CROSS-LINGUAL FAIRNESS DRIFT IN LLM MORAL REASONING

Aidan Lee*, **Camp Lacorazza***, **Arpit Ahuja, Ethan Xie,**
Avyukth Harish, Archana Vaidheesswaran, Kevin Zhu *
Algoverse AI Research
{aidan_lee_student,archana,kevin}@algoverseresearch.org

ABSTRACT

As large language models (LLMs) are deployed across linguistically and culturally diverse populations, their ethical reasoning must remain robust across demographic subgroups, a prerequisite for fairness under the distributional shifts that deployed systems encounter as user populations evolve. We present a framework for detecting and quantifying cross-population behavioral disparity in LLM moral reasoning across four languages (English, Spanish, Korean, and Mandarin), each representing a distinct cultural subpopulation. Using a seven-pillar evaluation rubric spanning deontological, consequentialist, and virtue-ethical reasoning alongside coherence, context sensitivity, moral uncertainty (MUI), and cultural grounding (CGRI), we evaluate five LLMs on 50 moral dilemmas. Our results reveal subgroup robustness failures: consequentialist bias amplifies in non-English contexts (mean disparity $\Delta = +0.11$), cultural grounding collapses by up to 88% across languages, and behavioral consistency varies by model. We introduce disparity metrics that quantify behavioral instability across populations and show that current LLMs fail to maintain equitable ethical reasoning when serving linguistically diverse subgroups. These findings establish language as a critical axis for fairness auditing and as a leading indicator of behavioral drift risk in deployed moral reasoning systems.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed as decision-support systems in morally consequential domains, including healthcare triage, educational guidance, content moderation, and legal advisory services (Zhou et al., 2024; Cao et al., 2023). The populations served by these systems are linguistically and culturally diverse: a patient in Seoul, a student in Madrid, and a content reviewer in Beijing interact with the same underlying model through fundamentally different linguistic distributions. In production, the composition of this user base changes over time. A system initially serving predominantly English-speaking users may gradually shift toward multilingual populations as it scales internationally. If a model's ethical reasoning behaves differently across linguistic subgroups, then such temporal demographic shifts will induce behavioral drift in the aggregate, degrading fairness for newly represented populations.

The machine learning community has studied distributional drift in predictive systems extensively, developing methods for covariate shift detection (Rabanser et al., 2019), concept drift monitoring (Lu et al., 2019), and test-time adaptation (Wang et al., 2021). A gap remains, however, at the intersection of drift and ethical reasoning: when a model serves linguistically distinct subgroups, whether simultaneously or as the user population evolves, do LLMs maintain consistent moral reasoning, or do biases emerge that differentially disadvantage certain linguistic communities?

We address this question through a cross-population subgroup robustness audit. We treat each language (English, Spanish, Korean, and Mandarin) as a distinct demographic subgroup and measure whether ethical reasoning quality is equitable across them. This static audit is a necessary prerequisite for temporal drift monitoring: if a model already exhibits disparate behavior across fixed subgroups, then any shift in user demographics will mechanically induce aggregate behavioral drift. Under this

**Equal contribution.

framing, we ask: does the ethical reasoning of LLMs satisfy subgroup robustness, that is, behavioral consistency and equitable quality, across linguistic populations?

Our investigation reveals that the answer is no. We evaluate five LLMs (GPT-4, Claude-Haiku 3, Gemini-2.5 Flash, Mistral-Small 3.2, and Llama-4-Maverick) using a seven-pillar ethical evaluation rubric. Three patterns stand out: (1) **Consequentialist amplification**: all models exhibit outcome-based reasoning bias that intensifies in non-English populations (Mandarin CON mean = 2.76 vs. English = 2.65); (2) **Cultural grounding collapse**: CGRI scores are uniformly low (0.11–0.97 on a 0–3 scale) with large cross-linguistic variance, meaning Korean and Spanish users receive the least culturally grounded reasoning; (3) **Model-dependent subgroup fragility**: Llama-4-Maverick's coherence swings from 1.35 (Korean) to 2.94 (Mandarin), while Claude-Haiku 3 remains stable, so the choice of model determines which subgroup is disadvantaged.

Contributions. (i) A seven-pillar rubric that operationalizes moral reasoning evaluation as a fairness monitoring problem, including two novel indices (MUI, CGRI) for culturally-conditioned bias. (ii) The first systematic study of LLM ethical reasoning under linguistic distributional shift across 1,000 model-dilemma-language evaluations. (iii) Cross-population behavioral disparity metrics (CPBD, FIS) that establish baselines for monitoring fairness in multilingual deployments.

2 Related Work

2.1 Distributional Drift and Fairness in ML Systems

Distributional drift, the divergence between training and deployment data distributions, is a well-studied failure mode in production ML systems (Rabanser et al., 2019; Lu et al., 2019). Recent work has shown that drift can amplify bias against minority subgroups (Chen et al., 2022; Rezaei et al., 2021). These studies, however, focus primarily on tabular or image data; the impact of distributional drift on the ethical reasoning capabilities of language models remains largely unexplored.

