Do LLMs Give Consistent Opinions? Evaluating Response Reliability Under Varying Likert-Scale Formulations in Survey-Style MCQA

Anonymous ACL submission

Abstract

Survey methodology requires that equivalent question formulations elicit stable responses from the same respondent, regardless of how the Likert-type answer scale is presented. As LLMs are increasingly deployed as synthetic survey respondents, a critical but underexplored question arises: do they satisfy this basic *parallel-forms reliability* standard when answer scales vary in wording, cardinality, polarity, or option order? We extend the OpinionQA into two stress-test benchmarks, 1,235 subjective questions each paired with five reworded scale variants and six positional permutations, evaluated across four prompt-label formats. Using a four-metric consistency framework, we benchmark eight open-weight models from four families in both Base and Instruction-Tuned configurations. Three findings emerge. First, Base models fail basic reliability requirements: they exhibit severe mechanical instability driven by primacy bias, label sensitivity, and formatting failures, including full polarity inversions for the same question. Second, instruction tuning acts as a reliability filter, reducing response variance by approximately 50% and enabling models to treat semantically equivalent scales as equivalent. Third, for instruction-tuned models under randomized option ordering, removing explicit labels paradoxically *increases* consistency by eliminating symbol-binding conflicts. These findings establish that survey reliability is an engineered property of LLMs requiring both alignment training and deliberate prompt design, with concrete implications for synthetic survey pipelines. [1]

1 Introduction

The reliability of survey measurement instruments is a cornerstone of empirical social science. Classical psychometric theory formalizes this through the concept of *parallel-forms reliability*: equivalent versions of a survey question, differing only in surface-level scale properties, should produce comparable responses from the same respondent (Lord et al., 1968; Cronbach, 1951). As LLMs are increasingly deployed as synthetic survey respondents (Aher et al., 2023; Santurkar et al., 2023), this standard becomes directly testable. A reliable LLM respondent should express the same underlying opinion regardless of whether the scale uses four or five points, whether agreement is worded as "Strongly" or "Fully", or whether the positive end of the scale appears first or last. Figure 1 shows an example of LLM in responding to a question with option scale different setups.

Figure 1: Example of LLMs in responding to a question with different scales. Are they consistent?

Evidence on LLM format sensitivity suggests this requirement is far from guaranteed. Models exhibit systematic positional biases (Zheng et al., 2024; Wang et al., 2024a), fail to robustly bind answer content to option labels (Xue et al., 2024), and produce volatile outputs under minor prompt perturbations (Ngweta et al., 2025; Lunardi et al., 2025). Several studies document that LLM survey responses shift substantially with scale redesign (Rupprecht et al., 2025; Khan et al., 2025; Lee et al., 2025). However, these studies typically examine a single model family, do not manipulate scale design

[1] Data and code are available at: https://anonymous.4open.science/r/SurveyScaleQA-C875, which will be made public upon paper publication.

as systematically as survey researchers would, and rarely isolate the role of alignment training as a driver of reliability.

This paper addresses these gaps through a controlled, multi-factor evaluation grounded in survey methodology. We vary three dimensions of Likert-scale design: (1) semantic wording and cardinality, (2) option order, and (3) prompt-label format. We evaluate all combinations across eight models from four families, comparing Base and Instruction-Tuned variants.

Contributions. This paper makes the following contributions:

- We release an extended OpinionQA benchmark with systematic Likert-scale variations (6,175 Reworded and 7,410 Shuffled question-scale pairs, each across four prompt formats), designed as a reusable consistency evaluation resource.
- We introduce a four-metric consistency framework that disentangles average drift (APD, MSD), worst-case divergence (MPD), and cross-format structural stability (Pearson r), grounded in the psychometric concept of parallel-forms reliability.
- We provide the first large-scale comparison of Base vs. Instruction-Tuned models across Likert scale design dimensions, establishing that alignment is the primary reliability determinant and reduces response variance by ≈50%.
- We identify and explain the *Label-Free Consistency Paradox*: under randomized option ordering, removing labels increases consistency for instruction-tuned models by bypassing symbol-binding errors, revealing an underappreciated interaction between prompt format and option-order perturbation.

2 Background and Related Work

Survey reliability and Likert-scale design. Classical Test Theory decomposes an observed survey response into a true latent score and a measurement error component (Lord et al., 1968). A key source of measurement error in Likert-scale surveys is scale design itself: the number of response options, the directionality (unipolar vs. bipolar), and verbal labels attached to scale points have all been shown to shift human responses independently of the underlying attitude (Menold, 2021; Höhne et al., 2022). Human respondents also exhibit well-documented response styles, including acquiescence bias (Couch and Keniston, 1960), satisficing (Krosnick, 1991), and order effects (Krosnick and Alwin, 1987). Parallel-forms reliability addresses this challenge by requiring that two equivalent instrument versions, differing only in surface properties, yield correlated results (Cronbach, 1951). This is precisely the standard we apply to LLMs: a model queried with semantically equivalent scales should express the same underlying opinion, and violations indicate that observed responses reflect properties of the measurement instrument rather than stable model-internal judgments.

LLM format sensitivity and symbol binding. Accumulating evidence shows that LLM responses in multiple-choice settings are highly sensitive to presentation format. Models exhibit primacy and recency biases, with GPT-3.5's accuracy on MMLU dropping by over 6 percentage points when the correct answer is consistently placed last (Zheng et al., 2024). First-token probabilities and text-output responses frequently disagree for the same prompt, with mismatch rates exceeding 60% for instruction-tuned models (Wang et al., 2024b), and minor formatting changes can cause accuracy fluctuations exceeding 76% (Ngweta et al., 2025). Xue et al. (2024) trace these failures to inadequate *symbol binding*, namely the inability to reliably associate option content with its assigned label, a failure mode that persists through fine-tuning. Molfese et al. (2025) further show that evaluation procedures can confound reasoning failures with formatting failures, underscoring the need to evaluate response consistency separately from accuracy.

LLMs as synthetic survey respondents. When LLMs act as virtual survey respondents, format sensitivity becomes a validity threat. Rupprecht et al. (2025) systematically perturb World Values Survey items and find consistent human-like biases, including strong recency effects, in LLM responses. Khan et al. (2025) show that apparent cultural alignment signals are largely methodological artifacts of prompt variation rather than stable model traits. Tjuatja et al. (2024) find that LLMs generally fail to replicate human response biases in controlled survey settings, and Lee et al. (2025) demonstrate that even high-performing models are not consistent evaluators across rating scales. Santurkar et al. (2023) found that base models tend toward liberal-leaning opinions and are difficult to steer toward specific demographic views, a finding whose im-

plications for opinion stability this paper directly addresses. An extended literature survey is provided in Appendix A.

3 Methodology

3.1 Dataset Construction and Variation Engine

Source Data. We build on the OpinionQA corpus (Santurkar et al., 2023), derived from the Pew American Trends Panel (Pew Research Center, 2024). After manually filtering questions requiring specific respondent personas or responses that cannot be mapped to Likert-type scales, we retain $N = 1{,}235$ highly subjective questions. Each question is classified into one of 15 semantic *content dimensions* (e.g., Agreement, Importance, Frequency, Better/Worse), which determine which family of Likert scales is appropriate for generating variations. The full distribution is in Appendix B.

Reworded Dataset. For each question, we design five semantically equivalent answer-option sets that vary along three survey-design dimensions: (1) **cardinality** (4, 5, or 6 response points), (2) **phrasing** (synonymous verbal labels, e.g., "Strongly" vs. "Fully" vs. "Completely"), and (3) **polarity** (bipolar scales from negative to positive, vs. unipolar scales from zero to high). This yields a Reworded Dataset of $6{,}175$ question-scale pairs ($1{,}235 \times 5$), where the question stem is fixed and only the answer scale varies.

