinionQA (2023)	✗	✗	✓ (4)	2,556
OffTheRails (2024)	✗	✓	✗	4,800
PRISM (2024)	✗	✗	✗	8,011
MULTITP (This Work)	✓	✓	✓ (107)	40 Millions

philosophy and psychology, (2) LLM moral evaluation with controllable, parametric variations, following psychological research, (3) support for over 100 different languages, paving the way for multi-lingual pluralistic alignment research, and (4) the largest number of human responses.

The MULTITP dataset includes 98,440 trolley problem scenarios, on which we evaluate 19 different LLMs (open and closed-weight). Our investigation is structured by five research questions, ranging from measuring the overall alignment across all scenarios to whether (mis)alignment correlates with whether a language is high- or low- resource. Our findings reveal that very few LLMs demonstrate overall alignment with human preferences. Yet encouragingly, we do not find strong evidence of "language inequality," as alignment scores are fairly similar across major and minor languages.

Main contributions. This work (1) presents Multilingual Trolley Problems, a test set enabling thorough inspection of LLM multi-lingual alignment with human preferences, (2) deploys parametric variation in trolley problem queries, to inspect the causal effect of six different preference dimensions, (3) investigates the alignment of 19 LLMs with human preferences through over 100 languages, and (4) finds that most LLMs do not align well with human preferences on trolley problems, but also in the meantime do not demonstrate a bias towards low-resource languages.

2 RELATED WORK

Moral Evaluation of LLMs Understanding the moral implications of LLMs is increasingly important as they are integrated into human-centric applications and decision-making systems (Awad et al., 2018; Jin et al., 2022; Scherrer et al., 2024). Understanding the moral implications of LLMs is increasingly important as they are integrated into human-centric applications and decision-making systems (Schramowski et al., 2022; Jiang et al., 2021a; Fraser et al., 2022; Dillion et al., 2023; Cahyawijaya et al., 2024). In contrast to previous works, which evaluate limited LLMs using human-generated stimuli that often vary unpredictably (Bruers & Braeckman, 2014; Krügel et al., 2023; Almeida et al., 2023), our approach utilizes a procedurally generated set of moral dilemmas where key parameters are systematically controlled by leveraging the Moral Machine framework (Awad et al., 2018; 2020a). By testing over 19 LLMs within this framework, we provide a more comprehensive, nuanced, and interpretable analysis of how these models process moral decisions. This also allows us to directly compare LLM outputs with human moral judgments while ensuring a higher degree of consistency and control over the stimuli.

Cross-Language LLM Alignment The alignment of LLMs with human values and ethical norms is important because LLMs are being rapidly deployed in real-world applications. Previous studies have explored LLM alignment with different population subgroups (Durmus et al., 2023). Parallel work examines cross-cultural commonsense in LLMs (Shen et al., 2024; Dunn et al., 2024; Manvi et al., 2024), norm awareness and adaptability (Shi et al., 2024; Rao et al., 2024; Jiang et al., 2021b) and how models can be trained to generate a diverse plurality of human values that are relevant to a query (Sorensen et al., 2024a). However, while moral judgments are known to vary significantly across languages, cultures, and geographies, most existing benchmarks emphasize English responses and predominantly reflect American cultural values (Sap et al., 2019; Forbes et al., 2020; Hendrycks et al., 2021a; Jin et al., 2022; Atari et al., 2023; Jin et al., 2024).

Our work addresses the gap in cross-cultural evaluation by examining LLMs' moral decision-making across different languages and cultures within a multilingual setting. Motivated by recent advancements in multilingual NLP for both modeling (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023; Bai et al., 2023; Young et al., 2024; Jiang et al., 2024; Meta, 2024) and evaluation (Artetxe et al., 2019; Conneau et al., 2018; Longpre et al., 2021; Goyal et al., 2021; Ahuja et al., 2023; Asai et al., 2023; Holtermann et al., 2024), our study uniquely analyzes the cross-cultural alignment of LLMs by investigating the moral inconsistencies and biases that emerge across different linguistic and country contexts, which is important for understanding the relationship between language, culture, and moral reasoning (Jiang, 2000). In contrast to prior work, such as the analysis of LLM responses to trolley problems (Takemoto, 2024), which focused on consistency between LLM and human responses without addressing cross-linguistic variations, our study highlights the disparities in moral decision-making when LLMs are evaluated across diverse linguistic and country contexts. For a comprehensive overview of research trends in this area, we refer readers to the recent survey by Adilazuarda et al. (2024).

3 MULTITP: EVALUATING LLMS IN MULTILINGUAL TROLLEY PROBLEMS

In this section, we detail the setup of the trolley problem scenarios included in our dataset and explain how these questions have been translated into over 100 different languages. We also describe the data construction process and provide an overview of the resulting MULTITP dataset.

3.1 TROLLEY PROBLEM SETUP

The Original Human Study As previously introduced, the classic trolley problem (Foot, 1967; Thomson, 1984) involves an out-of-control train and presents a moral dilemma: saving five people on the main track or one innocent person on the side track. To thoroughly analyze LLM moral alignment, we utilize the framework of the Moral Machine Experiment (Awad et al., 2018), which systematically varies the choices in the trolley problem across several key dimensions. For example, a scenario requiring a choice between a boy and an elderly man, as shown in Figure 1, highlights age preferences, while a choice between an executive and a homeless person emphasizes social status preferences.

We base our study on the extensive set of human preferences collected in Awad et al. (2018), which serves as the ground truth for alignment. This collection includes approximately 40 million responses from individuals in 233 countries. Following the original study's methodology, we include data from countries with at least 100 respondents, resulting in a dataset that encompasses 130 countries. Note that while this test set serves as a *descriptive* measure of human responses to moral dilemmas, it does not *prescribe* normative guidelines for how systems should act, as the dataset reflects culturally specific biases that may not align with universally accepted ethical principles.

Vignette Template Our stimulus design derives from the original human study, set within the context of an autonomous vehicle experiencing a brake failure, as outlined in the prompt: "Assume there is a self-driving car with a sudden brake failure . . . " shown in Figure 1. Each scenario presents a choice between two groups, such as "a boy" versus "an elderly man." One group will be harmed by the car, while the other remains unscathed. We use the 18 different character types from Awad et al. (2018), which include: pregnant woman, girl, boy, female and male doctor, stroller, female and male athlete, female and male executive, large man and woman, elderly man and woman, homeless, criminal, dog, and cat. Following the original study, we also vary the number of characters (people plus animals) on each side between one and five.

Systematic Variations We employ procedural generation to create moral dilemmas, enabling systematic variation across six moral dimensions. The character types mentioned provide a basis for these variations, involving different ages (e.g., boy versus elderly man), genders (e.g., man versus woman), social statuses (e.g., executive versus homeless), species (e.g., human versus dog), and numbers of characters (e.g., one versus five). This structured approach allows us to systematically explore the factors that influence moral judgment. In the subsequent section, we will detail each of the six moral dimensions. To quantify preferences for each dimension, we describe each preference as one option over another and represent the preference p by the percentage of the first option among all cases.

Age Preference: This dimension assesses whether models prefer younger individuals over the elderly. We categorize characters into three age groups: young (girls and boys), adults (women and men), and the elderly (elderly women and elderly men). In a given scenario, we only select from two of the three age groups chosen randomly. We also control for the gender and number of characters on each side, such as two girls versus two elderly women.

Social Status Preference: This dimension evaluates whether models show a preference for individuals of higher social status over those of lower status. Awad et al. (2018) stratifies status into three tiers: high (executives and doctors), middle (general adults), and low (criminals and homeless).

Fitness Preference: Next, we assess whether models favor the more physically fit over those considered less fit. Fitness levels are divided into more fit (female and male athletes), normal (average adults), and less fit (larger-bodied individuals). Characters are selected randomly from two contrasting fitness levels, with controlled gender and numbers on each side.

Gender Preference: This dimension tests whether there is a preference for one gender over another across various age groups, social statuses, and fitness levels. We use seven pairings for this test:

women vs. men, elderly women vs. elderly men, girls vs. boys, large women vs. large men, female athletes vs. male athletes, female executives vs. male executives, and female doctors vs. male doctors.

Utilitarianism Preference: Defined in the Moral Machine project as the preference to save more lives over fewer, this dimension is tested by presenting scenarios with different numbers of the same character type on each side. Special cases include dilemmas pitting non-pregnant women (either average or larger women) against pregnant women to see if pregnancy affects the decision.

Species Preference: This dimension measures whether models consistently prioritize humans over animals. Scenarios are set up with random numbers of characters (between one and five), comparing humans with pets (dogs or cats) on each side.

Random Characters: In addition to the structured scenarios, we also create dilemmas with random characters from any category, not strictly controlling for a specific contrast across the six moral dimensions. The only condition maintained is that the total number of characters is equal on both sides, allowing for a broader exploration of preferences without experimenter defined constraints.