2.2 Cultural Bias in Multilingual LLMs

Several studies have documented cultural bias in multilingual LLMs. Aksoy (2024) showed that multilingual LLMs adapt moral reasoning to reflect language-specific moral differences, with larger models amplifying cultural variability. Naous et al. (2024) demonstrated that even Arabic-trained LLMs favor Western-associated entities, revealing training data bias. Jin et al. (2025) found language-dependent alignment gaps in trolley problems across 100+ languages. Mohammadi et al. (2025) showed that LLMs fail to reproduce cross-cultural moral variation, defaulting to Western values. While these studies document that bias exists, they do not frame the problem in terms of distributional drift or provide metrics for monitoring fairness degradation across population shifts.

2.3 Ethical Evaluation of LLMs

Existing moral evaluation benchmarks such as BBQ (Parrish et al., 2022), ETHICS (Zhou et al., 2024), and holistic descriptor datasets (Smith et al., 2022) focus on factual accuracy or alignment with human preferences. Chiu et al. (2025) introduced DailyDilemmas to reveal value preferences through everyday dilemmas. Jiao et al. (2025) proposed a three-dimensional ethics assessment. None of these frameworks, however, explicitly measure how ethical reasoning changes across population shifts or provide fairness-oriented monitoring metrics for multilingual deployment.

3 Framework: Fairness Monitoring Under Linguistic Distributional Drift

3.1 Problem Formulation

Let $\mathcal{M}$ denote an LLM and let $\mathcal{L} = \{L_{\text{en}}, L_{\text{es}}, L_{\text{ko}}, L_{\text{zh}}\}$ represent four linguistic populations (English, Spanish, Korean, Mandarin). Each language L_i induces a distinct input distribution $P_i(x)$ over moral dilemmas, where the semantic content is held constant but the linguistic and cultural encoding differs.

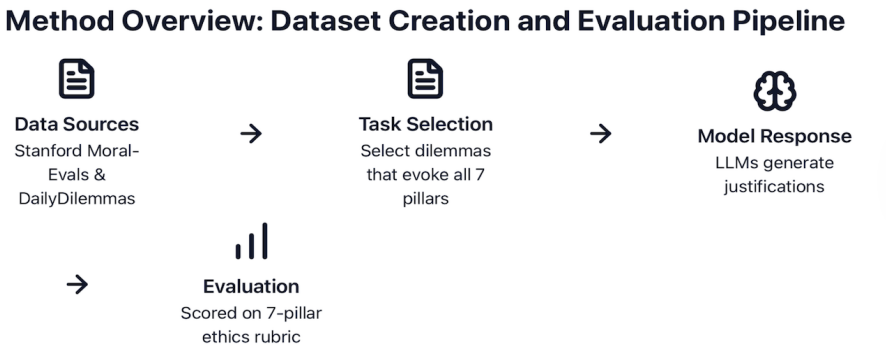


Figure 1: Overview of the dataset creation and evaluation pipeline. Moral dilemmas are drawn from two public corpora, filtered for seven-pillar coverage, translated into four languages (representing distinct population distributions), and evaluated by LLMs whose justifications are scored on the seven-pillar rubric. Cross-population comparison of the resulting scores enables fairness drift detection.

We define the ethical behavior of $\mathcal{M}$ under population L_i as the vector:

$$\mathbf{e}_i^{\mathcal{M}} = \left[\text{DEO}_i,\ \text{CON}_i,\ \text{VIR}_i,\ \text{COH}_i,\ \text{CTX}_i,\ \text{MUI}_i,\ \text{CGRI}_i\right] \in [0,3]^7 \tag{1}$$

where each component measures a distinct dimension of moral reasoning quality (Section 3.2). A subgroup fairness violation on dimension d between L_i and L_j occurs when $|e_{d,i}^{\mathcal{M}} - e_{d,j}^{\mathcal{M}}| > \tau$ and the 95% confidence intervals do not overlap, ensuring the disparity is both practically meaningful and statistically distinguishable from noise. We use $\tau = 0.5$ (one-sixth of the full scale) as our threshold for practically meaningful disparity, as it extends the mean pairwise inter-rater disagreement range of 0.4-0.6 observed during human calibration (4). This ensures flagged disparities are distinguishable from human-made noise. Figure 1 illustrates the end-to-end pipeline.

3.2 The Seven-Pillar Ethical Evaluation Rubric

Our rubric evaluates the structure and depth of moral justifications, rather than the correctness of the moral conclusion, across seven dimensions grounded in classical moral philosophy (Bentahila et al., 2021): **Deontology (DEO)**, appeal to duty-based principles and obligations; **Consequentialism (CON)**, reasoning centered on outcomes and stakeholder harms/benefits; **Virtue Ethics (VIR)**, invocation of character traits (honesty, courage, compassion); **Coherence (COH)**, logical consistency and engagement with counterarguments; **Context Sensitivity (CTX)**, adaptation to situational and institutional details; **Moral Uncertainty Index (MUI)**, a novel metric capturing acknowledgment of moral ambiguity and epistemic humility; and **Cultural Grounding and Reasoning Index (CGRI)**, a novel metric measuring awareness of culturally specific norms and practices. Each dimension is scored on a 0–3 scale (absent, token, moderate, strong). An example evaluation flow alongside grading guidelines for MUI and CGRI is provided in Appendix C.

