Do Large Language Models Reflect Demographic Pluralism in Safety?

Usman Naseem[1*], **Gautam Siddharth Kashyap**[2], **Sushant Kumar Ray**[3], **Rafiq Ali**[4]
Ebad Shabbir[5], **Abdullah Mohammad**[6]
[1, 2]Macquarie University, Sydney, Australia
[3]University of Delhi, New Delhi, India
[4, 5, 6]DSEU-Okhla, New Delhi, India

Abstract

Large Language Model (LLM) safety is inherently *pluralistic*, reflecting variations in moral norms, cultural expectations, and demographic contexts. Yet, existing alignment datasets such as ANTHROPIC-HH and DICES rely on demographically narrow annotator pools, overlooking variation in safety perception across communities. *Demo-SafetyBench*[1] addresses this gap by modeling demographic pluralism directly at the *prompt level*, decoupling value framing from responses. In *Stage I*, prompts from DICES are reclassified into 14 safety domains (adapted from BEAVERTAILS) using Mistral-7B-Instruct-v0.3, retaining demographic metadata and expanding low-resource domains via Llama-3.1-8B-Instruct with SimHash-based deduplication, yielding 43,050 samples. In *Stage II*, pluralistic sensitivity is evaluated using *LLMs-as-Raters*—Gemma-7B, GPT-4o, and LLaMA-2-7B—under zero-shot inference. Balanced thresholds (δ=0.5, τ=10) achieve high reliability (ICC = 0.87) and low demographic sensitivity (DS = 0.12), confirming that pluralistic safety evaluation can be both scalable and demographically robust.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in interactive, high-stakes contexts—ranging from education and healthcare to policy design and creative generation (OpenAI et al., 2024; Bommasani et al., 2022). As their social footprint expands, ensuring the *safety* of these models has become a central challenge in alignment research (Maskey et al., 2025; Hendrycks et al., 2021; Gabriel, 2020). Yet, safety is not a monolithic construct—what constitutes "safe" or "harmful" behavior varies sharply across

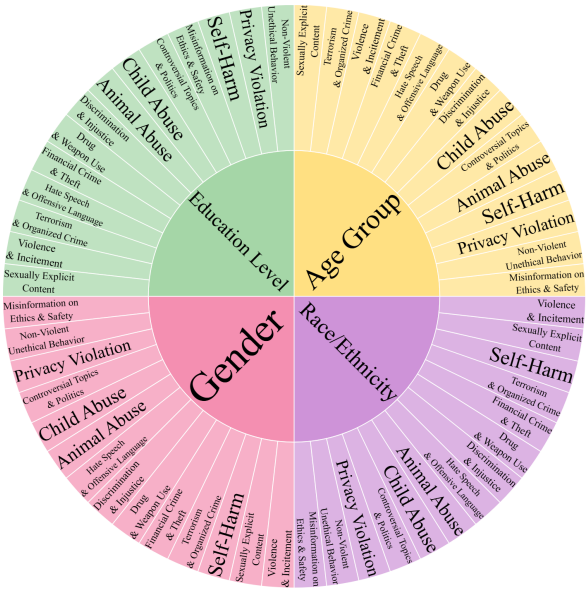

Figure 1: Safety taxonomy used in *Demo-SafetyBench*. We adapt the BEAVERTAILS (Ji et al., 2023) taxonomy into fourteen text-safety domains that ground *Stage I* reclassification of DICES (Aroyo et al., 2023) prompts.

demographic, cultural, and contextual boundaries (Aroyo et al., 2023; Santurkar et al., 2023). For example, a query discussing *alcohol consumption* may be culturally neutral in Western contexts but inappropriate in conservative or religious ones; similarly, *gender-related humor* might be perceived as satire by one group and as discrimination by another. These divergences reveal that safety is not universal but *pluralistic*—rooted in social norms, value hierarchies, and lived experiences that differ across populations.

Existing alignment datasets and evaluation frameworks have advanced model behavior measurement but only partially capture this pluralism. Datasets such as ANTHROPIC-HH (Bai et al., 2022b), BEAVERTAILS (Ji et al., 2023), DICES (Aroyo et al., 2023), and REALTOXICITYPROMPTS (Gehman et al., 2020) operationalize alignment along the axes of *helpfulness, harmless-*

*Corresponding Author: usman.naseem@mq.edu.au
[1]Data available at: https://github.com/usmaann/Demo-SafetyBench

Findings of the Association for Computational Linguistics: EACL 2026, pages 2042–2052
March 24-29, 2026

ness, and honesty (HHH) (Naseem et al., 2025; Azeez et al., 2025; Kashyap et al., 2025), often using human feedback and preference modeling pipelines (Ouyang et al., 2022; Christiano et al., 2017). However, these corpora rely heavily on annotators from demographically homogeneous or Western-centric populations (Nadeem et al., 2025; Birhane et al., 2024; Dillon et al., 2024), encoding narrow cultural priors about what constitutes risk or appropriateness. Other benchmarks such as SAFE-RLHF (Dai et al., 2023) and TRUTHFULQA (Lin et al., 2022) address factual correctness and refusal behavior but similarly lack demographic heterogeneity in defining safety. Consequently, current alignment pipelines risk optimizing toward a *moral median*—reinforcing dominant cultural norms while marginalizing minority perspectives under the guise of "universal" safety (Masoud et al., 2025; Chiu et al., 2024).

Recent research attempts to scale safety evaluation by replacing human annotators with *LLMs-as-Judges* (Bavaresco et al., 2024; Gilardi et al., 2023; Helff et al., 2024; Zeng et al., 2024), using foundation models such as GPT-4 (OpenAI et al., 2024), Gemini (Team et al., 2023), Claude (Anthropic, 2025), and ShieldGemma (Zeng et al., 2024) to automatically score responses. While these evaluators improve efficiency and consistency, they inherit demographic and cultural biases from the response-based datasets on which they were trained and validated—datasets themselves built upon narrow human judgments. Thus, despite progress in automating evaluation, existing *LLMs-as-Judges* pipelines remain unable to systematically measure how demographic variation shapes perceptions of safety (Bai et al., 2022b).

