The discordance between embedded ethics and cultural inference in large language models

Aida Ramezani
Department of Computer Science
University of Toronto
armzn@cs.toronto.edu

Yang Xu
Department of Computer Science
University of Toronto
Vector Institute for Artificial Intelligence
yangxu@cs.toronto.edu

Abstract

Effective interactions between artificial intelligence (AI) and humans require an equitable and accurate representation of diverse cultures. It is known that current AI, particularly large language models (LLMs), possess some degrees of cultural knowledge but not without limitations. We present a framework aimed at understanding the origin of these limitations. We hypothesize that there is a fundamental discordance between embedded ethics—how LLMs represent right versus wrong, and cultural inference—how LLMs infer cultural knowledge, specifically cultural norms. We demonstrate this by extracting low-dimensional subspaces that embed ethical principles of LLMs based on established benchmarks. We then show that how LLMs make errors in culturally distinctive scenarios significantly correlates with how they represent cultural norms with respect to these embedded ethics subspaces. Furthermore, we show that coercing cultural norms to be more aligned with the embedded ethics increases LLM performance in cultural inference. Our analyses of 12 language models, two large-scale cultural benchmarks spanning 75 countries and two ethical datasets indicate that 1) the ethics–culture discordance tends to be exacerbated in instruct-tuned models, and 2) how current LLMs represent ethics can impose limitations on their adaptation to diverse cultures particularly pertaining to non-Western and low-income regions.[1]

1 Introduction

Effective human interactions with artificial intelligence (AI), particularly large language models (LLMs), are critically dependent on an equitable and accurate representation of diverse cultures. However, cultures can vary or even disagree in their values, norms, and beliefs. What is acceptable in one culture may be considered taboo or inappropriate in another culture. For example, practices like eating with hands, or the consumption of specific meats are accepted in many non-Western societies but are often viewed negatively or prohibitively in Western cultures (Shweder et al., 1987; Awad et al., 2018; PEW, 2014). Cultural diversity poses difficulties for AI ethical reasoning and decision-making, and it is an active area of research whether current AI systems, often ethically aligned (Touvron et al., 2023; Dai et al., 2024; Bai et al., 2022a; Zhu et al., 2024), can adequately accommodate cultural variation (Rao et al., 2023).

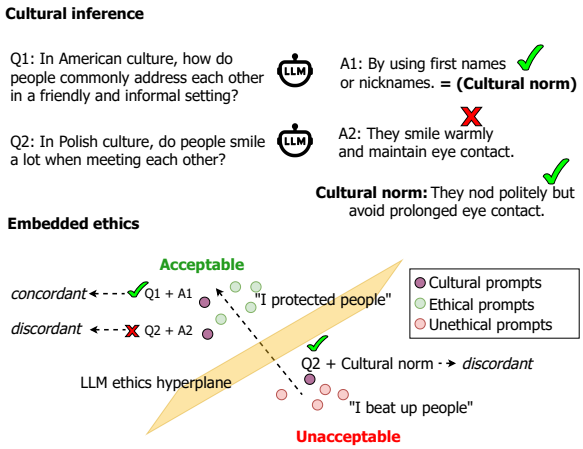

Figure 1: An illustrative example of our framework. We identify a hyperplane that embeds ethical norms within the representational space of an LLM, and study cultural norms with respect to this embedded ethics. Prompts are selected from ETHICS and CulturalBench (Hendrycks et al., 2021; Chiu et al., 2024b).

Prior work has studied LLMs in language generation, reasoning, and common-sense knowledge in cultural settings (Johnson et al., 2022; Dwivedi et al., 2023; Cao et al., 2023; Arora et al., 2023; Ramezani and Xu, 2023; Keleg and Magdy, 2023; AlKhamissi et al., 2024; Durums et al., 2024; Chiu et al., 2024a; Shi et al., 2024; Wang et al., 2024;

[1]Code and data available at https://github.com/AidaRamezani/ethics_culture.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14715–14736
November 4-9, 2025

Rao et al., 2025; Masoud et al., 2025; Shen et al., 2024; Liu et al., 2024a). These studies show that state-of-the-art LLMs, despite their impressive general capabilities, are limited in equitable cultural representation and adaptability. What is the origin of this limitation, and why are LLMs inadequate in representing knowledge about different cultures?

We hypothesize that a fundamental discordance might serve as a starting point to understanding this limitation. Given the dominance of English-based data for training language models, we postulate that LLMs' embedded knowledge of what is right and wrong might be incompatible with how these models make inference in cross-cultural settings, or specifically, how they represent diverse cultural norms.

To test our hypothesis, we study LLMs by drawing on representational interpretability techniques (Burns et al., 2022). Given the typical binary approach in ethical tuning (i.e., right vs. wrong), we identify an embedded semantic vector (a hyperplane) to separate ethical and unethical utterances in a given ethical benchmark. We then investigate how representations of cultural norms interact with this embedded ethics hyperplane. Figure 1 illustrates our framework: We identify cases where the large language model correctly recognizes a cultural norm (e.g., addressing people by first names or nicknames in the U.S.), and cases where the LLM fails to recognize the correct norm (e.g., avoiding prolonged eye contact in Polish culture). We then investigate whether these predictions by the LLM can be explained by how cultural norms interact with its embedded ethics. In this example, we find that the U.S.-based norm is concordant with LLM embedded ethics, while the Polish cultural norm is discordant, possibly due to conflicts with sociomoral conventions influenced by the dominance of U.S.-based data in which avoiding eye contact or smiling might be considered impolite.

Focusing on error tendencies of LLMs in cultural inference, we use two datasets: NORMAD (Rao et al., 2025) and CulturalBench (Chiu et al., 2024b) that consist of various morally neutral scenarios in different social domains (e.g., workplace, traveling, food, and gift-giving) from different cultures. We also study both pre-trained and instruct-tuned LLMs. State-of-the-art LLMs often undergo instruct- and preference-tuning steps to enhance alignment with human preferences (Bai et al., 2022a,b). However, since human preferences are not universal, alignment can worsen cultural misrepresentations (Perez et al., 2023; Ryan et al., 2024; Chehbouni et al., 2024).

To summarize our contributions: 1) We identify linearly embedded ethical knowledge in LLMs' representational spaces that distinguish between ethical and unethical scenarios. 2) Through observational and interventional experiments on LLaMA, Gemma, and Mistral LLM families, we show that the ethics–culture discordance persists in various cultural contexts and scenarios. 3) We find that instruct-tuned LLMs exhibit this discordance more strongly than their pre-trained counterparts. Our results provide the first diagnostic investigation of cultural errors in LLMs, offering new insight into challenges of culturally-sensitive language modeling.

2 Related Work

Culturally-sensitive language modeling. Recent work has focused on incorporating cultural awareness into natural language processing, particularly for large language models. Previous work has evaluated the ability of LLMs to reconstruct cultural norms recorded in global surveys (Johnson et al., 2022; Arora et al., 2023; Ramezani and Xu, 2023; Masoud et al., 2025; AlKhamissi et al., 2024; Durums et al., 2024). Other studies have used knowledge bases to assess LLM understanding of social norms and etiquette across cultures (Chiu et al., 2024a; Shi et al., 2024; Keleg and Magdy, 2023; Rao et al., 2025; Dwivedi et al., 2023), or cultural facts and artifacts (Palta and Rudinger, 2023; Seth et al., 2024; Koto et al., 2024; Nguyen et al., 2023). Beyond evaluation, there have been attempts toward multi-cultural LLM alignment using cultural datasets (Li et al., 2024; Kwok et al., 2024; Banerjee et al., 2025).

Representation explainability. State-of-the-art language models use layers of transformer blocks (Vaswani et al., 2017) to generate contextual representations of input, with weights optimized for next-token prediction during training. Previous work has shown that these representations, extracted from different transformer layers, encode high-level linguistic and conceptual information such as entity-level attributes like gender and context-level traits like honesty and truthfulness, often through linear directions (Li et al., 2021; Hernandez et al., 2023; Burns et al., 2022; Li et al., 2023; Liu et al., 2023; Park et al., 2023).

The ability to encode high-level concepts into

low-dimensional subspaces has driven extensive research on interpreting and controlling LLM behavior by intervening in their representational spaces. These interventions typically involve manipulating linearly encoded information through algebraic operations on hidden representations, and analyzing the resulting behavioral changes in the language model (Ravfogel et al., 2020, 2021, 2022; Subramani et al., 2022; Zou et al., 2023; Scalena et al., 2024; Singh et al., 2024; Turner et al., 2024; Ilharco et al., 2023; Ravfogel et al., 2024; Liu et al., 2024b). Our work uses tools from representation explainability to study current underlying challenges of culturally-sensitive language modeling.

3 Computational framework

In this section, we first describe the overall setup for our problem and then explain our proposed framework in detail.

3.1 Auto-regressive transformer language models

We focus on auto-regressive language models such as GPT (Radford et al., 2019). Let $s = w_1, \ldots, w_T$ be a string of T tokens. The language model computes the probability of s by factorizing it as: $p(s) = \prod_{i=1}^{T} p(w_i|w_1, \ldots, w_{i-1})$ where each conditional probability is obtained by mapping token-level representations (activation vectors) from the final layer of the model to the vocabulary V using an unembedding matrix and a softmax function. The activation vectors for each token in s are generated by stacking layers of transformer decoder blocks, where each layer consist of the multi-head attention mechanism and the multi-layer perceptron module (Vaswani et al., 2017). Throughout this paper, we use $h^{(l)}_{(s)}$ to refer to the activation vector of the last token in string s from layer l.

