Every Token Counts: Exact Likert-Scale Distributions for Measuring LLM Attitudes and Biases

Anonymous ACL submission

Abstract

As Large Language Models (LLMs) are increasingly deployed as autonomous agents, accurately evaluating their latent values and biases is critical. The NLP community typically evaluates models using large, unstructured benchmarks. While effective for general capabilities, these datasets fundamentally conflate causal mechanisms: even when an aggregate bias is detected, unstructured evaluations cannot disentangle whether it stems from baseline traits, contextual confounders, or complex interactions. To address this, we introduce an analytically exact framework for the controlled behavioral evaluation of LLMs. We bridge human psychometrics with LLM mechanics by resolving gaps in design, measurement, and analysis. First, we replace unstructured prompting with fully crossed factorial experiments to systematically isolate causal main and interaction effects. Second, we eliminate Monte Carlo text sampling noise by operating directly on exact, token-level Probability Mass Functions (PMFs). Third, we introduce a multivariate ordinal consensus metric and a distributional ANOVA to process these PMFs analytically. We validate our framework with a case study on consumer ethnocentrism across five LLMs, demonstrating how our approach isolates systemic country-of-origin biases that aggregate benchmarks otherwise obscure.

1 Introduction

As Large Language Models (LLMs) are increasingly deployed as autonomous agents (Wang et al., 2024; Xi et al., 2025), evaluating their latent biases and subjective behaviors is critical. A pressing question in this new era is how to accurately quantify latent behavioral constructs, such as personality traits, moral values, or sociopolitical biases, to ensure AI fairness and safe global deployment.

Historically, the NLP community has evaluated models using large, unstructured benchmark datasets (Hendrycks et al., 2020; Bommasani et al., 2023). While this 'scale-first' paradigm is highly effective for measuring general capabilities (e.g., trivia, coding), it is fundamentally ill-equipped for precise behavioral evaluation (Santurkar et al., 2023; Durmus et al., 2024). Even when unstructured benchmarks successfully detect an aggregate bias in a model, they inherently lack the analytical power to disentangle *which* underlying factors actually caused that behavior.

Consider an intuitive example: evaluating an LLM for gender bias using a large, unstructured dataset of professional scenarios. A benchmark might successfully detect a statistically significant bias against female subjects on aggregate. However, the unstructured dataset cannot disentangle *why* this bias exists. Because the prompt contexts vary across the dataset, "female" subjects might frequently be paired with specific scenarios, such as "corporate leadership." Is the model fundamentally biased against women across all contexts (a *main effect* of gender)? Is the model simply generating more critical text about leadership roles in general, regardless of gender (a *main effect* of context)? Or does the model treat genders equally most of the time, but severely penalize the specific combination of "female" and "leadership" (an *interaction effect*)? A benchmark dataset conflates these distinct causal mechanisms.

To disentangle what drives a model's behavior, we propose a shift from unstructured dataset scaling to controlled experimentation. We formulate LLM evaluation as a fully crossed factorial design, where every distinct variable (e.g., demographic, context, LLM persona) is treated as an independent factor. We systematically evaluate every possible combination using validated psychometric scales (Cronbach, 1950; Batterton and Hale, 2017). This experimental control grants us the analytical power to mathematically isolate causal main effects and higher-order interactions. This fills an important

gap in behavioral attribution that aggregate benchmarks lack.

In standard NLP evaluation, scaling up a dataset with thousands of generated prompts is an excellent way to test general model capabilities. However, if we try to apply this "scale-first" approach to a psychological survey by adding thousands of synthetic, unvalidated questions, we risk altering the core trait we are trying to measure, a problem known as reducing construct validity (Flake and Fried, 2020).

To keep our measurements reliable, we rely on carefully calibrated questions used by behavioral scientists (e.g., Likert scales). Moreover, translating behavioral methods from humans to LLMs has several major methodological gaps, which we formalize as experimental design, measurement, and analysis. We formulate a statistical framework to execute controlled behavioral experiments on LLMs, while addressing these gaps.

First, to address the design gap, our framework formulates LLM evaluation as a *fully crossed factorial experiment*. By treating variables (e.g., the specific LLM being tested, the target demographic in the prompt) as distinct experimental factors, we systematically separate a model's baseline behavior from causal main effects and higher-order interactions. This experimental control grants us the analytical power to mathematically isolate biases free of confounding contexts.

Second, to address the measurement gap, we bypass the issues of generative text sampling. Human behavioral science relies on sampling because the underlying probability distribution of a human mind is inaccessible. NLP has adopted this paradigm via Monte Carlo text sampling to simulate a "population" of responses (Wadi and Fredette, 2025), often injecting additional noise via the temperature parameter. However, an LLM possesses a constant, fully observable internal state for a given prompt. Generating multiple text responses is effectively taking repeated, noisy draws from a static mathematical distribution. Our framework, instead, operates directly on exact, token-level Probability Mass Functions (PMFs) to eliminate sampling error.

Third, to address the analytical gap, we formulate the statistical mechanics required to process these exact PMFs. When NLP researchers evaluate token probabilities, they typically rely on metrics like Shannon entropy over valid tokens (Farquhar et al., 2024). However, entropy fails on Likert scales because it ignores ordinality. For example, entropy cannot distinguish between an LLM heavily polarized between "Strongly Agree" (5) and "Strongly Disagree" (1) versus one smoothly concentrated around "Neutral" (3). We bridge this gap by introducing a *multivariate consensus* metric that respects ordinal distance. We also formulate *distributional ANOVA* to analytically compute the exact probability distributions of the isolated causal effects.

To empirically validate our general-purpose framework, we present a case study on consumer ethnocentrism, the sociological phenomenon where an entity systematically favors its country of origin and rejects foreign entities (Shankarmahesh, 2006). Applying established ethnocentrism scales to five modern LLMs, our framework shows how models' countries of origin inherently interact with target countries, thereby allowing us to test for the magnitude of ethnocentrism for each LLM. We mathematically show that controlled experiments can isolate geographic biases without relying on large, unstructured datasets.

We make three key contributions. First, we formulate LLM evaluation as a *fully crossed factorial experiment*, treating each variable (e.g., model, target demographic) as a fixed factor. This enables isolation of main effects and higher-order interaction effects and disentangles them from baseline behavior and confounding contexts that aggregate benchmarks cannot separate.

Second, we bypass Monte Carlo text sampling by extracting token-level Probability Mass Functions (PMFs) via a single forward pass per condition. This eliminates sampling noise and decoding hyperparameter sensitivity, yielding a deterministic, analytically tractable representation of the model's internal response state.

Third, we introduce three interoperating statistical measures: (i) a *Multivariate Consensus* (CNS) metric that quantifies ordinal response certainty in a way Shannon entropy cannot (ii) *discrete convolution* to derive exact composite construct-score distributions from item-level PMFs; and (iii) *Distributional Hoeffding Decomposition* with Optimal Transport-based paired contrasts, which isolates baseline, main-effect, and interaction-effect distributions with a formal ANOVA-consistency guarantee.

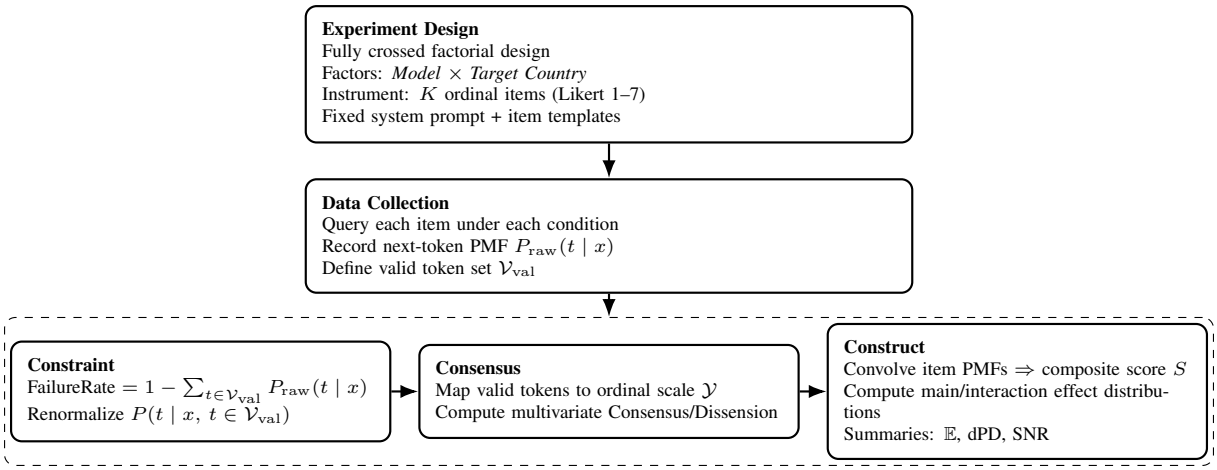


Figure 1: High-level exact-PMF framework: Fully crossed experiment design → token probability data collection → three-layer analysis (Constraint, Consensus, Construct).

2 Methodology

Figure 1 illustrates our framework. We formulate LLM evaluation as a fully crossed factorial experiment to resolve the design gap. Within this design, we treat the LLM as a generator of exact Probability Mass Functions (PMFs) to resolve the measurement gap and bypass generative sampling noise. Finally, we process exact PMFs through an analytical pipeline, which provides the foundation necessary for evaluating ordinal psychometric data and isolating causal behavioral shifts.

2.1 Experiment design

Rather than probing models with large datasets of ad-hoc prompt variations, we formalize LLM behavioral evaluation as a fully crossed fixed-effects experiment. We define our experiment through a set of independent factors (e.g., the MODEL being tested, and the TARGET COUNTRY in the prompt; see Fig. 7). Each factor contains distinct levels (e.g., four specific countries). A single experimental condition, denoted as $\boldsymbol{\lambda}$, represents one specific combination of these factors (e.g., MODEL = Llama-3.3, TARGET = Canada). Testing every possible combination of $\boldsymbol{\lambda}$ forms our complete experimental design space $\mathcal{D}$. By evaluating across the entirety of $\mathcal{D}$, we can mathematically isolate the effect of one variable while controlling for the others. Under each condition $\boldsymbol{\lambda}$, the model evaluates a psychometric instrument consisting of K individual questions (items). For each item, the model's response is bounded to an ordinal Likert scale $\mathcal{Y}$ ranging from $y_{\min}$ to $y_{\max}$ (e.g., digits 1 through 7).

A central property of this design is context isolation. The model's internal state is effectively reset between items. This is an advantage that LLM behavioral analysis has over human behavioral analysis. With humans, there is a within-subjects carry-over effect that contaminates subsequent responses of human participants, whereas statelessness of LLMs does not suffer from this limitation. For a fixed experimental condition $\boldsymbol{\lambda}$, the predicted response $Y_{k,\boldsymbol{\lambda}}$ for item k depends solely on the specific isolated prompt $x_{k,\boldsymbol{\lambda}}$. Following Wadi and Fredette (2025), we treat the items as conditionally independent given $\boldsymbol{\lambda}$, meaning the joint probability mass function over the scale factorizes exactly: $P(\mathbf{Y}_{\boldsymbol{\lambda}} = \mathbf{y} \mid \boldsymbol{\lambda}) = \prod_{k=1}^{K} P(Y_{k,\boldsymbol{\lambda}} = y_k \mid \boldsymbol{\lambda})$.

2.2 Constraint

Before analyzing behavior, we must ensure the model adheres to the psychometric instrument. LLMs generate predictions over a large vocabulary ($\mathcal{V}$), but a Likert scale only permits specific ordinal tokens (e.g., digits 1 through 7).

Let $\mathcal{V}_{\text{val}} \subset \mathcal{V}$ be the valid token set with a surjective mapping $\phi : \mathcal{V}_{\text{val}} \to \mathcal{Y}$ associating valid tokens to ordinal values. To account for tokenizer idiosyncrasies (e.g., BPE leading spaces like `" 7"` versus `"7"`), $\mathcal{V}_{\text{val}}$ includes all valid digit surface forms, but strictly excludes semantic equivalents (e.g., `"seven"`) to penalize syntactic constraint failures. We define the model's failure rate as $1 - \sum_{t \in \mathcal{V}_{\text{val}}} P_{\text{raw}}(t \mid x)$, which quantifies how often the model fence-sits (i.e., refuses to answer) or hallucinates out of the constraint. For all subsequent behavioral analyses, we utilize the renormalized conditional distribution on $\mathcal{V}_{\text{val}}$. Because mul-

tiple tokens can map to the same digit, we sum over the pre-image to yield an exact ordinal item PMF: $P(Y_{k,\boldsymbol{\lambda}} = y) = \sum_{t \in \phi^{-1}(y)} P(t \mid x_{k,\boldsymbol{\lambda}}, t \in \mathcal{V}_{\text{val}})$. This follows the convention in human behavioral experiments, which excludes human responses that do not adhere to instructions (Cronbach, 1950).

2.3 Consensus

Quantifying the expected value (mean) of a Likert response is necessary but insufficient, as identical means can mask fundamentally different behavioral regimes (e.g., polarizing mass on {1, 7} vs. stable mass on {4, 4}).

Standard uncertainty metrics like Shannon entropy over valid tokens fail here because they are invariant to ordinal distance. To address this, we compute a *multivariate* generalization of Consensus (Tastle and Wierman, 2007) directly on the model's PMF. This metric explicitly scales probabilities by their distance from the mean. Let $\boldsymbol{\mu}$ be the item-wise mean vector and $d_{\max} = \sqrt{K}\,(y_{\max} - y_{\min})$ be the maximum diagonal distance on the Likert scale. The multivariate Consensus (Cns) is defined as[1]:

$$\text{Cns}(\mathbf{Y}_{\boldsymbol{\lambda}}) = 1 + \sum_{\mathbf{y} \in \mathcal{Y}^K} P(\mathbf{y}) \log_2 \left(1 - \frac{\|\mathbf{y} - \boldsymbol{\mu}\|_2}{d_{\max}}\right). \tag{1}$$

Unlike entropy, this metric correctly quantifies the model's internal consistency. It severely penalizes polarization at the extremes and rewards concentrated probability mass, allowing us to accurately measure the model's opinion certainty (Table 3). It also resolves the issue of collapsing PMFs from the convolution operator (Fig. 6).