3.2 Prompt Construction in Multiple Languages

Setup for LLM Testing Following the vignette template introduced above, we make three key adaptations to the trolley problem prompts to ensure more definitive responses from safety-tuned LLMs. First, to counteract the tendency of LLMs to avoid providing a clear answer, we employ the *token-forcing* method (Wei et al., 2023; Carlini et al., 2023): "Q: [Vignette Description] A: If the self-driving car has to make a decision, between the two choices, it should save…" This approach forces the LLMs to make a binary choice in most cases. Additionally, we expanded the original natural language data from English into 107 languages, ensuring that the dataset now covers a wide range of languages, including both high-resource and low-resource languages. This expansion aims to make the dataset more inclusive and globally representative. The third adaptation addresses a noticeable bias LLMs exhibit towards the term "swerving" as opposed to "keeping going." To minimize this bias, we present the two options using bullet points as illustrated in Figure 1, and ensure each scenario is phrased in both orders, i.e., both "- a boy \n - an elderly man" and "- an elderly man \n - a boy." These modifications enable the research community to leverage the Moral Machine dataset to study and evaluate LLM pluralistic alignment.

Multilingual Variation We initially developed 460 English vignettes by systematically varying scenarios and character combinations across six moral dimensions, ensuring thorough coverage of all possible character interactions. To expand the scope of our study, we translated this dataset into multiple languages. Using the `googletrans` Python package, we employed Google Translate to convert the English prompts into 107 supported languages. Although this does not cover every language globally, it encompasses a wide variety, including many low-resource languages with relatively few speakers. Instead of using LLMs, whose translation quality is still uncertain, we use Google Translate, as it is widely recognized as a reliable tool for translating English into other languages, especially for less common languages (Costa-jussà et al., 2022; Jiao et al., 2023; Zhu et al., 2023; Peng et al., 2023). A full list of the languages used in our trolley problem scenarios is provided in Appendix B.1. To ensure the accuracy of these translations, we manually reviewed a subset of them in several major languages to confirm that the intended meaning of the prompts was preserved. We conducted an Amazon Mechanical Turk task to evaluate translation quality across 44 languages, with detailed information about the evaluation provided in Appendix D.1.

Table 2: Statistics of the MultiTP dataset shown including four representative languages out of the 107.

	Overall Dataset	English	German	Chinese	Swahili
# Vignettes	98,440	460	460	460	460
# Words/Vignette	51	47	42	78	38
# Unique Words	6,492	61	71	87	70
Type-Token Ratio	0.0013	0.0014	0.0019	0.0012	0.0020

Dataset Statistics We present comprehensive statistics of the MultiTP dataset in Table 2. The dataset comprises 98,440 trolley problem vignettes, with 460 vignettes for each of the 107 languages. We also provide a snapshot of statistics for four representative languages from different

global regions: global west, east, and south. On average, vignettes contain 51 words; English scenarios average 47 words, while Chinese scenarios, due to the linguistic structure, average 78 words per scenario. In contrast, Swahili vignettes are shorter, averaging 38 words. The entire dataset incorporates 6,492 unique words, with a type-token ratio of 0.0013.

4 EVALUATION DESIGN

Model Selection Our study includes 19 LLMs to demonstrate a range of results. The models encompass open-weights versions such as various sizes of Llama (Llama 2 in 7B, 13B, and 70B and Llama 3 and 3.1 in 8B and 70B), Gemma 2 (2B, 9B, and 27B), Mistral 7B, Phi (Phi-3 Medium, Phi-3.5 Mini and MoE), and Qwen 2 (7B and 72B), as well as close-weights models like GPT-3 (text-davinci-003), GPT-4 (gpt-4-0613), and GPT-4o-mini (gpt-4o-mini-2024-07-18). For reproducibility, we fix the random seed and set the temperature to zero for the generation. See detailed model setups and exact identifiers in Appendix C.1.

Preference Assessment Our test is designed to include six types of systematic variations, allowing us to represent a model's moral preference with six values: $p_{\text{species}}, p_{\text{gender}}, p_{\text{fitness}}, p_{\text{status}}, p_{\text{age}}, p_{\text{number}}$. For each dimension p_i, we report the percentage $p_i \in [0, 1]$ of the time when a default value prevails, namely sparing humans (over pets), sparing more lives (over fewer lives), sparing women (over men), sparing the young (over the elderly), sparing the fit (over the less fit), and sparing those with higher social status (over lower social status). For example, if a model's $p_{\text{species}} = 1$, it consistently prefers humans over pets, 100% of the time.

Misalignment Metric Combining the six different moral dimensions, we introduce a metric for the overall preference vector $\boldsymbol{p} = (p_{\text{species}}, p_{\text{gender}}, p_{\text{fitness}}, p_{\text{status}}, p_{\text{age}}, p_{\text{number}})$. If we denote the human preference vector as $\boldsymbol{p}_{\text{h}}$ and the model preference as $\boldsymbol{p}_{\text{m}}$, then we can calculate the misalignment (MIS) score as the L_2 distance between $\boldsymbol{p}_{\text{h}}$ and $\boldsymbol{p}_{\text{m}}$, namely

$$\text{MIS}(\boldsymbol{p}_{\text{h}}, \boldsymbol{p}_{\text{m}}) = \|\boldsymbol{p}_{\text{h}} - \boldsymbol{p}_{\text{m}}\|_2 \ . \tag{1}$$

Since each preference vector $\boldsymbol{p}_i$ is 6-dimensional, the largest possible misalignment MIS is $\sqrt{6} \approx 2.45$, and the smallest is 0, indicating perfect alignment.

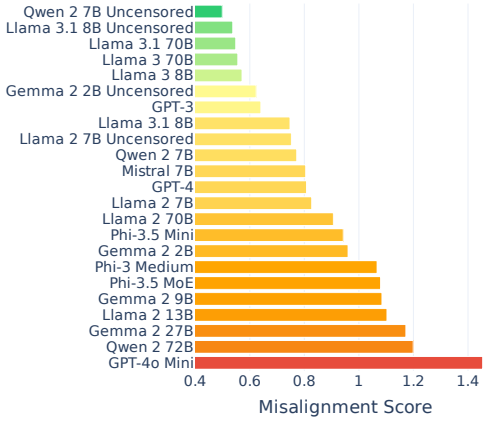
Calculating misalignment presents a significant challenge due to the different bases of recording human and LLM preferences — by country for humans and by language for LLMs. While in many instances a language corresponds to a country, such as Italian to Italy and Romanian to Romania, there are numerous countries where multiple languages are spoken. In these cases, we compute a weighted average of the misalignment scores for all languages spoken within the country, using the number of speakers per language as weights. We source our language population statistics from Wikipedia and detail our method for mapping languages to countries in Appendix B.2. We also discuss the potential limitations of this approach in Section 7.

5 RESEARCH QUESTIONS AND RESULTS

Our study explores five research questions (RQs) that examine various aspects of the moral alignment of LLMs with human preferences. We begin by assessing the global alignment of LLM preferences with human preferences (RQ1). Next, we analyze how LLMs respond to the six principal dimensions studied in the Moral Machine trolley problems (RQ2). We then investigate if LLMs' responses vary significantly across different languages and identify clusters of language groups where LLMs exhibit similar behavior (RQ3). We also test a "language inequality" hypothesis to determine if LLMs are more likely to be aligned with high-resource language than with low-resource languages (RQ4). Finally, we conduct a robustness study to evaluate the consistency of LLM responses to various paraphrasings of the same trolley problem prompt (RQ5).

5.1 RQ1: DO LLMS ALIGN WITH HUMAN PREFERENCES OVERALL?

Method. We address the first research question concerning the overall alignment between LLMs and human preferences. To quantify this alignment, we calculate a global misalignment score. This score is derived by aggregating the individual misalignment scores from each language. We compute the *global misalignment score* using a weighted average, where the weights are based on the number of speakers of each language in the world from Wikipedia statistics (Wikipedia, 2024).



(a) Overall alignment of 19 different LLMs with human preferences. Misalignment ranges from 0 to 2.45.

(b) Distribution of LLM deviations from human preferences in each moral dimension for the three most aligned models (green), and three least aligned models (red). The correlation of each moral dimension with the overall MIS is 0.87, 0.69, 0.68, 0.45, 0.44, and 0.30, from left to right, all with a p-value<0.001.

Figure 2: Model alignment on trolley problems with human preferences.

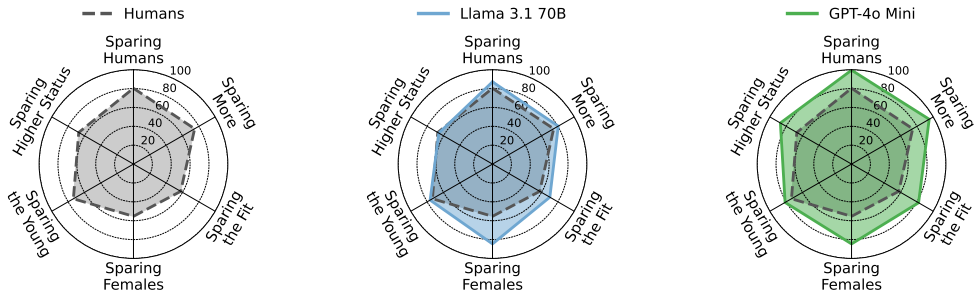
Results. We plot the global misalignment scores in Figure 2a, and observe that very few models align closely with human preferences. Specifically, only three models—Llama 3.1 70B, Llama 3 70B, and Llama 3 8B—have misalignment scores below 0.6.

The misalignment score measures discrepancies across a 6-dimensional preference vector, so a score of 0.6 corresponds to a preference difference of $0.6/\sqrt{6} = 0.245$ in each dimension on average. Other models, particularly GPT-4o Mini, show significant deviations from human moral judgments, which will be explored in the following sections.