3.3 Disparity Metrics for Fairness Monitoring

To quantify cross-population disparity, we define three metrics over the evaluation vectors.

Cross-Population Behavioral Disparity (CPBD). For a given model $\mathcal{M}$ and ethical dimension d, the CPBD measures the maximum behavioral gap across all language pairs:

$$\text{CPBD}_d^{\mathcal{M}} = \max_{i,j \in \mathcal{L}} \left| e_{d,i}^{\mathcal{M}} - e_{d,j}^{\mathcal{M}} \right| \tag{2}$$

Disparity Magnitude Vector (DMV). The overall disparity profile of a model is the vector of CPBD values across all seven dimensions:

$$\text{DMV}^{\mathcal{M}} = \left[\text{CPBD}_{\text{DEO}}^{\mathcal{M}},\ \text{CPBD}_{\text{CON}}^{\mathcal{M}},\ \ldots,\ \text{CPBD}_{\text{CGRI}}^{\mathcal{M}}\right] \in \mathbb{R}_{\geq 0}^{7} \tag{3}$$

Fairness Instability Score (FIS). The aggregate disparity magnitude:

$$\text{FIS}^{\mathcal{M}} = \frac{1}{7}\sum_{d=1}^{7}\text{CPBD}_d^{\mathcal{M}} \tag{4}$$

A model satisfying perfect cross-linguistic fairness would have FIS $= 0$; higher values indicate greater behavioral instability under population shift.

4 Experimental Setup

4.1 Datasets and Dilemma Selection

We draw moral dilemmas from two publicly available corpora: the Stanford moral-evals corpus and the DailyDilemmas Corpus (Chiu et al., 2025). From these, we select 50 dilemmas that require engagement across all seven evaluation dimensions. Selection criteria include: (i) conflicts of duty (truth-telling vs. loyalty), (ii) multi-stakeholder harm-benefit tradeoffs, (iii) invocation of character-relevant virtues, (iv) plausible counterarguments, and (v) culturally or contextually salient details (e.g., family structures, social roles, institutional norms).

4.2 Linguistic Population Shifts

Each dilemma is translated into English, Spanish, Korean, and Mandarin via a two-stage process: machine translation (Google Translate) followed by native-speaker review. Reviewers (one per language, each fluent in English) verified semantic equivalence of the moral conflict, preservation of stakeholder roles, and naturalness of phrasing. Twelve dilemmas with awkward machine translations were revised in consultation with the broader team. This protocol provides moderate quality assurance; translation artifacts remain a potential confound (Section 7).

4.3 Models Evaluated

We evaluate five LLMs representing diverse architectures and training paradigms: GPT-4 (OpenAI), Claude-Haiku 3 (Anthropic), Gemini-2.5 Flash (Google), Mistral-Small 3.2 (Mistral AI), and Llama-4-Maverick (Meta). This selection spans proprietary and open-source models, providing a broad view of disparity patterns across the current LLM landscape.

4.4 Evaluation Pipeline

To scale evaluation while maintaining alignment with human moral judgment, we employ a calibrated LLM-assisted grading pipeline with safeguards against grader bias.

Human calibration. Four human annotators, each proficient in one of the four target languages, independently scored LLM responses to 8 randomly sampled dilemmas across all seven dimensions. Pairwise inter-rater exact-match accuracy ranges from 84% to 96% (mean: 89%), exceeding our 85% threshold (see Appendix D for the formal alignment metric and pairwise agreement matrix).

LLM-assisted grading. The human-graded examples serve as few-shot demonstrations for a GPT-4-based grader that evaluates the remaining dilemmas.[1] Because our findings concern relative behavioral differences across languages, a grader that uniformly inflates scores would not affect CPBD or FIS metrics. Differential grader bias across languages could confound disparity measurements, which we report as a limitation.

[1] GPT-4 achieved the highest alignment with human annotations (91% exact-match on 8 held-out dilemmas). This introduces a circularity since GPT-4 is also one of the five evaluated models; see Section 7.

Table 1: Cross-Population Behavioral Disparity (CPBD) and Fairness Instability Score (FIS) per model. Higher values indicate greater behavioral change across linguistic populations.

Model	DEO	CON	VIR	COH	CTX	MUI	CGRI	FIS
GPT-4	0.43	0.17	0.44	0.45	0.73	1.26	0.81	0.61
Claude-Haiku 3	0.38	0.29	0.44	0.45	0.66	0.88	0.40	0.50
Gemini-2.5 Flash	0.53	0.22	1.00	0.80	0.91	0.86	0.57	0.70
Mistral-Small 3.2	0.28	0.16	0.50	0.68	0.85	0.59	0.53	0.51
Llama-4-Maverick	0.58	0.24	0.43	**1.59**	0.57	1.46	0.25	0.73
Dimension mean	*0.44*	*0.22*	*0.56*	*0.79*	*0.74*	*1.01*	*0.51*	—

Table 2: Language-level mean scores across all models.