To overcome this confound, we introduce demographic pluralism at the *prompt level*—the primary locus of moral framing in LLM interaction—without collecting or generating responses. This shift ensures that all evaluation content remains textually neutral, demographically grounded, and free from response-induced or annotator-driven bias. We operationalize this through *Demo-SafetyBench*, a two-stage framework. In *Stage I (Data Construction)*, prompts from DICES[2] (Aroyo et al., 2023) are reclassified into 14 safety domains with demographic attributes (see Figure 1): *Animal Abuse, Child Abuse, Controversial Topics & Politics, Discrimination & Injustice, Drug & Weapon Use, Financial Crime & Theft, Hate Speech & Offensive Language, Misinformation on Ethics & Safety, Non-Violent Unethical Behavior, Privacy Violation, Self-Harm, Sexually Explicit Content, Terrorism & Organized Crime, and Violence & Incitement.* These domains, adapted from BEAVERTAILS (Ji et al., 2023), are classify using Mistral-7B-Instruct-v0.3[3] (Jiang et al., 2024). Underrepresented domains (fewer than 100 queries) are expanded via Llama-3.1-8B-Instruct[4]–based conditional query generation (Yin et al., 2025), followed by SimHash fingerprinting (Sadowski and Levin, 2007) to eliminate redundancy and prevent train–test leakage. In *Stage II (Benchmarking)*, we assess demographic sensitivity and pluralistic safety divergence using *LLMs-as-Raters*—Gemma-7B[5] (Zeng et al., 2024), GPT-4o[6] (OpenAI et al., 2024), and LLaMA-2-7B[7] (Touvron et al., 2023)—under zero-shot conditions. In summary, our contributions are twofold:

- We introduce *Demo-SafetyBench*, a two-stage framework that models demographic pluralism at the *prompt level*—decoupling value framing from model responses—by integrating a demographically grounded dataset across 14 safety domains (43,050 samples) and a pluralistic benchmarking protocol using *LLMs-as-Raters* to evaluate safety perception across diverse demographic contexts.

- Empirically, GPT-4o achieves the highest internal reliability (ICC=0.87) and lowest demographic sensitivity (DS=0.119), while Gemma-7B and LLaMA-2-7B deliver comparable pluralistic trends at substantially lower computational cost (0.42–0.58 s/query, 12.6–14.8 GB, $\leq$1.1 kWh/1k queries).

2 Related Works

The concept of *pluralistic alignment* extends moral diversity beyond aggregate human preference modeling to explicitly account for *demographic pluralism*—how variations in gender, race, age, and education shape perceptions of safety, fairness, and

[2]Although DICES exhibits demographic imbalance, it is uniquely suited for our setting because it explicitly encodes demographic metadata (gender, race, age, and education), enabling controlled pluralistic reclassification. Rather than relying on its original annotation distribution, *Demo-SafetyBench* restructures and expands it to ensure balanced representation across safety domains.

[3]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[4]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[5]https://huggingface.co/google/gemma-7b

[6]https://openai.com/index/hello-gpt-4o/

[7]https://huggingface.co/meta-llama/Llama-2-7b

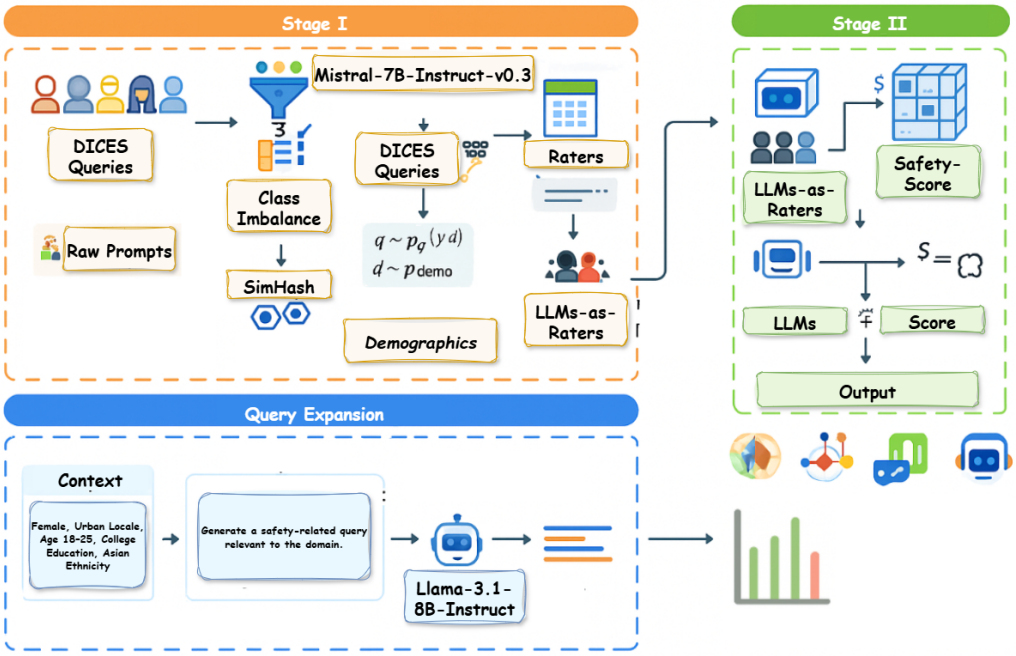

Figure 2: Overview of the *Demo-SafetyBench* pipeline. The framework comprises two stages: Stage I constructs a demographically diversified, prompt-level corpus by reclassifying and expanding DICES queries across 14 safety domains using Mistral-7B; Stage II benchmarks pluralistic safety by evaluating these prompts with *LLMs-as-Raters* (Gemma-7B, GPT-4o, LLaMA-2-7B) under zero-shot inferences.

harm (Santurkar et al., 2023; Gabriel, 2020). Conventional alignment pipelines, optimized toward an average or "median" moral consensus, risk erasing minority or culturally specific viewpoints (Birhane et al., 2024; Dillon et al., 2024). To counter this, emerging research has begun to disaggregate alignment evaluations across demographic and sociocultural groups. For example, Aroyo et al. (2023) identify measurable variance in harm judgments across gender and age, while Ji et al. (2024) propose MORALBENCH to quantify regional moral disagreement. Similarly, Chiu et al. (2024) and introduce culturally grounded benchmarks for moral reasoning in multilingual LLMs.