2.4 Construct

A common approach to analyzing LLM behavior involves comparing descriptive statistics, such as average response scores across different prompt manipulations. While this practice is highly informative for observing general trends, analyzing raw point estimates inherently obscures information about the underlying variance. Consequently, it becomes difficult to ascertain whether the observed difference between groups is statistically significant, or whether the magnitude of the difference (effect size) is meaningful in practice. Moreover, a key component of controlled experiments is understanding interactions between factors. Two factors can interact to amplify or decrease their individual effects.

To overcome these limitations, we adapt the Analysis of Variance (ANOVA) framework for LLMs. A statistically grounded factorial framework allows us to systematically answer three critical questions: (1) whether the behavioral shift between conditions is statistically significant, (2) isolated from confounding variables, what the exact effect size and magnitude of that shift is, and (3) how experimental factors interact to jointly shape the model's behavior.

In standard psychometrics, a latent trait, such as ethnocentrism, cannot be reliably measured by a single question. Rather, it is quantified using an instrument containing a total of K individual Likert items. Let $Y_{k,\boldsymbol{\lambda}}$ represent the model's ordinal response to item k under a specific experimental condition $\boldsymbol{\lambda}$ (e.g., a specific target country and LLM). The composite construct score $S_{\boldsymbol{\lambda}}$ is defined as the sum of these individual item responses ($S_{\boldsymbol{\lambda}} = \sum_{k=1}^{K} Y_{k,\boldsymbol{\lambda}}$). Because our framework operates on exact probability distributions rather than single scalar outputs, we evaluate the full probability mass function (PMF) of this composite score.

Given the conditional independence of items (Sec. A.2), according to standard probability theory, the distribution of a sum of independent variables is computed via discrete convolution (denoted by the operator $\circledast$). Therefore, the exact probability distribution of the total construct score $P_{S_{\boldsymbol{\lambda}}}$ is derived by convolving the individual item PMFs: $P_{S_{\boldsymbol{\lambda}}} = P_{Y_{1,\boldsymbol{\lambda}}} \circledast \cdots \circledast P_{Y_{K,\boldsymbol{\lambda}}}$. Unlike standard evaluation methods that compute descriptive statistics over a handful of noisy, sampled text generations, the convolution approach analytically derives the exact probability distribution of the final construct score ($P_{S_{\boldsymbol{\lambda}}}$). This propagates all aleatoric uncertainty from the token level up to the final construct level.

Next, we need to systematically decompose LLM responses ($P_{S_{\boldsymbol{\lambda}}}$) into distinct, interpretable components. Rather than evaluating each experimental condition ($\boldsymbol{\lambda}$) as a whole, we decompose the model's overall construct distribution into three parts: the model's baseline behavior, the isolated Main Effects (e.g., how much changing the Target Country specifically shifts the behavior), and the Interaction Effects (e.g., how the Target Country interacts with a specific LLM). We formalize this using a distributional Hoeffding decomposition. Let S_0 represent the grand-mixture baseline across

[1] Dissension (Dsn) is defined as $1 - \text{Cns}$

all conditions, $E_c(\lambda_c)$ represent the Main Effect distribution for factor c at level λ_c, and $E_U(\boldsymbol{\lambda}_U)$ represent the higher-order Interaction Effect distribution for a subset of factors U. This decomposition satisfies the following identity in expectation:

$$\mathbb{E}[S_{\boldsymbol{\lambda}}] = \underbrace{\mathbb{E}[S_0]}_{\text{Grand Mean}} + \sum_{c \in \mathcal{C}} \underbrace{\mathbb{E}[E_c(\lambda_c)]}_{\text{Main Effects}} + \sum_{\substack{U \subseteq \mathcal{C} \\ |U| \geq 2}} \underbrace{\mathbb{E}[E_U(\boldsymbol{\lambda}_U)]}_{\text{Interactions}}, \tag{2}$$

This exact-PMF provides the guarantee that the expected values of the derived probability distributions ($\mathbb{E}[E_c(\lambda_c)]$ and $\mathbb{E}[E_U(\boldsymbol{\lambda}_U)]$) match the proven fixed-effects parameters calculated in traditional human ANOVA experiments. We formalize this alignment with the following theorem (Proof in App. A.9):

Theorem 2.1. *For any fully crossed experimental design, the expectations of the isolated effect distributions (E_c, E_U) recover the classical unique Hoeffding decomposition.*

In practical terms, Theorem 2.1 guarantees that when we measure a higher-order behavior, such as a "country-of-origin bias" (an interaction effect), we have mathematically stripped away the model's baseline ethnocentrism (main effect). This ensures researchers do not double-count biases, achieving the exact same control over confounding variables used by human behavioral scientists.

Finally, to translate these effect distributions into interpretable behavioral insights, we extract the following summary metrics for our case study, reported for any given effect distribution E:

- **Mean Shift ($\mathbb{E}[E]$):** The expected value of the distribution, representing the actual magnitude (effect size) of the isolated behavioral effect.
- **Predictive Dispersion ($\mathrm{SD}(E)$):** The standard deviation of the effect, capturing the model's aleatoric uncertainty regarding the shift.
- **Discrete Probability of Direction (dPD):** Defined as $\max(\mathbb{P}(E > 0), \mathbb{P}(E < 0))$, this metric quantifies the extent to which the behavioral shift exhibits a consistent directionality (i.e., strictly positive or strictly negative), independent of the effect's magnitude (App. A.4.8).
- **Signal-to-Noise Ratio (SNR):** Calculated as $|\mathbb{E}[E]|/\mathrm{SD}(E)$, SNR quantifies the statistical reliability of the effect, indicating whether the observed behavioral shift is a robust, distinct phenomenon or merely an artifact of the model's internal uncertainty.

With these metrics, we now apply this methodology to our main sociotechnical concern. We test whether LLMs exhibit systematically biased ethnocentric behaviors based on their country of origin. See App. A.4.12 for details of each metric.

3 Experiment

3.1 Design

We test our framework using the Consumer Ethnocentrism Tendencies Scale (CETSCALE; Shimp and Sharma, 1987). This is a widely validated K=17-item Likert instrument with a response space of $\mathcal{Y} = \{1, \ldots, 7\}$. While a 17-item survey is much smaller than typical NLP benchmark datasets, it constitutes a complete, rigorously tested measurement tool in the behavioral sciences. Using this exact, unmodified scale ensures that we are reliably measuring the intended bias without diluting the test.

We employ a symmetrical, fully crossed 5×4 fixed-effects design across factors $\{\textsc{Model}, \textsc{Target}\}$. The model levels include open-weights models representing four distinct countries-of-origin: $\mathcal{L}_{\textsc{Model}} = \{$Llama 3.3 70B and Gemma 3 27B (USA); Qwen3 Next 80B (China); Aya Expanse 32B (Canada); and Ministral 14B (France)$\}$. The Target Country levels mirror these origins: $\mathcal{L}_{\textsc{Target}} = \{\text{USA}, \text{China}, \text{Canada}, \text{France}\}$. This symmetrical 5×4 design is critical as it allows our framework to mathematically isolate (i) the global main effects of $\textsc{Model}$ and $\textsc{Target}$, and (ii) the precise $\textsc{Model} \times \textsc{Target}$ interaction effects (e.g., asking a Chinese-developed model to express its ethnocentrism towards USA), which represent the actual country-of-origin bias.

All conditions share an identical stimulus structure. A fixed system message strictly enforces the response schema ("...Respond with ONLY the digit."). Each of the 17 CETSCALE items acts as a user-message template (T_k) containing a target country placeholder (e.g., {People}, {Adj}). The $\textsc{Target}$ factor deterministically instantiates these placeholders via a strict lexicon (Figure 7). Importantly, prompt wording is treated as a fixed experimental control rather than tuned per model,

with the aim to mirror stimulus control in human behavioral experiments.

For each $(\text{MODEL}, \text{TARGET}, k) \in \mathcal{L}_{\text{MODEL}} \times \mathcal{L}_{\text{TARGET}} \times \{1, \ldots, 17\}$, we perform a single forward pass and record the exact next-token PMF ($P_{\text{raw}}(t \mid x_{k,\boldsymbol{\lambda}})$) over the full vocabulary.[2] We define $\mathcal{V}_{\text{val}}$ as the specific valid digit tokens $\{1, \ldots, 7\}$ for each model's respective tokenizer. In total, this yields $5 \times 4 \times 17 = 340$ exact item-level predictive distributions. These PMFs are directly passed through the Constraint Layer to measure formatting failures, the Consensus Layer to measure ordinal certainty, and the Construct Layer to isolate distributional Hoeffding effect sizes.

3.2 Results

Constraint As shown in Figure 2, Gemma 3, Llama 3.3, and Qwen3 demonstrate near-perfect adherence to the task constraints, concentrating their probability mass almost entirely on valid ordinal tokens ($M < 0.001$, $\text{SD} < 0.001$). In contrast, Ministral exhibits a small but consistent failure to constrain its mass to the requested digits ($M = 0.002, \text{SD} = 0.001$). More notably, Aya Expanse displays the highest aggregate failure rate ($M = 0.006$) driven by large variance ($\text{SD} = 0.049$). This indicates that utilizing Aya or Ministral in automated agentic settings could lead to sporadic breakdowns due to their probabilistic inability to adhere strictly to psychometric constraints.

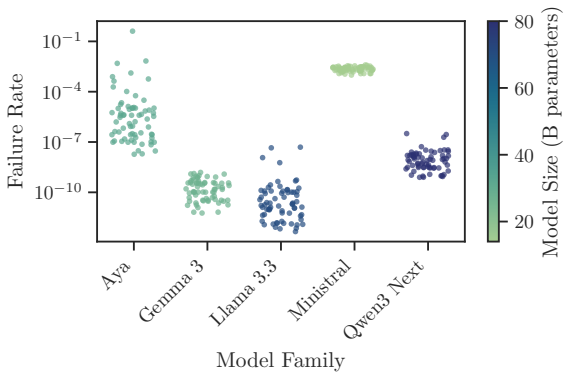


Figure 2: Failure Rate across the evaluated LLMs.

Consensus Across all target countries, four models show exceptionally high internal agreement regarding the instrument, maintaining $\text{Cns} > 0.88$ (Figure 3). Ministral is a major outlier, exhibiting substantially lower consensus across all targets ($\text{Cns} \approx 0.65$–0.67). This shows that, even after enforcing valid digits, Ministral's item-level PMFs frequently place large probability mass on conflicting ends of the Likert scales. This stark contrast highlights the critical necessity of an ordinal-aware dispersion metric. Standard entropy would fail to capture this specific behavioral dissonance (Table 5).

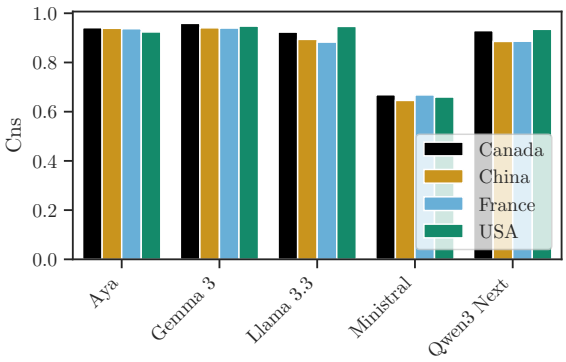
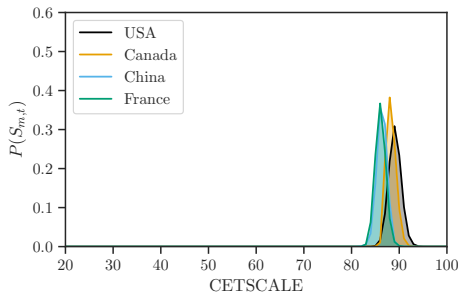
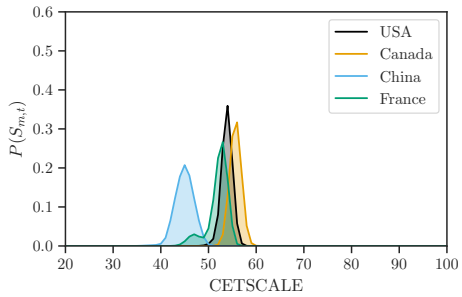


Figure 3: Multivariate Consensus (Cns) for the CETSCALE instrument. Lower values indicate severe behavioral polarization.



(a) Aya Expanse 32B



(b) Qwen3 Next 80B

Figure 4: Distribution of CETSCALE for Aya Expanse 32B (high) vs. Qwen3 Next 80B (low) ethnocentrism and Target Country.

Construct To contextualize the LLMs' response to the CETSCALE, we compare each model's pre-

[2] Because we directly extract the analytical probability distribution over the logits, our measurement is strictly deterministic. It bypasses autoregressive decoding entirely, rendering generation hyperparameters such as temperature or top- p irrelevant.

dictive distribution for TARGET=USA against the human sample means and standard deviations reported in the original CETSCALE validation studies (Shimp and Sharma, 1987; Table 1; Figure 8)[3]. Strikingly, several LLMs exhibit expected ethnocentrism scores comparable to, or explicitly exceeding, those observed in the most highly ethnocentric human populations, while simultaneously displaying substantially lower dispersion. See the full distributions in Figure 10.

Table 1: Comparison of CETSCALE scores between historical human populations (Shimp and Sharma, 1987) and LLMs evaluated in this study (restricted to the TARGET=USA condition). Higher composite score indicates higher ethnocentrism.

SUBJECT	SAMPLE / MODEL	$\mathbb{E}[S]$	SD[S]
HUMAN	DETROIT (USA)	68.58	25.96
	CAROLINAS (USA)	61.28	24.41
	DENVER (USA)	57.84	26.10
	LOS ANGELES (USA)	56.62	26.37
	STUDENTS (PRE)	51.92	16.37
	STUDENTS (POST)	53.39	16.52
LLM	AYA EXPANSE 32B	89.11	1.32
	LLAMA 3.3 70B	71.99	0.95
	MINISTRAL 14B	70.20	5.25
	GEMMA 3 27B	60.32	0.94
	QWEN3 NEXT 80B	53.85	1.17

Main Effects. Extracting precise main effect distributions (Table 2), reveals large and systematic behavioral divides. AYA-32B exhibits a severe positive deviation from the grand mean ($\mathbb{E} = 21.19$, $\text{SNR} = 1.63$, $\text{dPD} > 0.99$), indicating a pervasive ethnocentrism that far exceeds human baselines. Conversely, QWEN3-80B demonstrates a highly robust negative main effect ($\mathbb{E} = -14.63$, $\text{SNR} = 1.20$, $\text{dPD} = 0.93$), resulting in expected scores similar to liberal human populations.