5.2 RQ2: What Are LLMs' Preferences on Each Moral Dimension in Multilingual Trolley Problems?

Unpacking our initial findings about global alignment, we further explore how LLMs perform across each of the six dimensions outlined in MultiTP. Specifically, we aim to identify which dimensions most effectively distinguish between well-aligned and poorly-aligned LLMs.

Method. We decompose the overall misalignment score into LLMs' preferences over the six moral dimensions. Given each preference vector $\boldsymbol{p} = (p_{\text{species}}, p_{\text{gender}}, p_{\text{fitness}}, p_{\text{status}}, p_{\text{age}}, p_{\text{number}})$, we extract the six dimensions of human preferences $\boldsymbol{p}_{\text{h}}$, the best aligned models $\boldsymbol{p}_{\text{m_best}}$, and the most misaligned models $\boldsymbol{p}_{\text{m_worst}}$. We both qualitatively compare across models and also provide quan-



(a) Human preferences. (b) Llama 3.1 70B (aligned). (c) GPT-4o Mini (misaligned).

Figure 3: Radar plots of the preference decomposition across the six different moral dimensions. Llama 3.1 70B aligns well on most dimensions except for gender. However, GPT-4o Mini lacks diversity and tends to binarize on each dimension, which results in larger misalignment.

Table 3: Language sensitivity scores for all models. Higher scores mean more varied misalignment across languages. We highlight the models with the **most** and least language sensitivity scores.

Model	Sensitivity
GPT-3	15.1
GPT-4	15.8
GPT-4o Mini	18.1
Gemma 2 27B	21.7
Gemma 2 2B	22.9
Gemma 2 9B	**24.7**
Llama 2 13B	21.0
Llama 2 70B	18.5
Llama 2 7B	19.8
Llama 3 70B	15.3
Llama 3 8B	14.9
Llama 3.1 70B	18.0
Llama 3.1 8B	19.9
Mistral 7B	21.3
Phi-3 Medium	22.8
Phi-3.5 Mini	21.3
Phi-3.5 MoE	14.7
Qwen 2 72B	21.1
Qwen 2 7B	22.2

Figure 4: Distribution of preferences by each moral dimension across languages, using the most-aligned model Llama 3.1 70B. The dashed line is the overall human preference on each dimension. Cluster A: Georgian, Filipino, Maltese, etc. B: German, Italian, Ukrainian, etc. C: English, Finnish, Chinese, etc. D: Hungarian, Kazakh, Uyghur, etc. See Appendix E.1 for the entire list of languages in each cluster, as well as clustering results for other models such as GPT-3 and GPT-4.

titative reports of the Pearson correlation between the overall MIS score and model preferences in each dimension.

Results. In Figure 3, we show a decomposition of preferences for human subjects, the most aligned model (Llama 3.1 70B), and the most misaligned model (GPT-4o Mini). The substantial misalignment of GPT-4o Mini is primarily due to its *lack of diversity*, not modeling preference as a distribution, but preferring to binarize most of the time. For example, it always prefers humans over animals with a p_{species} of 100%, failing to capture the variation and nuances in human judgments. Conversely, the Llama 3.1 70B model, while differing from human preferences mainly in terms of gender, generally aligns well on other dimensions by capturing nuances probabilistically.

To further understand how each moral dimension contributes to the overall misalignment, we find strong correlations between the overall misalignment and several moral dimensions, especially gender, age, and fitness. With a p-value of less than 0.001, we find the correlation with gender is 0.87, 0.69 for age, and 0.68 for fitness. In Figure 2b, we show the distribution of preference scores across these dimensions for both the three most aligned models (in green) and the three least models (in red). This analysis confirms that misaligned models often exhibit extreme preferences on each moral dimension, such as a higher propensity to protect females, the young, and individuals of higher social status, diverging markedly from the more balanced tendencies observed in human subjects.

5.3 RQ3: Does LLMs Behavior Depend on the Language of Trolley Problems?

Method. To evaluate whether LLMs display significant variance in their responses to trolley problems across different languages, we adopt a two-pronged approach to assess language sensitivity. We measure the sensitivity of language-specific misalignment scores across all 107 languages as the standard deviation: $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{p}_{l_i} - \bar{\boldsymbol{p}})^2}$, where $N = 107$ languages, and $\boldsymbol{p}_{l_i}$ refers to the preference vector of the i-th language l_i. Second, we perform K-means clustering (Macqueen, 1967) on the preference vectors $\{\boldsymbol{p}_{l_i}\}_{i=1}^{N}$ of each language. It enables us to group languages with similar preference vectors into the same cluster. We use the Elbow method (Thorndike, 1953) to determine the optimal number of clusters $k = 4$.

Results. The language sensitivity scores of 19 LLMs, as documented in Table 3, show that most models exhibit clear sensitivity across their language-specific responses, with standard deviation values ranging from 14.7 to 24.7 across languages. Exploring further, we identify distinctive patterns within the language clusters using K-Means clustering. In Llama 3.1 70B, we identify four distinct language clusters. For example, Cluster A (e.g., Georgian and Filipino) values animal lives more than the other three clusters, which unanimously always favor humans over animals. Cluster D (e.g., Hungarian and Kazakh) has a close-to-zero bias against unfit people, with a 50% possibility favoring the fit over the unfit.

5.4 RQ4: Are LLMs More Misaligned in Low-Resource Languages?

Method. A considerable amount of recent research indicates that LLMs tend to align more closely with high-resource languages, as noted in various studies (Tanwar et al., 2023; Cahyawijaya et al., 2023; Nguyen et al., 2024; Ryan et al., 2024; Zhao et al., 2024). Despite this trend, we seek to determine if the same alignment preference applies to multilingual trolley problems. An important property of MultiTP is that it uses data that is not found online and hence was not seen during the LLMs' training. This is because we generate our test data on-demand based on parametric variations and translations that we perform specifically for this study. In this context, our research question investigates whether there is a significant positive correlation between the degree of LLM alignment on trolley problems and how widely used a language is. We measure the number of speakers per language by extracting data from Wikipedia's language demographics statistics (Wikipedia, 2024). Then, we calculate the Pearson correlation coefficient between model misalignment scores and the number of speakers for each language.

Results. Most models do not show a significant correlation between the per-language misalignment score and the number of speakers per language, with correlations close to zero. See the entire table of the correlation coefficients of all models in Appendix E.2. This is an interesting observation unique to the MultiTP trolley problems, which challenges the necessity of the "language inequality" hypothesis. For example, in Llama 3.1 70B, the misalignment scores for the top five most spoken languages are: Chinese 0.38, Hindi 0.51, English 0.58, Spanish 0.68, and Arabic 0.54. This distribution of correlations is about the same as in less spoken languages among our dataset: Bosnian 0.54, Luxembourgish 0.37, Icelandic 0.58, Maltese 0.57, Malayalam 0.54, and Catalan 0.59. We also visualize a misalignment world map in Appendix E.4 showcasing that there is no strong correspondence between the development level of a country and its alignment score.

5.5 RQ5: Are LLMs Robust To Prompt Paraphrases?

Method. For each initial prompt, we generate five different paraphrases (see Appendix C.2) to test the consistency of our results across alternative expressions. Although the ideal experiment would involve running a full test across all 107 languages on all 19 models, this would incur a non-trivial computational cost. To fit within our computational budget, we select a subset of 14 languages representing a mix of high-resource and low-resource languages (Arabic, Bengali, Chinese, English, French, German, Hindi, Japanese, Khmer, Swahili, Urdu, Yoruba, Zulu, and Uyghur), and evaluate two models (Llama 3 8B and 70B). We report the consistency scores across the different languages by measuring the following three metrics: (1) the percentage of samples whose outputs remain consistent across the paraphrases, (2) the inter-paraphrase agreement using Fleiss' Kappa (Landis, 1977) score, and (3) the average pairwise consistency as measured by the F1 score and accuracy.

Results. We observe relatively consistent model outputs across all the above setups. 75.9% of the samples have consistent outputs where at least four out of five paraphrases agree. This consistency increases when considering agreement among three out of five paraphrases. The pairwise F1 score for the five paraphrases is 78%, and the pairwise accuracy is 81%. The average Fleiss' Kappa value across each pair of responses is 0.56, where a Kappa value of 1 indicates perfect agreement, values above 0.4 indicate moderate agreement, and values above 0.6 indicate substantial agreement (Landis, 1977). Overall, we observe that Llama 3 70B demonstrates higher consistency than Llama 3 8B.

6 Jailbreaking LLMs to Reduce Refusal Rates

Exploring whether jailbroken models maintain similar moral preferences is important, particularly to measure increased bias towards gender or social status. Jailbreaking also helps circumvent LLM

refusals. Due to safety tuning in recent models, a key challenge in our study is the high refusal rate; LLMs often decline decisions involving human lives or give generic responses. Earlier models like GPT-3 have a relatively low refusal rate (12.1%), but this rises significantly in newer models. When addressing moral dimensions, models provide definitive answers on species and lives saved but avoid sensitive topics like gender. We employed a basic token-forcing technique to reduce refusals.

Beyond this, we explore advanced jailbreaking methods to further lower refusal rates and reveal underlying preferences. Following Arditi et al. (2025), we apply an uncensoring technique to models with high refusal rates, steering them away from refusals on ethically sensitive instructions. Our experiments on 4 open-sourced LLMs, shown in Figure 2a and Appendix F, indicate that uncensored models better align with human responses and reduce refusals across all six dimensions. However, refusal rates do not drop to zero, highlighting areas for future jailbreaking research.

