"Construct Validity" in LLMs: Metrics to Measure Consistency & Alignment in Multi-Turn Likert and Free-Text Scenarios

Simon Münker[1], Charlott Jakob[2], Achim Rettinger[1], Vera Schmitt[2]
[1]Trier University, {muenker,rettinger}@uni-trier.de
[2]Technical University Berlin, {c.jakob,vera.schmitt}@tu-berlin.de

Abstract

As LLMs are increasingly utilized as generative proxies for human participants in survey-based studies, evaluating their behavioral consistency becomes critical. Current assessments typically rely on single-turn interactions, which ignore sequential dependencies. We address this gap by introducing two novel evaluation techniques: Multi-Turn Decision Tracing (MTDT), which maps similarity in Likert-scale responses across branching decision paths by tracking probability distributions, and Multi-Turn Reward Consistency (MTRC), which assesses alignment stability in free-text responses through variance in reward model scores. We evaluate ten open-weight LLMs across three psychological instruments and four reward classifiers. Our results show cross-metric correlations in both response variability and internal consistency, providing evidence that both metrics capture a shared dimension of architectural stability despite measuring fundamentally different behavioral aspects. However, no model achieves acceptable psychometric thresholds. Thus, while our findings challenge the validity of LLMs as reliable human proxies, we establish strong cross-metric convergence for validity measurements in sequential LLM interactions.

1 Introduction

As LLMs evolve from static question-answering tools into general-purpose agents capable of complex sequential inference used in decision-making processes (Chiang et al., 2024) and human simulation (Argyle et al., 2023), we see the need to pivot from stateless single-turn evaluation to robust multi-turn evaluation methodologies. Current paradigms are narrow and often fail to capture the dynamic nature of interactions (Yi et al., 2025). Thus, a fundamental challenge emerges at the intersection of artificial intelligence and behavioral assessment: How can we reliably measure decision-response consistency in LLMs across numerical and textual outputs in multi-turn interactions?

This stability/consistency question is not solely methodological. If LLMs should serve as synthetic research participants (Argyle et al., 2023; Motoki et al., 2024), the consistency of their simulated attitudes across sequential interactions is a prerequisite for any substantive interpretation. A simulated personality that shifts chaotically as context accumulates cannot *"hold"* attitudes in a psychometrically meaningful sense (Sandhan et al., 2025; Röttger et al., 2024). Without accounting for sequential dependencies and the probabilistic nature of generation, it remains unclear whether observed orientation patterns represent stable behavioral tendencies or statistical artifacts sensitive to perturbation (Sclar et al., 2024).

Often, research relies on surface-level observation of model outputs aggregated across turns (Rozado, 2023; Faulborn et al., 2025; Exler et al., 2025; Ueda and Suwa, 2025). However, these observations lack measuring the underlying stability of the constructs (Cronbach, 1951). Without considering the probabilistic nature of the generation process and the influence of the context window, it remains unclear whether the recorded artificial behavior resembles human traits or depicts statistical artifacts sensitive to perturbations (Sclar et al., 2024; Shu et al., 2024).

We address this gap by proposing a shift from analyzing discrete outputs to analyzing decision traces and reward trajectories (see Fig. 1). Concretely, we introduce two complementary metrics designed to expose different facets of sequential stability: one operating on Likert-scale probability distributions across branching decision trees, the other on multiple reward model score trajectories extracted from free-text responses. Their deliberate complementarity is a methodological feature: if they capture a shared underlying dimension of architectural stability despite measuring structurally