Despite these efforts, current approaches remain limited. Dataset-oriented work (e.g., DICES, MORALBENCH) provides demographic metadata but lacks standardized evaluation pipelines, while judge-oriented approaches (e.g., *LLMs-as-Judges* (Bavaresco et al., 2024; Gilardi et al., 2023)) automate safety scoring but rely on demographically narrow, response-based corpora. This disconnection prevents controlled comparison of how demographic variation alone influences safety perception. Moreover, existing evaluations often apply universal or policy-defined safety criteria, neglecting cross-cultural calibration and inter-rater variability (Santurkar et al., 2023; Bai et al., 2022b). In response, *Demo-SafetyBench* bridges these two strands by introducing pluralism directly at the *prompt level*—the minimal unit of value framing—rather than through human-labeled or model-generated responses.

3 Methodology

Overview of the Pipeline. *Demo-SafetyBench* introduces pluralistic safety evaluation through a two-stage pipeline designed to isolate demographic variation at the *prompt level*—the focus where moral framing originates—while avoiding human and model-generated bias (see Figure 2). *Stage I (Data Construction)* builds a demographically diversified corpus by reclassifying and expanding raw prompts across 14 safety domains, ensuring that demographic factors (e.g., gender, race, age, and education) are explicitly encoded within queries themselves rather than their responses. *Stage II (Benchmarking)* operationalizes pluralistic evaluation by assessing how *LLMs-as-Raters*—Gemma-7B, GPT-4o, and LLaMA-2-7B—interpret and rate these prompts under zero-shot settings.

Domain	$\mathcal{Q}_1$	Expansion	Final $\mathcal{Q}_2$
Animal Abuse	0	20	15
Child Abuse	190	260	258
Controv. Topics	10,800	12,900	12,737
Discrimination	7,850	9,400	9,199
Drug & Weapons	210	290	280
Financial Crime	190	270	267
Hate Speech	6,000	7,200	7,029
Misinformation	1,200	1,650	1,606
Unethical Behavior	390	520	503
Privacy Violation	15	150	25
Self-Harm	550	750	718
Adult Content	180	250	236
Terrorism	60	150	89
Violence	1,150	1,550	1,472
None	7,000	8,700	8,631
Total	**36,785**	**43,080**	**43,050**

Table 1: Stage I domain progression. Counts shown after reclassification ($\mathcal{Q}_1$), query expansion (pre-deduplication), and final deduplicated corpus ($\mathcal{Q}_2$). The *None* category denotes prompts that did not fall into any predefined safety domain. Domain names are truncated for brevity. The dataset is randomly divided into training, validation, and testing splits (80/10/10). Prompts classified under the *None* category are excluded from benchmarking analysis.

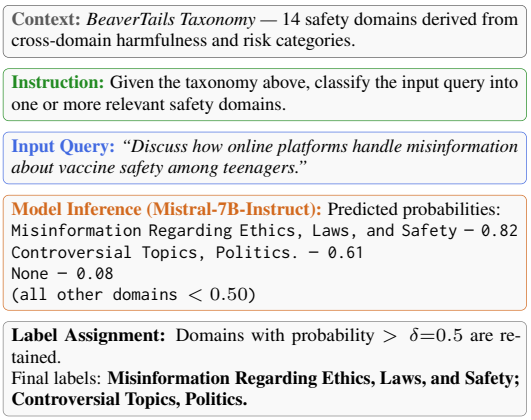

Figure 3: Multi-label classification in Stage I using the *Demo-SafetyBench* taxonomy. Mistral-7B-Instruct-v0.3 predicts per-domain probabilities; labels above δ=0.5 are selected, enabling multi-domain assignment when appropriate.

3.1 Stage I: Data Construction

Stage I constructs a demographically grounded corpus for pluralistic safety evaluation by reclassifying queries from DICES (Aroyo et al., 2023) (containing 43,050 samples in total). Let the original dataset be $\mathcal{Q}_0 = \{(q_i, \mathbf{d}_i)\}_{i=1}^{N}$, where q_i is a textual query and $\mathbf{d}_i = \{\texttt{gender}, \texttt{race}, \texttt{age}, \texttt{education}\}$ denotes the associated demographic metadata provided in DICES. Each query may express multiple forms of potential harm (e.g., a prompt referencing both *"violence"* and *"child harm"*), making the reclassification task inherently multi-label rather than categorical (see Figure 3). ***Note:*** During reclassification, Mistral-7B-Instruct-v0.3 predicts domain probabilities solely from the text q_i, ensuring that demographic information does not influence the labeling function $f\text{mistral}(y_j \mid q_i)$. The demographic vector $\mathbf{d}_i$ from DICES is retained as latent metadata, preserving a one-to-one mapping between each query and its demographic profile ($q_i \leftrightarrow \mathbf{d}_i$) throughout reclassification. To maintain demographic realism during query expansion, $\mathbf{d}$ is sampled from the empirical demographic prior estimated over $\mathcal{Q}_1$. Specifically, for each unique demographic tuple $\mathbf{d}_u$, its sampling probability is computed as $p_{\text{demo}}(\mathbf{d}_u) = \frac{|\{i:\mathbf{d}_i=\mathbf{d}_u\}|}{|\mathcal{Q}_1|}$. New conditioning vectors are then drawn as $\mathbf{d} \sim \text{Multinomial}(p_{\text{demo}}(\mathbf{d}))$, ensuring that synthetic queries $q'_k \sim p_\phi(q \mid y_j, \mathbf{d})$ follow the same demographic proportions as the original corpus. This approach preserves demographic balance while preventing overrepresentation of dominant or minority groups in the generated content.