Shuffled Dataset. To isolate positional bias, we select one canonical scale per question and generate five random permutations of option order using fixed seeds (2, 7, 10, 67, 101), combined with the original order. This yields a Shuffled Dataset of $7{,}410$ pairs ($1{,}235 \times 6$), where option wording and cardinality are held constant while only option position varies.

Response Normalization. To enable cross-scale comparison, all valid substantive responses are mapped to a unified latent interval $I \in [-1, 1]$, where $+1.0$ represents the most positive extreme (e.g., "Strongly Agree") and -1.0 the most negative. Options are mapped equidistantly: a 4-point scale maps to $\{-1.0, -0.33, 0.33, 1.0\}$, a 5-point scale to $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$, and so on. This normalization allows distance-based consistency metrics to be computed across scales of different cardinalities.

Illustrative Example. Table 1 traces the full pipeline for one OpinionQA question. The question stem is held constant throughout. Three reworded variants illustrate how cardinality, phrasing, and polarity are varied: V1 and V2 are both 4-point bipolar scales with synonymous labels, while V3 extends to five points, adding a neutral midpoint. All three map semantically equivalent top options to $+1.0$, so a consistent model should select $+1.0$ across all variants regardless of the specific wording. The final row shows a shuffled permutation of V1: here, the option previously in position 1 ("Strongly agree") now occupies position 3. A model relying on positional cues would select option 1 in both cases, yielding $+1.0$ in V1 but -0.33 in S1, and an APD of 1.33 for this single question.

3.2 Model Configurations

We evaluate four open-weight model families, including Llama 3.1 (Grattafiori et al., 2024), Mistral (Jiang et al., 2023), Qwen 2.5 (Qwen et al., 2025), Gemma 2 (Team et al., 2024), each in both a pretrained Base and an Instruction-Tuned (IT) variant. All models range from 7B to 9B parameters, balancing computational feasibility with broad representativeness across pre-training recipes and tokenization strategies. Comparing Base and IT variants within each family isolates the effect of alignment training while holding architecture and scale constant.

All inference uses greedy decoding (`do_sample=False`) to eliminate sampling noise, ensuring that any observed response variation is attributable to prompt design rather than stochastic generation. Details are shown in Appendix B.

3.3 Prompting Strategies

We evaluate four zero-shot prompt formats as a cross-cutting experimental factor, designed to test robustness to structural label choices that vary across real survey instruments:

- **P1 – Numeric:** Options labeled "1), 2), 3), . . . "
- **P2 – Alphabetic:** Options labeled "A), B), C), . . . "
- **P3 – Unlabeled:** Options listed as plain text lines, with no labels.
- **P4 – Extended:** Numeric labels augmented with two non-substantive options: "Don't Know" and "Refused", mimicking real survey instruments that include refusal codes.

Variant	Options (in presentation order)	Normalized values
Question: "How much do you agree that the government should do more to protect the environment?"		
V1 (4-pt, bipolar)	Strongly agree / Somewhat agree / Somewhat disagree / Strongly disagree	$+1.0$ / $+0.33$ / -0.33 / -1.0
V2 (4-pt, bipolar)	Fully agree / Partially agree / Partially disagree / Fully disagree	$+1.0$ / $+0.33$ / -0.33 / -1.0
V3 (5-pt, bipolar)	Completely agree / Mostly agree / Neither / Mostly disagree / Completely disagree	$+1.0$ / $+0.5$ / 0.0 / -0.5 / -1.0
S1 (V1, shuffled)	Somewhat disagree / Strongly disagree / **Strongly agree** / Somewhat agree	-0.33 / -1.0 / $+1.0$ / $+0.33$

Table 1: Variation Engine example. V1 and V2 use different labels for the same 4-point bipolar structure; V3 adds a neutral midpoint. All three map the top option to $+1.0$. S1 shuffles V1: a primacy-biased model selecting option 1 receives -0.33 instead of $+1.0$, yielding APD $= 1.33$ for this question.

IT models receive all instructions via a single `user` turn in the model's chat template, with no separate system prompt (to standardize across model families). Base models receive a raw text-completion prompt ending with a forcing suffix (e.g., "The best answer is option number:"). Full prompt templates are given in Appendix C.

3.4 Consistency Metrics

Rather than measuring accuracy against a ground truth, which is undefined for subjective opinions, we quantify *self-consistency* at the question level. This mirrors the psychometric concept of parallel-forms reliability: a model expressing a stable underlying opinion should produce responses that cluster tightly regardless of which scale formulation is used. For a set of valid normalized responses $V = \{v_1, \ldots, v_n\}$ ($n = 5$ for Reworded, $n = 6$ for Shuffled), we compute four metrics:

Average Pairwise Distance (APD). The mean absolute difference across all unique response pairs. APD $= 0$ indicates perfect consistency; higher values indicate semantic drift across scale variants.

$$\mathrm{APD} = \frac{2}{n(n-1)} \sum_{i<j} |v_i - v_j| \qquad (1)$$

Maximum Pairwise Distance (MPD). The worst-case divergence, $\max_{i<j} |v_i - v_j|$. An MPD value of 2.0 represents a complete polarity inversion, i.e., a shift from the most positive response extreme to the most negative for the same question.

Mean Standard Deviation (MSD). The mean of per-question standard deviations across all questions in the dataset. MSD captures average response dispersion and is more sensitive to general instability than to individual extreme cases.

Pearson Correlation (r). Computed between model response vectors across prompt-format pairs (e.g., P1 vs. P2). A high r indicates that the model preserves the relative semantic ranking of questions even when the absolute response format changes, reflecting structural stability of the model's internal opinion ordering.

Responses failing strict output parsing and non-substantive P4 selections ("Don't Know", "Refused") are excluded. A question is included only if all n variations yielded valid substantive responses, to avoid artificially deflating variance through missing data.

4 Results

4.1 Global Response Patterns: Mechanical Artifacts vs. Semantic Stability

Before computing consistency metrics, we examine the raw response distributions to characterize the systematic biases present in each model class. Figure 2 shows the percentage of responses in each option category for Base models on the Reworded Dataset.

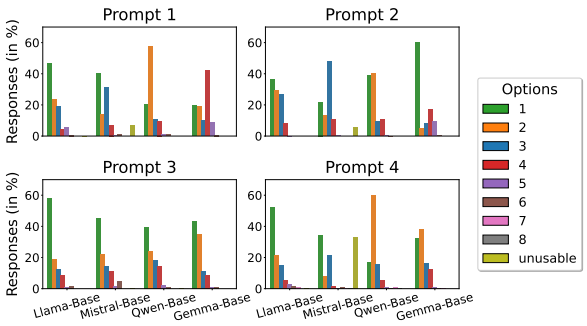

Figure 2: Response category distributions for Base models on the Reworded Dataset. The four subplots within each model panel correspond to the four prompt formats (P1–P4). Valid option indices range from 1 to 6 (up to 8 in P4). Llama-3.1-8B shows consistent primacy bias; Gemma-2-9B shows radical label sensitivity (strong preference for Option 4 in P1, shifts to Option 1 in P2); Mistral-7B-v0.3 accumulates high rates of unusable outputs in P4; Qwen2.5-7B shows moderate and relatively stable preference patterns.