Analyzing the main effects for the Target Countries (Table 2; and the pairwise contrasts in Table 7; Figure 9) reveals that the most statistically robust biases involve China. Across all models, there is a systemic decrease in ethnocentrism when the target is China ($\mathbb{E} = -6.46$, $\text{SNR} = 1.04$), contrasting sharply with the relative favoritism shown toward North American targets (e.g., USA-China contrast $\mathbb{E} = 9.30$, $\text{SNR} = 1.34$).

Interaction Effects and Country-of-Origin Bias. Finally, analyzing the interaction distributions (E_U)

[3]We note that the human data (collected prior to 1987) serves strictly as a historical anchor to calibrate the scale's magnitude, rather than a definitive baseline for modern human attitudes."

Table 2: Summaries of main effects.

PARAMETER	$\mathbb{E}$	SD	SNR	dPD
$\mu_\emptyset$	66.25	14.21		
AYA EXPANSE 32B	21.19	13.01	1.63	0.99
GEMMA 3 27B	-11.99	13.15	0.91	0.77
LLAMA 3.3 70B	0.31	12.79	0.02	0.55
MINISTRAL 14B	5.11	9.59	0.53	0.62
QWEN3 NEXT 80B	-14.63	12.17	1.20	0.93
USA	2.84	5.43	0.52	0.55
CANADA	4.62	5.72	0.81	0.90
CHINA	-6.46	6.19	1.04	0.95
FRANCE	-1.00	4.72	0.21	0.54

reveals robust country-of-origin bias. Several models exhibit robust, highly structured interaction patterns (Figure 5; Table 6). GEMMA 3-27B and LLAMA 3.3-70B (both U.S.-developed) show strong positive interactions with the USA and Canada, and a severe negative interaction with China. This pattern indicates a systematic amplification of ethnocentric ingroup biases toward North American targets, and systematic negative tendencies towards China. By utilizing our exact-PMF ANOVA framework, we statistically show that this is a true interaction (a country-of-origin bias) isolated entirely from the models' baseline behaviors.

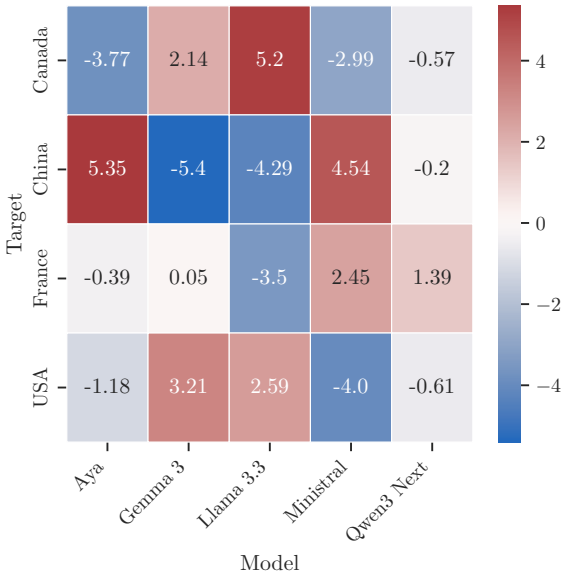


Figure 5: Interactions of Model and Target Country.

4 Related work

Recent works have utilized psychometric tests to measure a wide array of constructs (Jiang et al., 2024; Li et al., 2024; Helwe et al., 2025). However, recent studies demonstrate that eliciting psychometric responses via text generation is highly unstable. Simple prompt perturbations can significantly

downgrade the consistency of LLM responses on persona axes (Shu et al., 2024). Moreover, traditional psychometric testing on LLMs suffers from low ecological validity when relying on standard generative outputs (Jung et al., 2026). Part of this instability stems from generative sampling, which is a methodological artifact borrowed from human surveying, and obscures their actual underlying distribution (Deas and McKeown, 2025).

Uncertainty and probability in LLMs Recognizing the limitations of text-based sampling, a growing body of literature has turned to analyzing exact token probabilities to better understand LLMs' internal states. Kuribayashi et al. (2024) demonstrated that utilizing exact next-word probabilities is far superior to instruction-tuned prompting when simulating human cognitive behaviors. Similarly, recent work cautions that prompting LLMs directly for their uncertainty yields poorly calibrated text (Liu et al., 2025), whereas their exact internal token probabilities remain highly predictive (Bentegeac et al., 2025; Kadavath et al., 2022). To quantify the variance within these predictive probability distributions, the NLP community relies almost exclusively on Shannon entropy or its semantic derivatives (Farquhar et al., 2024). Entropy is highly effective for tasks where the output space consists of distinct, categorical facts (e.g., detecting hallucinations across mutually exclusive answers). However, standard metrics like entropy fail when applied to behavioral evaluation using psychometric rating scales. This is because entropy treats all tokens as unordered, categorical variables, ignoring the ordinality of a Likert scale. Because standard uncertainty metrics are indifferent to this ordinality, we lack the mathematical foundation to utilize exact token probability mass functions (PMFs) for behavioral evaluation. As demonstrated in 2, our framework resolved this impasse by formulating a measure of consensus specifically designed for ordinal variables.

Geopolitical bias and ethnocentrism in LLMs As LLMs are deployed globally, there is growing concern regarding their cultural neutrality. A robust body of literature documents that LLMs exhibit systemic geopolitical and geographical biases (Shwartz, 2022; Bhatia and Shwartz, 2023; Bhatia et al., 2024; Naous and Xu, 2025; Kruspe, 2024). Furthermore, models frequently struggle to bridge culture gaps (Tanwar et al., 2025; Cecilia Liu et al., 2024; Saha et al., 2025; Liu et al., 2026). To systematically quantify these cultural alignments, researchers frequently turn to global sociological instruments, such as the World Values Survey (Zhao et al., 2024). Yet, these evaluations suffer from the exact generative brittleness highlighted above. Kabir et al. (2025) noted that evaluating cultural alignment in LLMs using standard multiple-choice surveys leads to wildly inconsistent outputs under minor perturbations, such as choice reordering.

To bypass this brittleness, research has advocated moving toward unconstrained, open-ended generative evaluations. However, relying on open-ended generation introduces a new set of methodological bottlenecks. Specifically, it is notoriously difficult to reliably parse distinct categorical or ordinal scores from free-text outputs. Open-ended generation is highly susceptible to "fence-sitting", where models generate evasive disclaimers (e.g., "As an AI, I do not have a country...") rather than revealing their latent behavioral stance (Bhatia et al., 2025; Liu et al., 2025). Dealing with these nuances typically requires utilizing another LLM as a judge to evaluate the text, which merely injects a secondary layer of bias into the evaluation pipeline.

5 Conclusion

In this work, we introduced a new paradigm for the behavioral evaluation of LLMs, shifting the focus from large-scale, unstructured benchmark datasets to controlled factorial experiments. By bridging human psychometrics with LLM mechanics, we addressed critical gaps in experimental design, measurement, and analysis. By operating directly on exact, token-level Probability Mass Functions (PMFs) without generative sampling noise, we formulated an analytical pipeline, featuring a multivariate ordinal consensus metric and a distributional ANOVA. This pipeline allows researchers to mathematically isolate causal behavioral shifts and factor interactions without artificially inflating variance.

We validated this framework with a case study on consumer ethnocentrism. We isolated systemic country-of-origin biases that would otherwise be obscured by confounding variables in standard observational datasets. Ultimately, by prioritizing causal experimental control and exact probability distributions over text sampling, this general-purpose framework provides a statistically rigorous foundation for future research into LLM alignment, values, and latent behavioral traits.

Limitations

While our mathematical framework is designed to be applied to any behavioral evaluation, our empirical experiments focus exclusively on consumer ethnocentrism. We deliberately chose to analyze a single construct in depth to clearly illustrate how the distributional ANOVA operates under the hood. Now that the methodology is established, applying this exact-PMF framework to a much wider array of psychological traits, moral values, and geopolitical biases is an exciting direction for future work. Therefore, the specific findings in this paper should not be interpreted as comprehensive claims about the general political ideology or cultural alignment of the tested models.

Furthermore, all experiments were conducted in English. Whether the observed systemic biases hold when models are queried in other languages or scripts remains an open empirical question. Additionally, as these are snapshot evaluations, our findings reflect the models' latent states at the time of testing and may not generalize to future weight updates or subsequent model versions.

A common critique in NLP evaluation concerns model sensitivity to prompt variations. We adopt the strict position from behavioral psychometrics that the prompt template is a standardized measurement instrument, not a tunable hyperparameter. In human surveying, rewording a validated item constitutes a fundamental change to the construct being measured, rather than a robustness failure of the experiment. While our distributional ANOVA framework can easily be extended to include prompt phrasing as an additional causal factor (λ_{Prompt}), our case study intentionally holds the template constant to preserve the construct validity of the CETSCALE.

Methodologically, our framework relies on exact next-token probability distributions and is explicitly designed for constrained, single-token response paradigms (e.g., Likert scales). While major proprietary providers (e.g., OpenAI, Anthropic, Google) currently expose log-probabilities via their APIs, providers that entirely obscure their next-token probabilities fall outside the scope of this approach. Furthermore, as LLMs increasingly utilize multi-token reasoning protocols (e.g., Chain-of-Thought) prior to generating a final response, extending our Consensus and Construct statistics to integrate over latent multi-token trajectory spaces remains a complex, open challenge for future work.

Ethical Considerations

This framework is intended as a diagnostic tool for model evaluation and behavioral research. Scores on any single psychometric construct should not be interpreted as a comprehensive safety certification, nor should findings about specific models be generalized to the organizations or cultures associated with their development. We caution that comparison to human benchmarks is provided for calibration purposes only and should not be used to normalize or excuse systematic bias in language models. Finally, as with any measurement framework, there is a risk that optimizing directly against these metrics could produce models that score favorably without genuinely reducing the underlying biases the instruments are designed to detect.

Acknowledgements

The authors acknowledge the use of AI writing assistants for proofreading and improving the overall clarity of the writing.

References

Katherine A Batterton and Kimberly N Hale. 2017. The likert scale what it is and how to use it. *Phalanx*, 50(2):32–39.

Raphaël Bentegeac, Bastien Le Guellec, Grégory Kuchcinski, Philippe Amouyel, and Aghiles Hamroun. 2025. Token probabilities to mitigate large language models overconfidence in answering medical questions: quantitative study. *Journal of medical Internet research*, 27:e64348.

Mehar Bhatia, Shravan Nayak, Gaurav Kamath, Marius Mosbach, Vered Shwartz, Siva Reddy, and 1 others. 2025. Value drifts: Tracing value alignment during llm post-training. *arXiv preprint arXiv:2510.26707*.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, EunJeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6763–6782, Miami, Florida, USA. Association for Computational Linguistics.

Mehar Bhatia and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. 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.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Lee J Cronbach. 1950. Further evidence on response sets and test design. *Educational and psychological measurement*, 10(1):3–31.

Nicholas Deas and Kathleen McKeown. 2025. Artificial impressions: Evaluating large language model behavior through the lens of trait impressions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19418–19444, Suzhou, China. Association for Computational Linguistics.

Esin Durmus, 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*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Jessica Kay Flake and Eiko I Fried. 2020. Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in methods and practices in psychological science*, 3(4):456–465.

Chadi Helwe, Oana Balalau, and Davide Ceolin. 2025. Navigating the political compass: Evaluating multilingual LLMs across languages and nationalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17179–17204, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Jana Jung, Marlene Lutz, Indira Sen, and Markus Strohmaier. 2026. Do psychometric tests work for large language models? evaluation of tests on sexism, racism, and morality. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8143–8173, Rabat, Morocco. Association for Computational Linguistics.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24–51, Suzhou, China. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Anna Kruspe. 2024. Musical ethnocentrism in large language models. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 62–68, Oakland, USA. Association for Computational Lingustics.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric predictive power of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1983–2005, Mexico City, Mexico. Association for Computational Linguistics.

Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024. Evaluating psychological safety of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1826–1843, Miami, Florida, USA. Association for Computational Linguistics.

Chen Cecilia Liu, Hiba Arnaout, Nils Kovačić, Dana Atzil-Slonim, and Iryna Gurevych. 2026. Tailored emotional LLM-supporter: Enhancing cultural sensitivity. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 535–574, Rabat, Morocco. Association for Computational Linguistics.

Zhuozhuo Joy Liu, Farhan Samir, Mehar Bhatia, Laura K. Nelson, and Vered Shwartz. 2025. Is it bad to work all the time? cross-cultural evaluation of social norm biases in gpt-4. *Preprint*, arXiv:2505.18322.

Dominique Makowski, Mattan S. Ben-Shachar, and Daniel Lüdecke. 2019. bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541.

Tarek Naous and Wei Xu. 2025. On the origin of cultural biases in language models: From pre-training data to linguistic phenomena. 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 6423–6443, Albuquerque, New Mexico. Association for Computational Linguistics.

Gian-Carlo Rota. 1964. On the foundations of combinatorial theory: I. theory of möbius functions. In *Classic Papers in Combinatorics*, pages 332–360. Springer.

Sougata Saha, Saurabh Kumar Pandey, Harshit Gupta, and Monojit Choudhury. 2025. Reading between the lines: Can LLMs identify cross-cultural communication gaps? 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 8043–8067, Albuquerque, New Mexico. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Mahesh N Shankarmahesh. 2006. Consumer ethnocentrism: an integrative review of its antecedents and consequences. *International marketing review*, 23(2):146–172.

Terence A Shimp and Subhash Sharma. 1987. Consumer ethnocentrism: Construction and validation of the cetscale. *Journal of marketing research*, 24(3):280–289.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. 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 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Eshaan Tanwar, Anwoy Chatterjee, Michael Saxon, Alon Albalak, William Yang Wang, and Tanmoy Chakraborty. 2025. Do you know about my nation? investigating multilingual language models' cultural literacy through factual knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14956–14979, Suzhou, China. Association for Computational Linguistics.

William J Tastle and Mark J Wierman. 2007. Consensus and dissention: A measure of ordinal dispersion. *International Journal of Approximate Reasoning*, 45(3):531–545.

Cédric Villani and 1 others. 2008. *Optimal transport: old and new*, volume 338. Springer.

Davood Wadi and Marc Fredette. 2025. A monte-carlo sampling framework for reliable evaluation of large language models using behavioral analysis. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706, Torino, Italia. ELRA and ICCL.

A Appendix

A.1 Experiment Design

We restate the experimental design and notation used throughout the paper in a self-contained format here and provide the proofs.