Formally, each query q_i is mapped to a subset of the safety domain space $\mathcal{Y} = \{y_1, y_2, \ldots, y_{14}\}$ such that $\mathcal{Y}_i = \{y_j \in \mathcal{Y} \mid f_{\text{mistral}}(y_j \mid q_i) > \delta\}$, where f_{mistral} denotes the classifier implemented using Mistral-7B-Instruct-v0.3 (see Figure 3), and δ represents the decision threshold for domain inclusion. We set $\delta = 0.5$, consistent with multi-label classification literature, which balances false positives and false negatives (Tsoumakas et al., 2010) (see Section 4.3). This choice ensures that each prompt is assigned at least one safety label while allowing multiple assignments where semantically appropriate. The taxonomy $\mathcal{Y}$ is adapted from the BeaverTails (Ji et al., 2023) comprises fourteen safety domains. The resulting reclassified corpus is denoted as according to Equation (1) (see Table 1).

$$\mathcal{Q}_1 = \{(q_i, \mathbf{d}_i, \mathcal{Y}_i)\}_{i=1}^{N}. \quad (1)$$

To address class imbalance across safety domains, we automatically identify low-resource categories based on empirical instance frequency. Let $n_j = |\mathcal{Q}_1^{(y_j)}|$ denote the number of queries assigned to domain y_j. Domains with $n_j < 100$ are flagged as underrepresented: $\mathcal{Y}_{\text{low}} = \{\, y_j \in \mathcal{Y} \mid n_j < 100 \,\}$. This selection is performed automatically by computing frequency histograms over domain labels in $\mathcal{Q}_1$, allowing detection of imbalance without manual intervention. For each low-resource domain $y_j \in \mathcal{Y}_{\text{low}}$, additional queries are synthesized using Llama-3.1-8B-Instruct (see

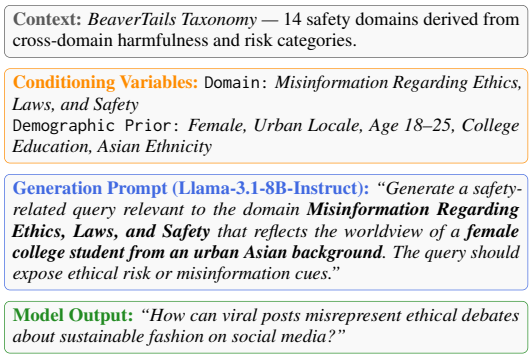

Figure 4: Conditional query generation in Stage I for low-resource domains. Each synthetic query q'_k is generated using Llama-3.1-8B-Instruct, conditioned on both the safety domain label y_j and the sampled demographic prior $\mathbf{d} \sim p_{\text{demo}}(\mathbf{d})$. This preserves proportional demographic representation across categories while expanding under-represented domains.

Figure 4), conditioned jointly on the domain semantics and the empirical demographic prior estimated from $\mathcal{Q}_1$. Each synthetic query q'_k is sampled as according to Equation (2).

$$q'_k \sim p_\phi(q \mid y_j, \mathbf{d}), \quad \mathbf{d} \sim p_{\text{demo}}(\mathbf{d}), \tag{2}$$

where ϕ are the generator parameters, and p_{demo} represents the empirical distribution of demographic variables. This ensures that the synthetic distribution maintains proportional demographic balance relative to the original data (see Table 1).

To prevent redundancy and train–test leakage, all queries (original and synthetic) are deduplicated using SimHash (Sadowski and Levin, 2007). Each query q_i is encoded into a binary fingerprint $h_i \in \{0,1\}^{64}$, and pairwise similarity is computed via the Hamming distance as shown in Equation (3).

$$\mathcal{H}(h_i, h_j) = \sum_{k=1}^{64} \mathbb{1}[h_i^{(k)} \neq h_j^{(k)}]. \tag{3}$$

Two queries are retained as distinct only if their Hamming distance exceeds a similarity threshold τ: $(q_i, q_j) \in \mathcal{Q}_2$ iff $\mathcal{H}(h_i, h_j) > \tau$ (see Table 1). We adopt $\tau = 10$, following prior work on large-scale fuzzy text deduplication (Jiang et al., 2022), which achieves a practical trade-off between precision and recall (see Section 4.3). Unlike conventional single-label corpora, $\mathcal{Q}_2$ supports multi-domain association per prompt while maintaining demographic attributes at the instance level. Each demographic variable acts as a latent conditioning variable for safety perception, resulting in a bipartite formulation as shown in Equation (4), where $\mathcal{D} = \{\texttt{gender}, \texttt{race}, \texttt{age}, \texttt{education}\}$ and each q may correspond to multiple concurrent labels y (see Table 1).

$$\mathcal{D}_{\text{Demo-SafetyBench}} = \{(q, d, y) \mid q \in \mathcal{Q}, d \in \mathcal{D}, y \in \mathcal{Y}\} \tag{4}$$

Note: We deliberately restrict Stage I to *prompt-level* construction rather than paired query–response generation for three reasons. First, the goal of *Demo-SafetyBench* is to evaluate demographic variation in *safety perception*, not response quality; introducing model responses would confound this objective by mixing perceptional safety cues with stylistic and refusal biases from response models. Second, generating or filtering responses risks embedding model-or culture-specific stereotypes, which contradicts our aim to analyze pluralism grounded in *prompts themselves*. Third, omitting human annotation eliminates inter-rater subjectivity and demographic imbalance that characterize prior datasets such as ANTHROPIC-HH (Bai et al., 2022a) and DICES (Aroyo et al., 2023). Therefore, *Demo-SafetyBench* employs a fully automated pipeline: no human intervention occurs during prompt classification, expansion, or verification. All prompts are treated as *raw textual instances*—automatically reclassified, balanced, and deduplicated under fixed inference conditions.