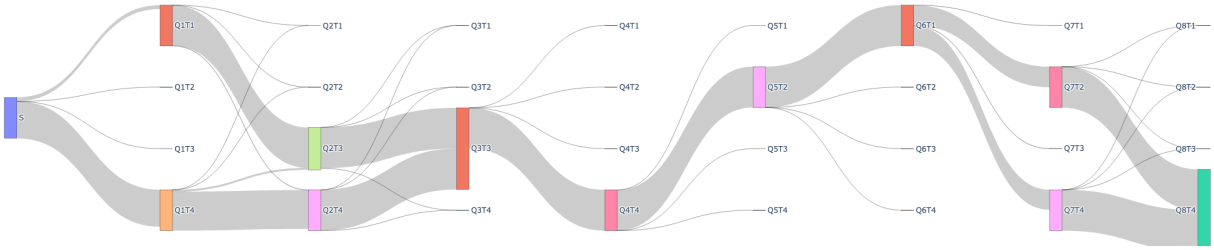

Figure 1: Probabilistic decision trace (M) of Gemma 3 12B on the Symbolic Racism 2000 Scale, visualized as a Sankey diagram. The visualization demonstrates how our proposed MTDT (Sec. 3.4) metric captures the dependency structure in LLM responses to sequential Likert-scale items. Each flow represents a response trajectory (horizontal: $q_0, \ldots, q_8$), with width proportional to the probability mass (P) of that path. Node size indicates the aggregated probability ($\bar{P}$) of selecting each response (vertical: $r_0, \ldots, r_4$) option at a given question (q_i).

different behavioral aspects, their cross-metric convergence constitutes evidence for construct validity that neither metric could provide alone.

Research Questions

RQ_1 How can we enhance the reliability of measuring in multi-turn surveys by leveraging logit probabilities and reward modeling?

RQ_2 To what extent do the behavior of LLMs in Likert-based evaluation and free-text answers on a psychological instrument correlate?

Contributions

We introduce two novel metrics and demonstrate that both correlate with variability and internal consistency across our model and instrument selection. We publish the data, calculations, and pipelines for reproducibility (see App. A).

1. **Multi-Turn Decision Tracing (MTDT),** a technique that maps the similarity of Likert-rating response patterns by tracking probability distributions across statistically significant decision paths (see Sec. 3.4).

2. **Multi-Turn Reward Consistency (MTRC),** a method for assessing alignment stability in free-text responses by analyzing the variance of reward model scores across multi-turn interactions (see Sec. 3.5).

2 Background

2.1 Multi-Turn Analysis of LLM Behavior

While single-turn evaluations started LLM personality and attitude assessment (Serapio-García et al., 2025; Rozado, 2023; Münker, 2025b), a growing body of work recognizes that sequential interaction fundamentally alters model behavior (Bai et al., 2024; Yi et al., 2025; Sandhan et al., 2025). Multi-turn settings introduce compounding dependencies: each response conditions the probability distribution of subsequent outputs, a mechanism absent from independent item administration. This has motivated recent frameworks (Lee et al., 2025; Sandhan et al., 2025) that treat conversational context as a first-class variable rather than a confound to be controlled away.

Rather than administering items in isolation, Sandhan et al. (2025) embeds questionnaire prompts within rich conversational scenarios and measures whether personality signals remain stable as context evolves. The central finding, that LLMs exhibit pronounced *"context-driven personality drift"*, directly motivates our methodological choices: stability cannot be assessed without explicitly modeling the branching structure of conversational histories. Similarly, work on multi-turn dialogue evaluation (Bai et al., 2024; Yi et al., 2025) has demonstrated that response coherence degrades non-linearly as conversation length increases, suggesting that the accumulated context window introduces systematic variance not captured by single-turn benchmarks.

2.2 Psychometric Assumptions

The core challenge in evaluating LLMs lies in the friction between classical measurement theory and the architecture of Transformer models (Sparrenberg et al., 2024; Liu et al., 2025). Tjuatja et al. (2024) demonstrate that while LLMs exhibit response biases similar to humans, such as sensitivity to ordering and framing, the underlying mechanisms differ fundamentally. Classical psychometric theory assumes a human cognitive architecture defined by working memory limits and biological fatigue (Raykov and Marcoulides, 2011). LLMs

violate these assumptions while introducing unique sources of systematic variance that traditional methods fail to capture (Demszky et al., 2023; Shu et al., 2024; Petrov et al., 2024). Our proposed methods are designed specifically to address these non-human characteristics:

Deterministic vs. Probabilistic Output Unlike humans who show test-retest variability due to internal state changes, LLMs can exhibit perfect consistency or wild divergence based on sampling strategies. This requires tracking the probability distribution rather than just the final token output to understand the true confidence of the model.