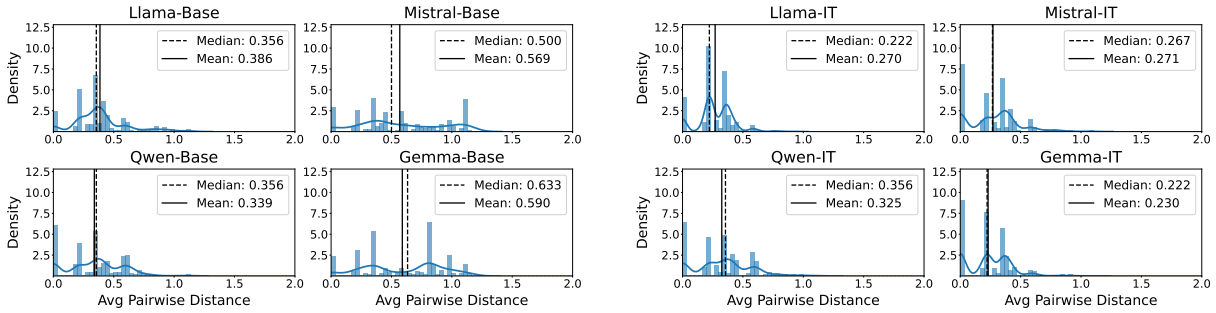

Figure 3: APD distributions on the Shuffled Dataset under P1. **Left**: Base models show highly diffuse distributions, reflecting continuous semantic drift across option permutations. **Right**: Instruction tuning compresses distributions sharply toward 0. The improvement is most dramatic for Gemma-2-9B, which transforms from the most volatile Base model to the most stable IT model.

The figure reveals four qualitatively distinct failure modes in Base models, each reflecting a different source of measurement error from a survey methodology perspective:

Primacy bias (Llama-3.1-8B). This model defaults to Option 1 in 35–60% of all responses, depending on the prompt format, regardless of question content. This mirrors the primacy effects documented in human survey research (Krosnick and Alwin, 1987), but in an extreme, mechanically consistent form.

Label sensitivity (Gemma-2-9B). This model strongly favors Option 4 (>40%) under numeric labels (P1), but shifts dramatically to Option 1 (>60%) under alphabetic labels (P2). The preference is not driven by semantic content but by how option labels interact with the model's learned probability distribution over tokens.

Formatting failures (Mistral-7B-v0.3). This model produces unusable responses at elevated rates in P1 and P2 (≈8%) and catastrophically in P4 (≈35%), where the unusable category rivals the most selected valid option. This reflects a mismatch between the model's text-completion training and the structured output format required by survey questions.

Aggregate distributional instability. A reliable survey respondent should express a similar aggregate opinion profile regardless of whether options are presented in their standard order or shuffled. In contrast, Gemma-2-9B Base shows a near-complete inversion of its aggregate opinion distribution between the Reworded (fixed order) and Shuffled conditions: under the Reworded Prompt 1, density is concentrated at the negative extreme (-1.0); under the Shuffled condition, density spikes at the positive extreme ($+1.0$). This polarity flip, driven purely by option reordering, constitutes a fundamental reliability failure under the parallel-forms standard.

In contrast, IT models show near-perfect format adherence (unusable rates <0.1%), maintain stable response profiles across semantic rewordings, and produce aggregate opinion distributions that overlap substantially between Reworded and Shuffled conditions, demonstrating that alignment successfully grounds outputs in semantic content rather than presentation artifacts.

4.2 Intra-Prompt Consistency: Volatility Analysis

We now quantify how much a model's response to a given question changes across scale variations within a fixed prompt format.

Average and worst-case divergence. Figure 3 shows APD distributions for all models on the Shuffled Dataset under P1. Base model distributions are markedly diffuse: Mistral-7B-v0.3 spreads density nearly uniformly from APD 0.0 to 1.25 (mean ≈0.57), and Gemma-2-9B has a distribution shifted further right, with a median of ≈0.63 and a distinct density spike near 0.8. For comparison, an APD of 0.67 corresponds to a model that, on average, moves one full step on a 4-point scale between scale variants. Llama-3.1-8B and Qwen2.5-7B show somewhat better clustering in the 0.2–0.5 range, but retain substantial high-distance tails.

The worst-case metric (MPD) adds further evidence of Base model fragility. Across all Base models on the Reworded Dataset, a non-trivial fraction of questions receive MPD $= 2.0$, representing a complete polarity inversion where the model gives diametrically opposite responses (e.g., "Strongly Agree" and "Strongly Disagree") to the

same question under different scale formulations. For Gemma-2-9B, a bimodal MPD distribution is observed, with density spikes at both 0.0 (perfect stability for some questions) and 2.0 (full inversion for others), indicating that the model's stability is entirely content-dependent with no general consistency principle.

The reliability filter effect of instruction tuning. Figure 3 (right) and Table 2 together quantify the stabilization produced by instruction tuning. IT model APD distributions concentrate sharply near 0 across all four families. Table 2 shows that across all models and prompt formats, IT models achieve MSD values roughly half those of their Base counterparts (e.g., Llama-3.1-8B drops from 0.357 to 0.207 under P1; Gemma-2-9B drops from 0.344 to 0.174). Gemma-2-9B-IT achieves the lowest MSD of all models (≈0.17 on average), a particularly striking result given that Gemma-2-9B (Base) is among the most volatile, confirming that stability is a consequence of alignment rather than an inherent architectural property.

Model	P1	P2	P3	P4
Base Models				
Llama-3.1-8B	0.357	0.315	0.411	0.340
Mistral-7B-v0.3	0.302	0.254	0.453	0.294
Qwen2.5-7B	0.254	0.304	0.415	0.224
Gemma-2-9B	0.344	0.325	0.421	0.229
IT Models				
Llama-3.1-8B-IT	0.207	0.229	0.248	0.203
Mistral-7B-IT	0.221	0.227	0.287	0.206
Qwen2.5-7B-IT	0.232	0.249	0.244	0.241
Gemma-2-9B-IT	0.174	0.172	0.178	0.164

Table 2: Mean Standard Deviation (MSD) on the Reworded Dataset. P1 (Numeric), P2 (Alphabetic), P3 (Unlabeled), P4 (Extended). IT models achieve approximately half the MSD of Base models across all conditions.

Prompt format and label effects. An important secondary finding from Table 2 concerns the Unlabeled format (P3). For Base models, removing option labels causes MSD to spike substantially (Mistral-7B-v0.3: 0.453; all others: $\geq$0.41), confirming that labels function as structural anchors that support mapping semantic intent to a discrete output token. Without these anchors, Base models lose the ability to produce consistent selections. IT models are significantly less affected (P3 MSD: 0.178–0.287), reflecting their stronger semantic grounding.

4.3 The Label-Free Consistency Paradox

Comparing results across the Reworded and Shuffled datasets reveals a counter-intuitive interaction between prompt format and option-order perturbation that we term the *Label-Free Consistency Paradox*.

On the Reworded Dataset, where option order is fixed, Prompt P3 produces the highest MSD for both model classes, as discussed above. However, on the Shuffled Dataset, where option order is randomized, P3 yields the *lowest* MSD for all four IT model families.

We interpret this paradox as follows. Under labeled prompts (P1, P2) with shuffled options, IT models face a conflict between two competing signals: a learned label preference (e.g., a tendency to output "A" or "1" across many training examples) and a semantic preference for the content that represents their actual opinion (e.g., "Agree"). When the option expressing their semantic preference is not assigned the preferred label, this conflict manifests as instability. Removing labels entirely (P3) eliminates the competing signal and forces the model to attend directly to semantic content, yielding higher consistency.

This account is consistent with the Multiple Choice Symbol Binding (MCSB) framework of Xue et al. (2024): symbol binding failures arise precisely at the interface between label and content, and bypassing the binding step (by removing labels) eliminates the failure mode for semantically capable models. The paradox does not appear in Base models, because they lack the semantic grounding to exploit label removal, and P3 remains their worst condition throughout.

4.4 Inter-Prompt Structural Robustness

We now assess whether models preserve a consistent logical stance when the structural format of the prompt itself changes, independent of option order and wording variations.

Categorical consistency. Figure 4 shows the distribution of categorical agreement levels across all four prompt formats on the Reworded Dataset. Perfect agreement ("4 Same") means the model selects the same response category for all four prompt formats for a given question.