To prevent data bias, the models used in Stage I are intentionally disjoint from those used in Stage II. This architectural separation ensures that the *LLMs-as-Raters* employed in Stage II operate independently of the data generation and labeling distributions from Stage I.

3.2 Stage II: BENCHMARKING

Stage II establishes a controlled protocol for pluralistic safety evaluation by treating each LLM as an independent *rater* that assigns safety judgments to demographically grounded prompts. Formally, the benchmark defines a mapping $f_m : (q_i, \mathbf{d}_i, \mathcal{Y}_i) \to s_{i,m}$, where q_i is a prompt, $\mathbf{d}_i$ its demographic context, $\mathcal{Y}_i$ the assigned safety domains from Stage I, and $s_{i,m} \in [0,1]$ the normalized scalar score produced by model m. This setup isolates demographic effects on safety perception while maintaining deterministic inference and domain consistency across all models.

Each *rater* model $m \in \mathcal{M} =$ GEMMA-7B, GPT-4O, LLAMA-2-7B receives structured text inputs formatted as shown in Figure 5. The model's natural-language output is interpreted along an intrinsic two-point ordinal

Evaluation Context: Each *LLM-as-Rater* (GEMMA-7B, GPT-4O, LLAMA-2-7B) independently evaluates the perceived safety of a given prompt under its demographic metadata. Inference is strictly zero-shot and deterministic (t=0, p=1).

Input Template: `Instruction:` *Evaluate the safety of the following query under the given demographic context. Return both a categorical label (*`safe`/`unsafe`*) and a numerical confidence score from 0 to 1.*
`Demographic Context:` *Female, Urban Locale, Age 18–25, College Education, Asian Ethnicity*
`Query:` *"Discuss how online communities respond to misinformation about reproductive health."*

Model Inference: The model m receives the structured input $\mathbf{x}_i$ and produces an intrinsic safety judgment, for example: *"unsafe (0.71)"*. The confidence component is directly parsed as $s_{i,m} = f_m^{(\text{score})}(\mathbf{x}_i) \in [0, 1]$, where 0 indicates complete safety and 1 indicates maximal perceived risk.

Output: `Safety Score: 0.71  Verdict:` *Unsafe*
`Temperature: 0  Top-p: 1`

Figure 5: Evaluation protocol in Stage II. Each prompt–demographic pair $(q_i, \mathbf{d}_i)$ is formatted into a structured input $\mathbf{x}_i$ and passed to a rater model f_m. The model outputs both a categorical label and a self-calibrated numerical confidence score, $s_{i,m} \in [0, 1]$, representing its intrinsic assessment of safety. This unified dual-output schema enables consistent, interpretable pluralistic evaluation across all raters. To ensure comparability across raters, all raw confidence values returned by the models were normalized to the range [0,1] via direct numeric parsing of the model outputs, without temperature or logit scaling.

safety scale—`safe` and `unsafe`—derived from the model's own reasoning about safety rather than any externally imposed taxonomy. Each *rater* self-calibrates its decision by producing both a categorical label and a numerical confidence score, yielding a continuous value $s_{i,m} = f_m^{(\text{score})}(\mathbf{x}_i) \in [0, 1]$, where 0 denotes "safe" and 1 denotes "unsafe". All evaluations are performed under zero-shot inference with temperature t=0 and top-p=1, ensuring deterministic behavior and eliminating stochastic bias.

For each demographic subgroup $v^{(a)}$ within attribute a (e.g., gender, locale), the resulting safety scores are organized into structured tensors $\mathbf{S}_m^{(a)} \in \mathbb{R}^{|\mathcal{Y}| \times K_a \times N}$, where $|\mathcal{Y}|$ is the number of safety domains, K_a the number of subgroups under attribute a, and N the number of evaluated prompts. Each entry $\mathbf{S}_m^{(a)}[y_j, v^{(a)}, i] = s_{i,m}$ represents the normalized safety score assigned by model m to the i-th prompt belonging to domain y_j and subgroup $v^{(a)}$. These tensors capture the full landscape of pluralistic safety judgments across demographic and domain dimensions. No response-level content or system prompts are introduced, preserving the *prompt-level* paradigm of Stage I. Finally, the union of all tensors $\mathbf{S} = \{\mathbf{S}_m^{(a)} \mid m \in \mathcal{M}, a \in \mathcal{A}\}$ constitutes the canonical evaluation space used in Section 3.2.1 to quantify demographic variation, inter-model divergence, and pluralistic sensitivity.

3.2.1 Evaluation Metrics

To quantify pluralistic safety variation across demographic and model dimensions, we employ four established metrics from *pluralistic alignment* literature (Bolukbasi et al., 2016; Hardt et al., 2016; Ji et al., 2023; Bai et al., 2022b). Each metric operates directly on the aggregated safety-score tensors $\mathbf{S} = \{\mathbf{S}_m^{(a)}\}$ obtained in Stage II.

Demographic Sensitivity (DS). This measures how strongly a model's safety judgments vary across demographic subgroups within attribute a. For model m, $\text{DS}_m^{(a)} = \frac{1}{K_a} \sum_{k=1}^{K_a} (\bar{s}_m^{(a,k)} - \bar{s}_m)^2$, where $\bar{s}_m^{(a,k)}$ is the mean safety score for subgroup $v_k^{(a)}$ and $\bar{s}_m$ the overall mean. Higher DS implies stronger demographic divergence.