Context Window Dependencies LLMs maintain *"memory"* through preceding interactions, denoted as a chat-like structure (Bai et al., 2024). This creates a dependency chain where previous answers mathematically condition future possibilities, requiring evaluation methods that trace the full decision path rather than treating questions as isolated events.

2.3 LLMs as Human Simulacra

The use of LLMs as substitutes for human participants in social science research has grown rapidly (Motoki et al., 2024; Bisbee et al., 2024; Larooij and Törnberg, 2025; Thapa et al., 2025), driven by advantages in cost, scalability, and experimental control. Models can complete thousands of questionnaires instantaneously, enable perfect experimental manipulation, and eliminate concerns about participant fatigue, dropout, or demand characteristics. However, these advantages come with fundamental questions about external validity and generalizability (Agnew et al., 2024; Münker et al., 2026; Wang et al., 2025; Bean et al., 2025; Taubenfeld et al., 2024).

Without establishing this level of structural validity, conclusions about LLMs' utility as agents or research proxies rest on uncertain foundations. A model that scores "low" on a specific trait cannot be meaningfully described as having that personality if the underlying probability distribution shifts chaotically with context. The question extends to the deployment of AI as decision-support systems: if LLMs cannot maintain consistent response patterns across multi-turn interactions, this raises theoretical concerns about their behavioral reliability in contexts requiring stable response behavior, and thus, they cannot reliably represent coherent attitudes, and their reliability as human simulacra, a multiverse of simulated human personalities (Baudrillard, 1994; Shanahan, 2026; Sandhan et al., 2025), remains in question.

3 Methods

3.1 Psychological Instruments

We select three psychological questionnaires (Ho et al., 2015; Henry and Sears, 2002) to assess a diverse range of constructs. To ensure a robust evaluation of psychometric properties, we prioritized instruments with well-established human baselines that show sufficient internal consistency (see Sec. 3.6). Appendix B contains the complete description of the items, scoring keys, and original validation results. All questionnaires are administered in the original published item order with verbatim option wording. Both factors are known sources of LLM sensitivity (Sclar et al., 2024; Tjuatja et al., 2024): reordering items changes the accumulated conversational context before each question, and rephrasing response labels can shift logit distributions for semantically equivalent options. We do not investigate the impact of these variables in this work.

3.2 Language Models Selection

We select open-weight language models that span multiple model families. Our selection criteria prioritized: (1) public availability to enable full reproducibility and future work on investigating circuit activation (Dunefsky et al., 2024), (2) a range of model sizes within feasible computational constraints, and (3) documented performance on general language tasks. We restrict our experiments to open-weight models with parameter counts ranging from 1B to 12B. This range ensures accessibility for researchers with moderate computational resources, approx. 32GB GPU memory for 16-bit precision models with up to 14B parameters. While excluding proprietary models (e.g., GPT-5 or Claude) represents a limitation, our focus is on analyzing the intrinsic capabilities of LLMs to align with psychological constructs, rather than benchmarking specific commercial endpoints (Ollion et al., 2024). Understanding construct validity patterns in open-weight models provides first insights that may be applicable to the broader LLM ecosystem. We select the following models: Llama 3.2 (1/3B) (Grattafiori et al., 2024), Qwen 3 (4/8/14B) (Yang et al., 2025), and Gemma 3 (4/12B) (Kamath et al., 2025), Ministral 3 (3/8/14B).

3.3 Prompting

We use only minimal prompting to obtain raw model responses without optimizing for specific outcomes. Thus, we present the original questionnaire instruction combined with an optional persona prompt. In the case of free-text experiments, we replace the scale with the statement *"Write a comment that reflects your opinion."*. For persona conditioning, we sample 500 items from the Nemotron Personas USA dataset (Meyer and Corneil, 2025). The data set is a collection of synthetically generated personas grounded in real-world demographic, geographic, and personality trait distributions based on US citizens. We additionally conduct an ablation study examining the effect of sampling temperature on both metrics across all instruments. The results are reported in Appendix D.