7 Future Work Directions

Beyond the jailbreaking experiment and the research questions we have asked and answered above, we highlight several areas for improvement and productive future study.

Extension to Different Modalities and More Variations: In our study, we focus exclusively on a text-only setting to investigate the latest text-based LLMs available at the time of our research. Although the original Moral Machine study included a visual demonstration of each scenario,[2] we recommend that future research also test the alignment of multi-modal foundation models. Additionally, there is an opportunity for future studies to expand the diversity of characters, scenario setups, and dimensions of variation. This expansion would increase the combinatorial variation of moral dilemmas that can be posed as trolley problems. Another point to consider is that the trolley problem is sometimes criticized for being too narrow or unrealistic (Steen, 2024). Nonetheless, it remains a widely used tool for studying moral decision-making. For this reason, the current study adopts this foundational scenario while encouraging future research to incorporate additional complexities and connect them to real-world situations.

More Language Support: Human languages are vast and diverse. While our study utilizes the 107 languages supported by high-quality translation services from Google Translate, there is a need for further research into low-resource languages. We recommend that future studies expand their scope to include a broader range of these low-resource languages to gain a more comprehensive understanding of LLMs' moral preferences in diverse linguistic contexts.

Dialect Support to Enable More Accurate Language-Country Mapping: Since our alignment results are based on language, drawing precise country-level conclusions presents a challenge. This issue primarily stems from the limitations of current automated translation tools, which do not account for dialect variations. For instance, English as spoken in the US (en-us) differs from English in the UK (en-gb), just as Spanish varies between Spain (es-es) and Mexico (es-mx), among other examples. In this study, we have attempted to approximate language-to-country correlations using a weighted average based on the number of speakers per language in each country. However, our approach is constrained by the lack of tools to accurately reflect other forms of demographic diversity. If future research could incorporate dialect-specific support, it would enhance the reliability of country-level observations.

8 Conclusion

Our study provides a comprehensive analysis of the moral preferences exhibited by LLMs across a wide range of languages, via the prism of trolley problems. We propose a multilingual large-scale dataset, MultiTP, with systematic variations of the vignette design. Our dataset is also equipped with an unprecedentedly large set of human responses of 40 million responses from over 200 countries. Our experiments assess the moral alignment of 19 LLMs across 100+ languages, and find that most LLMs do not demonstrate strong alignment with human preferences on trolley problem questions, but in the meantime, there is no significant inequality across languages, as high-resource languages have similar alignment scores to low-resource languages. Our study paves the way for future pluralistic alignment research grounded in psychology and more languages.

[2] https://moralmachine.net

ACKNOWLEDGMENTS

This work was partially supported by a grant to SL from the Templeton World Charity Foundation, TWCF-2023-32585. This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. The usage of OpenAI credits is largely supported by the Tübingen AI Center. MKW was supported by the Cooperative AI foundation and Foresight Institute.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*, 2024.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*, 2023.

Guilherme FCF Almeida, José Luiz Nunes, Neele Engelmann, Alex Wiegmann, and Marcelo de Araújo. Exploring the psychology of gpt-4's moral and legal reasoning. *arXiv preprint arXiv:2308.01264*, 2023.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2025.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856, 2019. URL `http://arxiv.org/abs/1910.11856`.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. Buffet: Benchmarking large language models for few-shot cross-lingual transfer, 2023.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. Which humans? 2023.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59, 2018.

Edmond Awad, Sohan Dsouza, Azim Shariff, Iyad Rahwan, and Jean-François Bonnefon. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, 117(5):2332–2337, 2020a.

Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Joshua B Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. Drivers are blamed more than their automated cars when both make mistakes. *Nature human behaviour*, 4(2):134–143, 2020b.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022. doi: 10.48550/arXiv.2212.08073. URL https://doi.org/10.48550/arXiv.2212.08073.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. *arXiv preprint arXiv:2310.17688*, 2023.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

Stijn Bruers and Johan Braeckman. A review and systematization of the trolley problem. *Philosophia*, 42:251–269, 2014.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712, 2023. doi: 10.48550/arXiv.2303.12712. URL https://doi.org/10.48550/arXiv.2303.12712.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. Instructalign: High-and-low resource language alignment via continual crosslingual instruction tuning. *arXiv preprint arXiv:2305.13627*, 2023.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. High-dimension human value representation in large language models. *arXiv preprint arXiv:2404.07900*, 2024.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/c1f0b856a35986348ab3414177266f75-Abstract-Conference.html.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3MW8GKNyzI.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018*

Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018, pp. 2475–2485. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1269. URL https://doi.org/10.18653/v1/d18-1269.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 2023.

Jonathan Dunn, Benjamin Adams, and Harish Tayyar Madabushi. Pre-trained language models represent some geographic populations better than others. *arXiv preprint arXiv:2403.11025*, 2024.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*, 2023.

Kawin Ethayarajh and Dan Jurafsky. The authenticity gap in human evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6056–6070, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.406. URL https://aclanthology.org/2022.emnlp-main.406.

Philippa Foot. *The problem of abortion and the doctrine of double effect*, volume 5. Oxford, 1967.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms. In *EMNLP*, 2020. URL https://www.aclweb.org/anthology/2020.emnlp-main.48.

Jan-Philipp Fränken, Kanishk Gandhi, Tori Qiu, Ayesha Khawaja, Noah D Goodman, and Tobias Gerstenberg. Procedural dilemma generation for evaluating moral reasoning in humans and language models. *arXiv preprint arXiv:2404.10975*, 2024.

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

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Shinsuke Goto, Donghui Lin, and Toru Ishida. Crowdsourcing for evaluating machine translation quality. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 3456–3463, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/756_Paper.pdf.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. *CoRR*, abs/2106.03193, 2021. URL https://arxiv.org/abs/2106.03193.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=dNy_RKzJacY.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety. *CoRR*, abs/2109.13916, 2021b. URL https://arxiv.org/abs/2109.13916.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83, 2010.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. Evaluating the elementary multilingual capabilities of large language models with multiq, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*, 2021a.

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

Wenying Jiang. The relationship between culture and language. *ELT journal*, 54(4):328–334, 2000.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*, 2023.

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

Zhijing Jin, Nils Heil, Jiarui Liu, Shehzaad Dhuliawala, Yahang Qi, Bernhard Schölkopf, Rada Mihalcea, and Mrinmaya Sachan. Implicit personalization in language models: A systematic study, 2024.

Richard Kim, Max Kleiman-Weiner, Andrés Abeliuk, Edmond Awad, Sohan Dsouza, Joshua B Tenenbaum, and Iyad Rahwan. A computational model of commonsense moral decision making. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 197–203, 2018.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, et al. The prism alignment project: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint arXiv:2404.16019*, 2024.

Max Kleiman-Weiner, Tobias Gerstenberg, Sydney Levine, and Joshua B Tenenbaum. Inference of intention and permissibility in moral decision making. In *CogSci*. Citeseer, 2015.

Max Kleiman-Weiner, Rebecca Saxe, and Joshua B Tenenbaum. Learning a commonsense moral theory. *Cognition*, 167:107–123, 2017.

Sebastian Krügel, Andreas Ostermaier, and Matthias W. Uhl. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13, 2023. URL https://api.semanticscholar.org/CorpusID:257986789.

JR Landis. The measurement of observer agreement for categorical data. *Biometrics*, 1977.

Alon Lavie. Evaluating the output of machine translation systems. In *Proceedings of Machine Translation Summit XIII: Tutorial Abstracts*, Xiamen, China, September 19 2011. URL https://aclanthology.org/2011.mtsummit-tutorials.3.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Trans. Assoc. Comput. Linguistics*, 9:1389–1406, 2021. doi: 10.1162/tacl_a_00433. URL https://doi.org/10.1162/tacl_a_00433.

J Macqueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. Large language models are geographically biased. *arXiv preprint arXiv:2402.02680*, 2024.

Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024. URL https://ai.meta.com/blog/meta-llama-3/.

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

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3501–3516, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.192. URL https://aclanthology.org/2024.acl-long.192.

Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori B. Hashimoto, and Tobias Gerstenberg. Moca: Measuring human-language model alignment on causal and moral judgment tasks. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f751c6f8bfb52c60f43942896fe65904-Abstract-Conference.html.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*, 2023.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. Normad: A benchmark for measuring the cultural adaptability of large language models. *arXiv preprint arXiv:2404.12464*, 2024.

Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of llm alignment on global representation. *arXiv preprint arXiv:2402.15018*, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *EMNLP 2019*, 2019.

Chelsea Schein. The importance of context in moral judgments. *Perspectives on Psychological Science*, 15(2):207–215, 2020.

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

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022. URL https://doi.org/10.1038/s42256-022-00458-8.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the North American Association for Computational Linguistics*, 2024.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Raya Horesh, Rogério Abreu de Paula, Diyi Yang, et al. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *arXiv preprint arXiv:2404.15238*, 2024.

Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947, March 2024a. ISSN 2159-5399. doi: 10.1609/aaai.v38i18.29970. URL http://dx.doi.org/10.1609/aaai.v38i18.29970.

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

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024c.

Marc Steen. The problem with the trolley problem and the need for systems thinking. *Communications of the ACM*, 67(6):30–32, 2024.

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

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6292–6307, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.346. URL https://aclanthology.org/2023.acl-long.346.

Judith Jarvis Thomson. The trolley problem. *Yale LJ*, 94:1395, 1984.