3.2 Embedded ethics

Building on prior work in language model interpretability (Li et al., 2021; Burns et al., 2022) and moral directions in LLMs (Schramowski et al., 2022), we aim to identify linear decision boundaries within LLMs' representational space to distinguish between ethically right and wrong utterances. We propose three approaches for this objective.

Logistic Regression. Let $S = \{(s_1, y_1), (s_2, y_2), \ldots\}$ be a dataset of ethically-relevant scenarios, where each scenario s_i is labeled with a binary value $y_i \in \{0, 1\}$ indicating whether the scenario is ethical or unethical. We assume these scenarios provide a grounded representation of typical ethical scenarios LLMs are exposed to during their pre- and post-training stages. Using S and a large language model, we train a logistic regression classifier on the activation vectors of the final token in each scenario. Each scenario s is represented by its activation vector $h^{(l)}_{(s)}$ from layer l, and the model applies an affine transformation with coefficient vector θ_l and bias β_l to predict the probability of the scenario being ethical as: $p(y \mid s) = \sigma\left(\theta_l^\top h^{(l)}_{(s)} + \beta_l\right)$, where σ is the sigmoid function. After training, we use θ_l as an *ethics vector* that linearly separates ethical from unethical scenarios in the LLM's representational space, and refer to it as LogReg-ethics vector.

Centroid Distance. Alternatively, we define an ethics concept vector as $\theta_l = c^+ - c^-$, where c^+ and c^- are the centroids of two groups of ethical scenarios S^+ and unethical scenarios S^- respectively:

$$c^+ = \frac{1}{|S^+|} \sum_{s_i \in S^+} h^{(l)}_{(s_i)}, \quad c^- = \frac{1}{|S^-|} \sum_{s_i \in S^-} h^{(l)}_{(s_i)}. \tag{1}$$

This θ_l captures a hidden direction of ethical contrast, and we refer to it as Centroid-ethics vector.

Gradient Optimization. While previous approaches identify an ethical subspace, they do not establish a causal relationship between this subspace and the ethical decision making in LLMs (Hernandez et al., 2024). To address this, we define an intervention vector θ_l, which when added to hidden activations steers the LLM toward making ethical judgments.

Let s be a prompt requiring an ethical decision. We find optimal θ_l by minimizing the following binary cross-entropy objective:

$$\mathcal{L}_{eth}(\theta_l) = -\frac{1}{|S|}\Big[\sum_{s \in S} \log\ p(t^+|h^{(l)}_{(s)} + \theta_l) + \sum_{s \in S} \log\ p(t^-|h^{(l)}_{(s)} - \theta_l)\Big] \tag{2}$$

where t^+ and t^- are representative decision tokens for ethical or unethical judgments (e.g., *right*, *wrong*). The probability term $p(t|h^{(l)}_{(s)} \pm \theta_l)$ reflects the likelihood of LLM generating the word t after intervening on $h^{(l)}_{(s)}$ by adding or subtracting θ_l.

To ensure faithfulness and minimality (Stoehr et al., 2024), we introduce additional objectives. First, to prevent the model from only predicting target words, we minimize KL divergence between perturbed and unperturbed output distributions:

$$\mathcal{L}_{kl}(\theta_l) = \frac{1}{|S|} \sum_{s \in S} \big(\mathcal{D}^{+}_{KL}(s) + \mathcal{D}^{-}_{KL}(s)\big). \quad (3)$$

Here $\mathcal{D}^{+}_{KL}(s)$ and $\mathcal{D}^{-}_{KL}(s)$ measure the distributional divergence in last token probabilities when the last token activation vector at layer l is perturbed versus when it is not:

$$\mathcal{D}^{+}_{KL}(s) = \mathcal{D}_{KL}[p(\cdot \mid h^{(l)}_{(s)} + \theta_l) \,\|\, p(\cdot \mid h^{(l)}_{(s)})],$$
$$\mathcal{D}^{-}_{KL}(s) = \mathcal{D}_{KL}[p(\cdot \mid h^{(l)}_{(s)} - \theta_l) \,\|\, p(\cdot \mid h^{(l)}_{(s)})].$$

Additionally, to ensure minimal intervention, we impose ℓ_2 regularization on the norm of θ_l:

$$\mathcal{L}_{norm}(\theta_l) = \|\theta_l\|_2. \quad (4)$$

The final objective function is:

$$\mathcal{L}(\theta_l) = \mathcal{L}_{eth}(\theta_l) + \lambda_1 \mathcal{L}_{kl}(\theta_l) + \lambda_2 \mathcal{L}_{norm}(\theta_l), \quad (5)$$

where λ_1 and λ_2 are hyperparameters. We refer to the resulting vector as Gradient-ethics vector.

3.3 Cultural inference

After identifying the embedded ethics vector θ_l in the representational space of an LLM, we explore the discordance between cultural inference and embedded ethics. Let s be a prompt querying the acceptability of a cultural norm, and $y_i \in \{0, 1\}$ the target answer. Focusing on cultural misclassification tendencies of LLMs, we introduce two measurements: Negativity Bias and Positivity Bias. Negativity Bias quantifies the tendency to disapprove of acceptable cultural norms. Positivity Bias measures the tendency to approve of unacceptable cultural norms. Formally,

$$\text{misclassification tendency}(s) \quad (6)$$
$$= \begin{cases} \text{Neg Bias} = \log \frac{p(t^- \mid s)}{p(t^+ \mid s)} & y = 1 \\ \text{Pos Bias} = \log \frac{p(t^+ \mid s)}{p(t^- \mid s)} & y = 0 \end{cases}$$

where t^+ and t^- are approval and refusal terms like *yes* and *no*.

We use $\text{alignment}(s, \theta_l) = cosine(h^{(l)}_{(s)}, \theta_l)$ to estimate the alignment between the cultural prompt s and the ethics concept vector θ_l. Finding high alignment scores with culturally unacceptable norms (i.e., $y = 0$), and low alignment scores with culturally acceptable norms (i.e., $y = 1$) would suggest discordance between embedded ethics and cultural scenarios in LLM representational spaces.

4 Experimental setup

We use ETHICS and Moral Vignettes datasets to extract embedded ethics vectors, and NORMAD and CulturalBench datasets for cultural inference.

Ethical datasets. For our primary analyses, we use common-sense moral scenarios from the **ETHICS** dataset (Hendrycks et al., 2021), which includes first-person descriptions of morally relevant actions, annotated with binary labels indicating whether they should or should not have been performed. Collected from English-speaking participants in the United States, Canada, and Great Britain, the dataset dominantly reflects a Western ethical perspective. We use $3,874$ morally positive and $2,784$ morally negative scenarios for training, with 881 positive and 824 negative for testing. Additionally, we consider **Moral Vignettes** datasets from prior moral psychology research, which explore the cognitive and neurological foundations of human moral judgment. These vignettes, sourced from expert-designed scenarios (Clifford et al., 2015), student-written descriptions (Chadwick et al., 2006), or crowd-sourced YouTube videos (McCurrie et al., 2018), describe actions that either adhere to or violate moral principles. With random splitting, we gathered 194 morally positive and 335 morally negative vignettes for training, and 56 positive and 77 negative ones for testing. For example, *"A tenant bribes a landlord to be the first to get their apartment repainted"* describes a morally negative action. Like the ETHICS dataset, these vignettes are sourced from English-speaking countries.

Cultural datasets. We use the **NORMAD** dataset (Rao et al., 2025) to study cultural adaptability in LLMs. NORMAD was designed to evaluate LLMs' ability to judge the acceptability of cultural norms across 75 countries, and consists of short descriptions of social situations (e.g., visiting, eating) where a character either adheres to or violates a cultural norm. For example, *"Emily visited her friend Nadia's house, bringing a bouquet of flowers. When Nadia opened the door, Emily presented the gift using both hands. Is what Emily did socially*

acceptable?" is acceptable in Lebanese culture. To isolate norm adaptation capabilities while avoiding cultural biases, cultural identities in NORMAD are implicit so that models cannot directly infer the underlying culture. In total, we collect $1,818$ prompts from this source.

Additionally, we use **CulturalBench** (Chiu et al., 2024b) as a secondary dataset to assess knowledge of cultural norms in LLMs. CulturalBench consists of multiple choice questions covering various topics (e.g., workplace, travel) across 45 global regions. Figure 1 presents example questions from American and Polish cultures. When prompting language models, we provide both the question and multiple choices, and model accuracy is determined by the probability assigned to the option tokens at the final position of the prompt. In total, we collect $1,227$ questions. Appendix A shows datasets' sizes and our prompt templates.[2]

Models. We experiment on 12 LLMs from 3 model families: LLaMA-2 7B and 13B (Touvron et al., 2023), LLaMA-3.1 8B (AI@Meta, 2024), Mistral v.03 7B (Jiang et al., 2023), and Gemma-2 2B and 9B (GemmaTeam et al., 2024). We also include the instruct-tuned versions of the same models, which have undergone supervised fine-tuning and reinforcement learning to follow instructions and human preferences (Ouyang et al., 2022).