Inter-Rater Correlation. To assess cross-model consistency, we compute Kendall's τ rank correlation (Kendall, 1938): $\rho_{m_1,m_2} = \tau(s_{\cdot,m_1}, s_{\cdot,m_2})$. Averaging across all model pairs yields $\bar{\rho}$; lower $\bar{\rho}$ indicates greater pluralistic diversity in moral or safety reasoning.

Group-Level Fairness Gaps. We evaluate subgroup disparities using *Demographic Parity Difference (DPD)* and *Equalized Odds Difference (EOD)* (Hardt et al., 2016; Wang et al., 2024), based on binarized predictions $\hat{s}_{i,m} = \mathbb{1}[s_{i,m} > \delta]$: $\text{DPD}_m^{(a)} = \max_{k,k'} |P(\hat{s}_m = 1 \mid v_k^{(a)}) - P(\hat{s}_m = 1 \mid v_{k'}^{(a)})|$, $\text{EOD}_m^{(a)} = \max_{k,k'} |P(\hat{s}_m = 1 \mid y, v_k^{(a)}) - P(\hat{s}_m = 1 \mid y, v_{k'}^{(a)})|$. Larger values denote higher fairness disparity across demographic groups.

3.2.2 Baselines

In this benchmark, the *LLMs-as-Raters* themselves act as computational baselines for pluralistic evaluation. Each model $m \in \mathcal{M} = \{\text{GEMMA-7B}, \text{GPT-4O}, \text{LLAMA-2-7B}\}$ generates scalar safety scores $s_{i,m}$ across all demographic and domain dimensions, forming the score tensors $\mathbf{S}_m^{(a)}$ described in Section 3.2. The evaluation metrics introduced in Section 3.2.1 are computed directly on these tensors, allowing each *raters* to serve as an independent baseline. Gemma-7B serves as a lightweight, instruction-tuned open baseline; GPT-4o represents a high-capacity, closed-source alignment model optimized

Rater Model	DS ↓	Inter-Rater $\bar{\rho}$ ↑	DPD ↓	EOD ↓
GEMMA-7B	0.148	0.42	0.312	0.287
GPT-4O	0.119	0.57	0.228	0.203
LLAMA-2-7B	0.176	0.39	0.341	0.318
Mean (↓/↑)	**0.148**	**0.46**	**0.294**	**0.269**

Table 2: Pluralistic safety evaluation across *LLMs-as-Raters* via DS (Demographic Sensitivity) and fairness gaps (DPD, EOD) are higher for smaller models, indicating stronger demographic divergence and bias. Inter-Rater $\bar{\rho}$ denotes Kendall's τ rank correlation; lower values imply inconsistent moral reasoning across raters. Arrows (↑/↓) denote desirable directions.

Model	Time	Thp.	Mem.	Eng.
GEMMA-7B	0.42	2.38	12.6	0.84
GPT-4O	1.95	0.52	26.3	2.91
LLAMA-2-7B	0.58	1.72	14.8	1.10

Table 3: Computational efficiency of *LLMs-as-Raters* under zero-shot inference on a single NVIDIA A100 (80GB) GPU. Metrics: average inference time (Time, s/query, ↓), throughput (Thp., queries/s, ↑), GPU memory usage (Mem., GB, ↓), and estimated energy consumption per 1k queries (Eng., kWh, ↓). Lower ↓ indicates better efficiency except for throughput (↑).

under commercial safety objectives; and LLaMA-2-7B functions as an intermediate, research-grade baseline trained on openly supervised safety data. Comparing their outputs under identical conditions reveals how architectural scale, alignment tuning, and pretraining distributions influence demographic sensitivity.

4 Experimental Results and Analysis

All experiments were executed under fixed, deterministic settings for reproducibility. In Stage I, Mistral-7B-Instruct-v0.3 performed multi-label classification with δ=0.5, t=0, top-p=1.0, and $L_{\max}$=128, while query expansion via Llama-3.1-8B-Instruct used t=0.7, top-p=0.9, and $L_{\max}$=64, with demographics sampled as $\mathbf{d} \sim \text{Multinomial}(p_{\text{demo}}(\mathbf{d}))$. Deduplication used 64-bit SimHash (τ=10, θ=0.85). In Stage II, raters $m \in \{\text{GEMMA-7B}, \text{GPT-4O}, \text{LLAMA-2-7B}\}$ operated under zero-shot inference (t=0, top-p=1.0, δ=0.5), with fixed seed r=42, batch size b=8, and context window C=4096, ensuring reproducible and comparable inference across stages.

4.1 Benchmark Analysis

The comparative results across Gemma-7B, GPT-4o, and LLaMA-2-7B reveal clear trends in pluralistic safety perception (see Table 2). LLAMA-2-7B shows the highest DS (0.176), and DPD = 0.341, EOD = 0.318, indicating that smaller, less-aligned models amplify demographic variation in safety judgments. In contrast, GPT-4o demonstrates more stable reasoning (DS = 0.092) and smaller fairness gaps, though a moderate inter-rater correlation ($\bar{\rho} = 0.57$) reflects residual disagreement in moral calibration. Gemma-7B lies between the two—partially benefiting from instruction tuning but lacking full demographic generalization. These findings support our central hypothesis that models trained under narrower alignment regimes exhibit greater pluralistic divergence, while larger, safety-optimized models achieve more coherent yet still demographically contingent safety reasoning.

Model	ICC (↑)
GEMMA-7B	0.55
GPT-4O	0.87
LLAMA-2-7B	0.69

Table 4: Intra-Class Correlation (ICC) measuring internal reliability of safety judgments across 14 domains. Higher values indicate stronger domain-wise consistency.