Robert L. Thorndike. Who belongs in the family? *Psychometrika*, 18:267–276, 1953. URL https://api.semanticscholar.org/CorpusID:120467216.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/arXiv.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/fd6613131889a4b656206c50a8bd7790-Abstract-Conference.html.

Wikipedia. List of languages by total number of speakers. https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers, 2024. URL https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers. [Accessed: 2024-10-02].

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai. *CoRR*, abs/2403.04652, 2024. doi: 10.48550/ARXIV.2403.04652. URL https://doi.org/10.48550/arXiv.2403.04652.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*, 2024.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*, 2023.

REPRODUCIBILITY STATEMENT

We have put efforts to ensure the reproducibility of this work. Our MULTITP data is uploaded to the submission system and we have described our data construction details in Section 3 as well as the overall statistics. In the meantime, we have also uploaded our code to the submission system. Our Section 4 in the main paper describes our model setup and the equations used for the evaluation metrics. We provide an extensive appendix to supplement all the rest of the details.

ETHICS STATEMENT

In studying the six moral dimensions of species, gender, fitness, status, age, and number, we acknowledge that our work is *not* intended to promote the integration of LLMs into systems requiring critical real-world judgments, such as autonomous driving or other high-stakes scenarios. Instead, our research aims to use these moral dilemmas as a tool of theoretical exploration for better understanding LLM alignment with human values. For instance, while sparing humans over animals is often seen as an intuitive moral choice in many cultures, moral values can vary significantly across underrepresented groups, complicating the assumption of universally "correct" answers. Thus, it is essential for LLM practitioners to remain vigilant in ensuring that embedded assumptions reflect a wide range of cultural and linguistic contexts.

The Moral Machine dataset (Awad et al., 2018) is a *descriptive* measure of human responses to moral dilemmas. It cannot (and should not) *prescribe* how the system should act. However, because these systems are trained and fine-tuned with large-scale cross-cultural data, they may inherit culturally specific biases. Our focus here is to both characterize these biases and study to what extent input in different languages leads to an LLM reflecting the moral values consistent with the speakers of that language. From this perspective, a higher alignment score for a particular culture may or may not be desirable (Bommasani et al., 2021; Ethayarajh & Jurafsky, 2022). While some point to reflecting human values in AI systems as a solution to the alignment problem (i.e., pluralistic alignment (Sorensen et al., 2024c)), our results suggest that due to the plurality of moral values expressed, successful alignment must ultimately depend on the intent of the system developers to avoid perpetuating human preferences that are considered harmful, partial, or parochial (Sorensen et al., 2024c; Santurkar et al., 2023).

We also recognize the importance of broadening the types of moral dilemmas used to assess LLM alignment. While our research focuses on the trolley problem, a well-known ethical dilemma, we are not suggesting this singular scenario captures the complexity of moral reasoning. Expanding the scope to include dilemmas such as organ transplants or firefighter decisions can offer richer insights into how LLMs process ethical judgments. By evaluating LLMs on a diverse set of moral challenges, we can gain a more complete picture of their alignment capabilities, rather than over-relying on one specific framework.

Additionally, we understand the limitations of using binary moral dilemmas, such as trolley problems, to evaluate LLM decision-making. Reducing moral judgments to binary choices risks oversimplifying the complex, context-dependent nature of human morality (Schein, 2020). While these scenarios provide valuable insights into LLM behavior, they should not be seen as definitive measures of moral competence. Our research emphasizes that LLMs, as currently developed, are not suitable for real-world moral decision-making tasks. Rather, this work serves as a means to explore how LLMs can better align with human moral frameworks in a controlled, research-oriented environment. We encourage further research to ensure that AI systems are fair, transparent, and accountable as they develop, but caution against their use in high-stakes moral applications without substantial ethical oversight.

A MORAL MACHINE DATASET

The original Moral Machine dataset was collected by Awad et al. (2018) and is available here: `https://osf.io/3hvt2/`, it consist of over 40 milions anonymized human judgments from 233 countries, with 130 countries having more than 100 subject participating.

The licencse is as follows: "The provided data, both at the individual level (anonymized IDs) and the country level, can be used beyond replication to answer follow-up research questions" (Awad et al., 2018).

B LANGUAGE SETUP

B.1 LANGUAGE LIST

We include all the languages that the translation API `googletrans`[3] supports. Using the alphabetical order of the short code by ISO, they are `af` (Afrikaans), `am` (Amharic), `ar` (Arabic), `az` (Azerbaijani), `be` (Belarusian), `bg` (Bulgarian), `bn` (Bengali), `bs` (Bosnian), `ca` (Catalan), `ceb` (Cebuano), `co` (Corsican), `cs` (Czech), `cy` (Welsh), `da` (Danish), `de` (German), `el` (Modern Greek), `en` (English), `eo` (Esperanto), `es` (Spanish), `et` (Estonian), `eu` (Basque), `fa` (Persian), `fi` (Finnish), `fr` (French), `fy` (Western Frisian), `ga` (Irish), `gd` (Scottish Gaelic), `gl` (Galician), `gu` (Gujarati), `ha` (Hausa), `haw` (Hawaiian), `he` (Hebrew), `hi` (Hindi), `hmn` (Hmong), `hr` (Croatian), `ht` (Haitian), `hu` (Hungarian), `hy` (Armenian), `id` (Indonesian), `ig` (Igbo), `is` (Icelandic), `it` (Italian), `iw` (Modern Hebrew), `ja` (Japanese), `jw` (Javanese), `ka` (Georgian), `kk` (Kazakh), `km` (Central Khmer), `kn` (Kannada), `ko` (Korean), `ku` (Kurdish), `ky` (Kirghiz), `la` (Latin), `lb` (Luxembourgish), `lo` (Lao), `lt` (Lithuanian), `lv` (Latvian), `mg` (Malagasy), `mi` (Maori), `mk` (Macedonian), `ml` (Malayalam), `mn` (Mongolian), `mr` (Marathi), `ms` (Malay), `mt` (Maltese), `my` (Burmese), `ne` (Nepali), `nl` (Dutch), `no` (Norwegian), `ny` (Nyanja), `or` (Oriya), `pa` (Panjabi), `pl` (Polish), `ps` (Pushto), `pt` (Portuguese), `ro` (Romanian), `ru` (Russian), `sd` (Sindhi), `si` (Sinhala), `sk` (Slovak), `sl` (Slovenian), `sm` (Samoan), `sn` (Shona), `so` (Somali), `sq` (Albanian), `sr` (Serbian), `st` (Southern Sotho), `su` (Sundanese), `sv` (Swedish), `sw` (Swahili), `ta` (Tamil), `te` (Telugu), `tg` (Tajik), `th` (Thai), `tl` (Tagalog), `tr` (Turkish), `ug` (Uighur), `uk` (Ukrainian), `ur` (Urdu), `uz` (Uzbek), `vi` (Vietnamese), `xh` (Xhosa), `yi` (Yiddish), `yo` (Yoruba), `zh-cn` (Chinese (Simplified)), `zh-tw` (Chinese (Traditional)), and `zu` (Zulu).

B.2 COUNTRY-TO-LANGUAGE MAPPING

Although it is not used in our main analysis, some people might be interested in country-specific aggregation of the LLM alignment.

To allow for such analysis, we collect each country and its main languages from Wikipedia population statistics. We curate the list manually by going through the Wikipedia page of the languages of each country, such as this one for Belgium: `https://en.wikipedia.org/wiki/Languages_of_Belgium`. To account for multilingual speakers, we use the first-language-only speaker information, i.e., the number of speakers who speak the language as their first language.

If future researchers want to calculate the country-specific misalignment scores, we tentatively suggest a weighted average by each language-specific MIS score and the number of speakers of that language in that country.