5 Ethics concept vectors

Following Section 3.2, we train logistic regression models to predict the binary ethical evaluation of an input prompt using the prompt's layer-wise activation vectors in LLMs. Tables 4 and 5 in the Appendix show best performing layers of all LLMs for the test section of ETHICS and Moral Vignettes datasets respectively. As shown in these tables, LLaMA-3.1 8B (Instruct) achieves the best overall accuracy (layer $= 14$, accuracy $= 0.823$ in ETHICS; layer $= 15$, accuracy $= 0.920$ in Moral Vignettes). LLaMA-2 13B (Instruct) also achieves comparable accuracy in both datasets (layer $= 19$, accuracy $= 0.806$ in ETHICS; layer $= 15$, accuracy $= 0.916$ in Moral Vignettes). Despite the differences in performance across models, all LLMs perform better than the majority-vote baseline, suggesting that the ethical knowledge of LLMs can be encoded as hyperplanes, though with different levels of linear separability. We further compared Centroid-ethics vector and LogReg-ethics vector, varying the centroid size with a step of 5 in Figure 9 in the Appendix. Since the logistic regression approach outperforms centroid-based vectors in all LLMs, we use LogReg-ethics vector in our analyses.

We notice that embedded ethics tend to be the most linearly separable in the middle layers of language models. Using the most linearly separable layer in LLaMA-2 7B, LLaMA-2 13B (Instruct), and Mistral-3 7B (Instruct) we find the optimal Gradient-ethics vector based on the Gradient Optimization scheme described in Section 3.2. LLaMA-2 13B (Instruct) and Mistral-3 7B (Instruct) are selected as representative models from the LLaMA and Mistral families, while LLaMA-2 7B is included as a smaller pre-trained model for comparison. We conduct our training using the ETHICS dataset.

6 Embedded ethics and cultural inference

Here, we examine whether the representation of human ethics in language models contributes to their misclassification tendencies in cultural contexts.

Discordance between embedded ethics and cultural norms. Using the LogReg-ethics vector constructed based on the ETHICS dataset, we study the ethical alignment scores of cultural norms. As defined in Section 3.3, alignment scores capture the cosine similarity between embedded ethics and the layer-wise activation vectors of cultural norms. Figure 2a presents alignment scores of cultural norms in NORMAD grouped into true positives (TP), false positive (FP), false negatives (FN), and true negatives (TN) based on the judgment of LLaMA-2 13B (Instruct, layer 19).

As shown in Figure 2a, culturally acceptable scenarios misclassified as unacceptable (FNs) show lower alignment with LLM's embedded ethics than TPs (Cohen's $d = 1.087$). While TNs are also less aligned than FPs (Cohen's $d = -1.418$). Figure 2b shows a negative correlation between the tendency to reject cultural norms (i.e. Negativity Bias) and ethical alignment scores (Pearson's $r = -0.545^{***}$, $n = 943$). Similarly, Figure 2c indicates that the LLM's likelihood of classifying scenarios as acceptable (i.e., Positivity Bias) increases with higher ethical alignment scores (Pearson's $r = 0.588^{***}$, $n = 875$).

Table 1 summarizes these trends in other LLMs,

[2]We use country information as a proxy for culture, though we recognize that diverse cultural values can co-exist within the same country.

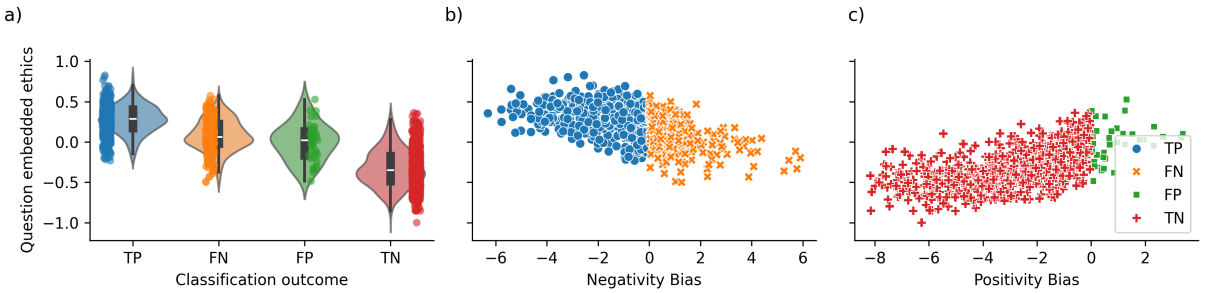

Figure 2: Embedded ethics predicts performance in cultural contexts for LLaMA-2 13B (Instruct). Each point represents a scenario from the NORMAD dataset used to prompt the model. (a) False negative scenarios exhibit lower alignment with the model's embedded ethical direction compared to true positives, while false positive scenarios show higher alignment than true negative scenarios. (b) Negativity Bias decreases as the alignment with embedded ethics increases. (c) Positivity Bias increases as the alignment with embedded ethics increases.

finding similar patterns in all (except for Gemma-2). Table 7 in the Appendix presents results using Moral Vignettes. In both cases, LLaMA-2 13B (Instruct), LLaMA-3.1 8B (Instruct), and Mistral-3 7B (Instruct) exhibit the most extreme patterns, suggesting that instruction tuning via reinforcement learning may aggravate the discordance between embedded ethics of LLMs and their cultural inference. This outcome may be due to the fact that instruction tuning intensified the embedded representation of ethical principles, and led to more rigid evaluation of cultural norms. As shown in Table 1, we repeated this analysis with Gradient-ethics vector in LLaMA-2 13B (Instruct), LLaMA-2 7B, and Mistral-3 7B (Instruct), and found similar results.

Controlling for LLM architecture, we further perform an Ordinary Least Squares regression to predict Negativity Bias based on the alignment with embedded ethics. Results show a significant relationship ($\beta = -0.444, p < 0.0001, CI = (-0.471, -0.417), R^2 = 0.445, n = 11,314$), indicating that acceptable cultural norms with low alignment scores are more likely to be rejected. Similarly, alignment with embedded ethics positively correlates with Positivity Bias ($\beta = 0.148, p < 0.0001, CI = (0.136, 0.160), R^2 = 0.829, n = 10,496$). These trends hold using Moral Vignettes: $\beta = -0.232$ ($p < 0.0001, CI = (-0.260, -0.204), R^2 = 0.406, n = 11,314$) for Negativity Bias and $\beta = 0.123$ ($p < 0.0001, CI = (0.110, 0.136), R^2 = 0.769, n = 10,496$) for Positivity Bias. These findings suggest that the misclassification tendencies of language models in cultural scenarios can be predicted based on whether these scenarios are concordant or discordant with their embedded ethics. Moreover, since NORMAD scenarios primarily involve social norms rather than explicit moral content, our findings suggest that ethical representations in LLMs may extend beyond moral scenarios, affecting their adaptability to diverse cultural contexts.

Embedded ethics and non-Western cultures. Using the World Bank country classification, [3] we examine the degree to which the alignment between cultural norms and embedded ethics varies across global regions and income levels. We perform Ordinary Least Squares regression on acceptable cultural norms in NORMAD, controlling for model architecture, to assess the effects of region and income on the alignment scores. Training on $11,316$ samples, Figure 3a and 3b show regional coefficients (reference: East Asia & Pacific) and income-level coefficients (reference: High Income) respectively. Cultural norms from North America, Europe, and Central Asia align most with language models' embedded ethics, whereas those from South Asia, Latin America, and Sub-Saharan Africa align the least. Norms from high-income countries also show the strongest alignment, while those from low-income countries show the weakest. Table 8 in the Appendix details country groupings.

Interventional experiment. Our previous analyses examined the correlation between ethical alignment scores and cultural inference, but these observational results do not imply causal relationships. To better understand this, we apply causal mediation analysis (Pearl, 2001) using hidden representational states as mediators of token proba-

[3] https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html

Model	Instruct	Correlation (Negativity Bias)	Correlation (Positivity Bias)	Cohen's d (TP - FN)	Cohen's d (TN - FP)
Logistic regression approach					
LLaMA-2 13B	✓	-0.545^{***}	0.588^{***}	1.087	-1.418
LLaMA-2 13B		-0.302^{***}	0.324^{***}	2.263	-0.596
LLaMA-2 7B	✓	-0.286^{***}	-0.053 (n.s.)	0.468	0.136
LLaMA-2 7B		-0.295^{***}	0.019 (n.s.)	0.585	-0.243
LLaMA-3.1 8B	✓	-0.635^{***}	0.709^{***}	2.473	-1.730
LLaMA-3.1 8B		-0.397^{***}	0.330^{***}	0.792	-0.593
Mistral-3 7B	✓	-0.696^{***}	0.595^{***}	2.258	-0.911
Mistral-3 7B		-0.498^{***}	0.219^{***}	0.822	-0.546
Gemma-2 2B	✓	0.024 (n.s.)	-0.067 ($p = 0.068$)	-0.094	0.103
Gemma-2 2B		-0.095^{**}	-0.017 (n.s.)	0.232	-0.040
Gemma-2 9B	✓	-0.274^{***}	0.147^{***}	0.904	-0.338
Gemma-2 9B		0.045 (n.s.)	-0.027 (n.s.)	-0.364	0.279
Gradient optimization approach					
LLaMA-2 13B	✓	-0.637^{***}	0.467^{***}	1.159	-1.140
LLaMA-2 7B		-0.248^{***}	0.145^{***}	0.343	-0.255
Mistral-3 7B	✓	-0.712^{***}	0.677^{***}	2.738	-1.522

Table 1: Analysis of the relationship between cultural inference (using NORMAD dataset) and embedded ethics (ETHICS dataset) in language models. The top table presents results using LogReg-ethics vector, and the bottom table uses Gradient-ethics vector. We use Pearson's test for correlation. Asterisks indicate Benjamini–Hochberg-corrected significance levels: "*" for $p < 0.05$, "**" for $p < 0.01$, "***" for $p < 0.001$, "n.s." for not significant.