Base models exhibit clear logical fragmentation. Gemma-2-9B is the most fragile: "2 Same" is its dominant category (≈45%), and it records the highest rate of total disagreement ("4 Different":

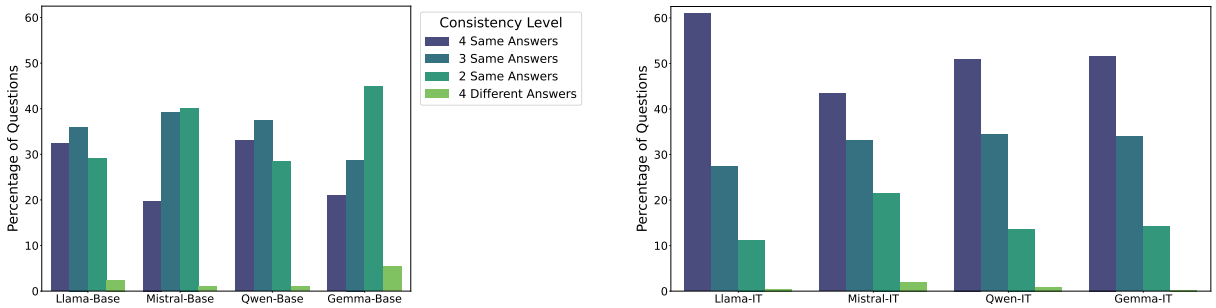

Figure 4: Categorical consistency across four prompt formats on the Reworded Dataset. **Left**: Base models exhibit logical fragmentation, with perfect agreement rarely dominant. **Right**: Instruction tuning shifts the dominant outcome to perfect agreement across all four families. The "3 Same" residual in IT models largely reflects structurally valid selections of "Don't Know" or "Refused" in P4.

≈5%). Mistral-7B-v0.3 achieves perfect categorical agreement in fewer than 20% of questions. Llama-3.1-8B and Qwen2.5-7B perform somewhat better, with "4 Same" around 32%, but "3 Same" or fewer is still the majority outcome.

Instruction tuning induces a decisive shift toward structural stability. "4 Same" becomes the dominant outcome for all four IT families, ranging from 43% (Mistral-7B-IT) to 62% (Llama-3.1-8B-IT). The "4 Different" category drops to <1% for all IT models. The residual "3 Same" category (≈27–34%) primarily reflects structurally valid behavior: when a model appropriately selects "Don't Know" or "Refused" in P4, this prevents a four-way match even though the model's substantive response in P1–P3 is perfectly consistent.

Correlational stability. Table 3 presents Pearson's r between response vectors across all prompt-format pairs on the Shuffled Dataset, measuring whether the model's relative opinion ranking across questions is preserved under format changes.

IT models maintain high structural stability ($r \approx$ 0.70–0.92) regardless of format pair. Base models' response vectors are substantially scrambled when format changes ($r \approx$ 0.15–0.85). Two consistent structural patterns emerge. First, P1–P4 yields the highest correlations across all models: for IT models, r ranges from 0.796 to 0.921, and even for Base models Qwen2.5-7B achieves 0.849. This confirms that models recognize the structural similarity between numeric and extended-numeric formats and maintain opinions when refusal options are added. Second, prompt pairs involving P3 consistently yield the lowest correlations, often 0.1–0.2 points below labeled pairs, confirming that unlabeled prompts require a qualitatively different processing pathway.

Model	1–2	1–3	1–4	2–3	2–4	3–4
Base Models						
Llama-3.1-8B	0.545	0.354	0.594	0.441	0.511	0.413
Mistral-7B-v0.3	0.309	0.183	0.496	0.191	0.269	0.244
Qwen2.5-7B	0.608	0.428	0.849	0.431	0.572	0.400
Gemma-2-9B	0.468	0.155	0.499	0.312	0.557	0.378
IT Models						
Llama-3.1-8B-IT	0.797	0.636	0.796	0.686	0.778	0.644
Mistral-7B-IT	0.730	0.709	0.901	0.627	0.728	0.718
Qwen2.5-7B-IT	0.886	0.783	0.921	0.777	0.882	0.780
Gemma-2-9B-IT	0.849	0.733	0.856	0.784	0.824	0.717

Table 3: Pearson's r between prompt-format pairs on the Shuffled Dataset. IT models maintain high logical ranking stability ($r \approx$ 0.70–0.92). Base models suffer severe degradation, especially in pairs involving P3. The P1–P4 pair is consistently the most stable, reflecting structural similarity.

5 Discussion

5.1 Psychometric Interpretation: Reliability as an Engineered Property

From the perspective of Classical Test Theory, our results can be summarized concisely: Base models fail to meet the parallel-forms reliability standard, while IT models approach it but do not fully achieve it. The measurement error introduced by scale design is substantial for Base models (MSD 0.22–0.55), meaning that a large fraction of the observed variation in Base model survey responses reflects properties of the measurement instrument rather than the model's underlying stance. IT models reduce this error substantially, but residual scale sensitivity (MSD 0.16–0.30) means that even aligned models do not yet meet the reliability standards typically required for survey instruments in social science research.

Importantly, this distinction is not primarily a matter of model architecture or parameter count.

Within each of the four families, the IT variant dramatically outperforms the Base variant on every consistency metric. The effect of alignment training (SFT + RLHF) is a fundamental behavioral shift: Base models optimize for next-token prediction and are therefore sensitive to surface-level statistical regularities in their pre-training data, such as the frequency of particular option labels or the distribution of answers appearing at particular positions (Zhao et al., 2021). IT models are additionally trained to follow semantic instructions, which reduces sensitivity to these surface features and grounds outputs more directly in the semantic content of the question (Ouyang et al., 2022).

5.2 Model-Specific Behavior Profiles

Beyond the general Base/IT contrast, our results reveal instructive model-specific patterns that illuminate sources of instability. **Gemma-2-9B (Base)** represents the most dramatic case of format-driven instability. Under Prompt 1 (Numeric), responses concentrate at the negative extreme (-1.0) on the Reworded Dataset, but flip to the positive extreme ($+1.0$) when shuffled. This polarity inversion suggests a strong position bias when numerical labels no longer reliably identify options. After instruction tuning, Gemma-2-9B-IT becomes the most consistent model, confirming that alignment training effectively resolves this sensitivity.

Qwen2.5-7B (Base) exhibits atypically high intrinsic stability, with inter-prompt correlations ($r = 0.849$) approaching IT-level performance. Its aggregate distributions remain similar between Reworded and Shuffled conditions, likely reflecting high-density structured MCQA reasoning in pre-training. **Mistral-7B-v0.3 (Base)** is the least structurally consistent Base model, with inter-prompt correlations frequently below 0.25. Its IT variant shows improvement but remains the least consistent IT model, suggesting alignment training is less effective at resolving underlying symbol-binding tensions.

5.3 Symbol Binding, Labels, and Prompt Design

The Label-Free Consistency Paradox (Section 4.3) reveals a subtle but practically important interaction in how IT models process labeled vs. unlabeled prompts. Labels serve a dual function. For Base models, they are essential structural anchors that map the abstract output space to the semantic space of answer options. Removing them collapses this mapping. For IT models, however, labels introduce a secondary signal that can conflict with semantic preferences, particularly when option order is randomized. This tension between label-driven and content-driven selection is consistent with MCSB failures documented by Xue et al. (2024), but here we observe that the failure can be avoided by removing the binding target entirely, rather than by improving the binding mechanism.

This finding also connects to alignment artifacts. RLHF training may produce a bias toward outputting specific label tokens (e.g., "A" or "1") independently of the content they are paired with, as a learned response style rather than genuine reasoning. When option order is randomized, this stylistic preference conflicts with the model's semantic understanding, producing inconsistency. Removing labels eliminates the conflict by eliminating the trigger for the learned style. While IT models' stability may thus partially reflect genuine semantic reasoning and partially reflect a form of learned compliance (Sharma et al., 2024), our results show that the net effect of alignment is strongly positive for survey reliability in all conditions.