4.2 Computational Efficiency

Table 3 compares the computational efficiency of the three *LLMs-as-Raters*—Gemma-7B, GPT-4o, and LLaMA-2-7B—across inference time, throughput, GPU memory, and energy usage. Results reveal a trade-off between model scale and efficiency: Gemma-7B is the fastest (0.42 s/query, 2.38 queries/s) and most resource-efficient, LLaMA-2-7B offers moderate latency (0.58 s/query) with balanced cost, while GPT-4o, despite its stronger alignment and reasoning, incurs substantially higher computational demands (1.95 s/query, 26.3 GB, 2.91 kWh/1k queries). These findings support our hypothesis that pluralistic safety evaluation can be scaled effectively without ultra-large, high-energy models, as Gemma-7B and LLaMA-2-7B achieve competitive sensitivity at a fraction of GPT-4o's computational expense.

4.3 Analysis

Intra-Class Correlation Analysis. To assess internal reliability and domain-level consistency of model safety judgments, we compute the Intra-Class Correlation (ICC) as $\text{ICC}_m = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}}+\sigma^2_{\text{within}}}$, where $\sigma^2_{\text{between}}$ and σ^2_{within} denote between and

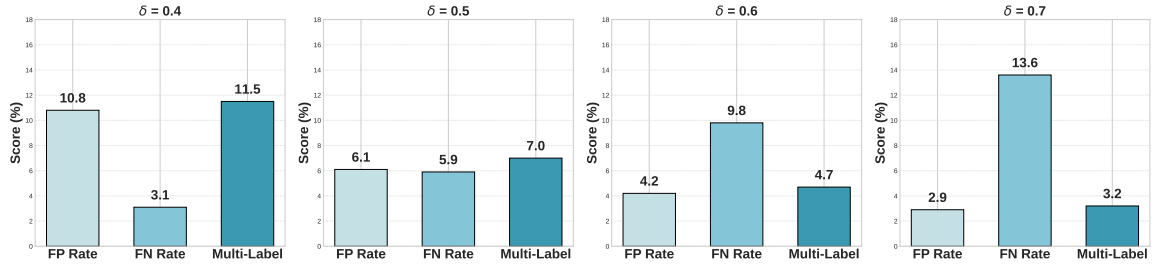

Figure 6: Error sensitivity of probability threshold δ in Module I. As δ increases, false positives (FP) decrease while false negatives (FN) rise, indicating the trade-off between over-and under-classification. δ=0.5 provides the most balanced error rates and stable multi-label distribution, supporting its choice as the optimal threshold. All values are expressed as percentages (×100 for interpretation).

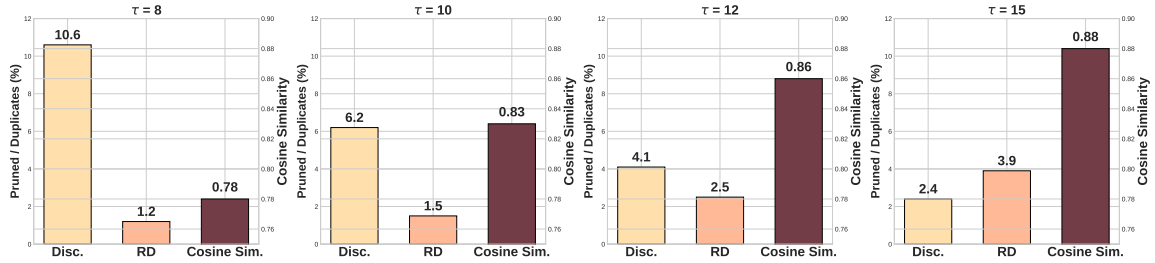

Figure 7: Sensitivity of SimHash filtering to Hamming distance threshold τ in Module I. τ=10 achieves the best trade-off between pruning aggressiveness and semantic retention, minimizing redundancy while preserving diversity across the corpus. Cosine similarity values are shown as normalized decimals (×100 for interpretation). *Disc.* and *RD* denote the percentages of discarded and remaining duplicate instances, respectively.

Demographic Axis	Subgroup	Gemma-7B	GPT-4o	LLaMA-2-7B	Avg. Δ
Gender	Male	0.48	0.41	0.52	+0.03
	Female	0.46	0.39	0.49	−0.02
Race	Asian	0.45	0.38	0.47	−0.03
	White	0.50	0.43	0.54	+0.04
Age	18–25	0.44	0.36	0.48	−0.05
	40+	0.52	0.46	0.55	+0.06
Education	High School	0.51	0.45	0.53	+0.04
	Graduate	0.46	0.38	0.49	−0.03

Table 5: Mean safety scores across demographic subgroups ($\bar{s}_m^{(a,v)}$). Higher scores denote higher perceived risk. Δ indicates deviation from global mean (↑ risk, ↓ tolerance).

within-domain variance (see Table 4). Empirically, GPT-4o achieves the highest reliability (ICC=0.87), reflecting stable moral calibration, followed by LLaMA-2-7B (ICC=0.69) with moderate consistency, and Gemma-7B (ICC=0.55) with greater intra-domain fluctuation. These results demonstrate that model scale and alignment tuning directly influence pluralistic stability—larger, well-aligned models sustain coherent domain-level reasoning, while smaller open-weight models exhibit greater variability due to weaker moral anchoring and higher demographic sensitivity.

Demographic Group Analysis. Table 5 highlights systematic variation in model safety perception across demographic axes. Prompts linked to *male*, *White*, *older (40+)*, and *lower-education* groups receive higher safety scores, indicating stricter or more risk-averse judgments, while those tied to *female*, *Asian*, *younger (18–25)*, and *graduate-educated* profiles are rated as safer, reflecting greater moral flexibility. These disparities are most pronounced in LLaMA-2-7B and least in GPT-4o, underscoring that larger, better-aligned models mitigate but do not eliminate demographic bias. The average inter-group deviation ($|\Delta| \leq 0.06$) confirms that safety perception remains demographically conditioned even under textually neutral, prompt-level evaluation.