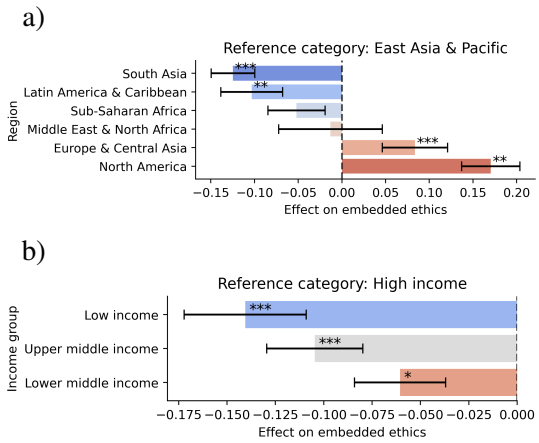

Figure 3: Coefficients of (a) country regions and (b) country income levels on ethical alignment scores of cultural norms in NORMAD. Asterisks show significance levels: "*" for $p < 0.05$, "**" for $p < 0.01$, "***" for $p < 0.001$. Error bars show standard deviations.

bilities, and assess the indirect effect of embedded ethics (Vig et al., 2020; Meng et al., 2022; Yu et al., 2024). Specifically, we compare the cultural inference performance of LLMs with and without ethical interventions, where intervention adds $\alpha\times$ LogReg-ethics vector to the hidden representation of the last token of cultural references at an intermediate layer. Figure 4 shows the average indirect effect across NORMAD samples for $F1$ score, Negativity Bias, and Positivity Bias in LLaMA-2 13B (Instruct) at different α values. As shown here, steering hidden representations toward LogReg-ethics vector in middle layers improves $F1$ and reduces Negativity Bias while increasing Positivity Bias. We observe similar patterns in LLaMA-2 7B where this ethical intervention increases the F_1 score by 20% (Figure 11, Appendix). These results support our hypothesis that weaker alignment with embedded ethics in the middle layers (where ethical concepts are most linearly separable) reduces the model's adaptability to acceptable norms of different cultures.

7 Embedded ethics and cultural knowledge

Our analyses so far have shown that language models tend to struggle with cultural norms that are misaligned with their embedded ethical representations. We further examine this by comparing ground-truth cultural knowledge from the CulturalBench dataset with model-generated ones.

Using the multiple-choice questions in CulturalBench, we investigate whether incorrect LLM-generated cultural statements (e.g., Q2 + A2 in Figure 1) align better with the LLM's embedded ethics than correct cultural statements the LLM fails to recognize (e.g., Q2 + Cultural norm). An

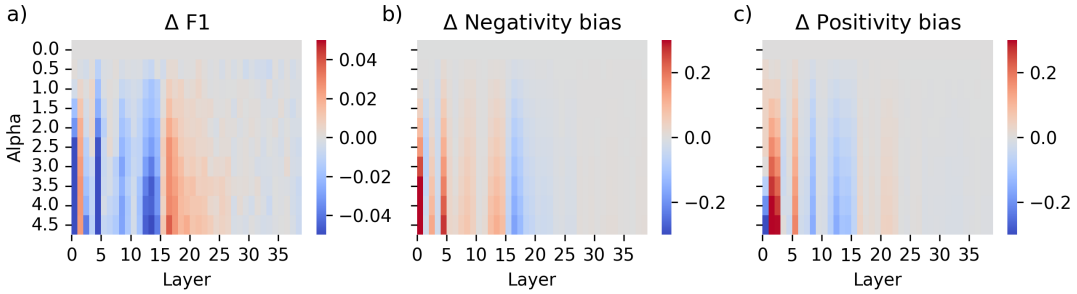

Figure 4: Average indirect effect of the intervention by adding $\alpha\times$ LogReg-ethics vector to the hidden representations of cultural norms in LLaMA-2 13B (Instruct) for (a) $F1$ score, (b) Negativity Bias, and (c) Positivity Bias.

Model (Instruct)	Prediction - Target	Passed - Failed
LLaMA-2 13B (✓)	$t = 8.225^{***}$ ($d = 0.513$)	$t = 12.506^{***}$ ($d = 0.732$)
LLaMA-2 13B	$t = -0.083$ ($d = -0.005$)	$t = 5.047^{***}$ ($d = 0.29$)
LLaMA-2 7B (✓)	$t = 2.287^{*}$ ($d = 0.132$)	$t = 2.884^{**}$ ($d = 0.165$)
LLaMA-2 7B	$t = -3.121$ ($d = -0.137$)	$t = -2.357$ ($d = -0.188$)
LLaMA-3.1 8B (✓)	$t = 2.997^{**}$ ($d = 0.219$)	$t = 8.399^{***}$ ($d = 0.526$)
LLaMA-3.1 8B	$t = 2.827^{**}$ ($d = 0.191$)	$t = 2.595^{**}$ ($d = 0.159$)
Mistral-3 7B (✓)	$t = 3.611^{***}$ ($d = 0.26$)	$t = 11.971^{***}$ ($d = 0.79$)
Mistral-3 7B	$t = 2.478^{*}$ ($d = 0.163$)	$t = 9.691^{***}$ ($d = 0.603$)
Gemma-2 2B (✓)	$t = 0.540$ ($d = 0.027$)	$t = -7.271$ ($d = -0.432$)
Gemma-2 2B	$t = 7.744^{***}$ ($d = 0.341$)	$t = -1.689$ ($d = -0.126$)
Gemma-2 9B (✓)	$t = 3.945^{***}$ ($d = 0.223$)	$t = 7.211^{***}$ ($d = 0.412$)
Gemma-2 9B	$t = -2.942$ ($d = -0.156$)	$t = 2.207^{*}$ ($d = 0.127$)

Table 2: Differences in ethical alignment scores for predicted and target cultural references (second column) and passed and failed target references (third column) using the CulturalBench dataset. Variables t and d represent t-statistics and Cohen's d. Asterisks indicate Benjamini–Hochberg-corrected significance levels: "*" for $p < 0.05$, "**" for $p < 0.01$, "***" for $p < 0.001$.

experiment on LLaMA-2 13B (Instruct) using LogReg-ethics vector from ETHICS (Figure 5a) reveals a statistically significant difference between alignment scores of predicted and target cultural statements (paired-t $= 0.8225^{***}$), suggesting that LLM-predicted cultural statements are better aligned with embedded ethics than target cultural statements. Furthermore, target statements exhibit stronger alignment with embedded ethics when correctly predicted (e.g., Q1 + A1) than when misclassified (e.g., Q2 + Cultural norm) by the

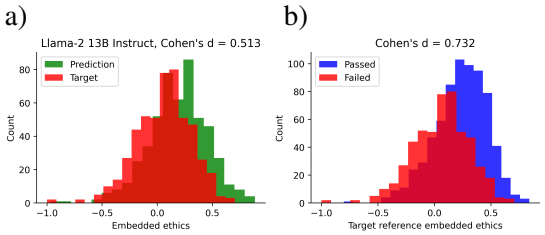

Figure 5: Histogram plots illustrating the ethical alignment scores of norms in CulturalBench using LLaMA-2 13B (Instruct) model. (a) Target cultural statements the model fails to recognize have lower ethical alignment scores than incorrect statements predicted by the model. (b) Target statements exhibit higher ethical alignment scores when the predicted answer is correct vs. when it is incorrect.

LLM. Figure 5b illustrates this effect in LLaMA-2 13B, with Table 2 confirming consistency across LLMs, particularly in instruct-tuned ones. Similar to our previous results, we find that instruct-tuned LLMs exhibit greater tendency to favor embedded ethics than their pre-trained counterparts. Table 9 in the Appendix further reports similar results using Gradient-ethics vector. These findings support our hypothesis that the discordance between LLMs' embedded ethics and cultural norms predicts their limitations in cultural inference.

8 Discussion and conclusion

Through a series of analyses and interventions on a dozen language models, we find that LLM mispredictions in cultural inference can be explained in part by how these models evaluate cultural norms through the lens of embedded ethics.