6 Conclusion

We presented a psychometrically grounded evaluation of LLM response consistency under systematic Likert-scale design variations, motivated by the need to determine whether LLMs can meet the parallel-forms reliability standard required of survey instruments. Our main finding is unambiguous: reliability in LLM-based survey research is not an inherent property of the models but an engineered one, produced by instruction tuning and sensitive to prompt label format. Base models fail basic reliability requirements and should not be used as synthetic survey respondents. IT models reduce variance by approximately 50% and show meaningful structural robustness, but residual scale sensitivity persists. The Label-Free Consistency Paradox reveals a non-obvious interaction between prompt format and option shuffling that has practical implications for survey design. Future work should extend this evaluation to question-stem paraphrasing, larger model scales, non-English survey corpora, and, crucially, should benchmark LLM reliability against human parallel-forms reliability coefficients to establish whether current IT models meet the standards applied to human respondents.

Limitations

Our evaluation is subject to several limitations. First, we rely on strict output parsing, which may conflate formatting failures with genuine reasoning failures (Molfese et al., 2025). Penalizing a model for selecting "Don't Know" in P4, for instance, ignores that expressing uncertainty can reflect appropriate calibration rather than inconsistency (Kadavath et al., 2022). Second, Mistral-7B-v0.3's high unusable rate in P4 means that the consistency metrics for that model are computed over a filtered, potentially non-representative subset of questions. Third, all inference uses greedy decoding; robustness under stochastic sampling and higher temperatures remains unexplored. Fourth, the study is limited to English-language questions and four 7–9B parameter models; consistency patterns may differ for larger models or non-English corpora. Fifth, and most critically, we lack human baselines answering the same reworded and shuffled variants, making it impossible to determine whether current LLM consistency levels are subhuman, comparable to human reliability, or in some cases superior ("hyper-accurate") but brittle in qualitatively different ways (Aher et al., 2023). Future work should also extend perturbations to question-stem paraphrasing (Elazar et al., 2021) and to a broader range of scale cardinalities and polarities (Menold, 2021).

Ethical Considerations

We use the English OpinionQA dataset (Santurkar et al., 2023), which is freely available for research. Our extended dataset is released exclusively for non-commercial research purposes and respects all intellectual property and privacy requirements of the original dataset. Experimental compute was performed on institutional GPU infrastructure. No personal information was collected. The authors acknowledge the use of ChatGPT solely for grammar correction and text coherence, and for assistance with coding.

References

Gati V Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 337–371. PMLR.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.

Arthur Couch and Kenneth Keniston. 1960. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology*, 60(2):151–174.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Jennifer Dykema, Nora Cate Schaeffer, Dana Garbarski, Nadia Assad, and Steven Blixt. 2022. Towards a reconsideration of the use of agree-disagree questions in measuring subjective evaluations. *Research in Social and Administrative Pharmacy*, 18(2):2335–2344.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jan Karem Höhne, Dagmar Krebs, and Steffen M Kühnel. 2022. Measuring income (in)equality: Comparing survey questions with unipolar and bipolar scales in a probability-based online panel. *Social Science Computer Review*, 40(5):108–123.

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The

unreliability of evaluating cultural alignment in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, page 2151–2165, New York, NY, USA. Association for Computing Machinery.

Jon A Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Jon A. Krosnick and Duane F. Alwin. 1987. An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51(2):201–219.

Noah Lee, Jiwoo Hong, and James Thorne. 2025. Evaluating the consistency of LLM evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10650–10659, Abu Dhabi, UAE. Association for Computational Linguistics.

Frederic M. Lord, Melvin R. Novick, and Allan Birnbaum. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley.

Riccardo Lunardi, Vincenzo Della Mea, Stefano Mizzaro, and Kevin Roitero. 2025. On robustness and reliability of benchmark-based evaluation of llms. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pages 4603–4610. IOS Press.

Natalja Menold. 2021. Response bias and reliability in verbal agreement rating scales: Does polarity and verbalization of the middle category matter? *Social Science Computer Review*, 39(1):130–147.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. 2025. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18477–18494, Vienna, Austria. Association for Computational Linguistics.

Sérgio Silva Mucciaccia, Thiago Meireles Paixão, Filipe Wall Mutz, Claudine Santos Badue, Alberto Ferreira de Souza, and Thiago Oliveira-Santos. 2025. Automatic multiple-choice question generation and evaluation systems based on LLM: A study case with university resolutions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2246–2260, Abu Dhabi, UAE. Association for Computational Linguistics.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs robustness to changes in prompt format styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 529–537, Albuquerque, USA. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 27730–27744, Red Hook, NY, USA. Curran Associates Inc.

Pew Research Center. 2024. The American Trends Panel (ATP). Dataset. Accessed via OpinionQA.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Jens Rupprecht, Georg Ahnert, and Markus Strohmaier. 2025. Prompt perturbations reveal human-like biases in large language model survey responses. *Preprint*, arXiv:2507.07188.

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.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pages 110–144.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings*

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. 2024. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4344, Bangkok, Thailand. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pages 19426–19454.

A Related Work (Extended)

Alignment and MCQA Evaluation. Base language models are trained as probabilistic text completers. Alignment via Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF) transforms them into instruction-following agents (Wei et al., 2022; Ouyang et al., 2022). MCQA is widely used to evaluate these capabilities efficiently (Balepur et al., 2025; Mucciaccia et al., 2025), but evaluation methods can misrepresent true model capabilities in systematic ways. Molfese et al. (2025) demonstrate that a seemingly minor choice between parsing strategies (regex string matching vs. log-probability extraction) can produce divergent accuracy estimates for the same model, since enforcing a strict answer format drives measured performance down while free-form generation inflates scores but requires brittle heuristics. Balepur et al. (2025) further note that MCQs appear in 71% of GPT-4 evaluation tasks despite the fact that over 90% of real user queries are free-form, raising concerns that benchmark scores reflect test-taking ability more than general capability.

Positional Bias and Prompt Brittleness. Evidence for position-driven LLM biases spans multiple model families and benchmarks, with documented accuracy swings that are large enough to materially change a model's ranking. Zheng et al. (2024) show that GPT-3.5's score on the MMLU benchmark drops from 67.2% to 60.9% when the correct answer is consistently placed last rather than first. Wang et al. (2024a) document a systematic primacy bias in GPT-4 when it acts as a pairwise evaluator, and also observe recency bias in models fine-tuned on instruction data, where the model gravitates toward the final option due to autoregressive generation priors. The bias extends beyond ordering: Ngweta et al. (2025) report accuracy fluctuations exceeding 76% between semantically equivalent prompts that differ only in surface formatting (punctuation, spacing, bracket style), and show that training on a mixture of prompt formats provides only partial mitigation. The practical implication is that a model's apparent "preference" on any single-shot MCQA item may be as much a function of presentation as of underlying judgment.

Survey Methodology and LLM Personas. Classical psychometrics establishes that reliability requires minimizing measurement error arising from instrument design rather than from the respondent's true attitude (Lord et al., 1968). Parallel-forms reliability assesses consistency across equivalent instrument versions (Cronbach, 1951), and human response styles such as satisficing (Krosnick, 1991), acquiescence (Couch and Keniston, 1960), and order effects (Krosnick and Alwin, 1987) are well-documented sources of such error. Emerging evidence shows that LLMs do not merely repli-

cate these patterns but can amplify them substantially. Rupprecht et al. (2025) apply systematic perturbations to World Values Survey items and find that some models favor the last-presented option approximately 20 times more frequently than the first under certain conditions, a recency effect far stronger than observed in human respondents. They also document strong sensitivity to semantic paraphrasing, with no tested model proving robust across all perturbation types. Khan et al. (2025) similarly report that LLM responses to cultural alignment surveys swing dramatically with minor prompt variations, suggesting that measured "opinions" often reflect methodological artifacts rather than stable model traits. Notably, alignment training does not eliminate these biases: Tjuatja et al. (2024) find that both base and instruction-tuned models display amplified acquiescence and order effects relative to human baselines.