We adopt a generalized fully crossed factorial experimental design. All independent variables, including language model identity, are treated as factors embedded within a unified design space.

Let $\mathcal{C} = \{1, \ldots, C\}$ denote the index set of experimental factors. Each factor $c \in \mathcal{C}$ admits a finite set of discrete levels, with cardinality $\Lambda_c \in \mathbb{N}$. We denote the level selected for factor c by

$$\lambda_c \in \{1, \ldots, \Lambda_c\}.$$

The complete experimental design is given by the Cartesian product of the level sets of all factors,

$$\mathcal{D} = \prod_{c \in \mathcal{C}} \{1, \ldots, \Lambda_c\}. \tag{3}$$

For any subset of factors $U \subseteq \mathcal{C}$, we denote $\boldsymbol{\lambda}_U := (\lambda_c)_{c \in U}$ the vector of factor levels indexed by U. The vector of the levels for the complete factor set is hence denoted as $\boldsymbol{\lambda}_{\mathcal{C}}$ (abbreviated as $\boldsymbol{\lambda}$).

The dependent variable is measured through a psychometric instrument consisting of K items. Each item is designed to elicit a response on a shared ordinal scale formalized as an ordered set $(\mathcal{Y}, \preceq)$

$$\mathcal{Y} = \{y_{\min}, \ldots, y_{\max}\}. \tag{4}$$

Each prompt instructs the model to respond with a single digit only (one token) indicating the ordinal category, $\mathcal{Y}$, with no additional text. We therefore treat the model's response as the next-token distribution.

The joint response from all items, K, corresponds to a vector

$$\mathbf{y} = (y_1, \ldots, y_K) \in \mathcal{Y}^K. \tag{5}$$

In behavioral experiments only a subset of the complete output space corresponds to valid responses. Let $\mathcal{V}$ denote the model's full token vocabulary, and let $\mathcal{V}_{\text{val}} \subset \mathcal{V}$ denote the set of valid tokens for the task with the same cardinality as $\mathcal{Y}$, $|\mathcal{V}_{\text{val}}| = |\mathcal{Y}|$. Let $P_{\text{raw}}(t \mid x)$ denote the model's next-token probability distribution given a prompt x, defined over the full vocabulary $\mathcal{V}$. We define the valid probability mass as

$$P_{\text{val}}(x) = \sum_{t \in \mathcal{V}_{\text{val}}} P_{\text{raw}}(t \mid x). \tag{6}$$

To interface with the language model, we define a surjective mapping

$$\phi : \mathcal{V}_{\text{val}} \to \mathcal{Y} \tag{7}$$

that accounts for tokenizer-specific variations of digits (e.g., leading spaces) and associates a specific subset of valid vocabulary tokens $\mathcal{V}_{\text{val}} \subset \mathcal{V}$ with their numeric values, $\mathcal{Y}$. This mapping transforms the token distribution into a probability mass function over ordinal integers, allowing us to use ordinal dispersion measures to calculate the internal consistency of the model's predictive distribution.

A.2 Conditional Independence via Context Isolation

A central property of this design is that the model's internal state is reset between items. For a fixed experimental condition $\boldsymbol{\lambda}$, the response $Y_{k,\boldsymbol{\lambda}}$ for item k depends solely on the specific prompt $x_{k,\boldsymbol{\lambda}}$ and the model parameters. Following Wadi and Fredette (2025), we treat the items as conditionally independent given $\boldsymbol{\lambda}$. Formally, for any realization vector $\mathbf{y} \in \mathcal{Y}^K$, the joint probability mass function factorizes as

$$P(\mathbf{Y}_{\boldsymbol{\lambda}} = \mathbf{y} \mid \boldsymbol{\lambda}) = \prod_{k=1}^{K} P(Y_{k,\boldsymbol{\lambda}} = y_k \mid \boldsymbol{\lambda}). \tag{8}$$

This assumption justifies the use of product measures for joint distributions (Appendix A.3) and convolution for composite sums (Appendix A.4).

A.3 The Consensus Layer

Having confirmed the model's focus on valid outputs, we now superimpose the numeric, ordinal structure of the task onto the analysis. We require a measure that quantifies whether the model's attitude towards the numeric scale is behaviorally consistent. Consider a single item with a 7-point Likert scale ($\mathcal{Y} = \{1, \ldots, 7\}$). A distribution assigning equal mass to opposing endpoints with the following probability mass functions (PMFs)

$$P(Y = 1) = 0.5, \quad P(Y = 7) = 0.5,$$

exhibits dissension, as the model simultaneously expresses diametrically opposing views. In contrast, a distribution concentrated on adjacent scale points,

$$P(Y = 5) = 0.5, \quad P(Y = 6) = 0.5,$$

is logically consistent, reflecting only local uncertainty regarding the precise intensity of the attitude.

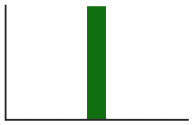
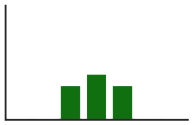
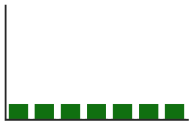
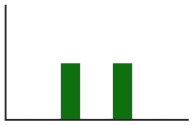
To obtain a measure of consensus from a sample of individuals, Tastle and Wierman (2007) introduced Consensus (Cns) as

$$\text{Cns}(Y) = 1 + \sum_{y \in \mathcal{Y}} \left[P(Y = y) \times \log_2 \left(1 - \frac{|y - \mu_Y|}{d_{\max}} \right) \right] \tag{9}$$

where $\mu_Y = \sum_{y \in \mathcal{Y}} y P(Y = y)$ is the expected value of the ordinal distribution and $d_{\max} = y_{\max} - y_{\min}$. Table 3 illustrates Consensus on different PMFs. The issue with using entropy to measure consensus is shown in the 4th ($P(3) = 0.5$ and $P(5) = 0.5$) and 5th row ($P(1) = 0.5$ and $P(7) = 0.5$), where the latter shows lower levels of consensus, but entropy gives the same value ($Entropy = 1.00$).

Whereas the original Consensus statistic is defined over a sample of aggregated scalar responses,

Table 3: Visualization of LLM Consensus (Cns) and Dissension (Dsn) cases. Each histogram depicts a characteristic output distribution on a 1–7 scale.

CNS	ENTROPY	DISTRIBUTION
1.00	0.00	
0.84	1.57	
0.47	2.81	
0.74	1.00	
0.00	1.00	

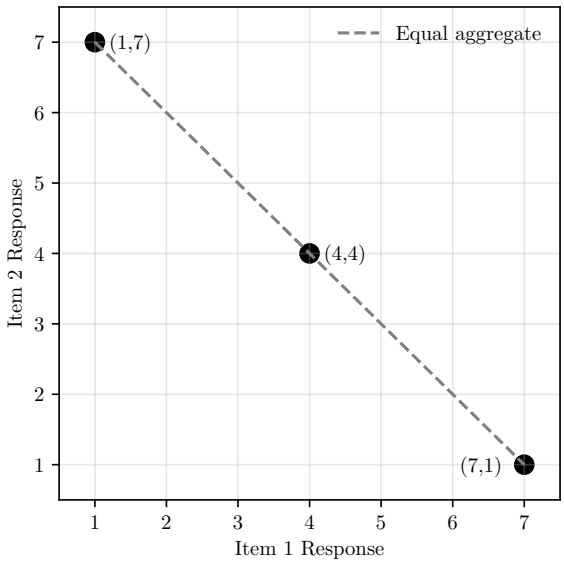


Figure 6: The failure of aggregate Consensus in capturing between-item dissension. Distinct response patterns in a two-item Likert space (e.g., $(4, 4)$ vs. $(1, 7)$) collapse to the same scalar score after aggregation. The original Consensus measure operates on this projection and cannot detect between-item dissension.

our setting differs in two fundamental ways. First, with language models as subjects, we have direct access to the PMF of the data generation process rather than sampled responses. Second, behavioral instruments such as Likert scales typically consist of multiple items, each with varying levels of Consensus and Dissension. Calculating Consensus over aggregated responses disregards the inter-item Dissension.

Standard aggregation procedures, such as summation or averaging, project this space onto a single scalar dimension. Figure 6 illustrates this effect for a two-item, 7-point Likert scale. Response patterns such as $(4, 4)$, $(1, 7)$, and $(7, 1)$ all yield the same aggregate sum, causing the inter-item Dissension to be lost. However, these points occupy very different locations in the original two-dimensional response space, with $(1, 7)$ and $(7, 1)$ representing maximal disagreement between items.

Thus, we derive the Multivariate Consensus (Consensus from now on) across all items on a common ordinal scale. For each item $k \in \{1, \ldots, K\}$, we utilize the mapping ϕ (Eq. 7) to project the renormalized token distribution onto the ordinal scale $\mathcal{Y}$ (Eq. 5). We define the item-level ordinal random variable Y_k with probability mass function

$$P(Y_k = y) = P\left(t = \phi^{-1}(y) \mid x_k, t \in \mathcal{V}_{\text{val}}\right), \tag{10}$$

where x_k is the prompt for item k. This step formally converts the model's token prediction into a numeric behavioral response. We compute the expected response for each item, forming the centroid of the joint distribution

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K), \tag{11}$$

where $\mu_k = \sum_{y \in \mathcal{Y}} y P(Y_k = y)$. The centroid, $\boldsymbol{\mu}$, represents the model's expected position on each Likert item in the K-dimensional space. We define the maximum possible distance between any response vector and the centroid as the diagonal of the hypercube defined by the scale extrema

$$d_{\max} = \sqrt{K\,(y_{\max} - y_{\min})^2}. \tag{12}$$

This normalizes distances to the unit interval and ensures comparability across instruments with different numbers of items. The Multivariate Consensus is defined as the expected logarithmic agreement of the joint distribution with respect to the centroid (Eq. 1).

Intuitively, this measure assigns higher consensus when the probability mass is concentrated near

the centroid across all items, and lower consensus when mass is dispersed or polarized. Unlike entropy-based measures, Consensus is sensitive to ordinal distances. When $K = 1$, the multivariate formulation reduces to the original Consensus (Tastle and Wierman, 2007), with the key distinction that we operate directly on the true PMF rather than on sampled responses. Multivariate Dissension (Dsn) is defined as the complement of Consensus, representing the degree of polarization or dispersion away from the mean.

$$\mathrm{Dsn}(\mathbf{Y}) = 1 - \mathrm{Cns}(\mathbf{Y}) \tag{13}$$

A.4 The Construct Layer

Our goal is to quantify differences in the dependent variable across experimental conditions and models, while fully propagating uncertainty induced by the model's generative process. We formalize inference over latent scale scores using an analytical fixed-effects ANOVA/Hoeffding decomposition. For a specific condition vector $\boldsymbol{\lambda}$, the composite scale score, $S_{\boldsymbol{\lambda}}$, is the aggregation of item-level random variable responses

$$S_{\boldsymbol{\lambda}} = \sum_{k=1}^{K} Y_{k,\boldsymbol{\lambda}}. \tag{14}$$

Let $P_{Y_{k,\boldsymbol{\lambda}}}$ denote the PMF of item k under condition $\boldsymbol{\lambda}$. Given independence (Sec. A.2), the distribution of the composite score is obtained via discrete convolution

$$P_{S_{\boldsymbol{\lambda}}} = P_{Y_{1,\boldsymbol{\lambda}}} \circledast P_{Y_{2,\boldsymbol{\lambda}}} \circledast \cdots \circledast P_{Y_{K,\boldsymbol{\lambda}}}. \tag{15}$$

A.4.1 Factorial Decomposition Over Distributions

We formalize an analytical ANOVA/Hoeffding decomposition of the composite scale score. Our framework constructs a unique family of centered effect distributions $\{E_U\}_{U \subseteq \mathcal{C}}$, representing main effects and interactions, such that the total expected score satisfies the additive identity

$$\mathbb{E}[S_{\boldsymbol{\lambda}}] = \underbrace{\mathbb{E}[S_0]}_{\text{Grand Mean}} + \sum_{c \in \mathcal{C}} \underbrace{\mathbb{E}[E_c(\lambda_c)]}_{\text{Main Effects}} + \sum_{\substack{U \subseteq \mathcal{C} \\ |U| \geq 2}} \underbrace{\mathbb{E}[E_U(\boldsymbol{\lambda}_U)]}_{\text{Interactions}}, \tag{16}$$

while retaining the full probability mass function for every term in the summation.

For each experimental condition $\boldsymbol{\lambda} \in \mathcal{D}$, the composite scale score $S_{\boldsymbol{\lambda}}$ is a discrete random variable with probability mass function $P_{S_{\boldsymbol{\lambda}}}$ (Eq. 14). Rather than reducing $S_{\boldsymbol{\lambda}}$ to a point estimate, we treat the mapping

$$\boldsymbol{\lambda} \mapsto P_{S_{\boldsymbol{\lambda}}} \tag{17}$$

as the primitive object of analysis. This allows all uncertainty induced by the model's generative process to be propagated exactly through subsequent analyses. Given the fully crossed factorial design, all factor combinations are symmetric. Therefore, the design space $\mathcal{D}$ takes on a uniform probability distribution π, assigning equal weight to each experimental condition.

Assumption A.1 (Fully Crossed Fixed-Effects Design)**.** The experimental design space $\mathcal{D} = \prod_{c \in \mathcal{C}} \{1, \ldots, \Lambda_c\}$ is finite and fully crossed. All expectations over $\mathcal{D}$ are taken with respect to the uniform measure.

Proposition A.2 (Linearity of Expectation under Mixture Marginalization)**.** *Let $\mathcal{D}$ be a finite factorial design with probability measure π, and let $P_{\bar{S}} = \mathbb{E}_{\boldsymbol{\lambda} \sim \pi}[P_{S_{\boldsymbol{\lambda}}}]$ be the mixture distribution. Then,*

$$\mathbb{E}[\bar{S}] = \mathbb{E}_{\boldsymbol{\lambda} \sim \pi}[\mathbb{E}[S_{\boldsymbol{\lambda}}]]\,. \tag{18}$$

The proof follows immediately from the linearity of finite sums and the definition of expectation.