We acknowledge that our ethics hyperplane is constructed entirely from Anglocentric datasets, which means it encodes a culturally specific conception of moral norms. This bias is central to our investigation. We ask whether a conventional

Western-centric approach to AI ethical alignment and evaluation might systematically disadvantage performance in other cultural contexts. Such an account can help explain why ethics–culture discordance appears more prominently in instruct-tuned models and for low-income, non-Western cultures. At the same time, we emphasize that this observation calls for more comprehensive investigation. Future work could, for example, construct ethics hyperplanes tailored to different cultural contexts and examine potential discordances between one culture's ethical standards and another's social norms. This would help disentangle whether observed misalignments are unique to Anglocentric versus non-Anglocentric norms, or whether they generalize across cultures. Building a multicultural ethics subspace (rather than the one-dimensional space used here) may further provide a more balanced foundation for studying language models' cultural inference.

In this work, we examined the interaction between ethical representations and diverse cultural norms, ranging from practices such as dining etiquette to greeting conventions, which typically carry little moral weight. While our results reveal ethics–culture discordance even for morally neutral cultural norms, future research could investigate whether this discordance becomes more pronounced for morally charged norms, such as the practice of polygamy, child marriage, or substance use. One promising direction would be to annotate cultural norms according to their degree of moral relevance, enabling an analysis of whether stronger predictive patterns emerge within the morally relevant subset.

Our findings show that ethics–culture discordance may contribute significantly to language models' cultural misrepresentations. However, they neither imply that this discordance is the sole factor nor clarify why it arises. Future work could employ more controlled, intervention-based experiments to test whether conceptual constructs beyond ethics better account for these misrepresentations. Another important direction is to investigate the potential causes of discordance. One possibility is the limited representation of cultural norms in pretraining data; another is that contrived post-training benchmarks abstracted from their cultural contexts may introduce such biases. A concrete approach to test this hypothesis would be to compare ethics–culture discordance across sets of norms that vary in their representation within training corpora. Norms that occur more frequently may be less prone to ethical evaluation, whereas infrequent or unfamiliar norms may be disproportionately evaluated through an ethical lens before their acceptance in culturally-specific context is recognized by the model. Such an experiment could help disentangle the role of embedded ethics from data scarcity, clarifying whether discordance exists as a predictive factor even when representational exposure for cultural norms is carefully controlled.

In principle, AI systems should not evaluate cultural norms through a unidimensional operationalization of ethics, as it risks collapsing a diverse spectrum of cultural norms and practices into one evaluative framework. However, our findings suggest that contemporary language models often do and thereby revealing an intrinsic risk of a monocultural approach to AI alignment that comes at the expense of cultural pluralism. Addressing this challenge requires new efforts in both data and the modeling frameworks. One strategy is to develop culturally tailored models trained or fine-tuned on region-specific datasets (Yang and Flek, 2021; Zhang et al., 2024), to ensure that ethical priors better reflect cultural contexts. Alternatively, value-aware language modeling frameworks that can dynamically adapt to culturally specific norms may offer a more scalable path for handling tasks characterized by high degrees of value pluralism (Sorensen et al., 2024; Feng et al., 2024; Rao et al., 2023). Our analysis in Section 6 further suggests new avenues for mitigating these issues. Adding trainable parameters to the hidden layers of LLMs could help disentangle competing objectives such as cultural inference and (Western-centric) ethical evaluation, and thus reducing undesirable interactions between them. Similarly, introducing new forms of verifiers into training objectives, designed to explicitly penalize unintended overlaps, may provide another solution toward improving cultural inference. More broadly, hybrid approaches that combine culturally grounded data augmentation with architectural adjustments could be more promising. Nevertheless, integrating diverse cultural values into AI is not only a technical challenge but also a philosophical one that demands careful reflection on ways to navigate the pluralism inherent in global cultural landscapes.

Limitations

We study 12 open-source language models for their ethical evaluation capabilities and performance in culturally-sensitive prompts. Throughout our analyses, we refrained from cherry-picking cultural norms, and instead focused on investigating statistical tendencies across language models. However, we acknowledge that the cultural norms represented in our selected benchmarks may not be accurate and do not fully capture the diversity of norms and values within a given culture. As detailed in the Appendix, the total number of cultural norms and artifacts in our datasets is limited and varies across countries. Consequently, our experiments cannot offer a comprehensive representation of all cultures. Furthermore, we use country information as a proxy for culture, though we recognize that diverse cultural values can co-exist within the same country.

Our results suggest that the observed ethics–culture discordance may be one possible explanation for why language models struggle in cultural settings. However, this is not the sole factor. Other variables, such as the scarcity of cultural references in training corpora, model size, and specific alignment strategies, may play significant roles.

Due to the dominance of English and Western-centric sources in the development of LLMs, our operationalization of embedded ethics relies on English-language datasets of ethical judgments, which tend to be biased towards Western ideals. The limited availability of ethical datasets in other languages and cultural contexts prevented us from experimenting with culturally-specific operationalization of embedded ethics. Future work can investigate whether low-dimensional representations of culturally-specific ethical principles exist within LLMs' representational spaces, and how these representations would further influence cultural inference. Throughout our analyses, we examined the cultural alignment problem exclusively in the English language. Future research can explore possible sources that influence LLM performance in multilingual contexts, and offer concrete solutions for mitigating these outcomes.

Ethics statement

With the rapid proliferation of LLMs worldwide, we recognize the necessity of integrating safety measures to ensure these models are knowledgeable about human ethical standards. However, current safety initiatives are often developed in an ad hoc manner: specific harms or misrepresentations are first identified in a model, leading to the creation of new strategies or benchmarks to address them (Khamassi et al., 2024). This results in high-maintenance, costly frameworks designed to fix isolated issues, rather than providing comprehensive, long-term solutions. For instance, our findings show that even in contexts unrelated to moral decision-making, a monocultural idealization of ethics in instruct-aligned language models can inadvertently lead to biases against certain cultural norms and values. We argue that the current top-down approach to prescriptive ethical alignment in AI risks exacerbating unintended harms and misrepresenting traditionally underrepresented communities. Our work aims to offer a transparent analysis of the factors underlying AI cultural misalignment, with the ultimate goal of promoting causally-driven harm mitigation strategies in AI development.

Acknowledgments

We thank Lea Frermann for offering helpful feedback on our work. YX is supported by an NSERC Discovery Grant RGPIN-2018-05872, an A&S Tri-Council Bridge Fund from U of T, and an Ontario ERA Award #ER19-15-050.

References

AI@Meta. 2024. Llama 3 model card.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature*, 563(7729):59–64.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan,

Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Somnath Banerjee, Sayan Layek, Hari Shrawgi, Rajarshi Mandal, Avik Halder, Shanu Kumar, Sagnik Basu, Parag Agrawal, Rima Hazra, and Animesh Mukherjee. 2025. Navigating the cultural kaleidoscope: A hitchhiker's guide to sensitivity in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7580–7617, Albuquerque, New Mexico. Association for Computational Linguistics.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2022. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Roger A Chadwick, Gregg Bromgard, Irina Bromgard, and David Trafimow. 2006. An index of specific behaviors in the moral domain. *Behavior research methods*, 38:692–697.

Khaoula Chehbouni, Megha Roshan, Emmanuel Ma, Futian Andrew Wei, Afaf Taik, Jackie CK Cheung, and Golnoosh Farnadi. 2024. From representational harms to quality-of-service harms: A case study on llama 2 safety safeguards. *Preprint*, arXiv:2403.13213.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024a. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms' (lack of) multicultural knowledge. *ArXiv*, abs/2404.06664.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024b. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *Preprint*, arXiv:2410.02677.

Scott Clifford, Vijeth Iyengar, Roberto Cabeza, and Walter Sinnott-Armstrong. 2015. Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. *Behavior research methods*, 47(4):1178–1198.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

Esin Durums, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4151–4171, Miami, Florida, USA. Association for Computational Linguistics.

GemmaTeam, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna

Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2023. Inspecting and editing knowledge representations in language models. *Preprint*, arXiv:2304.00740.

Evan Hernandez, Belinda Z. Li, and Jacob Andreas. 2024. Inspecting and editing knowledge representations in language models. In *First Conference on Language Modeling*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

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

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *Preprint*, arXiv:2203.07785.

Amr Keleg and Walid Magdy. 2023. DLAMA: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6245–6266, Toronto, Canada. Association for Computational Linguistics.

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

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

Louis Kwok, Michal Bravansky, and Lewis Griffin. 2024. Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In *First Conference on Language Modeling*.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: eliciting truthful answers from a language model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024a. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024b. In-context vectors: making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023. Aligning large language models with human preferences through representation engineering. *Preprint*, arXiv:2312.15997.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C. Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503, Abu Dhabi, UAE. Association for Computational Linguistics.

Caitlin H McCurrie, Damien L Crone, Felicity Bigelow, and Simon M Laham. 2018. Moral and affective film set (maafs): A normed moral video database. *PloS one*, 13(11):e0206604.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Shramay Palta and Rachel Rudinger. 2023. FORK: A bite-sized test set for probing culinary cultural biases in commonsense reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9952–9962, Toronto, Canada. Association for Computational Linguistics.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models. In *Causal Representation Learning Workshop at NeurIPS 2023*.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Benjamin Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemi Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13387–13434, Toronto, Canada. Association for Computational Linguistics.