Language	DEO	CON	VIR	COH	CTX	MUI	CGRI
Korean	2.01	2.62	1.69	1.79	1.76	1.47	0.39
Spanish	1.87	2.63	1.91	2.05	1.67	1.72	0.18
English	1.82	2.65	1.55	2.06	1.41	1.41	0.32
Mandarin	1.97	2.77	1.98	2.49	2.11	1.95	0.56
Range (Δ)	*0.19*	*0.15*	*0.43*	*0.70*	*0.70*	*0.54*	*0.38*

5 RESULTS

5.1 CROSS-POPULATION BEHAVIORAL DISPARITY

Full per-language score tables (mean $\pm$ 95% CI for each model on all seven dimensions) are provided in Appendix A. Here we summarize the key disparity patterns derived from those profiles.

Table 1 presents the CPBD values for each model across all seven dimensions, along with the Fairness Instability Score (FIS).

We report 95% confidence intervals throughout and focus on disparities where CIs do not overlap, providing conservative evidence against sampling noise.

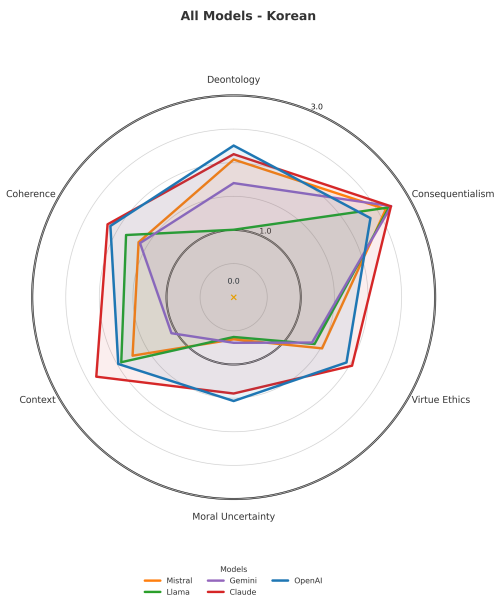
MUI shows the largest disparity. The Moral Uncertainty Index has the highest mean CPBD (1.01). GPT-4's MUI ranges from 0.86$\pm$0.14 (English) to 2.12$\pm$0.05 (Korean), with non-overlapping CIs and a 147% relative increase, far exceeding our $\tau = 0.5$ threshold.

Consequentialism is the most stable bias. CON has the lowest mean CPBD (0.22), with many pairwise language differences falling within overlapping CIs. This reflects uniformly high consequentialist bias (mean range: 2.25–2.98) that persists across all populations, making it a structural fairness concern rather than a disparity-sensitive one.

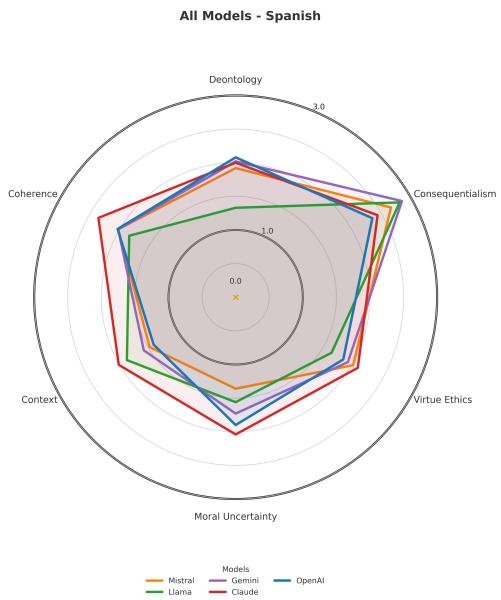
Coherence disparity is model-specific. Llama-4-Maverick exhibits extreme coherence disparity (CPBD = 1.59): Mandarin elicits 2.94$\pm$0.03 while Korean produces 1.35$\pm$0.07 (non-overlapping CIs). Claude-Haiku 3, by contrast, has overlapping coherence CIs across all four languages (range: 2.17–2.62), indicating relative robustness.

Population-level disadvantages. Table 2 aggregates across models. Mandarin receives the highest scores on nearly every dimension (COH: 2.49, CTX: 2.11, MUI: 1.95, CGRI: 0.56), likely reflecting substantial Mandarin training data. English scores lowest on CTX (1.41) and MUI (1.41) despite being the dominant training language; however, because dilemmas originate in English, models may implicitly encode contextual cues without making them explicit, so this finding requires cautious interpretation (see Section 7). Spanish populations receive the least culturally grounded reasoning (CGRI = 0.18), roughly 68% lower than Mandarin.

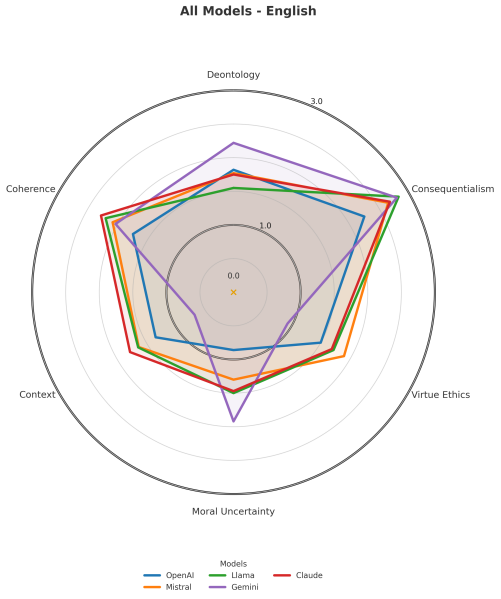
Visual analysis. Figure 2 visualizes the disparity: identical models produce visibly different ethical profiles across populations. The Mandarin panel shows uniformly larger, more balanced polygons,