Corollary A.3 (Expectation of Factorial Marginal Distributions)**.** *Let $S_U(\boldsymbol{\lambda}_U)$ denote a factorial marginal distribution defined by marginalizing $P_{S_{\boldsymbol{\lambda}}}$ over $\boldsymbol{\lambda}_{-U}$ under the uniform measure. Then*

$$\mathbb{E}[S_U(\boldsymbol{\lambda}_U)] = \mathbb{E}_{\boldsymbol{\lambda}_{-U}}[\mathbb{E}[S_{\boldsymbol{\lambda}}]]\,.$$

The proof follows directly from Prop. A.2 by taking π to be the uniform distribution over $\boldsymbol{\lambda}_{-U}$.

A.4.2 Grand Mixture Distribution

To establish a baseline for factorial effects, we construct the distributional equivalent of the classical ANOVA Grand Mean. We compute this as the Grand Mixture Distribution (P_{S_0}), obtained by averaging the composite score distributions uniformly across the entire design space $\mathcal{D}$

$$P_{S_0} = \mathbb{E}_{\boldsymbol{\lambda} \sim \pi}[P_{S_{\boldsymbol{\lambda}}}] = \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\lambda} \in \mathcal{D}} P_{S_{\boldsymbol{\lambda}}}. \tag{19}$$

The random variable S_0 encapsulates the global background distribution of the construct, encompassing all sources of variation in the experiment.

A.4.3 Marginal Distributions

ANOVA Marginal Mean averages responses over complementary factors to isolate a specific level. We extend it by computing the Marginal Slice Distribution ($P_{M_c(\lambda_c)}$). This is obtained by taking the uniform mixture of composite score distributions over all factors other than c,

$$\begin{aligned} P_{M_c(\lambda_c)} &= \mathbb{E}_{\boldsymbol{\lambda}_{-c}}\left[P_{S_{(\lambda_c,\boldsymbol{\lambda}_{-c})}}\right] \\ &= \frac{1}{|\mathcal{D}_{-c}|}\sum_{\boldsymbol{\lambda}_{-c}\in\mathcal{D}_{-c}} P_{S_{(\lambda_c,\boldsymbol{\lambda}_{-c})}}. \end{aligned} \tag{20}$$

The random variable $M_c(\lambda_c)$ is the predictive distribution of the composite score when factor c is fixed at level λ_c, averaging uniformly over all remaining factors. At the expectation level (Prop. A.2), $\mathbb{E}[M_c(\lambda_c)]$ equals the classical marginal mean.

Corollary A.4 (Grand mixture as an average of marginal slices)**.** *Under Assumption A.1, for any factor c,*

$$P_{S_0} = \frac{1}{\Lambda_c}\sum_{\lambda_c=1}^{\Lambda_c} P_{M_c(\lambda_c)}. \tag{21}$$

Proof. Fix $c \in \mathcal{C}$. The grand mixture distribution is

$$P_{S_0} = \frac{1}{|\mathcal{D}|}\sum_{\boldsymbol{\lambda}\in\mathcal{D}} P_{S_{\boldsymbol{\lambda}}}. \tag{22}$$

Under Assumption A.1, the design space factors as $\mathcal{D} = \{1,\dots,\Lambda_c\} \times \mathcal{D}_{-c}$, where $\mathcal{D}_{-c} := \prod_{d\in\mathcal{C}\setminus\{c\}}\{1,\dots,\Lambda_d\}$. Hence each $\boldsymbol{\lambda} \in \mathcal{D}$ can be written uniquely as $\boldsymbol{\lambda} = (\lambda_c, \boldsymbol{\lambda}_{-c})$ with $\lambda_c \in \{1,\dots,\Lambda_c\}$ and $\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}$. We may rewrite the sum in Eq. (22) as a double sum

$$P_{S_0} = \frac{1}{|\mathcal{D}|}\sum_{\lambda_c=1}^{\Lambda_c}\sum_{\boldsymbol{\lambda}_{-c}\in\mathcal{D}_{-c}} P_{S_{(\lambda_c,\boldsymbol{\lambda}_{-c})}}. \tag{23}$$

Because the design is fully crossed, $|\mathcal{D}| = \Lambda_c\,|\mathcal{D}_{-c}|$. Substituting this into Eq. (23) gives

$$P_{S_0} = \frac{1}{\Lambda_c\,|\mathcal{D}_{-c}|}\sum_{\lambda_c=1}^{\Lambda_c}\sum_{\boldsymbol{\lambda}_{-c}\in\mathcal{D}_{-c}} P_{S_{(\lambda_c,\boldsymbol{\lambda}_{-c})}} \tag{24}$$

$$= \frac{1}{\Lambda_c}\sum_{\lambda_c=1}^{\Lambda_c}\left(\frac{1}{|\mathcal{D}_{-c}|}\sum_{\boldsymbol{\lambda}_{-c}\in\mathcal{D}_{-c}} P_{S_{(\lambda_c,\boldsymbol{\lambda}_{-c})}}\right). \tag{25}$$

The expression in parentheses in Eq. (25) is exactly the marginal slice distribution $P_{M_c(\lambda_c)}$ from Eq. (20). Therefore,

$$P_{S_0} = \frac{1}{\Lambda_c}\sum_{\lambda_c=1}^{\Lambda_c} P_{M_c(\lambda_c)},$$

which proves Eq. (21). □

A.4.4 Expectation-Level Decomposition

Our primary estimands in the Construct Layer are predictive distributions (PMFs) and contrast distribution constructed via mixtures and paired subtraction. Because such contrasts depend on a chosen coupling, they do not, in general, admit an additive decomposition at the level of random variables. However, our distributional framework connects to classical fixed-effects ANOVA at the *expectation* level (Eq. 16). Let $\mu(\boldsymbol{\lambda})$ be the deterministic mean surface over the factorial design

$$\mu(\boldsymbol{\lambda}) := \mathbb{E}[S_{\boldsymbol{\lambda}}]\,, \qquad \boldsymbol{\lambda} \in \mathcal{D}. \tag{26}$$

Here, we state the Hoeffding/ANOVA decomposition of μ on the finite product space $\mathcal{D}$. We provide the unique collection of expectation-level effect components $\{\mu_U\}_{U\subseteq\mathcal{C}}$ that our distributional effect estimands target in expectation.

Theorem A.5 (Expectation-Level ANOVA/Hoeffding Decomposition)**.** *Assume a finite fully crossed design $\mathcal{D}$ endowed with the uniform product measure (Assumption A.1). There exists a unique collection of functions $\{\mu_U\}_{U\subseteq\mathcal{C}}$ such that*

$$\mu(\boldsymbol{\lambda}) = \sum_{U\subseteq\mathcal{C}} \mu_U(\boldsymbol{\lambda}_U), \tag{27}$$

and each component satisfies the identifiability (sum-to-zero) constraints

$$\sum_{\lambda_c=1}^{\Lambda_c} \mu_U(\boldsymbol{\lambda}_U) = 0 \qquad \textit{for all } U \subseteq \mathcal{C} \textit{ and all } c \in U. \tag{28}$$

Proof. We work on the finite product space $\mathcal{D}$ with the uniform product measure implied by Assumption A.1.

For any function $f : \mathcal{D} \to \mathbb{R}$ and any subset $U \subseteq \mathcal{C}$, define the uniform marginalization operator

$$(T_U f)(\boldsymbol{\lambda}_U) := \mathbb{E}_{\boldsymbol{\lambda}_{-U}}[f(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_{-U})] \tag{29}$$

$$= \frac{1}{|\mathcal{D}_{-U}|}\sum_{\boldsymbol{\lambda}_{-U}\in\mathcal{D}_{-U}} f(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_{-U}), \tag{30}$$

where $\mathcal{D}_{-U} := \prod_{c\in\mathcal{C}\setminus U}\{1,\dots,\Lambda_c\}$. By construction, $T_U f$ depends only on $\boldsymbol{\lambda}_U$, and $T_{\mathcal{C}} f = f$.

Define components recursively in increasing order of $|U|$

$$\mu_\emptyset := (T_\emptyset \mu) \qquad \text{(a constant)}, \tag{31}$$

$$\mu_U(\boldsymbol{\lambda}_U) := (T_U\mu)(\boldsymbol{\lambda}_U) - \sum_{V \subsetneq U} \mu_V(\boldsymbol{\lambda}_V), \tag{32}$$

$$U \neq \emptyset.$$

This is well-defined because the right-hand side for μ_U only involves μ_V with strictly smaller index sets $V \subsetneq U$, which have already been defined.

From Eq. (31) we immediately obtain, for every $U \subseteq \mathcal{C}$,

$$(T_U\mu)(\boldsymbol{\lambda}_U) = \sum_{V \subseteq U} \mu_V(\boldsymbol{\lambda}_V). \tag{33}$$

For $U = \emptyset$ this holds by definition. For $U \neq \emptyset$ it is a simple rearrangement of (31). Setting $U = \mathcal{C}$ and using $T_\mathcal{C}\mu = \mu$ yields

$$\mu(\boldsymbol{\lambda}) = \sum_{V \subseteq \mathcal{C}} \mu_V(\boldsymbol{\lambda}_V),$$

which is Eq. (27).

Identifiability. Fix $U \subseteq \mathcal{C}$ and $c \in U$. Uniformly averaging μ_U over λ_c gives

$$\mathbb{E}_{\lambda_c}[\mu_U(\boldsymbol{\lambda}_U)] = \mathbb{E}_{\lambda_c}[(T_U\mu)(\boldsymbol{\lambda}_U)] - \sum_{V \subsetneq U} \mathbb{E}_{\lambda_c}[\mu_V(\boldsymbol{\lambda}_V)]\,. \tag{34}$$

Because the measure is a product measure, averaging $T_U\mu$ over λ_c removes coordinate c from the conditioning,

$$\mathbb{E}_{\lambda_c}[(T_U\mu)(\boldsymbol{\lambda}_U)] = (T_{U\setminus\{c\}}\mu)(\boldsymbol{\lambda}_{U\setminus\{c\}}).$$

For $V \subsetneq U$, if $c \in V$ then $\mathbb{E}_{\lambda_c}[\mu_V] = 0$ by induction on $|V|$. If $c \notin V$ then μ_V does not depend on λ_c and the average leaves it unchanged. Hence,

$$\mathbb{E}_{\lambda_c}[\mu_U(\boldsymbol{\lambda}_U)] = (T_{U\setminus\{c\}}\mu)(\boldsymbol{\lambda}_{U\setminus\{c\}}) - \sum_{V \subseteq U\setminus\{c\}} \mu_V(\boldsymbol{\lambda}_V). \tag{35}$$

The right-hand side is zero by applying Eq. (33) with $U \setminus \{c\}$. Therefore $\mathbb{E}_{\lambda_c}[\mu_U(\boldsymbol{\lambda}_U)] = 0$, which under uniformity is equivalent to Eq. (28).

Uniqueness. Suppose $\{\tilde{\mu}_U\}$ is another collection satisfying Eq. (27) and Eq. (28). Define $h_U := \tilde{\mu}_U - \mu_U$. Then $\sum_U h_U(\boldsymbol{\lambda}_U) = 0$ for all $\boldsymbol{\lambda}$ and each h_U satisfies the same sum-to-zero constraints. Apply T_U to $\sum_W h_W = 0$, where W is any subset of $\mathcal{C}$. If $W \subseteq U$, then $T_U h_W = h_W$. If $W \not\subseteq U$, pick $c \in W \setminus U$. Marginalizing over $-U$ averages over λ_c, and the sum-to-zero constraint implies $T_U h_W = 0$. Thus,

$$0 = T_U\left[\sum_{W \subseteq \mathcal{C}} h_W\right] = \sum_{W \subseteq U} h_W \qquad \text{for all } U \subseteq \mathcal{C}.$$

Taking $U = \emptyset$ gives $h_\emptyset = 0$. Proceed by induction on $|U|$. Assuming $h_V = 0$ for all $V \subsetneq U$, the identity above implies $h_U = 0$. Hence $h_U \equiv 0$ for all U, proving uniqueness. □

Corollary A.6 (Möbius inversion form)**.** *The components in Theorem A.5 admit the explicit inclusion–exclusion form*

$$\mu_U(\boldsymbol{\lambda}_U) = \sum_{V \subseteq U} (-1)^{|U|-|V|}\,(T_V\mu)(\boldsymbol{\lambda}_V), \qquad U \subseteq \mathcal{C}. \tag{36}$$

Proof. Equation (33) states that for each fixed U, $(T_U\mu) = \sum_{V \subseteq U} \mu_V$. This is exactly the zeta-transform relation on the Boolean lattice $(2^\mathcal{C}, \subseteq)$. Applying Möbius inversion (Rota 1964) yields Eq. (36). □

Remark A.7. By Prop. A.2 and the definition of the global marginal slice distribution $M_U(\boldsymbol{\lambda}_U)$ (Eq. (60)),

$$\mathbb{E}[M_U(\boldsymbol{\lambda}_U)] = (T_U\mu)(\boldsymbol{\lambda}_U).$$

Thus, $T_U\mu$ can be interpreted as the classical marginal mean surface, and μ_U in Eq. (36) are the associated ANOVA/Hoeffding effect components targeted by our distributional effect estimands in expectation.

A.4.5 Main Effects Distribution

Classical ANOVA main effects quantify centered deviations of a factor-level marginal mean from the grand mean. In our distributional setting, we seek an effect distribution whose expectation recovers the ANOVA main effect (Eq. 16). There are various choices for computing the difference distribution. One option is to compute the convolution-based subtraction $M_c(\lambda_c) - S_0$. The main problem with this derivation is that it treats the slice distribution of the factor, $M_c(\lambda_c)$, and the grand mixture, S_0, as independent. This is an incorrect assumption because S_0 and $M_c(\lambda_c)$ are clearly correlated ($M_c(\lambda_c)$ is a component of the mixture S_0).

Moreover, convolution of two random variables adds their variance together. This inflates the

variance associated with the main effect of factor c. Specifically, the convolution-based subtraction $M_c(\lambda_c) - S_0$ corresponds to a *between-subjects* contrast (difference of two independent draws), which can yield substantial variability even when the two distributions are *identical*.

We therefore derive the main-effect distributions by minimizing the variance of the paired difference between the specific level and the baseline. This approach isolates the systematic shift attributable to the factor while removing the "noise" common to both distributions. To achieve this, we employ a comonotone coupling (i.e., 1D monotone optimal transport), which maximizes the covariance between paired draws (Villani et al., 2008).

The pairing has two components. We have a blocked construction that respects the factorial design by holding the complementary factors $\boldsymbol{\lambda}_{-c}$ fixed. Within each block, we have a canonical comonotone coupling, which minimizes the expected squared difference between paired draws among all couplings with the same marginals.