OpinionQA and Survey Simulation. The OpinionQA corpus (Santurkar et al., 2023) provides 1,498 opinion-based MCQA questions from the Pew American Trends Panel (Pew Research Center, 2024), covering topics from political ideology and social issues to consumer preferences and behavioral patterns. Originally designed to assess whether LLM "views" align with specific demographic groups, it found that base models tend to exhibit liberal-leaning tendencies on issues such as climate change and gun control. However, the original work focuses on the direction of model opinions rather than their structural stability across different scale formulations. We repurpose the corpus as a stress-test for inter-scale reliability, filtering out items that presuppose a specific respondent background (e.g., location-specific or retrospective questions) and standardizing all items into Likert-administrable forms. This adaptation yields 1,235 validated subjective prompts suited for systematic scale variation without requiring the model to adopt a specific demographic persona.

B Implementation Details

B.1 Data Preparation and Filtering

Starting from the 1,498 OpinionQA questions, a two-stage manual review reduced the dataset to 1,235 items suitable for scale-variation experiments. In the first stage, questions were excluded if they implicitly assumed a specific respondent background (e.g., "While growing up, how often did you . . . "; "How much . . . would you say is in your neighborhood"), because such items would require the model to adopt a demographic persona rather than express a generalized opinion. In the second stage, any remaining question whose options were structured as substantive statements rather than scalar degrees (e.g., "Which of the following statements comes closer to your view?" with two distinct policy positions) was reformulated into a Likert-administrable question. This involved converting the original stem into a statement and prefacing it with "How much do you agree or disagree with the following statement?", then treating each original option as a separate item. This adaptation preserved the substantive content of the original survey while enforcing a uniform Likert-scale structure.

Content dimension classification was performed with an automated Python script combining keyword matching and Regular Expressions. Question stems were matched against 15 dimension-specific keyword lists (e.g., stems containing *agree/disagree/support/oppose* were assigned to the "Agreement" dimension; those containing *how likely* or *probability* to "Likelihood"; those with comparative terms like *better/worse* to "Better/Worse"). The resulting distribution is shown in Table 4. Each question type was then paired with a set of five semantically equivalent but differently worded answer option sets. The Agree-Disagree (AD) format was adopted as the primary scale structure because it allows a uniform scale design across diverse question topics, isolating the effect of scale parameters (cardinality, phrasing, polarity) from question semantics, in line with the goal of studying instrument effects rather than item content effects. We acknowledge that AD scales impose an additional cognitive mapping step compared to item-specific scales (Dykema et al., 2022), but this limitation is secondary to the experimental goal of systematic scale variation.

For the Shuffled Dataset, five random permutations of the original option order were generated for each question using fixed random seeds (2, 7, 10, 67, 101), yielding six orderings per item (original plus five shuffled). The Shuffled Dataset uses scales of up to five options, as the factorial growth of 6-option permutations ($6! = 720$) makes systematic sampling impractical within the experimental budget.

B.2 Hardware and Decoding

All experiments were executed on a single NVIDIA A100 (80 GB) GPU, using the Hugging Face `transformers` library (Wolf et al., 2020). Models were loaded in native precision (`dtype="auto"`, typically bfloat16 or float16) to avoid performance degradation from quantization. Greedy decoding (`do_sample=False`) was used throughout, with `max_new_tokens=50`. Greedy decoding ensures that any observed variation across items is attributable solely to the changes in scale wording or option order, not to stochastic sampling.

B.3 Inference Pipelines

IT models received prompts via the model's native chat template in a single `user` turn (no system prompt). Omitting the system prompt was a deliberate choice: adding a persona-shaping system message (e.g., "You are a helpful assistant") would introduce an uncontrolled variable across model families, which implement system-prompt parsing differently. Base models received a raw text-completion prompt ending with a forcing suffix (e.g., "The best answer is option number:"). Without this suffix, Base models tend to continue the list of options as if completing a document, rather than selecting one. Model-specific end-of-sequence tokens were added to the stopping criteria to prevent the model from generating additional text after the answer token.

B.4 Content Dimensions

Table 4 shows the distribution of questions across the 15 content dimensions after filtering and adaptation. The "Agreement" dimension is the most prevalent (32.7%), reflecting the design of the original Pew Research surveys, which frequently ask respondents to express their degree of agreement with policy positions or factual statements. The next-largest dimension is "Quantity" (17.3%), capturing items about the appropriate amount or level of something (e.g., government spending, personal frequency of behavior). The remaining 13 dimensions each account for under 9% of the dataset, providing a diverse coverage of opinion-related response formats while keeping the classification scheme manageable.

Content Dimension	Count	%
Agreement	404	32.7
Quantity	214	17.3
Better / Worse	101	8.2
Importance	91	7.4
Likelihood	85	6.9
Positive / Negative	67	5.4
Good / Bad	50	4.0
Magnitude of Problem	48	3.9
Frequency	38	3.1
Performance ("How Well")	38	3.1
Priority	35	2.8
Acceptance	31	2.5
Reason	15	1.2
Concern	11	0.9
Increase / Decrease	7	0.6
Total	**1,235**	**100.0**

Table 4: Content dimension distribution ($N = 1{,}235$).

C Prompt Templates

Design Rationale. The prompting strategy pursues two goals simultaneously: reliably extracting a single answer token, and minimizing extraneous biases not under study. The opening framing "You are taking a survey. There are no right or wrong answers" serves a dual purpose. It signals that the task is opinion-based rather than factual, reducing the likelihood of content-moderation refusals on politically sensitive items, and it suppresses the tendency of instruction-tuned models to hedge or justify rather than commit to a scalar response.

A **zero-shot** approach was adopted throughout. Including few-shot demonstration examples would introduce an additional positional confound: the label positions in the example items could prime the model toward particular options (e.g., a demonstration always selecting option A could bias subsequent selections). Zero-shot evaluation ensures that any systematic option preference reflects the model's own prior rather than a learned heuristic from the context.

For **IT models**, all content was delivered in a single `user` turn without a system prompt, as described in Appendix B. The output format constraint ("Provide exactly one [label] only") was added after pilot experiments revealed that instruction-tuned models frequently returned verbose explanations or cited multiple options unless explicitly instructed to output a single token.

For **Base models**, the instruction block was removed and replaced by a forcing suffix ("The best answer is option [label]:") appended directly after the option list. This text-completion framing exploits the base model's training objective: it predicts the next token following the suffix, which is typically the answer label.

The four prompt format variants (P1–P4) differ

IT Models (Chat Template)	Base Models (Text Completion)

`You are taking a survey. There are no right or wrong answers.`
`Please choose the one option that best fits your view. The order of the options does not imply any ranking.`
`Output format: Provide exactly one [label] only (no words, no punctuation, no quotes).`
 	
`Question: [question]`
 	
`Options:`
`[options]`
 	
`Please answer with ONLY one of the answer options given above, and with only the [label] corresponding to the option you choose.`
	`Question: [question]`
	
	`Options:`
	
	`[options]`
	
	`The best answer is option [label]:`

Figure 5: Core prompt templates. `[label]` is substituted with "number" (P1), "letter" (P2), or "option" (P3). For P3 (Unlabeled), the IT prompt omits the output-format line and truncates at "*...given above.*"; no forcing suffix is used for Base models under P3. For P4, the option block is extended with numbered "Don't Know" and "Refused" entries.

only in how options are labeled and presented. **P1 (Numeric)** uses "1) ... 2) ..." enumeration; **P2 (Alphabetic)** uses "A) ... B) ..."; **P3 (Unlabeled)** lists options as plain newline-separated text without any prefix; **P4 (Extended)** augments P1 with two additional options ("Don't Know" / "Refused") labeled $N+1$ and $N+2$, mapped to exclusion codes during analysis. For P3, the output-format instruction for IT models is abbreviated to "Please answer with ONLY one of the answer options given above" without specifying a label type, since no labels exist. Base models receive no forcing suffix under P3. The detailed template is shown in Figure 5.