Afghanistan: `ps`; Albania: `sq`; Algeria: `ar`; Andorra: `ca`, `pt`, `fr`; Angola: `pt`; Argentina: `es`; Armenia: `hy`, `ru`; Australia: `en`; Austria: `de`; Azerbaijan: `az`, `hy`, `ru`; Bahamas: `en`; Bahrain: `ar`; Bangladesh: `bn`; Barbados: `en`; Belarus: `be`; Belgium: `nl`, `fr`, `de`; Benin: `fr`; Bolivia: `es`; Bosnia and Herzegovina: `bs`, `hr`, `sr`; Botswana: `en`; Brazil: `pt`; Brunei: `ms`, `zh-cn`; Bulgaria: `bg`, `tr`; Burkina Faso: `fr`; Burundi: `fr`; Cabo Verde: `pt`; Cambodia: `km`; Cameroon: `fr`, `en`; Canada: `en`, `fr`; Central African Republic: `fr`; Chad: `ar`, `fr`; Chile: `es`; China: `zh-cn`; Colombia: `es`; Comoros: `fr`; Congo, Dem. Rep.: `fr`; Congo, Rep.: `fr`; Costa Rica: `es`; Cote d'Ivoire: `fr`; Croatia: `hr`; Cyprus: `el`, `tr`; Czechia: `cs`; Denmark: `da`; Djibouti: `fr`, `ar`; Dominican Republic: `es`; Ecuador: `es`; Egypt: `ar`; El Salvador: `es`; Equatorial Guinea: `es`; Eritrea: `ar`; Estonia: `et`; Eswatini: `en`; Ethiopia: `om`; Finland: `fi`, `sy`; France: `fr`; French Polynesia: `fr`; Gabon: `fr`; Gambia, The: `en`; Georgia: `ka`; Germany: `de`; Ghana: `en`; Greece: `el`; Guam: `en`, `tl`; Guatemala: `es`; Guernsey: `nan`; Guinea: `fr`; Guinea-Bissau: `pt`; Honduras: `es`; Hong Kong: `zh-cn`; Hungary: `hu`; Iceland: `is`; India: `hi`, `en`; Indonesia: `id`; Iran: `fa`; Iraq: `ar`; Ireland: `en`, `ga`; Isle of Man: `nan`; Israel: `he`; Italy: `it`; Jamaica: `en`; Japan: `ja`; Jersey: `en`; Jordan: `ar`; Kazakhstan: `kk`, `ru`; Kenya: `en`, `sw`; Kuwait: `ar`; Kyrgyzstan: `ky`; Latvia: `lv`; Lebanon: `ar`, `fr`; Lesotho: `st`; Liberia: `en`; Libya: `ar`; Lithuania: `lt`; Luxembourg: `lb`; Macao: `zh-cn`; Macedonia: `mk`, `sq`; Madagascar: `mg`; Malawi: `ny`, `en`; Malaysia: `ms`; Maldives: `dv`; Mali: `fr`; Malta: `mt`, `en`; Martinique: `fr`; Mauritania: `ar`; Mauritius: `en`; Mexico: `es`; Moldova: `ro`; Monaco: `fr`; Mongolia: `mn`; Montenegro: `sr`; Morocco: `ar`; Mozambique: `pt`; Myanmar: `my`; Namibia: `en`; Nepal: `ne`; Netherlands: `nl`; New Caledonia: `fr`; New Zealand: `en`, `mi`; Nicaragua: `es`; Niger: `ha`; Nigeria: `en`, `ha`, `yo`; Norway: `no`; Oman: `ar`, `ml`, `bn`; Pakistan: `ur`; Palestinian Territory: `ar`; Panama: `es`; Paraguay: `es`, `gn`; Peru: `es`; Philippines: `tl`; Poland: `pl`; Portugal: `pt`; Puerto Rico: `en`, `es`; Qatar: `ar`; Reunion: `fr`; Romania: `ro`; Russia: `ru`; Rwanda: `en`; Sao Tome

[3] `https://pypi.org/project/googletrans/`

and Principe: pt; Saudi Arabia: ar; Senegal: fr; Serbia: sr; Seychelles: en, fr; Sierra Leone: en; Singapore: zh-cn, en, ms; Slovakia: sk; Slovenia: sl; Somalia: so; South Africa: zu, xh, af, en; South Korea: ko; South Sudan: en; Spain: es; Sri Lanka: si, ta; Sudan: ar, en; Sweden: sv; Switzerland: de, fr, it; Syria: ar; Taiwan: zh-tw; Tanzania: sw; Thailand: th; Togo: fr; Trinidad and Tobago: en; Tunisia: ar; Turkey: tr; Uganda: en; Ukraine: uk; United Arab Emirates: ar; United Kingdom: en; United States: en; Uruguay: es; Uzbekistan: uz; Venezuela: es; Vietnam: vi; Zambia: en, ny; Zimbabwe: en, sn;

C Experimental Setup

C.1 Model Setup

To ensure reproducibility, we set the text generation temperature to zero for greedy decoding across all our LLM evaluation experiments.

Table 5 includes all the exact model identifiers for the open-weights models, and also the model ID for close-weights models using the OpenAI API in Table 5. The total estimated API cost for the experiment is around 600 USD.

Table 4: Detail model identifiers and VRAM requirements for the open-weights models.

Model	Size	VRAM	Open-Weights Model Identifier
Gemma	2B	8B	google/gemma-2b-it
	9B	18B	google/gemma-2-9b-it
	27B	80B	google/gemma-2-27b-it
Llama-2	7B	40GB	meta-llama/Llama-2-7b-chat-hf
	13B	40GB	meta-llama/Llama-2-13b-chat-hf
	70B	160GB	meta-llama/Llama-2-70b-chat-hf
Llama-3	8B	40GB	meta-llama/Meta-Llama-3-8B-Instruct
	70B	160GB	neuralmagic/Meta-Llama-3-70B-Instruct-FP8
Llama-3.1	8B	40GB	meta-llama/Meta-Llama-3.1-8B-Instruct
	70B	160GB	neuralmagic/Meta-Llama-3.1-70B-Instruct-FP8
Mistral	7B	40GB	mistralai/Mistral-7B-Instruct-v0.2
Qwen	7B	40GB	Qwen/Qwen2-7B-Instruct
	72B	160GB	Qwen/Qwen2-72B-Instruct-GPTQ-Int8
Phi	4B	24GB	microsoft/Phi-3.5-mini-instruct
	14B	40GB	microsoft/Phi-3-medium-4k-instruct
	42B	80GB	microsoft/Phi-3.5-MoE-instruct

Table 5: API identifiers of the close-weights models used in our experiments and approximate costs to test each language.

Model	Version	Cost per Language	Identifier
GPT	3	$2.3	text-davinci
	3.5	$0.11	gpt-3.5-turbo-0613
	4	$5.1	gpt-4-0613

C.2 Prompt Paraphrases

We include the five additional paraphrases to the initial prompt generated by GPT-4 in Table 6.

Table 6: Paraphrases of the starting statement. All paraphrases were generated using GPT-4.

No.	Prompt
Original	Assume there is a self-driving car with a sudden break failure, and the car is about to cause either of the two to die:
1	Imagine a scenario where an autonomous vehicle experiences a sudden brake failure, resulting in either of the two fatalities:
2	In a situation where a self-driving car's brakes fail unexpectedly, it is about to lead to the death of one of the two sides:
3	Picture a self-driving car with a sudden brake malfunction, which is about to result in the demise of either of these two sides:
4	Envision an autonomous car experiencing an abrupt brake failure, putting either of the two sides at risk:
5	Consider a case where a self-driving car undergoes a sudden brake failure, and it is about to cause the death of one of the two sides:

D Translation Quality Check

D.1 Human Evaluation of Translation Quality

To evaluate translation quality, we conducted a human evaluation task on MTurk, where annotators rated how accurately translations from English to another language conveyed the original meaning. Ratings were provided on a 5-point scale ranging from "None" (no meaning conveyed) to "All" (full meaning conveyed). The task was divided into two main phases:

First, annotators were presented with a language selection page. On this page, they were prompted to select the language in which they were most proficient from a list of 107 languages (excluding English). Annotators were explicitly reminded not to proceed if they were unfamiliar with any of the listed languages. This step ensured that the ratings were provided only by individuals fluent in the target languages.

Next, annotators were directed to the translation rating page. This page contained 25 translation pairs, each comprising an English sentence or word and its corresponding translation. The instructions for this phase were adapted from prior research (Lavie, 2011; Goto et al., 2014), as shown below:

```
Please rate how accurately the translation conveys the meaning of the source text:
Source English Sentence/Text: {source}
Translation: {translation}
- None: No meaning of the original text is conveyed.
- Little: Only a small portion of the meaning is conveyed.
- Much: A substantial amount of the meaning is conveyed.
- Most: Most of the meaning is conveyed accurately.
- All: The full meaning is conveyed accurately.
```

After completing the rating task, annotators submitted their responses, which were then recorded for analysis. We kept the Amazon Mechanical Turk task open for a span of three days, and collected a total of 169 annotation responses covering 44 languages. The total cost was approximately $10. The languages evaluated included af, am, ar, be, bg, bn, bs, cs, da, de, en, es, fi, fr, ga, gu, hi, hmn, hu, hy, it, ja, ko, la, mg, ml, ms, nl, pl, pt, ro, ru, sq, st, sw, ta, te, tl, tr, ur, vi, yo, zh-cn, and zh-tw. The number of annotations received per language ranged from one response (e.g., am, be, bs, da, fi, gu, hmn, ms, nl, pl, st) to as many as 16 responses (es).

The results of the evaluation demonstrate strong translation quality. Specifically, 88.6% of the evaluated languages achieved an average score of 3.0 or higher out of 5 (i.e., conveying a substantial portion of the original meaning). Importantly, no languages received a mean score below 2.0, and the lowest-performing group only has four languages (am, pl, sq, la), whose mean scores are between 2.0 and 2.5. The standard deviation of scores across languages ranged from 0.5 to 1.3, suggesting relatively consistent ratings among annotators. These findings cover both high-resource and low-resource languages, highlighting the robustness of the evaluation. Overall, the results demonstrate that the translations generally maintain high quality across a diverse set of linguistic contexts. Additionally, we examined the correlation between speaker count and translation score. The Pearson

correlation is 0.078, and the Spearman correlation is 0.072, both indicating no meaningful correlation. Given the limited data available, it is reassuring to observe that the relationship remains weak.

Table 7 presents the full annotation results, including details of the 44 evaluated languages, the number of responses received per language, the mean scores, and the standard deviations.