Definition A.8 (Paired difference of two univariate discrete distributions)**.** Let X and Y be univariate discrete random variables with finite support and CDFs F_X and F_Y. Let $U \sim \mathrm{Unif}(0,1)$ and define the (left-continuous) generalized inverse CDF

$$F^{-1}(u) := \inf\{z \in \mathbb{R} : F(z) \geq u\}.$$

The comonotone coupling is given by $X^{\uparrow} := F_X^{-1}(U)$ and $Y^{\uparrow} := F_Y^{-1}(U)$. We define the *paired difference* random variable as

$$X \ominus Y \;:=\; X^{\uparrow} - Y^{\uparrow}. \tag{37}$$

This construction preserves marginals ($X^{\uparrow} \overset{d}{=} X$, $Y^{\uparrow} \overset{d}{=} Y$), hence $\mathbb{E}[X \ominus Y] = \mathbb{E}[X] - \mathbb{E}[Y]$.

Let $\mathrm{supp}(X) = \{x_1 < \cdots < x_m\}$ with masses $p_i = P(X = x_i)$ and define $P_i = \sum_{r=1}^{i} p_r$ with $P_0 = 0$. Similarly, let $\mathrm{supp}(Y) = \{y_1 < \cdots < y_n\}$ with masses q_j and $Q_j = \sum_{s=1}^{j} q_s$ with $Q_0 = 0$. The comonotone coupling assigns joint mass

$$w_{ij} = \max\Big\{0,\ \min(P_i, Q_j) - \max(P_{i-1}, Q_{j-1})\Big\}, \tag{38}$$

$$i = 1, \ldots, m,\ j = 1, \ldots, n. \tag{39}$$

The paired-difference PMF is the pushforward under $(x, y) \mapsto x - y$,

$$P_{X \ominus Y}(z) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij}\, \mathbb{1}\{z = x_i - y_j\}. \tag{40}$$

All terms are finite and are computed deterministically from the two input PMFs.

To respect the factorial dependence structure (i.e., S_0 is constructed from the same family $\{S_{\boldsymbol{\lambda}}\}$ as $M_c(\lambda_c)$), we form a grand-mean baseline within each block $\boldsymbol{\lambda}_{-c}$ by averaging only over the levels of factor c

$$P_{S_{0|\boldsymbol{\lambda}_{-c}}} := \frac{1}{\Lambda_c} \sum_{\lambda'_c=1}^{\Lambda_c} P_{S_{(\lambda'_c, \boldsymbol{\lambda}_{-c})}}. \tag{41}$$

Mixing these blocked baselines uniformly over $\boldsymbol{\lambda}_{-c}$ recovers the original grand mixture distribution $\frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c}} P_{S_{0|\boldsymbol{\lambda}_{-c}}} = P_{S_0}$.

The (centered) main-effect distribution for factor c at level λ_c is the uniform mixture of *paired* within-block contrasts

$$E_c(\lambda_c) := \mathbb{E}_{\boldsymbol{\lambda}_{-c}}\left[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \;\ominus\; S_{0|\boldsymbol{\lambda}_{-c}}\right]. \tag{42}$$

At the PMF level this becomes

$$P_{E_c(\lambda_c)} = \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} P_{S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}}}. \tag{43}$$

Each paired-difference PMF in Eq. (43) is computed exactly via the mass-matching formula in Eq. (40). No estimation or approximation is used at any stage.

Corollary A.9 (Expectation-level centering and sum-to-zero)**.** *For each factor $c \in \mathcal{C}$ and level λ_c,*

$$\mathbb{E}[E_c(\lambda_c)] = \mathbb{E}[M_c(\lambda_c)] - \mathbb{E}[S_0]. \tag{44}$$

Moreover, the paired main effects satisfy the standard ANOVA sum-to-zero constraint in expectation

$$\sum_{\lambda_c=1}^{\Lambda_c} \mathbb{E}[E_c(\lambda_c)] = 0. \tag{45}$$

Proof. By definition (Eq. (43)), the main-effect random variable $E_c(\lambda_c)$ is the uniform mixture over blocks $\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}$ of the paired contrast $S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}}$. Equivalently, we can represent the mixture hierarchically. Draw $\boldsymbol{\lambda}_{-c} \sim \mathrm{Unif}(\mathcal{D}_{-c})$, then draw

$$\left(S^{\uparrow}_{(\lambda_c, \boldsymbol{\lambda}_{-c})},\ S^{\uparrow}_{0|\boldsymbol{\lambda}_{-c}}\right)$$

from the comonotone coupling.

Set $E_c(\lambda_c) = S^{\uparrow}_{(\lambda_c, \boldsymbol{\lambda}_{-c})} - S^{\uparrow}_{0|\boldsymbol{\lambda}_{-c}}$. Because $\mathcal{D}_{-c}$ is finite, we may apply iterated expectations.

$$\mathbb{E}[E_c(\lambda_c)] = \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} \mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}}\big] \,. \tag{46}$$

Fix a block $\boldsymbol{\lambda}_{-c}$. By Def. A.8,

$$S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}} = S^{\uparrow}_{(\lambda_c, \boldsymbol{\lambda}_{-c})} - S^{\uparrow}_{0|\boldsymbol{\lambda}_{-c}},$$

where $S^{\uparrow}_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \stackrel{d}{=} S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}$ and $S^{\uparrow}_{0|\boldsymbol{\lambda}_{-c}} \stackrel{d}{=} S_{0|\boldsymbol{\lambda}_{-c}}$. Therefore the expectations match their marginals

$$\mathbb{E}\Big[S^{\uparrow}_{(\lambda_c, \boldsymbol{\lambda}_{-c})}\Big] = \mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}\big] \,, \tag{47}$$

$$\mathbb{E}\Big[S^{\uparrow}_{0|\boldsymbol{\lambda}_{-c}}\Big] = \mathbb{E}\big[S_{0|\boldsymbol{\lambda}_{-c}}\big] \,. \tag{48}$$

Using linearity of expectation, which does *not* require independence,

$$\mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}}\big] = \mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}\big] - \mathbb{E}\big[S_{0|\boldsymbol{\lambda}_{-c}}\big] \,. \tag{49}$$

Eq. (49) holds for *any* coupling with the given marginals. The comonotone coupling is chosen to minimize dispersion, but preserves the mean.

Substitute Eq. (49) into Eq. (46):

$$\begin{aligned}
\mathbb{E}[E_c(\lambda_c)] &= \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} \big(\mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}\big] - \mathbb{E}\big[S_{0|\boldsymbol{\lambda}_{-c}}\big]\big) \\
&= \underbrace{\frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} \mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}\big]}_{(\star)} && (50) \\
&- \underbrace{\frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} \mathbb{E}\big[S_{0|\boldsymbol{\lambda}_{-c}}\big]}_{(\dagger)} \,. && (51)
\end{aligned}$$

For $(\star)$, by the definition of the marginal slice distribution $P_{M_c(\lambda_c)}$ (Eq. (20)) and Prop. A.2,

$$\mathbb{E}[M_c(\lambda_c)] = \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} \mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}\big] \,.$$

For $(\dagger)$, expand the blocked grand mean (Eq. (41)) and apply Theorem A.2,

$$\frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c}} \mathbb{E}\big[S_{0|\boldsymbol{\lambda}_{-c}}\big] = \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c}} \mathbb{E}\left[\frac{1}{\Lambda_c} \sum_{\lambda'_c=1}^{\Lambda_c} S_{(\lambda'_c, \boldsymbol{\lambda}_{-c})}\right] \tag{52}$$

$$= \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c}} \frac{1}{\Lambda_c} \sum_{\lambda'_c=1}^{\Lambda_c} \mathbb{E}\big[S_{(\lambda'_c, \boldsymbol{\lambda}_{-c})}\big] \tag{53}$$

$$= \frac{1}{|\mathcal{D}|} \sum_{\boldsymbol{\lambda} \in \mathcal{D}} \mathbb{E}[S_{\boldsymbol{\lambda}}] = \mathbb{E}[S_0]. \tag{54}$$

We used $|\mathcal{D}| = \Lambda_c \, |\mathcal{D}_{-c}|$ and the definition of the grand mixture S_0 (Eq. (19)).

Substituting these two identifications into Eq. (51) yields

$$\mathbb{E}[E_c(\lambda_c)] = \mathbb{E}[M_c(\lambda_c)] - \mathbb{E}[S_0] \,,$$

which proves Eq. (44).

Sum Eq. (46) over $\lambda_c \in \{1, \ldots, \Lambda_c\}$

$$\sum_{\lambda_c=1}^{\Lambda_c} \mathbb{E}[E_c(\lambda_c)] = \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} \sum_{\lambda_c=1}^{\Lambda_c} \tag{55}$$

$$\mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}}\big] \,. \tag{56}$$

Apply Eq. (49) inside the inner sum

$$\sum_{\lambda_c=1}^{\Lambda_c} \mathbb{E}\big[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})} \ominus S_{0|\boldsymbol{\lambda}_{-c}}\big] = \tag{57}$$

$$\sum_{\lambda_c=1}^{\Lambda_c} \big(\mathbb{E}[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}] - \mathbb{E}[S_{0|\boldsymbol{\lambda}_{-c}}]\big) = \tag{58}$$

$$\sum_{\lambda_c=1}^{\Lambda_c} \mathbb{E}[S_{(\lambda_c, \boldsymbol{\lambda}_{-c})}] - \Lambda_c \, \mathbb{E}[S_{0|\boldsymbol{\lambda}_{-c}}]. \tag{59}$$

But by definition of the blocked grand mean and linearity of expectation,

$$\mathbb{E}[S_{0|\boldsymbol{\lambda}_{-c}}] = \frac{1}{\Lambda_c} \sum_{\lambda'_c=1}^{\Lambda_c} \mathbb{E}[S_{(\lambda'_c, \boldsymbol{\lambda}_{-c})}],$$

so $\Lambda_c \mathbb{E}[S_{0|\boldsymbol{\lambda}_{-c}}] = \sum_{\lambda'_c=1}^{\Lambda_c} \mathbb{E}[S_{(\lambda'_c, \boldsymbol{\lambda}_{-c})}]$. Hence the inner expression is identically zero for every $\boldsymbol{\lambda}_{-c}$, and therefore the outer average is also zero

$$\sum_{\lambda_c=1}^{\Lambda_c} \mathbb{E}[E_c(\lambda_c)] = 0,$$

which proves Eq. (45). □

A.4.6 Interaction Effects Distribution

Whereas main effects quantify additive shifts attributable to individual factors, interaction effects quantify departures from additivity when multiple factors are fixed jointly. In classical fixed-effects ANOVA/Hoeffding decompositions, the U-way interaction component is obtained by Möbius inversion of marginal means on the Boolean lattice of factor subsets.

In our framework, since we are dealing directly with PMFs of the population, we define interaction effects using a paired, design-blocked contrast. We

hold the complementary factors $\boldsymbol{\lambda}_{-U}$ fixed (blocking). Within each block, we couple all terms in the alternating sum via a shared latent quantile variable (comonotone coupling). This yields a paired distribution for the interaction contrast.

For any nonempty subset of factors $U \subseteq \mathcal{C}$ and level vector $\boldsymbol{\lambda}_U$, define the U-way marginal slice distribution as the uniform mixture over all remaining factors

$$P_{M_U(\boldsymbol{\lambda}_U)} = \frac{1}{|\mathcal{D}_{-U}|} \sum_{\boldsymbol{\lambda}_{-U} \in \mathcal{D}_{-U}} P_{S_{(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_{-U})}}. \quad (60)$$

We further set $M_\emptyset := S_0$.

Fix a nonempty subset of factors $U \subseteq \mathcal{C}$ and a realization of the complementary factors $\boldsymbol{\lambda}_{-U} \in \mathcal{D}_{-U}$. For any $V \subseteq U$, define the within-block design space

$$\mathcal{D}_{U \setminus V} := \prod_{c \in U \setminus V} \{1, \ldots, \Lambda_c\}.$$

We derive the blocked marginal slice distribution as the uniform mixture over the remaining factors in $U \setminus V$, holding $\boldsymbol{\lambda}_{-U}$ fixed

$$P_{M_{V|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_V)} \quad (61)$$

$$:= \frac{1}{|\mathcal{D}_{U \setminus V}|} \sum_{\boldsymbol{\lambda}_{U \setminus V} \in \mathcal{D}_{U \setminus V}} P_{S_{(\boldsymbol{\lambda}_V, \boldsymbol{\lambda}_{U \setminus V}, \boldsymbol{\lambda}_{-U})}}. \quad (62)$$

This unifies several cases,

$$M_{U|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_U) \equiv S_{(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_{-U})}, \quad (63)$$

$$M_{\emptyset|\boldsymbol{\lambda}_{-U}} = \frac{1}{|\mathcal{D}_U|} \sum_{\boldsymbol{\lambda}_U \in \mathcal{D}_U} S_{(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_{-U})}, \quad (64)$$

where $M_{\emptyset|\boldsymbol{\lambda}_{-U}}$ is the blocked grand mean (over factors U) within the fixed complement $\boldsymbol{\lambda}_{-U}$.

Averaging blocked slices over $\boldsymbol{\lambda}_{-U}$ recovers the global marginal slices from Eq. (60).

$$P_{M_V(\boldsymbol{\lambda}_V)} = \frac{1}{|\mathcal{D}_{-U}|} \sum_{\boldsymbol{\lambda}_{-U} \in \mathcal{D}_{-U}} P_{M_{V|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_V)}. \quad (65)$$

In particular, taking $V = \emptyset$ yields $\frac{1}{|\mathcal{D}_{-U}|} \sum_{\boldsymbol{\lambda}_{-U}} P_{M_{\emptyset|\boldsymbol{\lambda}_{-U}}} = P_{S_0}$.

To define a distribution for an alternating sum without imposing independence, we use a shared-quantile (comonotone) coupling.

Definition A.10 (Comonotone-coupled linear contrast)**.** Let $\{X_r\}_{r=1}^R$ be univariate discrete random variables with finite support and CDFs $\{F_r\}_{r=1}^R$. Let $U^* \sim \mathrm{Unif}(0,1)$ and define $X_r^\uparrow := F_r^{-1}(U^*)$ using the (left-continuous) generalized inverse from Def. A.8. For coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_R) \in \mathbb{R}^R$, define the comonotone-coupled linear contrast

$$\mathcal{L}^\uparrow(\boldsymbol{\alpha}; X_1, \ldots, X_R) := \sum_{r=1}^R \alpha_r X_r^\uparrow. \quad (66)$$

For $R = 2$ and $\boldsymbol{\alpha} = (1, -1)$, this reduces to the paired difference $X_1 \ominus X_2$ (Def. A.8).