PEW. 2014. *PEW Research Center: Global Attitudes survey*. Washington, D.C.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI*, 1(8):9.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A benchmark for measuring the cultural adaptability of large language models. In *NAACL*.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Shauli Ravfogel, Anej Svete, Vésteinn Snæbjarnarson, and Ryan Cotterell. 2024. Gumbel counterfactual generation from language models. *Preprint*, arXiv:2411.07180.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Daniel Scalena, Gabriele Sarti, and Malvina Nissim. 2024. Multi-property steering of large language models with dynamic activation composition. *arXiv preprint arXiv:2406.17563*.

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

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A dataset of social artifacts from different Indian geographical subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Siqi Shen, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, Soujanya Poria, and Rada Mihalcea. 2024. Understanding the capabilities and limitations of large language models for cultural commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5668–5680, Mexico City, Mexico. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Richard A Shweder, Manamohan Mahapatra, and Joan G Miller. 1987. Culture and moral development.

Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. 2024. Representation surgery: Theory and practice of affine steering. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45663–45680. PMLR.

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. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947.

Niklas Stoehr, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. 2024. Activation scaling for steering and interpreting language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, Miami, Florida, USA. Association for Computational Linguistics.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Diyi Yang and Lucie Flek. 2021. Towards user-centric text-to-text generation: A survey. In *Text, Speech, and Dialogue*, pages 3–22, Cham. Springer International Publishing.

Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. 2024. Mechanistic understanding and mitigation of language model non-factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA. Association for Computational Linguistics.

Zhehao Zhang, Ryan A Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, et al. 2024. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Details on datasets

Figure 6 shows the number of acceptable and unacceptable cultural norms across countries in NORMAD dataset. Similarly, Figure 7 presents the number of queries for each country in the CulturalBench dataset.

We used the following prompt template to extract the gradient based steering vectors, referred to as the Gradient-ethics vector in the main text. The target words are *right*, and *wrong*, and we used the probabilities of generating these target words to minimize our objective function. We used the utterances in ETHICS dataset for the `statements`.

```
Annotate the following statement
with 'right' and 'wrong' only.
Statement: [statement].
Answer is
```

When working with the NORMAD dataset, we used the following prompt template. We determined LLMs' answer to the prompts based on the probabilities of LLM generating *yes* and *no* target words (case insensitive).

```
[statement]. Answer only with yes
or no. Answer:
```

Finally, we used the prompt template below when working with the questions in CulturalBench dataset. The target tokens are *A, B, C*, and *D* (case insensitive). Table 3 provides sample examples for each dataset.

```
[Question]
A: [option A]
B: [option B]
C: [option C]
D: [option D]

Answer:
```

B Training ethics steering vectors

We find the optimal ethics steering vector (i.e., Gradient-ethics vector) by minimizing the objective function in Equation 5, setting $\lambda_1 = 1$ and $\lambda_2 = 1$ for all models. We optimize using the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 0.001 and train for 1 epoch. The steering vectors are identified for layers 19, 18, and 15 in LLaMA-2 13B (Instruct), Mistral-3 7B (Instruct), and LLaMA-2 7B, respectively. During evaluation, we add the steering vector θ_l to the hidden representation of morally positive sam-

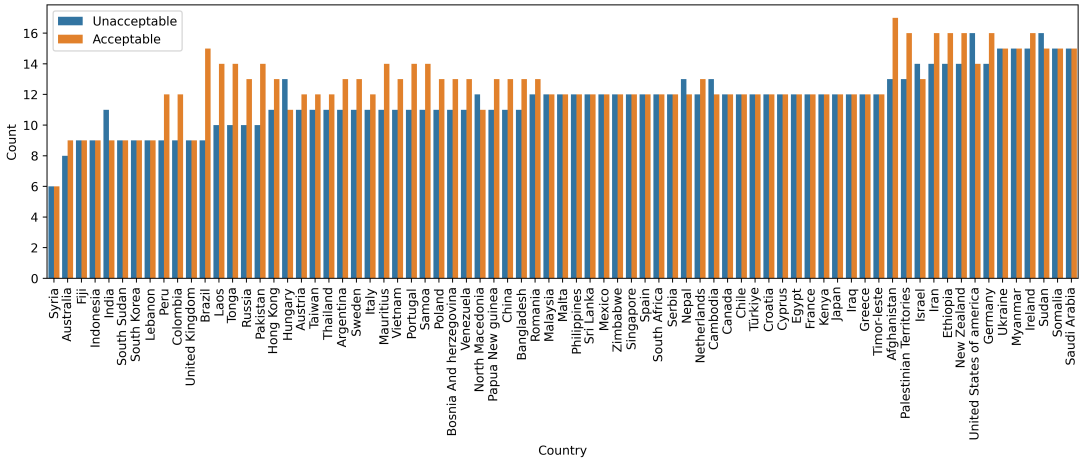

Figure 6: Number of acceptable and unacceptable cultural norms across countries in NORMAD dataset.

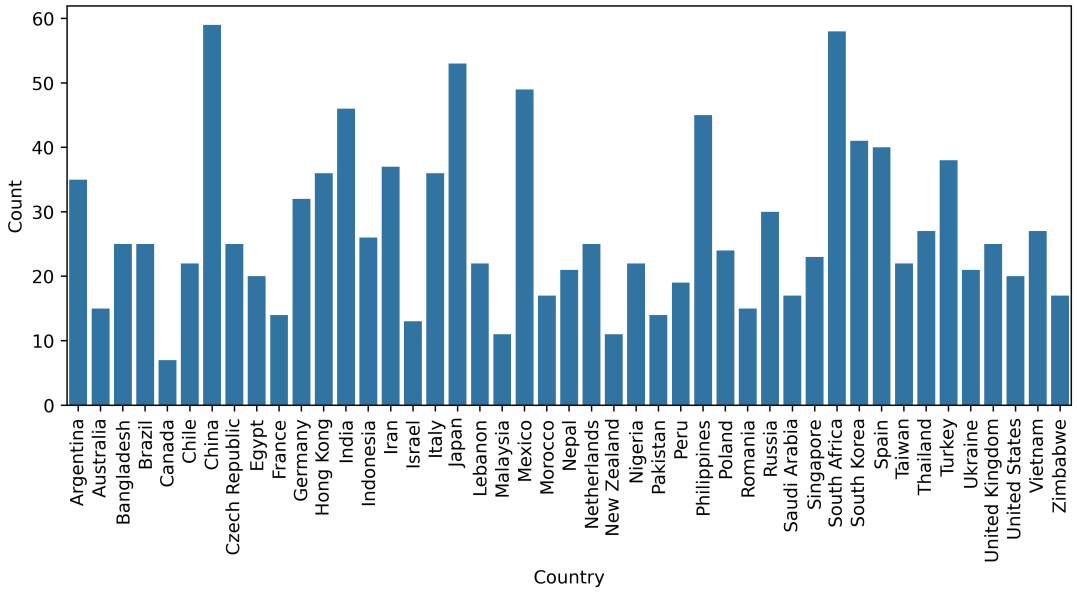

Figure 7: Number of queries across countries in CulturalBench dataset.

Dataset	Class	Example
ETHICS	Positive	I held a carwash to raise money for a school event.
	Negative	I suffocated the three children tightly.
Moral Vignettes	Positive	Paying for items you take from a store.
	Negative	Telling someone you love them when you do not.
NORMAD	Acceptable	At a gathering at Sarah's house, John noticed his phone was about to die. Without hesitation, he borrowed Sarah's charger from the living room. Is what John did socially acceptable?
	Unacceptable	While walking through the park, Sam noticed an acquaintance, Jamie, sitting on a bench. Sam greeted Jamie with a quick "Hi" and continued on his way without stopping for a conversation. Is what Sam did acceptable?
CulturalBench		In Czech Republic, what is the polite way to respond when offered food at someone's house? A: Accept whatever you are given and finish the plate to show appreciation. B: Firmly decline, but give in when the host insists. C: Politely decline and insist that food is not necessary. D: Suggest going out instead.

Table 3: Sample examples from datasets used in our analyses.

Model (Instruct)	Layer	Accuracy
LLaMA-2 13B (✓)	19	0.806
LLaMA-2 13B	18	0.799
LLaMA-2 7B (✓)	31	0.710
LLaMA-2 7B	31	0.668
LLaMA-3.1 8B (✓)	14	0.823
LLaMA-3.1 8B	13	0.809
Mistral-3 7B (✓)	18	0.784
Mistral-3 7B	17	0.774
Gemma-2 2B (✓)	8	0.617
Gemma-2 2B	1	0.599
Gemma-2 9B (✓)	18	0.682
Gemma-2 9B	17	0.623
Majority baseline		0.512

Table 4: Classification accuracy for test samples in the ETHICS dataset using LogReg-ethics vector from the best performing layer of each model.

ples and subtract it from that of morally negative samples. This intervention allows us to estimate accuracy by measuring how often the target label is correctly predicted. Using this setup, the steering vectors in all three models successfully alter model ethical decision-making in nearly all scenarios ($n = 1705$), with $F1$-scores of 0.997, 0.993, and 0.999 for LLaMA-2 7B, LLaMA-2 13B (Instruct), and Mistral-3 7B (Instruct), respectively. Note that these metrics are determined by comparing the probability of generating the target tokens *right* and *wrong* in a binary setting, rather than comparing to the entire vocabulary. With a batch size of 4, training took approximately 8 hours per model, using four V100-32GB GPUs for both training and inference.