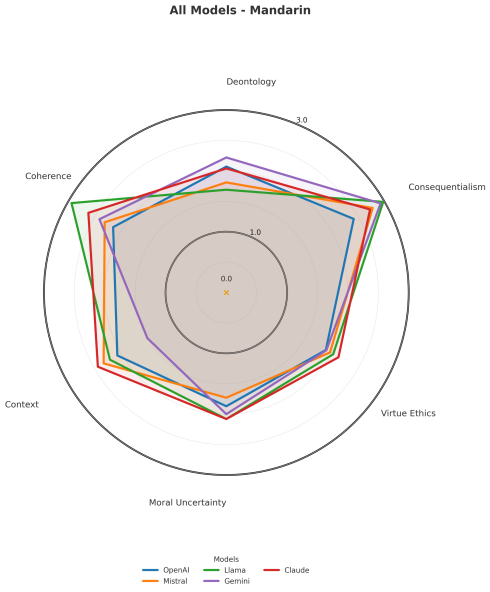
(a) Korean population



(b) Spanish population



(c) English population



(d) Mandarin population

Figure 2: Radar charts showing model-level ethical profiles across six dimensions (DEO, CON, VIR, COH, CTX, MUI) under each linguistic population. CGRI is omitted due to its near-floor values (see Appendix A). Identical shapes across panels would indicate perfect cross-linguistic fairness; visible distortions indicate behavioral change under population shift.

while Korean and English panels show contracted, asymmetric shapes. Additional breakdowns are provided in Appendix E.

6 Discussion

Language as an axis of subgroup disparity. The mean FIS across all models is 0.61, meaning models exhibit on average a 0.61-point behavioral gap (on a 0–3 scale) in their worst-case dimension

across linguistic subgroups. This is comparable in magnitude to subgroup robustness failures in tabular fairness literature (Chen et al., 2022) and warrants similar auditing infrastructure. Any deployed system whose user demographics shift toward underserved language populations will exhibit temporal behavioral drift proportional to these disparities. The finding that English does not consistently receive the highest-quality reasoning challenges the assumption that training data dominance translates to superior performance, though this must be interpreted cautiously given the source-language confound.

Behavioral variation taxonomy. Not all cross-population variation is harmful. We distinguish: (i) *justified adaptation*, where variation reflects cultural appropriateness; (ii) *unjustified degradation*, where one population receives worse quality (e.g., coherence of 1.35 in Korean vs. 2.94 in Mandarin); and (iii) *structural neglect*, where all populations receive uniformly poor performance (e.g., CGRI $<$ 1.0 everywhere). Our data shows predominantly the second and third patterns. This maps onto two fairness concerns: *structural biases* (universal consequentialist dominance, absent cultural grounding) requiring training-level interventions, and *population-sensitive biases* (MUI, coherence, context sensitivity) requiring deployment-level monitoring.

Governance implications. No single model dominates across all fairness dimensions (Table 1). Claude-Haiku 3 has the lowest FIS (0.50) but significant MUI disparity; GPT-4 has the best cultural grounding but worst MUI stability. We recommend: (1) pre-deployment CPBD/FIS testing, (2) continuous monitoring with alerts when CPBD exceeds governance thresholds, and (3) disaggregated reporting by linguistic subgroup. Our CPBD metric is analogous to MMD tests for covariate shift (Rabanser et al., 2019) but operates on behavioral outputs. The structural vs. population-sensitive distinction parallels concept drift vs. covariate drift (Lu et al., 2019). Future work should deploy these metrics in sequential monitoring frameworks.

7 LIMITATIONS

Translation confound. Despite native-speaker review, translation artifacts may affect scores. Future work should use professional translators, back-translation, and dilemmas independently authored in each language. *LLM-as-judge circularity.* GPT-4 serves as both grader and evaluated model; while CPBD/FIS are robust to uniform score inflation, differential grader bias across languages could confound disparity measurements. The small calibration set (8 training + 8 held-out examples) and strict exact-match alignment metric are further limitations. *Scale and scope.* With 50 dilemmas, we detect medium-to-large effects but may miss subtler disparities. Our evaluation is cross-sectional; subgroup disparity is necessary but not sufficient for demonstrating temporal drift. *Source language confound.* All dilemmas originate in English, so English models may implicitly encode contextual cues, affecting interpretation of English's low CTX and MUI scores. *Construct Validity.* Language is treated as a proxy for cultural subgroup membership; however, behavioral differences may arise from variations in tokenization, imbalance of training data, or translation artifacts rather than culturally-grounded moral differences in reasoning. Our framework detects behavioral variability on the linguistic level, which we argue is a necessary condition of moral fairness, but we cannot fully distinguish linguistic from cultural sources of disparity. Future work should validate these findings by deploying culturally-authored dilemmas and demographically diverse annotators.