3.4 Multi-Turn Decision Tracing (MTDT)

Standard psychometric evaluations of LLMs often treat items as independent events (Abdulhai et al., 2024; Ma et al., 2024). However, to account for a) the probability distributions over tokens and b) the influence of conversational context, we develop a novel evaluation method: Multi-Turn Decision Tracing (MTDT). MTDT tracks probability distributions across sequential questionnaire items to construct response pattern networks dependent on past model decisions. For each questionnaire item presented to a model, we extract the probability distribution over valid response options. Using the model's token-level logits, we compute:

$$P(r_i|q_i, h_{<i}) = softmax(logits_{r_i})$$

where r_i is a response to question q_i, and $h_{<i}$ is the conversation history of previous question-response pairs. We record *"how confident"* the model was across all plausible responses, conditioned on every plausible prior history. The full response profile is constructed as follows:

1. Present the first item with system instructions, including the questionnaire context and an optional persona prompt.

2. Extract the most likely responses based on the logits distribution. To manage the combinatorial explosion of the decision tree, we filter the response distribution using two complementary approaches: eliminating low-probability responses unlikely to represent genuine model tendencies ($\tau = 10^{-4}$) and limiting exploration to *top-k* most probable responses ($k = 3$).

3. For each likely response to question i, create new conversation branches for question $i + 1$. Continue until all items are processed.

A key design feature of MTDT is that context influence is not treated as noise to be controlled away, but as a structural property to be measured. By branching on every plausible response at each step, we obtain multiple conversation histories H_i leading to item q_i. The variance of $P(r_j|q_i, h)$ across $h \in H_i$ directly quantifies how sensitive a model's item response is to its prior conversational trajectory. Large within-item variance across branches indicates high context entanglement; small variance indicates robust, context-stable response patterns.

Similarity Evaluation For each experimental condition (questionnaire × model × persona), we create response profile matrices. The construction follows these steps:

1. For each item q_i and each response option r_j, we compute the marginal probability by aggregating over all conversation histories that could lead to this node. We define the aggregated probability as the mean likelihood across valid paths:

$$\bar{P}(r_j|q_i) = \frac{1}{|H_i|} \sum_{h \in H_i} P(r_j|q_i, h)$$

where H_i is the set of all valid conversation histories up to question i.

2. We construct matrix $M \in \mathbb{R}^{k \times n}$ where rows represent response options and columns represent questions. Each cell contains the aggregated probability:

$$M[r_j, q_i] = \bar{P}(r_j|q_i)$$

3. To compare two experimental conditions (e.g., different temperatures, different personas, or different models on the same instrument), we compute the similarity score s_h using the complement of the directed Hausdorff distance (Muller, 1959). We treat the columns of M and N as sets of vectors in metric space:

$$s_h(M, N) = 1 - \max\{\sup_{m \in M}(d(m, N)), \sup_{n \in N}(d(M, n))\}$$

where d quantifies the Euclidean distance between response distribution vectors for a single question across the two profile matrices. By employing the Hausdorff distance, we measure the worst-case divergence between two response profiles (Huttenlocher et al., 1993), ensuring that our similarity score is a conservative estimate of how well a model maintains its behavioral pattern under different conditions. Values close to 1 indicate highly congruent response patterns, while values near and below 0 indicate divergence.

3.5 Multi-Turn Reward Consistency (MTRC)

Although MTDT enables direct comparison of response distributions, it is inherently limited to single-vocabulary choices. However, the generative nature of LLMs suggests that analyzing free-text responses may better capture the expressiveness and reasoning abilities of the model (Röttger et al., 2024). To address this, we introduce a complementary metric: Multi-Turn Reward Consistency (MTRC). Instead of constraining models to predefined options, we prompt them to provide free-text explanations, using reward model scoring as a proxy for generative coherence across the questionnaire trajectory (Elle, 2025). To calculate the finished response profile, we conduct the following steps:

1. We follow a single greedy conversational trace along four reward models through all questionnaire items, maintaining the full conversation history. This mimics how humans complete questionnaires, while allowing models to develop internally consistent reasoning across items.