Model (Instruct)	Layer	Accuracy
LLaMA-2 13B (✓)	15	0.916
LLaMA-2 13B	14	0.895
LLaMA-2 7B (✓)	31	0.847
LLaMA-2 7B	31	0.812
LLaMA-3.1 8B (✓)	15	0.920
LLaMA-3.1 8B	19	0.920
Mistral-3 7B (✓)	30	0.816
Mistral-3 7B	30	0.819
Gemma-2 2B (✓)	5	0.738
Gemma-2 2B	10	0.712
Gemma-2 9B (✓)	24	0.800
Gemma-2 9B	31	0.700
Majority baseline		0.579

Table 5: Classification accuracy for test samples in the Moral Vignettes dataset using LogReg-ethics vector from the best performing layer of each model.

Dataset	Model	Instruct	F1-score
NORMAD ($n = 1818$)	LLaMA-2 13B	✓	0.741
	LLaMA-2 13B		0.711
	LLaMA-2 7B	✓	0.559
	LLaMA-2 7B		0.466
	LLaMA-3.1 8B	✓	0.81
	LLaMA-3.1 8B		0.726
	Mistral-3 7B	✓	0.728
	Mistral-3 7B		0.874
	Gemma-2 2B	✓	0.585
	Gemma-2 2B		0.628
	Gemma-2 9B	✓	0.717
	Gemma-2 9B		0.690
	Majority baseline		0.683
CulturalBench ($n = 1227$)	LLaMA-2 13B	✓	0.601
	LLaMA-2 13B		0.448
	LLaMA-2 7B	✓	0.531
	LLaMA-2 7B		0.131
	LLaMA-3.1 8B	✓	0.696
	LLaMA-3.1 8B		0.610
	Mistral-3 7B	✓	0.681
	Mistral-3 7B		0.621
	Gemma-2 2B	✓	0.347
	Gemma-2 2B		0.158
	Gemma-2 9B	✓	0.380
	Gemma-2 9B		0.394
	Majority baseline		0.374

Table 6: $F1$ scores of language models in cultural adaptability (NORMAD) and cultural knowledge tasks (CulturalBench). The scores are calculated based on target word token probabilities.

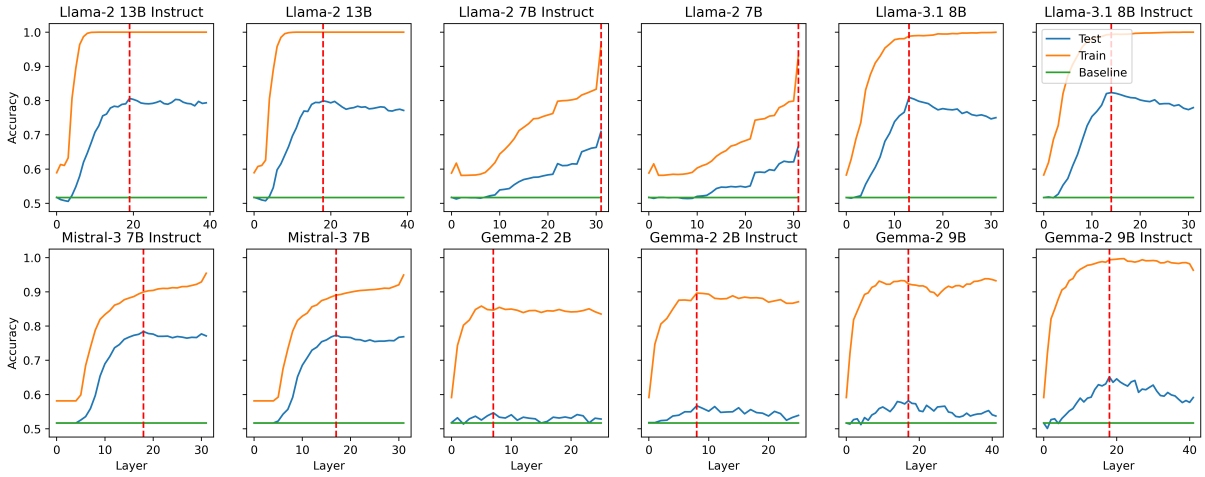

Figure 8: Evaluating embedded ethics concept vectors as decision boundaries using the ETHICS dataset and the logistic regression approach. The vertical dashed line indicates the layer with the best performance on the test set, while the green horizontal baseline represents the majority vote performance.

Model	Instruct	Correlation (Negativity Bias)	Correlation (Positivity Bias)	Cohen's d (TP - FN)	Cohen's d (TN - FP)
LLaMA-2 13B	✓	-0.556^{***}	0.290^{***}	1.098	-0.961
LLaMA-213B		-0.141^{***}	-0.008 (n.s.)	0.213	-0.208
LLaMA-2 7B	✓	-0.087^{*}	-0.094^{*}	0.158	0.358
LLaMA-2 7B		-0.087^{*}	0.070^{*}	0.125	-0.223
LLaMA-3.1 8B	✓	-0.736^{***}	0.790^{***}	2.345	-1.831
LLaMA-3.1 8B		-0.277^{***}	0.240^{***}	0.847	-0.289
Mistral-3 7B	✓	-0.241^{***}	0.267^{***}	0.287	-0.623
Mistral-3 7B		-0.037 (n.s.)	0.076^{*}	0.132	0.097
Gemma-2 2B	✓	-0.016 (n.s.)	0.082^{*}	0.013	-0.095
Gemma-2 2B		0.0002 (n.s.)	0.021 (n.s.)	0.006	0.006
Gemma-2 9B	✓	-0.172^{***}	0.441^{***}	0.526	-0.920
Gemma-2 9B		0.021 (n.s.)	0.072^{*}	-0.125	-0.042

Table 7: Analysis of the relationship between cultural adaptability (using NORMAD dataset) and embedded ethics (Moral Vignettes dataset) in language models. We use Pearson's correlation test, with asterisks indicating Benjamini–Hochberg-corrected significance levels: "*" for $p < 0.05$, "**" for $p < 0.01$, "***" for $p < 0.001$, and "n.s." stands for not significant.

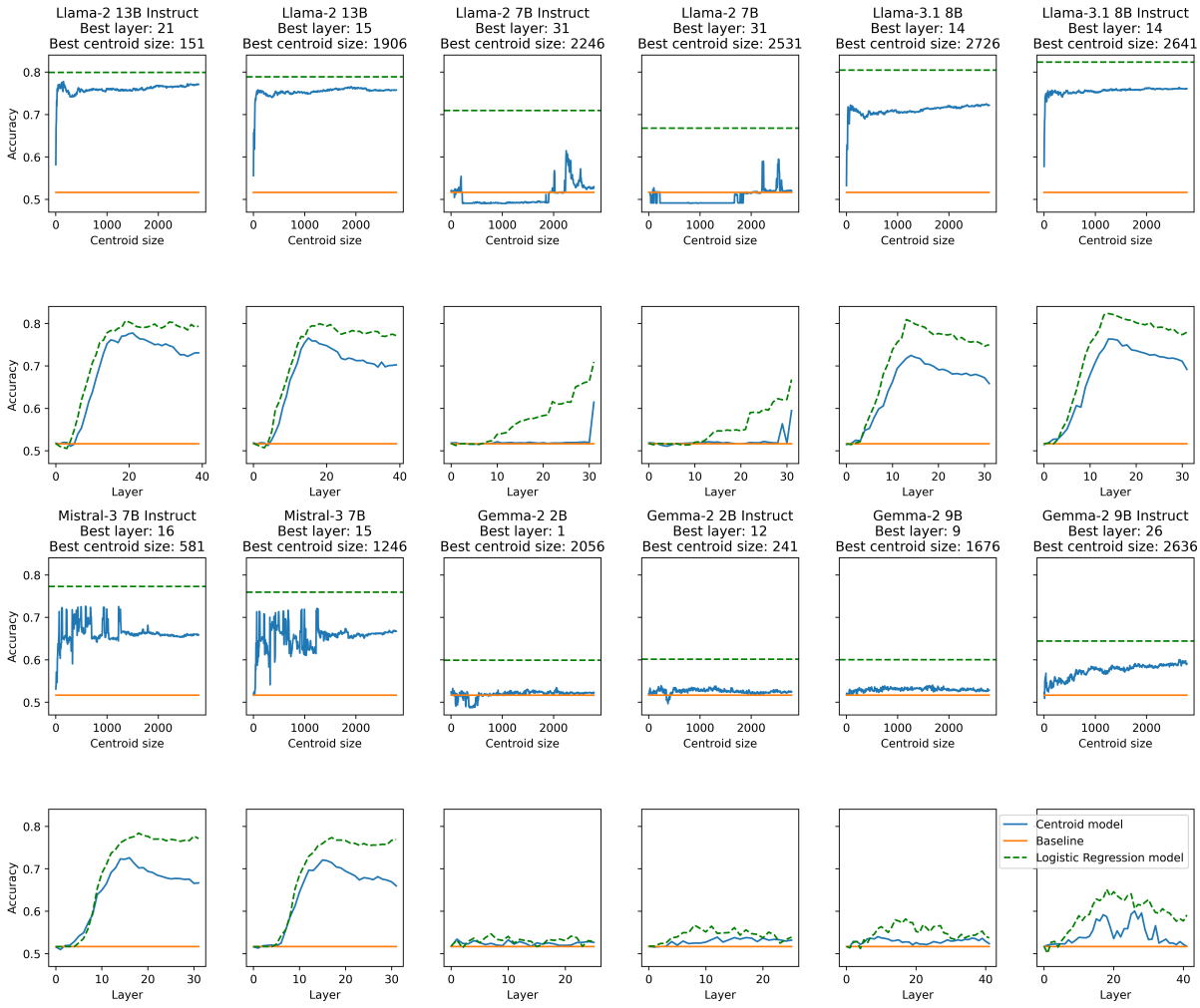

Figure 9: Comparing centroid-based distance ethics vectors from different layers and centroid sizes with logistic regression concept vectors on the test set of the ETHICS dataset. In all cases, the logistic regression vectors more effectively distinguish ethical samples from unethical ones.