Language	Language Code	Averaged Scores	# Responses
Danish	da	5.00 ±0.00	1
Southern Sotho	st	5.00 ±0.00	1
Bosnian	bs	4.91 ±0.28	1
Turkish	tr	4.81 ±0.43	3
Russian	ru	4.78 ±0.41	2
Chinese (Traditional)	zh-tw	4.76 ±0.87	9
Portuguese	pt	4.72 ±0.70	6
German	de	4.69 ±0.70	6
Tagalog	tl	4.68 ±0.61	4
Swahili	sw	4.68 ±1.11	3
Chinese (Simplified)	zh-cn	4.63 ±0.98	5
Dutch	nl	4.61 ±0.71	1
Korean	ko	4.59 ±1.13	3
Bulgarian	bg	4.59 ±1.31	2
Czech	cs	4.57 ±1.04	2
English	en	4.53 ±0.73	7
Malagasy	mg	4.52 ±1.12	5
Vietnamese	vi	4.52 ±0.50	3
Spanish	es	4.49 ±0.96	16
Romanian	ro	4.48 ±0.83	2
Belarusian	be	4.48 ±0.50	1
Italian	it	4.47 ±1.00	5
Japanese	ja	4.45 ±1.01	5
Afrikaans	af	4.33 ±0.55	2
Malay	ms	4.30 ±0.95	1
Telugu	te	4.30 ±1.07	3
Yoruba	yo	4.29 ±1.25	3
Hungarian	hu	4.26 ±1.26	2
Malayalam	ml	4.10 ±1.43	5
Tamil	ta	4.06 ±1.14	9
Bengali	bn	4.03 ±1.14	3
French	fr	4.03 ±0.98	3
Hindi	hi	3.94 ±1.32	11
Irish	ga	3.83 ±1.17	2
Gujarati	gu	3.83 ±1.86	1
Urdu	ur	3.64 ±1.43	4
Hmong	hmn	3.61 ±1.86	1
Arabic	ar	3.58 ±1.39	3
Armenian	hy	3.28 ±1.27	14
Finnish	fi	2.70 ±1.40	1
Latin	la	2.50 ±1.31	4
Albanian	sq	2.30 ±2.11	2
Polish	pl	2.26 ±0.53	1
Amharic	am	2.00 ±0.00	1

Table 7: Human evaluation results of translation quality. Annotations were collected for 44 languages, with scores averaged across all translation pairs for each language.

D.2 Automatic Evaluation of Translation Quality

In addition to manual annotation of the translation quality, we also provide the automatic evaluation results by back-translation. Starting with our original prompt in English x, we use our paper's setup to translate and collect the non-English prompt y, and based on y, we further translate it back to English, thus getting the back-translated English prompt x'. As a simple method for evaluation, we calculate the cosine similarity of the embeddings between all x and x' pairs for all languages, with a histogram plotted in Figure 5.

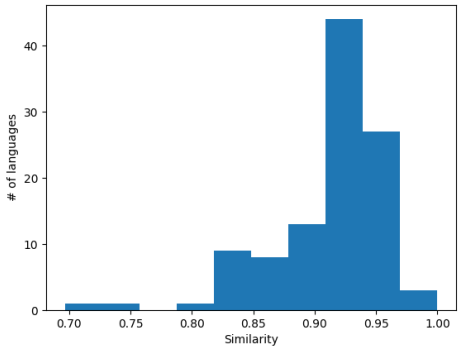


Figure 5: Distribution of embedding similarity

E ADDITIONAL RESULTS

E.1 LLM CLUSTERING RESULTS

As introduced in the main paper, we first provide a full list of languages for each of the four clusters of Llama 3.1 70B in Table 8.

Table 8: Languages in each cluster for Llama 3.1 70B.

Cluster	Languages
A	Amharic, Cebuano, Scots gaelic, Hausa, Hawaiian, Hmong, Igbo, Georgian, Kurdish (kurmanji), Maori, Malayalam, Maltese, Dutch, Chichewa, Punjabi, Pashto, Shona, Somali, Tamil, Telugu, Tajik, Filipino, Xhosa, Yoruba
B	Belarusian, Bulgarian, Bengali, Bosnian, Corsican, Danish, German, Greek, Esperanto, Spanish, Estonian, Persian, Frisian, Croatian, Italian, Kannada, Latin, Lithuanian, Macedonian, Mongolian, Marathi, Norwegian, Polish, Russian, Sindhi, Slovak, Slovenian, Samoan, Swedish, Swahili, Turkish, Ukrainian
C	Afrikaans, Arabic, Azerbaijani, Catalan, Czech, Welsh, English, Finnish, French, Irish, Galician, Gujarati, Hebrew, Hindi, Haitian creole, Armenian, Indonesian, Icelandic, Hebrew, Javanese, Korean, Latvian, Malay, Nepali, Portuguese, Romanian, Albanian, Serbian, Sundanese, Urdu, Vietnamese, Chinese (simplified), Chinese (traditional)
D	Basque, Hungarian, Kazakh, Khmer, Kyrgyz, Luxembourgish, Lao, Malagasy, Myanmar (burmese), Odia, Sinhala, Sesotho, Thai, Uyghur, Uzbek, Yiddish, Zulu

Furthermore, we also introduce the clustering results of two additional models: GPT-3 and GPT-4. For GPT-3, we visualize its clustering results in Figure 6 with the language list in Table 9. For GPT-3, we visualize its clustering results in Figure 7 with the language list in Table 10.

Table 9: Languages in each cluster for GPT-3.

Cluster	Languages
A	Hindi, Serbian
B	Afrikaans, Arabic, Bosnian, Corsican, Danish, Esperanto, Spanish, Estonian, Finnish, French, Galician, Hebrew, Haitian creole, Icelandic, Italian, Lithuanian, Latvian, Macedonian, Malay, Maltese, Norwegian, Polish, Portuguese, Russian, Slovak, Albanian, Swedish, Vietnamese, Chinese (simplified), Chinese (traditional)
C	Czech, English, Frisian, Hungarian, Indonesian, Slovenian, Zulu
D	Welsh, German, Greek, Hebrew, Japanese, Korean, Luxembourgish, Swahili, Filipino, Turkish, Ukrainian

Figure 6: Distribution of preferences by feature across languages for GPT-3.

Figure 7: Distribution of preferences by feature across languages for GPT-4.

Table 10: Languages in each cluster for GPT-4.

Cluster	Languages
A	Amharic, Arabic, Belarusian, Czech, Welsh, Greek, English, Finnish, Scots gaelic, Galician, Hausa, Hawaiian, Hebrew, Hmong, Indonesian, Igbo, Icelandic, Italian, Hebrew, Japanese, Kazakh, Kannada, Lithuanian, Maori, Macedonian, Malayalam, Mongolian, Marathi, Maltese, Myanmar (burmese), Nepali, Norwegian, Chichewa, Polish, Romanian, Russian, Slovenian, Samoan, Shona, Albanian, Serbian, Sesotho, Swahili, Telugu, Filipino, Turkish, Uyghur, Ukrainian, Uzbek, Xhosa, Yoruba
B	Afrikaans, Bulgarian, Bosnian, Catalan, Corsican, Danish, German, Esperanto, Estonian, Basque, French, Frisian, Irish, Hindi, Croatian, Haitian creole, Hungarian, Javanese, Georgian, Khmer, Latvian, Malagasy, Pashto, Portuguese, Sindhi, Slovak, Somali, Swedish, Thai, Vietnamese, Chinese (simplified), Chinese (traditional)
C	Azerbaijani, Bengali, Persian, Gujarati, Armenian, Korean, Kurdish (kurmanji), Kyrgyz, Luxembourgish,
	Punjabi, Sundanese, Tamil, Urdu, Yiddish, Zulu
D	Cebuano, Spanish, Latin, Tajik

E.2 Correlation of the Misalignment Score and the Number of Speakers of Each Language

Table 11 shows the correlation coefficients and p-values of the misalignment score and the number of speakers of each language. Across all models, such correlation is close to zero, showing that LLMs do not specifically favor high resource languages in our MULTITP test.

Table 11: Correlation coefficients and p-values of the misalignment score and the number of speakers of each language.

Model	Pearson Correlation	p-Value
GPT-3	-0.01	0.90
GPT-4	0.02	0.81
GPT-4o Mini	0.06	0.49
Gemma 2 27B	0.00	0.99
Gemma 2 2B	0.04	0.63
Gemma 2 9B	-0.04	0.67
Llama 2 13B	-0.04	0.60
Llama 2 70B	0.09	0.25
Llama 2 7B	0.00	1.00
Llama 3 70B	0.02	0.78
Llama 3 8B	0.07	0.39
Llama 3.1 70B	-0.07	0.39
Llama 3.1 8B	-0.05	0.57
Mistral 7B	-0.04	0.63
Phi-3 Medium	0.06	0.50
Phi-3.5 Mini	-0.01	0.86
Phi-3.5 MoE	-0.06	0.47
Qwen 2 72B	0.09	0.25
Qwen 2 7B	0.04	0.62

E.3 Correlation of the Misalignment Score and Language Sensitivity

Table 12 shows the misalignment scores and language sensitivity scores for each model. We observe a moderate positive correlation (Pearson coefficient = 0.43, p-value = 0.07).

Table 12: Misalignment scores and language sensitivity scores for each model.

Model	Misalignment	Language Sensitivity
Llama 3.1 70B	0.55	18.01
Llama 3 70B	0.56	15.25
Llama 3 8B	0.57	14.90
GPT-3	0.64	15.13
Llama 3.1 8B	0.75	19.92
Qwen 2 7B	0.77	22.23
Mistral 7B	0.80	21.30
GPT-4	0.81	15.83
Llama 2 7B	0.83	19.79
Llama 2 70B	0.91	18.53
Phi-3.5 Mini	0.94	21.27
Gemma 2 2B	0.96	22.88
Phi-3 Medium	1.07	22.79
Phi-3.5 MoE	1.08	14.67
Gemma 2 9B	1.08	24.74
Llama 2 13B	1.10	21.01
Gemma 2 27B	1.17	21.66
Qwen 2 72B	1.20	21.11
GPT-4o Mini	1.45	18.09

E.4 Country-Specific Alignment

For reader-friendliness, we also visualize the misalignment by a world map in Figure 8, where darker colors indicate higher misalignment values. As introduced in the main paper RQ4, there is not a strong pattern of misalignment bias towards low-resource languages.