2. For each generated response, we compute numeric reward scores using an ensemble of four reward models: Skywork-Reward-V2-Llama-3.1-8B and Skywork-Reward-V2-Qwen-3-8B (Liu et al., 2026), GRM-Gemma2-2B (Dong et al., 2023), and RM-Mistral-7B (Yang et al., 2024). These models differ in their base architectures and training corpora, spanning general preference alignment, instruction-following quality, and helpfulness (see C for more information). Critically, we treat MTRC as reward-model-agnostic: each scorer is applied independently, producing four parallel reward trajectories per model-instrument combination.

3. The result for each reward model is a vector $\hat{r} \in \mathbb{R}^n$ where each element represents the reward score for the model's response to the corresponding item. Consistency is defined as the variance and trend stability of $\hat{r}$ across the questionnaire trajectory. Final MTRC statistics (mean, SD, Cronbach's α) are computed separately per reward model and then averaged across the four scorers to yield a single ensemble estimate per model-instrument pair, used in all reported comparisons.

Reward model scores do not measure political attitudes or psychological constructs directly. Rather, they measure the coherence and quality of each generated response. A model maintaining a stable reasoning stance across questionnaire items will produce consistently scored outputs under any well-calibrated reward model; a model whose generation is destabilized by accumulating context will produce erratic reward trajectories. We therefore treat reward score variance as the operative signal: high trajectory variance indicates generative instability, independent of the ideological content of individual responses. This is consistent with prior findings that reward model scores correlate with response coherence and internal reasoning quality (Elle, 2025), making them a suitable ordinal signal for Cronbach's α analysis even when personality alignment is not the reward model's training objective.

3.6 Cronbach's Alpha: Internal Consistency

We assess internal consistency using Cronbach's alpha (α) (Cronbach, 1951) by measuring the degree to which items on a scale correlate with each other. The values range from 0 to 1, with higher values indicating greater internal consistency. For instruments containing reversed items (e.g., SDO-7, where items marked $[-]$ are scored inversely), we apply standard reverse-coding prior to computing α, consistent with the original validation procedures. Despite violations of classical test theory assumptions, we utilize α not as a measure of latent psychological traits but as a descriptive heuristic to quantify internal coherence. Following the evaluation convention in applied research (Nunnally, 1975), we consider $\alpha \geq 0.70$ as acceptable consistency/internal coherence.

Critically, α is computed separately for MTDT (from marginal probability profiles across the decision tree) and for MTRC (from reward-trajectory

vectors). The fact that both computations operate on the same items but on fundamentally different representations (probability distributions versus reward scalarsmeans) suggests that a significant correlation in α across models cannot be attributed to a shared computational artifact.

4 Results

4.1 Aggregate Construct Validity

Table 1 (MTDT, MTRC) presents psychometric properties for all model-instrument combinations across both response formats. The results reveal substantial variability in construct validity, with Cronbach's α values ranging from 0.155 (Llama 3.2 1B on MTRC) to 0.606 (Qwen 3 14B on MTDT). Critically, no model achieved the conventional threshold of $\alpha \geq 0.70$ for acceptable internal consistency across both metrics, and only Qwen 3 14B approached this threshold on MTDT ($\alpha = 0.606$). In contrast, the average human α is 0.74 (see Tab. B) across the reported instruments. MTRC statistics are ensemble averages over four reward models; per-scorer details are reported below.

4.2 Cross-Metric Correlation Analysis

The design logic of our dual-metric approach rests on a specific prediction: if MTDT and MTRC both capture architectural stability rather than surface-level behavioral regularities, they should agree not on the content of model responses (which they cannot, since one measures Likert probabilities and the other measures free-text reward scores) but on the structure of consistency across models.

Mean Response Patterns The near-zero correlation ($r = 0.118$, $p = 0.267$) between MTDT and MTRC mean values confirms that the two metrics measure different behavioral dimensions: higher MTDT means indicate more conservative positions on the political scales, while higher MTRC means reflect greater alignment with reward model preferences. These are orthogonal constructs, and the absence of mean correlation validates that the two metrics are not redundant proxies for the same surface pattern.