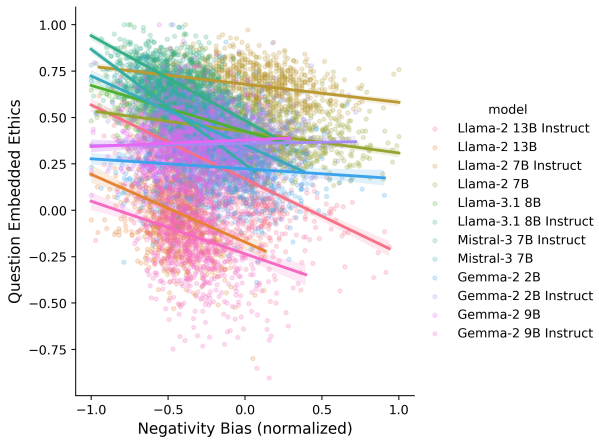

Figure 10: A negative correlation between ethical alignment scores and Negativity Bias is observed across models. The x- and y-axes are normalized to the range $[-1, 1]$ independently for each model. Each point represents a scenario from the NORMAD dataset used to prompt a language model.

Region	Number of Countries
East Asia & Pacific	21
Europe & Central Asia	22
Latin America & Caribbean	7
Middle East & North Africa	9
North America	2
South Asia	6
Sub-Saharan Africa	8
Income group	**Number of Countries**
High income	30
Low income	8
Lower middle income	17
Upper middle income	20

Table 8: Number of countries in NORMAD dataset falling into region and income group categories.

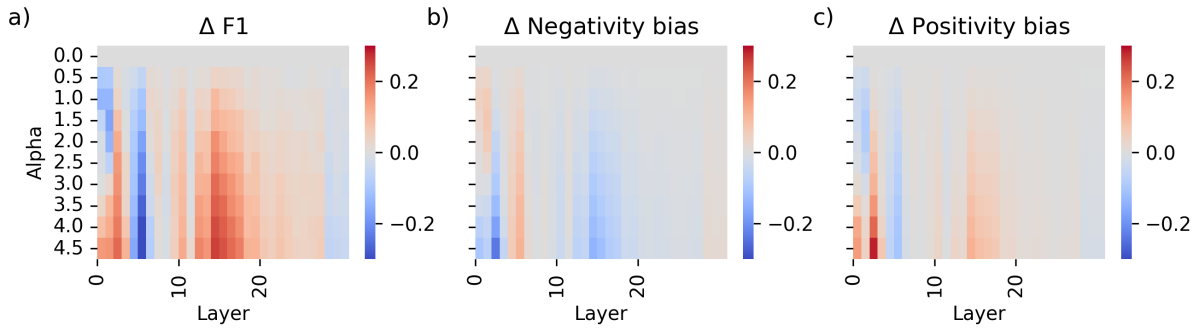

Figure 11: Average indirect effect of the intervention by adding $\alpha\times$ LogReg-ethics vector to the hidden representations of cultural norms in LLaMA-2 7B for (a) $F1$ score, (b) Negativity Bias, and (c) Positivity Bias. Samples are drawn from the NORMAD dataset.

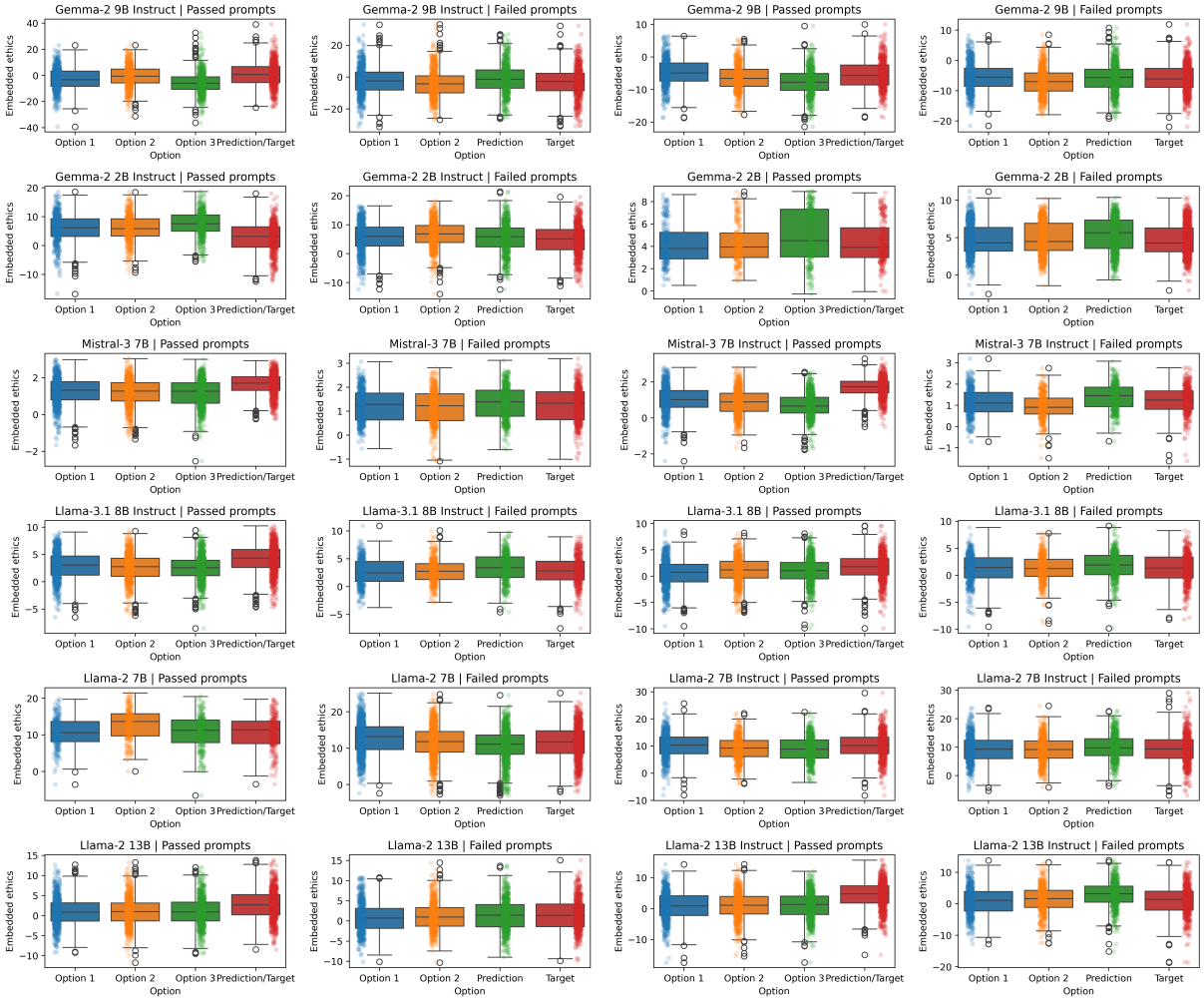

Figure 12: Differences in ethical alignment scores for all prompt options in CulturalBench across language models. Passed prompts compare the embedded ethics in cases where the model's prediction equals the target option. Failed prompts compare the model's prediction and target options with the rest of the options in the prompt.

Model (Instruct)	Prediction - Target	Passed - Failed
LLaMA-2 13B (✓)	$t = 5.698^{***}$ ($d = 0.355$)	$t = 7.294^{***}$ ($d = 0.417$)
LLaMA-2 7B	$t = 8.134^{***}$ ($d = 0.358$)	$t = -3.332$ ($d = -0.26$)
Mistral-3 7B (✓)	$t = 3.214^{***}$ ($d = 0.231$)	$t = 7.931^{***}$ ($d = 0.526$)

Table 9: Differences in ethical alignment scores for predicted and target cultural references (second column) and passed and failed target references (third column) using the CulturalBench dataset. The embedded ethics directions are derived using the $\mathrm{Gradient\text{-}ethics\ vector}$ from the ETHICS dataset. Variables t and d represent t-statistics and Cohen's d respectively. Asterisks indicate Benjamini–Hochberg-corrected significance levels: "*" for $p < 0.05$, "**" for $p < 0.01$, "***" for $p < 0.001$.