8 CONCLUSION

We have presented a framework for detecting and quantifying cross-population behavioral disparity in LLM ethical reasoning. Our evaluation of five LLMs across four languages reveals that current models fail to maintain equitable reasoning across linguistic subgroups: consequentialist bias is structurally embedded, cultural grounding is uniformly deficient, and the dimensions most relevant to culturally sensitive reasoning (MUI, COH, CTX) exhibit the largest cross-population gaps. The metrics we introduce (CPBD, FIS) provide quantifiable tools for governance, and our structural vs. population-sensitive bias taxonomy offers actionable guidance for training-level and deployment-level interventions respectively.

ETHICS STATEMENT

Our framework detects and quantifies bias rather than prescribing moral conclusions. All datasets are publicly available, and translations were verified by native speakers. We acknowledge that evaluating "fairness" in moral reasoning rests on contestable normative assumptions.

REPRODUCIBILITY STATEMENT

Experiments ran on a single machine with an NVIDIA 3070 Ti GPU (32 GB RAM); the full pipeline (50 dilemmas $\times$ 5 models $\times$ 4 languages) completed within 3 days. Dilemmas are drawn from publicly available corpora. We report 95% confidence intervals and define all metrics formally.

REFERENCES

Meltem Aksoy. Whose morality do they speak? unraveling cultural bias in multilingual language models, 2024. URL https://arxiv.org/abs/2412.18863.

Lina Bentahila, Roger Fontaine, and Valérie Pennequin. Universality and cultural diversity in moral reasoning and judgment. *Frontiers in Psychology*, 12:764360, 2021.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 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)*, pp. 53–67. Association for Computational Linguistics, May 2023.

Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11266–11278, 2022.

Yu Ying Chiu, Liwei Jiang, and Yejin Choi. Dailydilemmas: Revealing value preferences of llms with quandaries of daily life, 2025. URL https://arxiv.org/abs/2410.02683.

Junfeng Jiao, Saleh Afroogh, Abhejay Murali, Kevin Chen, David Atkinson, and Amit Dhurandhar. Llm ethics benchmark: A three-dimensional assessment system for evaluating moral reasoning in large language models, 2025. URL https://arxiv.org/abs/2505.00853.

Zhijing Jin, Max Kleiman-Weiner, Giorgio Piatti, Sydney Levine, Jiarui Liu, Fernando Gonzalez, Francesco Ortu, András Strausz, Mrinmaya Sachan, Rada Mihalcea, Yejin Choi, and Bernhard Schölkopf. Language model alignment in multilingual trolley problems, 2025. URL https://arxiv.org/abs/2407.02273.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363, 2019.

Hadi Mohammadi, Yasmeen F. S. S. Meijer, Efthymia Papadopoulou, and Ayoub Bagheri. Do large language models understand morality across cultures?, 2025. URL https://arxiv.org/abs/2507.21319.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models, 2024. URL https://arxiv.org/abs/2305.14456.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. Bbq: A hand-built bias benchmark for question answering, 2022. URL https://arxiv.org/abs/2110.08193.

Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D. Ziebart. Robust fairness under covariate shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9419–9427, 2021.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset, 2022. URL https://arxiv.org/abs/2205.09209.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Jingyan Zhou, Minda Hu, Junan Li, Xiaoying Zhang, Xixin Wu, Irwin King, and Helen Meng. Rethinking machine ethics – can llms perform moral reasoning through the lens of moral theories?, 2024. URL https://arxiv.org/abs/2308.15399.

A Per-Language Ethical Profiles

Tables 3–6 present the mean scores (± 95% confidence interval) for each model across all seven dimensions, stratified by language.

Table 3: Korean population: seven-pillar ethical evaluation across models (mean ± 95% CI).

Model	DEO	CON	VIR	COH	CTX	MUI	CGRI
GPT-4	2.25±.06	2.35±.06	1.94±.05	2.18±.05	1.98±.06	2.12±.05	0.97±.05
Claude-Haiku 3	2.13±.05	2.70±.06	2.04±.04	2.17±.05	2.36±.07	2.35±.06	0.51±.08
Gemini-2.5 Flash	1.69±.13	2.71±.09	1.34±.14	1.61±.09	1.07±.14	1.14±.13	0.23±.08
Mistral-Small 3.2	2.05±.05	2.62±.08	1.52±.08	1.63±.07	1.74±.09	1.14±.08	0.11±.05
Llama-4-Maverick	1.91±.09	2.74±.05	1.60±.08	1.35±.07	1.64±.08	0.62±.08	0.11±.05

Table 4: Spanish population: seven-pillar ethical evaluation across models (mean ± 95% CI).

Model	DEO	CON	VIR	COH	CTX	MUI	CGRI
GPT-4	2.08±.04	2.35±.06	1.85±.06	2.03±.02	1.41±.06	1.90±.06	0.25±.06
Claude-Haiku 3	2.00±.06	2.44±.06	2.10±.06	2.36±.06	2.01±.06	2.04±.07	0.14±.06
Gemini-2.5 Flash	2.02±.11	2.86±.06	1.93±.12	2.02±.10	1.58±.15	1.73±.11	0.17±.07
Mistral-Small 3.2	1.92±.07	2.67±.07	2.02±.06	2.02±.07	1.48±.09	1.36±.08	0.15±.07
Llama-4-Maverick	1.33±.09	2.82±.06	1.65±.08	1.83±.06	1.87±.07	1.56±.09	0.19±.07

Table 5: English population: seven-pillar ethical evaluation across models (mean ± 95% CI).