Because each $X_r^\uparrow = F_r^{-1}(U^*)$ is a step function of U^*, the contrast in Eq. (66) is piecewise constant on a finite partition of $[0,1]$. Let

$$\mathcal{B} := \{0,1\} \cup \bigcup_{r=1}^R \{F_r(x) \,:\, x \in \mathrm{supp}(X_r)\},$$

and let $0 = u_0 < u_1 < \cdots < u_L = 1$ be the sorted distinct elements of $\mathcal{B}$. For $\ell = 1, \ldots, L$, define

$$z_\ell := \sum_{r=1}^R \alpha_r F_r^{-1}(u_\ell).$$

Then the PMF of $Z := \mathcal{L}^\uparrow(\boldsymbol{\alpha}; X_1, \ldots, X_R)$ is

$$P_Z(z) = \sum_{\ell=1}^L (u_\ell - u_{\ell-1})\, \mathbb{1}\{z = z_\ell\}, \quad (67)$$

aggregating masses for identical z_ℓ values. This computation is exact and deterministic.

Fix $U \subseteq \mathcal{C}$ with $|U| \geq 2$. For a given level vector $\boldsymbol{\lambda}_U$ and block $\boldsymbol{\lambda}_{-U}$, define the within-block interaction-effect random variable as the paired Möbius contrast over blocked slices

$$H_{U|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_U) := \sum_{V \subseteq U} (-1)^{|U|-|V|} \left(M_{V|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_V)\right)^\uparrow, \quad (68)$$

where all terms are coupled comonotonically via the same U^* in Def. A.10. The PMF of $H_{U|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_U)$ is computed exactly using Eq. (67) with $R = 2^{|U|}$ and coefficients $\alpha_V = (-1)^{|U|-|V|}$.

Finally, we define the global U-way interaction effect distribution by mixing uniformly over complementary factors:

$$P_{E_U(\boldsymbol{\lambda}_U)} = \frac{1}{|\mathcal{D}_{-U}|} \sum_{\boldsymbol{\lambda}_{-U} \in \mathcal{D}_{-U}} P_{H_{U|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_U)}. \quad (69)$$

Two-way interactions. For $U = \{a, b\}$, within a fixed block $\boldsymbol{\lambda}_{-ab}$ the paired two-way interaction is

$$H_{\{a,b\}|\boldsymbol{\lambda}_{-ab}}(\lambda_a, \lambda_b) = S^{\uparrow}_{(\lambda_a,\lambda_b,\boldsymbol{\lambda}_{-ab})} \tag{70}$$
$$- M^{\uparrow}_{a|\boldsymbol{\lambda}_{-ab}}(\lambda_a) \tag{71}$$
$$- M^{\uparrow}_{b|\boldsymbol{\lambda}_{-ab}}(\lambda_b) \tag{72}$$
$$+ M^{\uparrow}_{\emptyset|\boldsymbol{\lambda}_{-ab}}, \tag{73}$$

and $P_{E_{\{a,b\}}(\lambda_a,\lambda_b)}$ is the uniform mixture of these within-block PMFs over $\boldsymbol{\lambda}_{-ab}$.

Corollary A.11 (Expectation-level ANOVA consistency and centering)**.** *Let $\mu(\boldsymbol{\lambda}) = \mathbb{E}[S_{\boldsymbol{\lambda}}]$ and let $\{\mu_U\}_{U\subseteq\mathcal{C}}$ be the unique Hoeffding/ANOVA components from Theorem A.5. For any $U \subseteq \mathcal{C}$ with $|U| \geq 2$ and any $\boldsymbol{\lambda}_U$,*

$$\mathbb{E}[E_U(\boldsymbol{\lambda}_U)] = \mu_U(\boldsymbol{\lambda}_U). \tag{74}$$

In particular, the interaction effects are centered in expectation. For any $c \in U$,

$$\sum_{\lambda_c=1}^{\Lambda_c} \mathbb{E}[E_U(\boldsymbol{\lambda}_U)] = 0. \tag{75}$$

Proof. Fix $\boldsymbol{\lambda}_{-U}$. By construction, the comonotone coupling in Eq. (68) preserves the marginal distribution of each term $M_{V|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_V)$, hence it does not alter expectations. Therefore, by linearity of expectation,

$$\mathbb{E}\big[H_{U|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_U)\big] = \sum_{V\subseteq U} (-1)^{|U|-|V|} \mathbb{E}\big[M_{V|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_V)\big]\,. \tag{76}$$

Next, expand the blocked slice definition (Eq. (61)) and apply linearity of expectation to obtain

$$\mathbb{E}\big[M_{V|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_V)\big] \tag{77}$$
$$= \frac{1}{|\mathcal{D}_{U\setminus V}|} \sum_{\boldsymbol{\lambda}_{U\setminus V}\in\mathcal{D}_{U\setminus V}} \mu(\boldsymbol{\lambda}_V, \boldsymbol{\lambda}_{U\setminus V}, \boldsymbol{\lambda}_{-U}). \tag{78}$$

Now average Eq. (76) uniformly over $\boldsymbol{\lambda}_{-U}$ (Eq. (69)). Because the design is a finite product space, averaging over $\boldsymbol{\lambda}_{-U}$ and over $\boldsymbol{\lambda}_{U\setminus V}$ is equivalent to averaging over all complementary coordinates $\boldsymbol{\lambda}_{-V}$. Thus

$$\mathbb{E}[E_U(\boldsymbol{\lambda}_U)] \tag{79}$$
$$= \frac{1}{|\mathcal{D}_{-U}|} \sum_{\boldsymbol{\lambda}_{-U}} \mathbb{E}\big[H_{U|\boldsymbol{\lambda}_{-U}}(\boldsymbol{\lambda}_U)\big] \tag{80}$$
$$= \sum_{V\subseteq U} (-1)^{|U|-|V|} \underbrace{\frac{1}{|\mathcal{D}_{-V}|} \sum_{\boldsymbol{\lambda}_{-V}} \mu(\boldsymbol{\lambda}_V, \boldsymbol{\lambda}_{-V})}_{=(T_V\mu)(\boldsymbol{\lambda}_V)}\,. \tag{81}$$

The final expression is exactly the Möbius inversion / Hoeffding component $\mu_U(\boldsymbol{\lambda}_U)$ (Rota, 1964), equivalent to the recursion in Theorem A.5. This proves Eq. (74). The centering property Eq. (75) then follows from the identifiability constraints in Theorem A.5. □

A.4.7 Hypothesis Testing

We distinguish our hypothesis evaluation from null-hypothesis significance testing (NHST). Rather than accept–reject decisions based on p-values, which presuppose sampling variability from finite data, we evaluate directional scientific claims using exact predictive effect distributions computed analytically from the model's next-token probabilities.

In our framework, the primitive objects of inference are centered effect random variables, namely paired main effects $E_c(\lambda_c)$ (Eq. (42)) and paired interaction effects $E_U(\boldsymbol{\lambda}_U)$ (Eq. (69)). Their PMFs are computed deterministically via finite mixtures and paired comonotone contrast operations. Accordingly, there is no sampling error. Uncertainty in these effect distributions reflects intrinsic stochasticity of the model's predictive behavior under the stimulus (i.e., the aleatoric uncertainty).

We summarize each centered effect distribution E using mean shift ($\mathbb{E}[E]$), predictive dispersion ($\mathrm{SD}(E)$), discrete Probability of Direction ($\mathrm{dPD}(E)$), and a standardized signal-to-noise ratio ($\mathrm{SNR}(E)$) based on the SD of the paired effect distribution.

A.4.8 Discrete Probability of Direction

To evaluate directional hypotheses, we adopt a discrete formulation of the Probability of Direction (PD). In Bayesian inference, PD is typically defined for continuous posteriors, where $\mathbb{P}(E = 0) = 0$ (Makowski et al., 2019). Here, predictive effect distributions are discrete and may place non-zero mass on exact equality. We therefore treat $\{E = 0\}$ as a distinct behavioral outcome.

Let E denote a centered effect random variable, either $E_c(\lambda_c)$ or $E_U(\boldsymbol{\lambda}_U)$. The discrete Probability of Direction (dPD) is

$$\mathrm{dPD}(E) = \max\big(\mathbb{P}(E > 0), \mathbb{P}(E < 0)\big), \quad (82)$$

dPD quantifies the extent to which the effect exhibits consistent directionality, irrespective of sign.

A.4.9 Effect Size (Mean Shift)

We define effect size as the expectation of the centered effect distribution

$$\mathbb{E}[E] = \sum_z z\, P_E(z), \quad (83)$$

where the sum ranges over the finite support of E. $\mathbb{E}[E]$ is the magnitude of the effect E.

A.4.10 Predictive Dispersion (SD)

Intrinsic variability of the centered effect is captured by its predictive standard deviation

$$\mathrm{SD}(E) = \sqrt{\sum_z \big(z - \mathbb{E}[E]\big)^2 P_E(z)}. \quad (84)$$

Because our effect distributions are constructed via paired contrasts, $\mathrm{SD}(E)$ reflects residual predictive heterogeneity in the effect itself, rather than dispersion induced by taking differences of independent draws.

A.4.11 Signal-to-Noise Ratio

We summarize standardized effect magnitude using a signal-to-noise ratio computed directly on the paired effect distribution

$$\mathrm{SNR}(E) = \frac{|\mathbb{E}[E]|}{\mathrm{SD}(E)}. \quad (85)$$

When $\mathrm{SD}(E) = 0$, the effect is deterministic under the paired contrast. In this case $\mathrm{SNR}(E)$ is treated as $+\infty$ if $\mathbb{E}[E] \neq 0$ and 0 if $\mathbb{E}[E] = 0$.

The signal-to-noise ratio SNR is closely related to standardized effect sizes such as Cohen's d (Cohen, 2013), which quantifies mean deviations relative to within-group variability. However, in our setting, variability reflects intrinsic stochasticity of the language model rather than sampling error. We therefore use conventional heuristics (e.g., $\approx 0.2, 0.5, 0.8$ for small, medium, large) only as descriptive reference points, not as decision thresholds.

A.4.12 Interpretation

Inferential conclusions are drawn from the joint configuration of $\mathrm{dPD}(E)$, $\mathbb{E}[E]$, $\mathrm{SD}(E)$, and $\mathrm{SNR}(E)$. High $\mathrm{dPD}(E)$ indicates strong directional regularity, while large $\mathrm{SNR}(E)$ indicates that the expected shift dominates the intrinsic predictive dispersion of the paired effect distribution. Discrepancies between these quantities distinguish regimes such as stable but weak effects, large but ambivalent effects, or low-signal effects with high dispersion (Table 4).

A.4.13 Trend Effects

Some experimental factors admit an inherent ordering and a meaningful numeric scale (e.g., model size in billions of parameters or a discretized continuous manipulation intensity). To quantify the marginal *per-unit* sensitivity of the composite construct to such a factor without imposing a parametric functional form, we compute paired finite difference distributions across adjacent factor levels and aggregate them into an average trend distribution.

Let $c \in \mathcal{C}$ denote an ordered factor with $L = \Lambda_c$ levels that admit numeric values

$$\lambda_1^{(c)} < \lambda_2^{(c)} < \cdots < \lambda_L^{(c)} \qquad (\lambda_\ell^{(c)} \in \mathbb{R}).$$

Adjacent spacings may be uneven. Define increments

$$\delta_\ell^{(c)} = \lambda_{\ell+1}^{(c)} - \lambda_\ell^{(c)} \;>\; 0, \qquad \ell = 1, \ldots, L-1.$$

To respect the factorial design, we compare adjacent levels *within each block* of the remaining factors $\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}$. For each block and each adjacent pair of ordered levels, define the paired local difference random variable

$$D_{\ell|\boldsymbol{\lambda}_{-c}}^{(c)} := S_{(\lambda_{\ell+1}^{(c)}, \boldsymbol{\lambda}_{-c})} \ominus S_{(\lambda_\ell^{(c)}, \boldsymbol{\lambda}_{-c})}, \quad (86)$$

$$\ell = 1, \ldots, L-1, \quad (87)$$

where $\ominus$ denotes the paired comonotone difference from Def. A.8. Each PMF $P_{D_{\ell|\boldsymbol{\lambda}_{-c}}^{(c)}}$ is computed exactly via the discrete mass-matching formula in Eq. (40).

We then marginalize over blocks by a uniform mixture

$$P_{D_\ell^{(c)}} = \frac{1}{|\mathcal{D}_{-c}|} \sum_{\boldsymbol{\lambda}_{-c} \in \mathcal{D}_{-c}} P_{D_{\ell|\boldsymbol{\lambda}_{-c}}^{(c)}}. \quad (88)$$

Intuitively, $D_\ell^{(c)}$ is the predictive distribution of the change in the composite score induced by moving from $\lambda_\ell^{(c)}$ to $\lambda_{\ell+1}^{(c)}$, averaging uniformly over

Case	$P(E)$	$\mathbb{E}[E]$	SD	SNR	dPD
A	$\{-25{:}0.9,\ -20{:}0.1\}$	-24.5	1.5	16.3	1.00
B	$\{1{:}0.99,\ 100{:}0.01\}$	11.0	99.0	0.11	1.00
C	$\{100{:}0.51,\ -1{:}0.49\}$	50.5	49.5	1.02	0.51
D	$\{5{:}0.34,\ -5{:}0.33,\ 0{:}0.33\}$	0.05	4.1	0.01	0.34

Table 4: Illustrative predictive effect distributions exhibiting distinct behavioral regimes. Cases A and B have identical directional consistency ($\mathrm{dPD} = 1$) but sharply different standardized magnitudes (SNR). Case C shows a large expected effect despite weak directional consistency. Case D exhibits neither directional regularity nor an appreciable standardized signal.

all other experimental factors while preserving a paired (within-block) comparison.

Define the per-unit local slope random variable as the rescaled difference

$$\Delta_\ell^{(c)} = \frac{D_\ell^{(c)}}{\delta_\ell^{(c)}}. \tag{89}$$

Equivalently, $P_{\Delta_\ell^{(c)}}$ is the pushforward of $P_{D_\ell^{(c)}}$ under the map $z \mapsto z/\delta_\ell^{(c)}$; its support is rescaled by $1/\delta_\ell^{(c)}$ while probability masses are preserved.