D Supplementary Results

D.1 Dataset Scale Characteristics

Table 5 shows the distribution of scale types across the two experimental datasets. The Reworded Dataset includes 6-point scales (bipolar and unipolar) that do not appear in the Shuffled Dataset. This asymmetry arises because 6-point scales generate 720 possible permutations (6!), making systematic random sampling computationally and analytically impractical. The five seeds used for the Shuffled Dataset are instead applied to scales of up to five options, which yield 120 permutations at most. 4-point scales dominate both datasets (58.6% of Reworded, 78.7% of Shuffled) because the majority of content dimensions are served by a standard 4-point bipolar or unipolar structure. The 5-point bipolar format accounts for the next largest share, capturing dimensions (such as Importance or Likelihood) for which a neutral midpoint is semantically appropriate. The co-occurrence of variable cardinalities in the Reworded Dataset is a key driver of the smoother APD distributions described in Appendix D.6: when a 4-point and a 5-point variant are compared, the normalized distance values can take on more intermediate steps than when comparing two variants of identical length.

	Reworded		Shuffled	
Scale Type	**N**	**%**	**N**	**%**
4-pt Unipolar	1,858	30.1	3,408	46.0
4-pt Bipolar	1,760	28.5	2,424	32.7
5-pt Bipolar	1,375	22.3	1,350	18.2
5-pt Unipolar	635	10.3	228	3.1
6-pt Bipolar	512	8.3	–	–
6-pt Unipolar	35	0.6	–	–

Table 5: Scale type distributions across the two datasets. P4 extends all scales by $+2$ options ("Don't Know" / "Refused"), not reflected in these counts.

D.2 IT Model Response Distributions

Figure 6 shows the response category distributions for IT models on the Reworded Dataset, and should be read alongside Figure 2 in the main paper, which covers Base models. The contrast is stark along two dimensions.

First, the **unusable rate** for IT models is negligible ($< 0.1\%$) across all four prompt formats,

whereas Base models exhibit unusable rates ranging from around 8% under standard formats to nearly 35% under P4 for Mistral. This confirms that instruction tuning reliably enforces output format constraints: IT models correctly interpret the label-selection instruction and respond with a single token, even under the more complex P3 (Unlabeled) and P4 (Extended) formats that confuse base models.

Second, IT models display markedly **stable semantic profiles** across P1–P4. Unlike Base models, which shift their modal preference dramatically depending on whether options are labeled numerically or alphabetically (e.g., Gemma-2-9B selects Option 4 most often under P1 but shifts to a sharp spike at Option 1 under P2), IT models maintain broadly consistent response distributions regardless of the label type used. This structural stability is the behavioral correlate of the lower MSD values reported in Table 2 of the main paper, and suggests that instruction tuning reduces sensitivity to surface-level label format rather than merely suppressing invalid outputs.

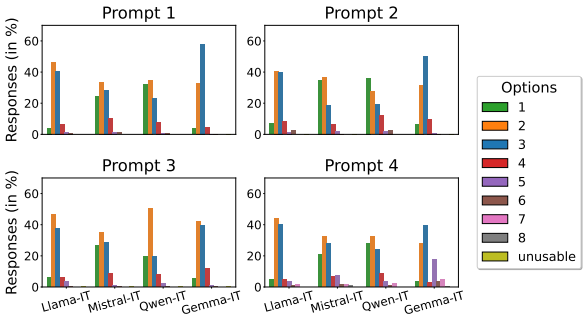

Figure 6: Response category distributions for IT models on the Reworded Dataset. Unusable rates are negligible (<0.1%) and response profiles are consistent across all four prompt formats, in sharp contrast to the Base model distributions in Figure 2.

D.3 The Qwen2.5 Base Model Anomaly

Qwen2.5-7B (Base) presents an atypical pattern among Base models. While other Base models show large distributional divergences between Reworded and Shuffled conditions, Qwen2.5-7B maintains relatively overlapping distributions across both datasets. Its inter-prompt correlations on the Reworded Dataset reach P1–P4: $r = 0.849$ and P1–P3: $r = 0.575$, values that are comparable to those of some IT models from other families (cf. Table 7) and substantially higher than the other Base models (e.g., Mistral Base P1–P3: $r = 0.408$, Gemma Base P1–P3: $r = 0.374$). Its MSD on the Reworded Dataset (P1: 0.254, lowest among all Base models) and the Shuffled Dataset (P1: 0.323, again lowest) further confirm this outlier status.

We attribute this stability to a high density of structured MCQA content in Qwen's pre-training data. Alibaba Cloud's Qwen series is trained on a corpus with a reportedly large fraction of question-answering and instruction-style data (Qwen et al., 2025), which may pre-train a rudimentary form of symbol binding even without explicit alignment. Figures 7 and 8 illustrate this: while other Base models show the blue (Reworded) and orange (Shuffled) projected-response distributions drifting substantially apart (indicating that option shuffling changes the apparent semantic opinion), Qwen's distributions overlap substantially across all four prompt formats.

The IT variant of Qwen further smooths and narrows the distributions but provides a smaller absolute improvement than seen in other model families. For instance, Llama's MSD drops by roughly 42% from Base to IT on P1 ($0.357 \rightarrow 0.207$), whereas Qwen's improvement is only around 19% ($0.254 \rightarrow 0.232$), suggesting diminishing marginal returns of alignment for a model that already exhibits partial stability at the pre-training stage.

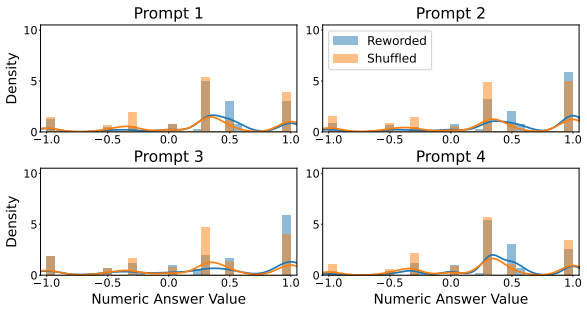

Figure 7: Projected response distributions for Qwen2.5-7B (Base). Blue: Reworded Dataset; Orange: Shuffled Dataset. The two distributions overlap substantially across all four prompt formats, unlike most Base models, which diverge sharply when options are shuffled.

D.4 MSD on the Shuffled Dataset

Table 6 reports the MSD on the Shuffled Dataset and provides the complete quantitative basis for the Label-Free Consistency Paradox discussed in Section 4.3 of the main paper.

For IT models, the P3 (Unlabeled) column is bolded because it achieves the lowest MSD for every model family. The reductions are substantial: Gemma-2-9B-IT drops from 0.232 (P1) to

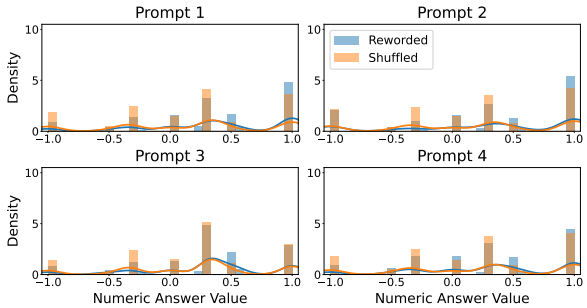

Figure 8: Projected response distributions for Qwen2.5-7B-Instruct (IT). Alignment further narrows the distributional spread, but the high baseline overlap observed in the Base model means the marginal gain from instruction tuning is smaller than for other families.