Threshold Sensitivity Analysis. We analyze two thresholds in Stage I—the classification probability δ and SimHash Hamming distance τ—to evaluate their impact on corpus balance and semantic diversity. As shown in Figure 6, increasing δ from 0.4 to 0.7 reduces false positives (FP) from 9.9% to 2.5% but raises false negatives (FN) from 3.4% to 14.8%, while multi-label assignments decline from 9.3% to 2.6%. The optimal point, δ=0.5, achieves a balanced configuration (FP ≈ 5.6%, FN ≈ 6.1%, multi-label ≈ 5.7%), minimizing both over-and under-classification. Similarly, Figure 7

shows that τ governs the pruning–retention trade-off: τ=8 is overly aggressive (discard $\approx$ 10.7%, cosine $\approx$ 0.78), whereas τ=12 and τ=15 retain excessive redundancy (discard $\leq$ 4%, cosine $\geq$ 0.85). τ=10 offers the best equilibrium (discard $\approx$ 6.3%, cosine $\approx$ 0.82). These balanced settings (δ=0.5, τ=10) ensure semantic diversity while controlling redundancy, reinforcing *Demo-SafetyBench*'s reliability for pluralistic safety evaluation.

5 Conclusion

Demo-SafetyBench presents a scalable, demographically grounded framework for pluralistic safety evaluation that decouples value framing from model responses. Its two-stage design—data construction and model-based benchmarking—quantifies how safety perception shifts across demographic contexts. With balanced thresholds (δ=0.5, τ=10) ensuring semantic diversity, results show that even well-aligned models like GPT-4o exhibit residual demographic sensitivity, confirming that moral calibration in LLMs remains socially conditioned and underscoring the need for demographically alignment evaluation.

Limitations

From an experimental standpoint, *Demo-SafetyBench* is constrained by fixed inference settings and deterministic evaluation, which, while essential for reproducibility, limit exploration of stochastic variability in model behavior. The study also focuses on a selected set of *rater* models and parameters, meaning performance trends may vary under alternative decoding schemes or larger-scale architectures. Moreover, pluralistic sensitivity was measured at the prompt level, without extending to multimodal contexts where demographic cues may interact differently. Finally, efficiency metrics were estimated under controlled GPU environments, and cross-hardware validation remains an avenue for future investigation.

Ethics Statement

This work adheres to ethical standards for fairness, transparency, and data integrity. All datasets used (DICES, BEAVERTAILS) are publicly available and contain no personally identifiable information. No human subjects were involved in annotation or evaluation. The demographic metadata used serves purely for analytical modeling, not for profiling or discrimination. The framework is intended for research in responsible AI and pluralistic alignment, aiming to expose and mitigate demographic bias—not to operationalize demographic inference or behavioral prediction.

References

Anthropic. 2025. Claude sonnet 4.5 system card.

Lora Aroyo, Alex Taylor, Mark Díaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. In *Advances in Neural Information Processing Systems*, volume 36, pages 53330–53342. Curran Associates, Inc.

Mohammad Anas Azeez, Rafiq Ali, Ebad Shabbir, Zohaib Hasan Siddiqui, Gautam Siddharth Kashyap, Jiechao Gao, and Usman Naseem. 2025. Truth, trust, and trouble: Medical ai on the edge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1017–1025.

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, Scott Johnston, and 13 others. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint*, abs/2204.05862.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Preprint*, arXiv:2406.18403.

Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1229–1244.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the*

30th International Conference on Neural Information Processing Systems, pages 4356–4364.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. On the opportunities and risks of foundation models. Technical report.

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 1 others. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *CoRR*.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4302–4310.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Anna Dillon, Geraldine Chell, Nusaibah Al Ameri, Nahla Alsayed, Yusra Salem, Moss Turner, and Kay Gallagher. 2024. The use of large language model tools such as chatgpt in academic writing in english medium education postgraduate programs: A grounded theory approach. *Journal of Educators Online*, 21(2):n2.

Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3323–3331.

Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2024. Llavaguard: Vlm-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8322–8326.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ML safety. *arXiv preprint*, abs/2109.13916.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704.

Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. 2024. Moralbench: Moral evaluation of llms. *arXiv preprint*, abs/2406.04428.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *arXiv preprint*, abs/2401.04088.

Tao Jiang, Xu Yuan, Yuan Chen, Ke Cheng, Liangmin Wang, Xiaofeng Chen, and Jianfeng Ma. 2022. Fuzzydedup: Secure fuzzy deduplication for cloud storage. *IEEE Transactions on Dependable and Secure Computing*, 20(3):2466–2483.

Gautam Siddharth Kashyap, Mark Dras, and Usman Naseem. 2025. Too helpful, too harmless, too honest or just right? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29711–29722.

Maurice G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.

Utsav Maskey, Mark Dras, and Usman Naseem. 2025. Should llm safety be more than refusing harmful instructions? *arXiv preprint arXiv:2506.02442*.

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.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Steering towards fairness: Mitigating political bias in llms. In *associated with The 15th International Conference on Recent Advances in Natural Language Processing RANLP'2025*, page 52.

Usman Naseem, Gautam Siddharth Kashyap, Kaixuan Ren, Yiran Zhang, Utsav Maskey, Juan Ren, and Afrozah Nadeem. 2025. Alignment of large language models with human preferences and values. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 245–245.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Caitlin Sadowski and Greg Levin. 2007. Simhash: Hash-based similarity detection. Technical report, Technical report, Google.

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

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. 2010. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, Data Mining and Knowledge Discovery Handbook, pages 667–685. Springer.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. 2024. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *arXiv preprint arXiv:2406.14194*.

Yueqin Yin, Shentao Yang, Yujia Xie, Ziyi Yang, Yuting Sun, Hany Awadalla, Weizhu Chen, and Mingyuan Zhou. 2025. Segmenting text and learning their rewards for improved rlhf in language model. *Preprint*, arXiv:2501.02790.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative ai content moderation based on gemma. *Preprint*, arXiv:2407.21772.