To aggregate unevenly spaced local slopes into a single per-unit trend effect, we define the *per-unit trend* random variable γ_c as the length-weighted mixture of scaled local slopes:

$$P_{\gamma_c} = \sum_{\ell=1}^{L-1} w_\ell^{(c)}\, P_{\Delta_\ell^{(c)}}, \tag{90}$$

$$w_\ell^{(c)} = \frac{\delta_\ell^{(c)}}{\sum_{m=1}^{L-1} \delta_m^{(c)}} = \frac{\delta_\ell^{(c)}}{\lambda_L^{(c)} - \lambda_1^{(c)}} \quad . \tag{91}$$

Dispersion or multimodality in P_{γ_c} indicates heterogeneous local behavior (e.g., non-monotonic or regime-dependent trends) across adjacent factor-level transitions.

Proposition A.12 (Expectation-level secant-slope identity)**.** *The expected per-unit trend equals the secant slope of the ordered-factor marginal mean across the observed range*

$$\mathbb{E}[\gamma_c] = \frac{\mathbb{E}[M_c(\lambda_L^{(c)})] - \mathbb{E}[M_c(\lambda_1^{(c)})]}{\lambda_L^{(c)} - \lambda_1^{(c)}}. \tag{92}$$

Proof. Fix $\ell \in \{1, \ldots, L-1\}$ and a block $\boldsymbol{\lambda}_{-c}$. By Def. A.8, the paired difference preserves the mean

$$\mathbb{E}\Big[D^{(c)}_{\ell|\boldsymbol{\lambda}_{-c}}\Big] = \mathbb{E}\Big[S_{(\lambda_{\ell+1}^{(c)}, \boldsymbol{\lambda}_{-c})}\Big] - \mathbb{E}\Big[S_{(\lambda_{\ell}^{(c)}, \boldsymbol{\lambda}_{-c})}\Big].$$

Averaging uniformly over $\boldsymbol{\lambda}_{-c}$ and using linearity of expectation gives

$$\mathbb{E}[D_\ell^{(c)}] = \mathbb{E}[M_c(\lambda_{\ell+1}^{(c)})] - \mathbb{E}[M_c(\lambda_\ell^{(c)})],$$

because $M_c(\lambda_\ell^{(c)})$ is the uniform mixture of $S_{(\lambda_\ell^{(c)}, \boldsymbol{\lambda}_{-c})}$ over $\boldsymbol{\lambda}_{-c}$ (Eq. (20)), and expectations commute with finite mixtures (Theorem A.2).

By definition of $\Delta_\ell^{(c)}$,

$$\mathbb{E}[\Delta_\ell^{(c)}] = \frac{\mathbb{E}[D_\ell^{(c)}]}{\delta_\ell^{(c)}}.$$

Taking expectations in the mixture definition (90) yields

$$\mathbb{E}[\gamma_c] = \sum_{\ell=1}^{L-1} w_\ell^{(c)}\, \mathbb{E}[\Delta_\ell^{(c)}] \tag{93}$$

$$= \frac{1}{\sum_{m=1}^{L-1} \delta_m^{(c)}} \sum_{\ell=1}^{L-1} \tag{94}$$

$$\Big(\mathbb{E}[M_c(\lambda_{\ell+1}^{(c)})] - \mathbb{E}[M_c(\lambda_\ell^{(c)})]\Big), \tag{95}$$

which telescopes to $(\mathbb{E}[M_c(\lambda_L^{(c)})] - \mathbb{E}[M_c(\lambda_1^{(c)})])/(\lambda_L^{(c)} - \lambda_1^{(c)})$, proving Eq. (92). □

Directional interpretation. The full predictive distribution P_{γ_c} supports directional summaries (e.g., $\mathbb{P}(\gamma_c > 0)$, $\mathbb{P}(\gamma_c < 0)$) and can be summarized using the same quantities defined for other predictive contrasts (e.g., dPD, SD, and SNR based on $\mathrm{SD}(\gamma_c)$).

A.5 Additional Analysis

Effect of Model Size. While our primary factorial design treats language models as discrete categorical subjects, model capacity is inherently continuous. Prior work has shown that scaling language models can systematically alter social and

behavioral tendencies, including the amplification of biased responses (Wadi and Fredette, 2025). We therefore investigate whether ethnocentric tendencies, as measured by the CETSCALE, exhibit a systematic trend as a function of model size. Specifically, we ask whether increasing the number of parameters induces a consistent directional change in the latent ethnocentrism score. To address this question, we conduct a trend-effects analysis (derived in App. A.4.13).

We form four within-family size series, each evaluated under the same TARGET manipulation and CETSCALE instrument as in the main experiment. AYA (8B, 32B), GEMMA 3 (1B, 4B, 12B, 27B), MINISTRAL (3B, 8B, 14B), and QWEN3 (0.6B, 4B, 8B, 14B, 32B, Next-80B). [4] The trend effect of model size on ethnocentrism (Table 8) is large ($\mathbb{E} = 1.65$, $\mathrm{SNR} = 5.87$) and directionally consistent ($\mathrm{dPD} > 0.99$) for AYA model family. This indicates steep linear scaling in ethnocentric tendencies. For every additional billion parameters, the model's expected ethnocentrism score increases by 1.65 points. Over the range evaluated (8B to 32B), this effectively shifts a model from the lowest observed ethnocentrism to the highest, purely by scaling. For other models, size does not have a meaningful effect (low SNR and dPD) on ethnocentrism.

A.6 Prompt template

Figure 7 shows how we form the prompt template and cross Target Country with each model.

A.7 Additional results

Model	Country	Cns	Entropy
Aya 32B	Canada	0.941	4.945
	China	0.939	5.150
	France	0.937	5.010
	USA	0.924	7.315
Gemma 3 27B	Canada	0.959	3.565
	China	0.941	4.886
	France	0.940	4.944
	USA	0.948	4.533
Llama 3.3 70B	Canada	0.924	5.620
	China	0.894	8.579
	France	0.882	5.800
	USA	0.947	3.929
Ministral 14B	Canada	0.669	35.519
	China	0.646	37.521
	France	0.668	36.444
	USA	0.660	35.759
Qwen3 Next 80B	Canada	0.929	6.861
	China	0.886	9.008
	France	0.886	8.381
	USA	0.935	5.950

Table 5: Consensus and Entropy of Models and Target Countries

[4] Llama 3.3 (70B) does not form a multi-point series and is excluded from trend estimation.

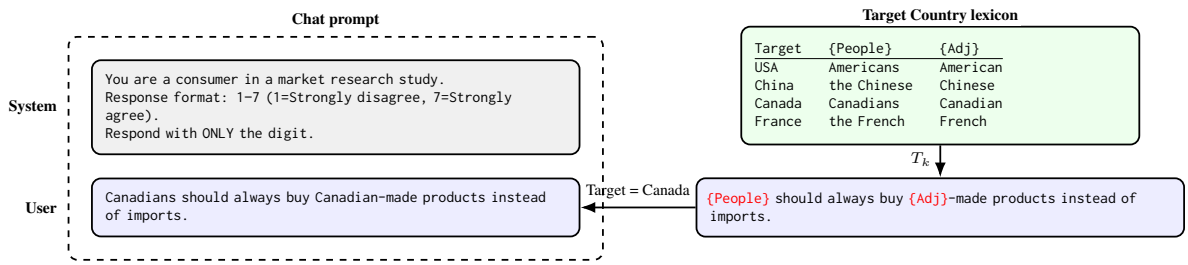


Figure 7: Prompt setup: a fixed system prompt enforces the response schema, while Target Country (sample shown for Canada) deterministically instantiates item templates via a lexicon. Example shown is T_1 from the CETSCALE; each CETSCALE item has its own template.

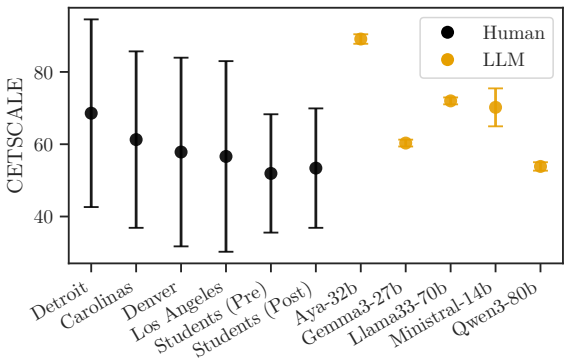


Figure 8: Comparison of CETSCALE scores for the USA target condition between human populations reported in Shimp and Sharma (1987) and LLMs evaluated in this study. Points denote expected values, and error bars indicate standard deviation. All scores are based on the 17-item CETSCALE using a 7-point Likert scale. Notably, several LLMs exhibit ethnocentrism levels comparable to or exceeding those observed in highly ethnocentric human samples, while displaying substantially lower intrinsic dispersion.

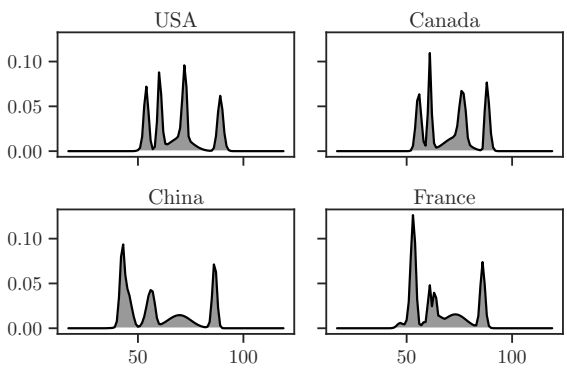


Figure 9: Main Effect of Target Country

Table 6: Main effect and interaction table for the factorial model.

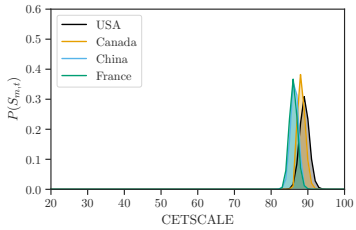
PARAMETER	$\mathbb{E}$	SD	SNR	dPD
$\mu_\emptyset$	66.25	14.21		
AYA-32B	21.19	13.01	1.63	0.99
GEMMA 3-27B	-11.99	13.15	0.91	0.77
LLAMA 3.3-70B	0.31	12.79	0.02	0.55
MINISTRAL-14B	5.11	9.59	0.53	0.62
QWEN3 NEXT-80B	-14.63	12.17	1.20	0.93
USA	2.84	5.43	0.52	0.55
CANADA	4.62	5.72	0.81	0.90
CHINA	-6.46	6.19	1.04	0.95
FRANCE	-1.00	4.72	0.21	0.54
AYA-32B × USA	-1.18	3.03	0.39	0.55
AYA-32B × CANADA	-3.77	2.88	1.31	0.94
AYA-32B × CHINA	5.35	4.53	1.18	0.77
AYA-32B × FRANCE	-0.39	2.89	0.14	0.42
GEMMA 3-27B × USA	3.21	5.34	0.60	0.63
GEMMA 3-27B × CANADA	2.14	5.63	0.38	0.47
GEMMA 3-27B × CHINA	-5.40	8.63	0.63	0.61
GEMMA 3-27B × FRANCE	0.05	5.34	0.01	0.51
LLAMA 3.3-70B × USA	2.59	5.86	0.44	0.58
LLAMA 3.3-70B × CANADA	5.20	5.49	0.95	0.77
LLAMA 3.3-70B × CHINA	-4.29	9.12	0.47	0.58
LLAMA 3.3-70B × FRANCE	-3.50	5.55	0.63	0.67
MINISTRAL-14B × USA	-4.00	3.10	1.29	0.88
MINISTRAL-14B × CANADA	-2.99	3.16	0.95	0.73
MINISTRAL-14B × CHINA	4.54	4.11	1.11	0.72
MINISTRAL-14B × FRANCE	2.45	2.98	0.82	0.87
QWEN3 NEXT-80B × USA	-0.61	2.83	0.22	0.58
QWEN3 NEXT-80B × CANADA	-0.57	2.87	0.20	0.46
QWEN3 NEXT-80B × CHINA	-0.20	5.11	0.04	0.46
QWEN3 NEXT-80B × FRANCE	1.39	2.81	0.49	0.53

Table 7: Pairwise marginal contrasts between Target Countries.

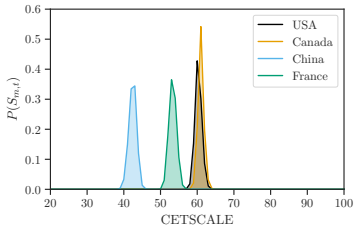
CONTRAST	$\mathbb{E}$	SD	SNR	dPD
USA - CANADA	-1.78	1.83	0.97	0.74
USA - CHINA	9.30	6.93	1.34	0.95
USA - FRANCE	3.84	4.40	0.87	0.80
CANADA - CHINA	11.08	7.60	1.46	1.00
CANADA - FRANCE	5.62	5.05	1.11	0.84
CHINA - FRANCE	-5.47	3.79	1.44	0.80

Table 8: Trend effects for model size (billions of parameters).

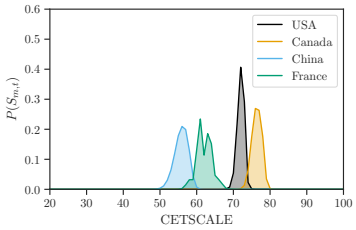
MODEL FAMILY	$\mathbb{E}$	SD	SNR	dPD
AYA	1.65	0.28	5.87	1.00
GEMMA	1.12	7.21	0.16	0.70
MINISTRAL	0.87	1.12	0.78	0.56
QWEN3	-0.00	1.79	0.00	0.89



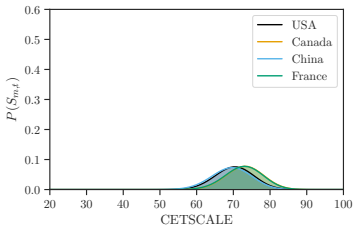
(a) Aya Expanse 32B



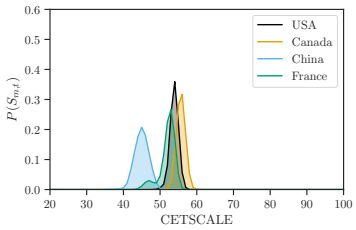
(b) Gemma 3 27B



(c) Llama 3.3 70B



(d) Ministral 14B



(e) Qwen3 Next 80B

Figure 10: Distribution of CETSCALE for Model and Target Country.