Response Variability Models exhibiting high variance in probability distributions across decision trees also show high variance in reward score trajectories ($r = 0.306$, $p = 0.050$). This cross-format consistency in inconsistency is the first evidence that both metrics are sensitive to the same underlying architectural property. A model that is sensitive to context perturbations shows this sensitivity both when forced to choose among Likert options and when generating unrestricted text despite the radically different generation mechanisms involved.

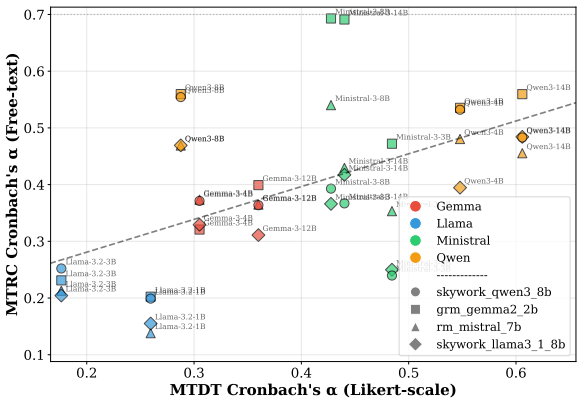

Figure 2: Correlation between MTDT and MTRC Cronbach's alpha values visualizing Table 1 main finding with a polynomial regression line.

Internal Consistency The strongest evidence for the dual-metric framework comes from the Cronbach's α correlation (Fig. 2): $r = 0.456$, $p = 0.006$. Models achieving higher internal consistency in structured Likert responses also tend toward higher consistency in free-text generation. This correlation is non-trivial: MTDT α is computed over probability-distribution co-variance across a branching decision tree, while MTRC α is computed over reward-score co-variance across a greedy sequential trace. The two quantities share no computational pathway. Their correlation therefore reflects a genuine property of the models, what we term architectural stability: the degree to which a model's internal representations support coherent, context-robust generation regardless of the output format imposed.

Per-Reward-Model Robustness To verify that the cross-metric correlation in Cronbach's α is not an artifact of any single scorer, we replicate the analysis for each of the four reward models independently. All four reward models yield significant positive correlations (GRM-Gemma2-2B: $r = 0.466, p = 0.005$; Skywork-Llama-3.1-8B: $r = 0.397, p = 0.015$; Skywork-Qwen-3-8B: $r = 0.337, p = 0.034$; RM-Mistral-7B: $r = 0.475, p = 0.004$).

Method	**MTDT**			**MTRC**			**Base Distance**	
Model	μ	σ	$\alpha \uparrow$	μ	σ	$\alpha \uparrow$	μ	σ
Gemma 3 4B	3.433	0.281	0.305	0.185	0.694	0.349	0.294	0.111
Gemma 3 12B	2.703	0.397	0.360	0.263	0.673	0.360	0.380	0.200
Llama 3.2 1B	1.410	0.374	0.260	-4.832	0.677	0.174	0.683	0.167
Llama 3.2 3B	3.450	0.417	0.176	-3.422	0.700	0.225	0.289	0.268
Ministral 3 3B	2.936	0.519	0.484	-2.022	0.862	0.329	0.539	0.073
Ministral 3 8B	3.075	0.298	0.428	-0.272	0.936	0.498	0.567	0.088
Ministral 3 14B	3.147	0.315	0.440	0.311	0.889	0.476	0.483	0.215
Qwen 3 4B	4.491	0.203	0.548	-4.049	0.887	0.486	0.788	0.073
Qwen 3 8B	2.049	0.423	0.288	-3.499	0.990	0.513	0.498	0.276
Qwen 3 14B	2.775	0.900	0.606	-2.590	1.077	0.496	0.356	0.194

Table 1: Psychometric evaluation of MTDT and MTRC across ten open-weight language models, alongside MTDT base distance between the base model without role-prompt and sampled cultural personas. Columns show mean response values (μ), response variability (σ), and internal consistency (Cronbach's α) for MTDT and MTRC; and mean similarity (μ) and variability (σ) for Base Distance, aggregated across all instruments.