Model	P1	P2	P3	P4
Base Models				
Llama-3.1-8B	0.367	0.403	0.445	0.328
Mistral-7B-v0.3	0.503	0.400	0.365	0.482
Qwen2.5-7B	0.323	0.361	0.363	0.330
Gemma-2-9B	0.553	0.346	0.415	0.378
IT Models				
Llama-3.1-8B-IT	0.277	0.233	**0.146**	0.274
Mistral-7B-IT	0.263	0.320	**0.240**	0.249
Qwen2.5-7B-IT	0.310	0.277	**0.206**	0.273
Gemma-2-9B-IT	0.232	0.187	**0.121**	0.220

Table 6: MSD on the Shuffled Dataset. Bold P3 values for IT models confirm the Label-Free Consistency Paradox: removing option labels yields the lowest inter-ordering variance for all four IT model families, even though P3 is the most volatile format in the Reworded Dataset.

0.121 (P3), a reduction of 47.8%; Llama-3.1-8B-IT drops from 0.277 (P1) to 0.146 (P3), a reduction of 47.3%; Qwen2.5-7B-IT drops from 0.310 (P1) to 0.206 (P3), a reduction of 33.5%. Even Mistral, which shows elevated P3 variance in the Reworded Dataset (Table 2), achieves its second-lowest MSD under P3 in the Shuffled condition (0.240 vs. 0.263 for P1).

The mechanism is as follows. Under labeled formats (P1, P2, P4), shuffling the options creates a conflict between the model's positional prior (e.g., a tendency to select option "A" or "1") and its semantic preference (e.g., a tendency to select content meaning "Agree"). Because these two cues point to different options after shuffling, the model's outputs become inconsistent across orderings. Under P3, no labels are present, so the positional prior is over raw text position rather than a symbol. Since IT models are better calibrated to attend to content, the removal of labels forces them to rely more directly on semantic information, which is invariant across orderings. The result is higher consistency under the very condition that most disrupts labeled formats.

For Base models, this paradox does not materialize systematically. Base model P3 values show no consistent pattern: Llama P3 (0.445) is the highest across its prompts, whereas Mistral P3 (0.365) is the lowest. This is expected: Base models lack reliable symbol binding in the first place, so removing labels does not alleviate a label-induced conflict but instead eliminates the only structural cue that anchored their responses.

D.5 Full Correlation Matrix (Reworded Dataset)

Table 7 reports all six pairwise Pearson correlations between prompt formats on the Reworded Dataset. Several structural patterns are noteworthy.

First, **P1–P4 correlations are consistently the highest** across all model families, for both Base and IT variants (e.g., Mistral-IT: 0.913; Llama-IT: 0.894; Gemma-IT: 0.897). This reflects the structural similarity between these two formats: P4 is a direct extension of P1 (numeric labels) with two additional options appended, so models that have converged on a numeric-label strategy under P1 apply essentially the same logic under P4, differing only in whether they ever select the "Don't Know" or "Refused" options.

Second, **pairs involving P3 (Unlabeled) tend to produce lower correlations**, particularly P1–P3 and P3–P4. For IT models, P1–P3 correlations range from 0.701 (Llama) to 0.778 (Gemma), compared to P1–P4 values of 0.894–0.897. This reflects the fundamentally different structural demands of P3, which requires the model to parse and select from plain-text options rather than binding a numeric or alphabetic symbol to content. The label-free format creates a distributional shift in response tendencies (visible in the IT MSD data of Appendix D.4), which lowers correlation with labeled formats.

Third, all correlations in Table 7 are substantially higher than their counterparts in the Shuffled correlation matrix (Table 3 in the main paper). This confirms at the level of individual-question response profiles that semantic rewording is a milder

perturbation than positional shuffling: models preserve the relative direction of their opinions across reworded variants better than across shuffled orderings. For IT models, many Reworded correlations exceed 0.80, whereas Shuffled correlations rarely approach this range.

Model	1–2	1–3	1–4	2–3	2–4	3–4
Base						
Llama-3.1	0.580	0.398	0.644	0.464	0.536	0.461
Mistral-7B	0.485	0.408	0.454	0.354	0.336	0.278
Qwen2.5-7B	0.608	0.575	0.750	0.467	0.560	0.481
Gemma-2-9B	0.534	0.374	0.564	0.473	0.627	0.504
IT						
Llama-3.1	0.820	0.701	0.894	0.712	0.797	0.683
Mistral-7B	0.707	0.722	0.913	0.646	0.709	0.703
Qwen2.5-7B	0.818	0.743	0.883	0.737	0.799	0.726
Gemma-2-9B	0.860	0.778	0.897	0.820	0.841	0.749

Table 7: Pearson's r across all prompt-format pairs on the Reworded Dataset. All values exceed their Shuffled counterparts (Table 3), confirming that semantic rewording is less disruptive than positional permutation. P1–P4 pairs are consistently the highest due to structural similarity between numeric-label and extended-numeric formats; pairs involving P3 (Unlabeled) are consistently lower.

D.6 APD Distributions: Reworded Dataset

Figures 9 and 10 show APD distributions on the Reworded Dataset under P1, and complement the Shuffled Dataset distributions in the main paper.

The most salient structural difference between the two datasets is the shape of the distributions. The **Shuffled Dataset produces discrete spikes** at a small set of APD values. This is a direct mathematical consequence of fixed scale cardinality: when all five or six variants of a question share the same scale length, the normalized values at which responses can land are constrained to a finite set (e.g., for a 4-point scale mapped to $\{-1.0, -0.33, 0.33, 1.0\}$, the only possible pairwise distances are 0, 0.67, 1.33, and 2.0). As a result, the APD for each question takes on one of a small number of discrete values, creating visible spikes in the density plot at those locations.

The **Reworded Dataset produces smoother, more continuous distributions**. Because the five reworded variants for a given question may draw from scales of different cardinalities (e.g., a 4-point variant paired with a 5-point variant), the set of possible normalized distances is richer. Cross-cardinality pairings generate intermediate values that do not appear in any single-cardinality comparison, effectively densifying the support of the APD distribution and smoothing out the discrete spikes.

For Base models (Figure 9), the distributions are wide and exhibit substantial density in the high-APD region, with a non-trivial fraction of questions producing APDs above 1.0, indicating near-maximal scale traversals triggered by mere wording changes. For IT models (Figure 10), the distributions are strongly compressed toward 0, with most questions achieving APD values below 0.5. The smooth shape is preserved because the variable scale cardinalities apply equally to Base and IT models; the compression reflects reduced sensitivity to wording rather than any change in scale structure.

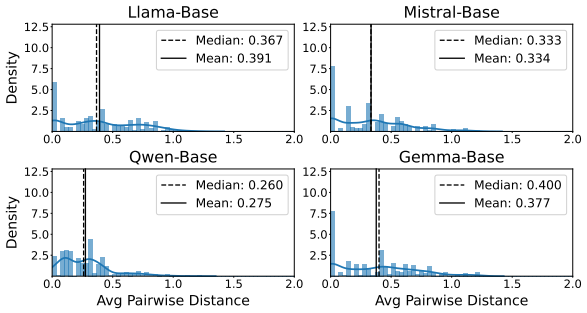

Figure 9: APD distributions for Base models on the Reworded Dataset (P1). The smooth, continuous shape arises from variable scale cardinalities across reworded variants. Wide dispersion and substantial density above APD $= 1.0$ indicate that wording changes alone can induce near-complete polarity reversals in Base models.

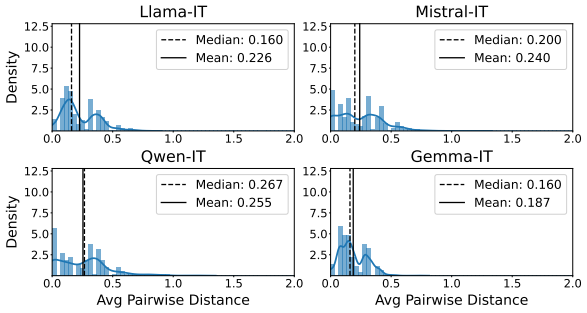

Figure 10: APD distributions for IT models on the Reworded Dataset (P1). The smooth curve shape is maintained by the same variable-cardinality structure, but distributions are strongly compressed toward 0 relative to Base models, reflecting substantially reduced sensitivity to semantic rewording.