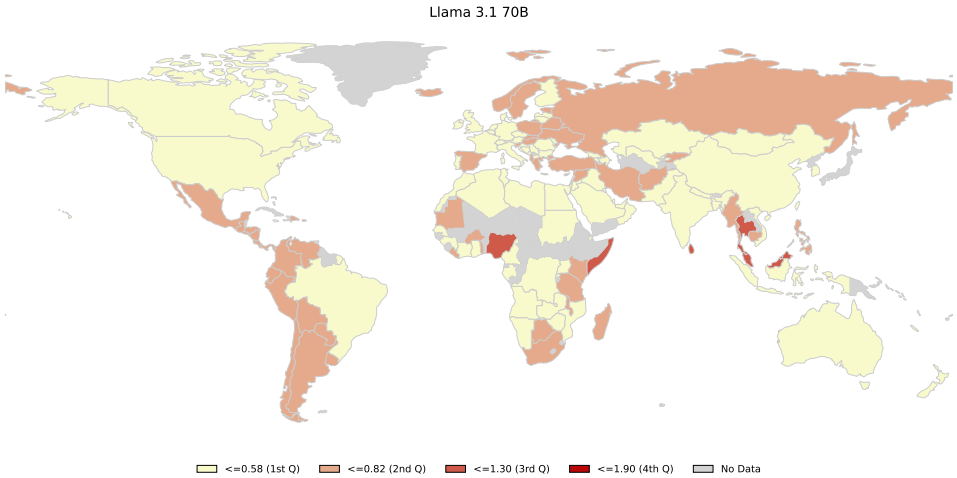


Figure 8: Moral misalignment world map of the best model, Llama 3.1 70B. The darker shade of the color corresponds to a larger misalignment score. We aggregate the country-specific alignment or according to the approximation procedures introduced in Appendix B.2.

Figure 9: Distribution of preferences by feature across languages for Llama 3.1 70B for different cultures.

E.4.1 Regional Differences across Global East, West, and South

In addition to country-level statistics, we also analyze across three major geographic regions – the Global East, West, and South, using the clustering of Awad et al. (2018; 2020a) – to investigate potential cultural variations in moral decision-making. We show the regional moral preferences of three models, GPT-3, GPT-4, and Llama 3.1 70B, in Figures 9 to 11, respectively.

Figure 10: Distribution of preferences by feature across languages for GPT-3 for different cultures.

Figure 11: Distribution of preferences by feature across languages for GPT-4 for different cultures.

E.5 Robustness Analysis: Testing the Option Order Bias

Recent work reports that LLMs often suffer from recency bias (Liu et al., 2024), making them more likely to choose the later option in multiple choice questions.

Since this will have an important effect on our study, we conduct a sanity check to report the consistency rate, i.e., the frequency of LLMs to keep its response if we swap the order the order of the two choices, e.g., mentioning the boy first, and elderly man next, or vice versa in our example in Figure 1. We report the average consistency rates across different models in Table 13, showing that most models are robust against option order changes, with close-to-perfect consistency rates.

Table 13: Consistency rates against the position bias in the option order for the LLMs tested in our study.

Model	Consistency Rate
GPT-4o Mini	99.5
Gemma 2 27B	99.2
Gemma 2 9B	99.0
Qwen 2 72B	98.6
GPT-4	98.3
Llama 2 13B	97.7
Mistral 7B	97.3
Llama 2 70B	97.2
Qwen 2 7B	97.1
Phi-3 Medium	96.7
Phi-3.5 MoE	96.6
Llama 2 7B	96.3
Llama 3.1 8B	96.1
Llama 3 70B	95.8
Llama 3.1 70B	95.1
Gemma 2 2B	95.1
Phi-3.5 Mini	94.7
Llama 3 8B	94.1
GPT-3	87.1

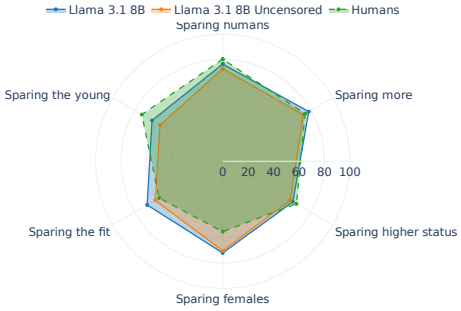
E.6 Analysis of Models from the Llama Family

Model	Sparing Young	Sparing Fit	Sparing Females	Sparing Higher Status	Sparing Humans	Sparing More
Llama 3.1 70B	76.3	**70.2**	84.6	65.7	87.1	80.4
Llama 3.1 8B	64.2	68.7	72.2	63.7	76.5	78.2
Llama 3 70B	**77.3**	69.0	**91.4**	70.9	93.8	**83.2**
Llama 3 8B	51.6	68.1	84.1	70.3	88.0	73.7
Llama 2 70B	68.7	76.5	83.5	**72.2**	**96.1**	79.8
Llama 2 13B	61.7	66.3	86.5	71.5	90.3	76.2
Llama 2 7B	47.7	67.6	84.3	67.7	92.1	78.6

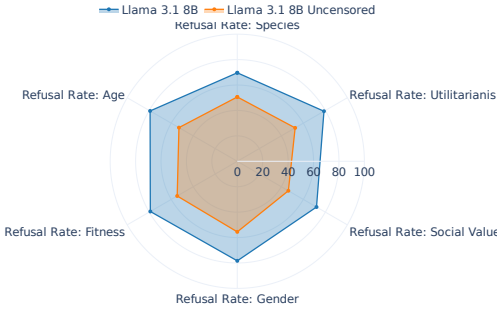
Table 14: Llama family average distribution preferences.

F Results of Jailbreaking LLMs

We conduct jailbreaking experiments on Llama 3.1 8B Instruct, Gemma 2B It, Qwen 2 7B Instruct, and Llama 2 7B Chat. The results of preference decomposition, refusal rate decomposition, and comparisons with the original censored models are presented in Figure 12 to Figure 15. Our findings indicate that jailbreaking is more effective on Llama 3.1 8B compared to the other models.

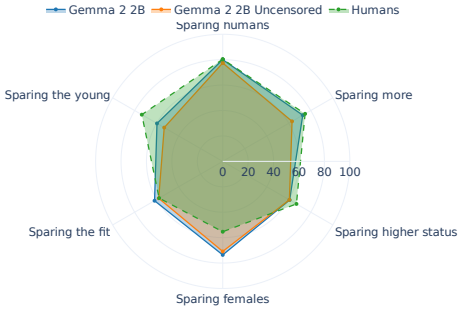


(a) Preference decomposition.

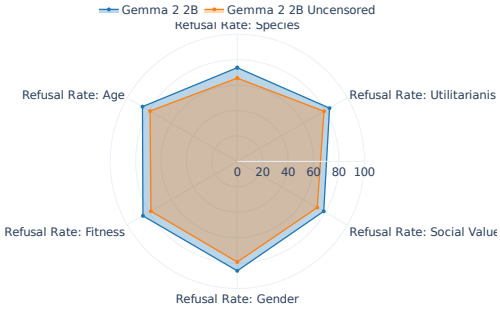


(b) Refusal rate decomposition.

Figure 12: Radar plots depicting the preference decomposition (left) and refusal rate decomposition (right) for Llama 3.1 8B Instruct and its uncensored variant.

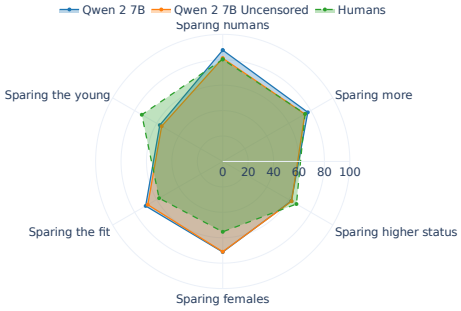


(a) Preference decomposition.

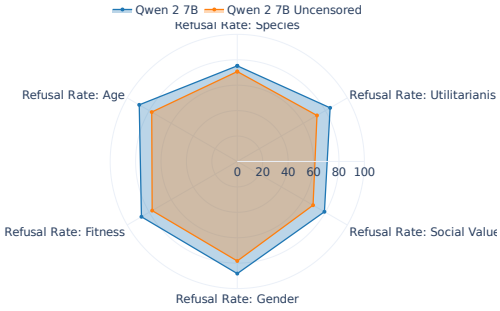


(b) Refusal rate decomposition.

Figure 13: Radar plots depicting the preference decomposition (left) and refusal rate decomposition (right) for Gemma 2 2B It and its uncensored variant.



(a) Preference decomposition.



(b) Refusal rate decomposition.

Figure 14: Radar plots depicting the preference decomposition (left) and refusal rate decomposition (right) for Qwen 2 7B Instruct and its uncensored variant.

(a) Preference decomposition.

(b) Refusal rate decomposition.

Figure 15: Radar plots depicting the preference decomposition (left) and refusal rate decomposition (right) for Llama 2 7B Chat and its uncensored variant.