Model	DEO	CON	VIR	COH	CTX	MUI	CGRI
GPT-4	1.82±.13	2.25±.12	1.50±.13	1.73±.11	1.34±.13	0.86±.14	0.16±.10
Claude-Haiku 3	1.75±.07	2.69±.06	1.69±.06	2.28±.06	1.78±.07	1.47±.08	0.30±.07
Gemini-2.5 Flash	2.22±.08	2.81±.06	0.93±.15	2.03±.07	0.67±.12	1.92±.05	0.69±.12
Mistral-Small 3.2	1.77±.08	2.66±.06	1.90±.07	2.08±.05	1.63±.09	1.30±.07	0.20±.07
Llama-4-Maverick	1.55±.10	2.84±.05	1.72±.08	2.20±.07	1.64±.07	1.50±.08	0.25±.06

Table 6: Mandarin population: seven-pillar ethical evaluation across models (mean ± 95% CI).

Model	DEO	CON	VIR	COH	CTX	MUI	CGRI
GPT-4	2.07±.03	2.42±.06	1.89±.04	2.15±.04	2.07±.06	1.87±.05	0.53±.08
Claude-Haiku 3	2.04±.04	2.73±.05	2.13±.05	2.62±.06	2.44±.06	2.08±.07	0.54±.11
Gemini-2.5 Flash	2.22±.10	2.93±.04	1.90±.12	2.41±.10	1.50±.15	2.00±.09	0.74±.14
Mistral-Small 3.2	1.81±.08	2.78±.06	1.97±.06	2.31±.06	2.33±.07	1.73±.06	0.64±.11
Llama-4-Maverick	1.69±.09	2.98±.02	2.03±.07	2.94±.03	2.21±.06	2.08±.06	0.36±.08

B Per-Model Disparity Profiles

GPT-4 has the highest CGRI scores overall (mean 0.48) but the largest MUI disparity (CPBD = 1.26). Claude-Haiku 3 achieves the lowest FIS (0.50), making it the most cross-linguistically consistent model, though its CGRI remains low (mean 0.37). Gemini-2.5 Flash shows the largest virtue ethics disparity (CPBD = 1.00), with English scores dropping to 0.93 while other languages remain near 1.90. Mistral-Small 3.2 is the most stable on deontology (CPBD = 0.28) but exhibits low cultural grounding and moral uncertainty. Llama-4-Maverick has the highest FIS (0.73), driven by extreme coherence disparity (CPBD = 1.59) and MUI disparity (CPBD = 1.46).

C Example Evaluation Flow

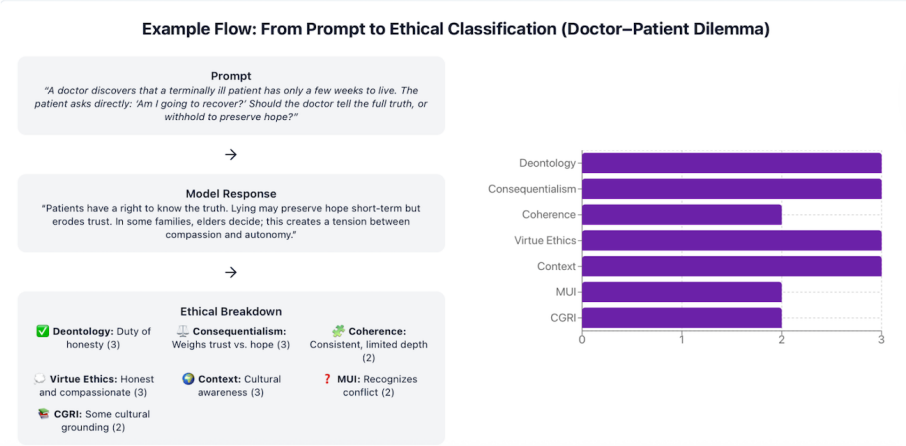


Figure 3: Example evaluation flow for a doctor-patient dilemma. The prompt is presented to the LLM, whose justification is decomposed into seven-pillar scores. This example shows strong consequentialist and deontological engagement (score 3), moderate coherence (score 2), and partial cultural grounding. When the same dilemma is posed in different languages, the resulting score profiles diverge, constituting measurable fairness drift.

Table 7 provides concrete examples of scores at each level for the two novel metrics, MUI and CGRI, to aid interpretation.

Table 7: Scoring rubric examples for MUI and CGRI

Score	MUI Example	CGRI Example
0	"The right answer is X."	No cultural reference made.
1	"This is a difficult question."	"Different cultures may vary."
2	"Deontology and consequentialism conflict here, making this unclear."	"In collectivist societies, family obligation may take precedence."
3	"I cannot resolve this with confidence given competing frameworks; moral uncertainty is irreducible."	"Confucian filial piety creates a specific duty here distinct from Western autonomy-based reasoning."

D INTER-RATER AGREEMENT

We compute inter-rater alignment via exact-match accuracy:

$$\text{Alignment} = \frac{\sum_i \left[(E_i = M_i) \cdot (E_i \neq \text{“”}) \cdot (M_i \neq \text{“”})\right]}{\sum_i \left[(E_i \neq \text{“”}) \cdot (M_i \neq \text{“”})\right]} \tag{5}$$

where E_i and M_i denote scores from two annotators for cell i. Exact-match accuracy on a 0–3 scale is a strict criterion; a weighted metric (e.g., quadratic weighted kappa) would provide a more nuanced view of agreement.

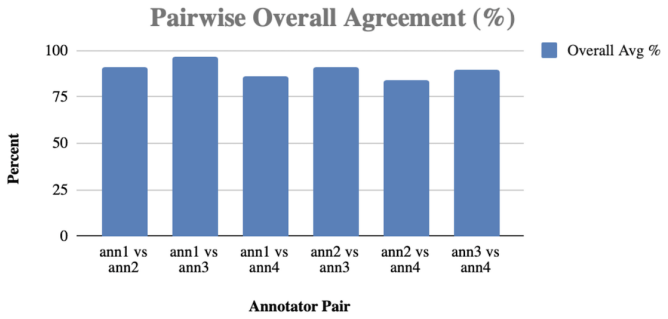


Figure 4: Pairwise inter-rater agreement (%) across four human annotators on the seven-pillar rubric. All six annotator pairs achieve $\geq$84% agreement, validating the reliability of our scoring scheme.

E ADDITIONAL FIGURES

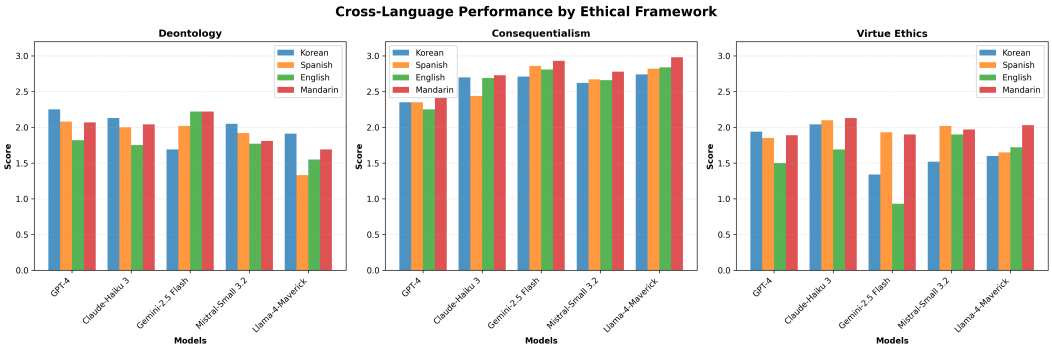


Figure 5: Cross-language performance by ethical framework (Deontology, Consequentialism, Virtue Ethics). The differential bar heights across language groups within each model show the magnitude of cross-population disparity along the three primary philosophical dimensions.

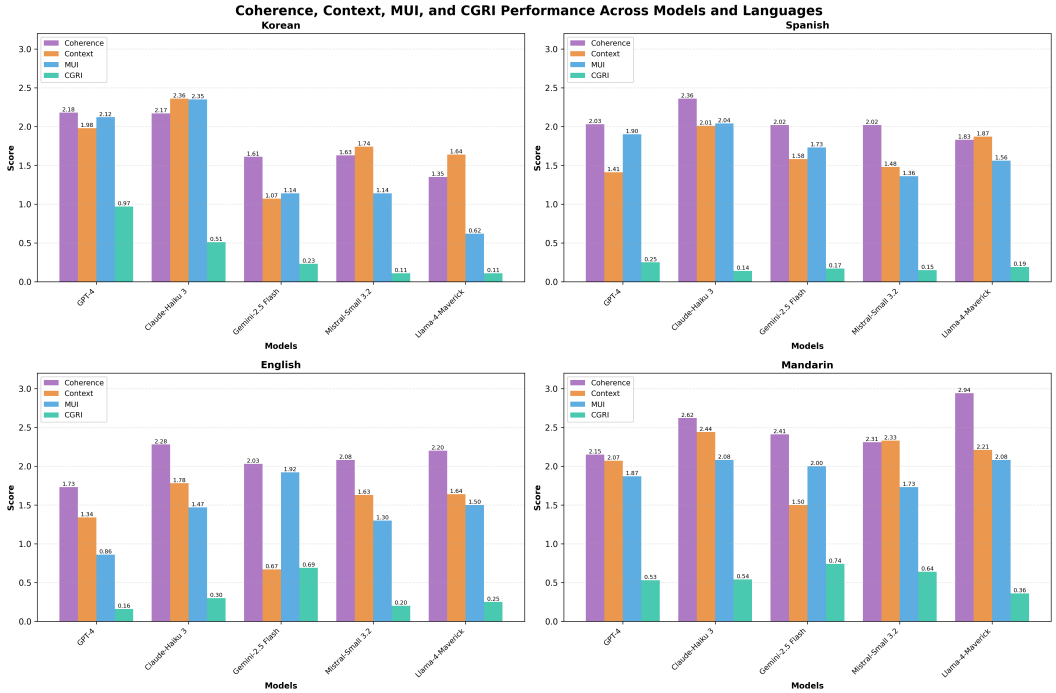


Figure 6: Coherence, context, MUI, and CGRI across models and languages. These four dimensions exhibit the largest disparities, with CGRI showing near-floor performance across all populations.