4.3 Base Model Similarity

Table 1 (Base Distance) presents the similarity between base models (without role prompts) and persona-conditioned responses. The results reveal substantial heterogeneity in persona sensitivity across model families. Qwen 3 4B exhibits the highest mean similarity ($\mu = 0.788$), indicating minimal differentiation between base and persona-conditioned responses. Conversely, Gemma 3 4B shows the lowest similarity ($\mu = 0.294$), suggesting greater sensitivity to persona conditioning. Notably, within-family patterns are inconsistent: larger Qwen models show decreasing similarity ($0.788 \rightarrow 0.356$), while Ministral 3 models maintain relatively stable high similarity (0.539-0.567).

5 Discussion

5.1 Cross-Metric Correlation

The central empirical claim of our paper is that MTDT and MTRC, despite measuring structurally different behavioral aspects (Likert-scale probability distributions versus free-text reward trajectories) converge significantly in their measures of internal consistency. This convergence is not guaranteed by construction.

A model that is highly consistent in its Likert choices could plausibly generate erratic free-text, and vice versa: the two output formats impose different generation constraints and draw on different parts of the model's learned representations. The fact that α correlates significantly across formats, robustly so across all four reward scorers, supports a specific theoretical interpretation: construct stability is a model-level property that manifests across output formats.

A naive single-turn evaluation that samples responses to each item independently and records only modal outputs cannot detect this property, because it collapses the distributional and sequential structure that both MTDT and MTRC exploit. Two models can produce identical aggregate scores on a naive evaluation while differing dramatically in their underlying stability, a difference that would only emerge under multi-turn, context-sensitive probing. The cross-metric α correlation is precisely the kind of signal that is invisible to naive evaluations and becomes measurable only when both dimensions of stability are assessed.

5.2 Reward Models as Consistency Proxies

A natural concern with MTRC is whether reward models, trained to evaluate response quality and human preference alignment, can serve as valid proxies for consistency in a psychological questionnaire. While reward scores do not measure political attitudes or personality traits. What they do measure is the coherence and quality of the generated reasoning trace, which we treat as a surface-level indicator of generative stability. To mitigate scorer-specific bias, we employ an ensemble of four reward models with different architectures and training corpora. The cross-metric α correlation is significant and consistent across all four scorers, strongly suggesting that the observed convergence reflects genuine model properties rather than artifacts of any individual reward model's preferences. Nonetheless, we stress that MTRC, as currently

operationalized, measures consistency of response quality, not consistency of expressed attitudes.

5.3 Persona Sensitivity and Identity Persistence

The base model similarity analysis reveals a critical dimension of construct validity: the degree to which models maintain stable response patterns when conditioned with demographic personas (Sclar et al., 2024). The wide range of similarity scores indicates that persona conditioning does not uniformly affect all models. High similarity scores suggest that these models exhibit strong default response patterns resistant to persona-based modulation, potentially indicating either robust safety alignment or insufficient capacity to represent diverse viewpoints. Conversely, low similarity scores indicate greater flexibility in adopting persona-specific response patterns, though this variability raises questions about whether such changes reflect genuine perspective-taking or superficial linguistic adjustment without coherent underlying attitude structures.

The inconsistency within-family scaling further complicates the assumption, common in LLM simulation studies, that larger models are better simulacra. If larger models simply learned more robust representations, we would expect monotonic relationships between size and persona sensitivity. The observed heterogeneity, Qwen 3 shows decreasing similarity with scale, Ministral 3 shows stable high similarity across sizes, suggests that persona sensitivity is shaped by training procedure and alignment strategy, not capacity alone. This has direct practical implications: deploying LLMs as human simulacra requires not only selecting models with appropriate base characteristics, but also understanding whether their apparent persona flexibility reflects genuine distributional re-weighting or is merely sampling noise on an unshifted distribution.

6 Conclusion

RQ_1 Our proposed methods successfully demonstrate efficient measurement approaches that leverage probability distributions and reward modeling: **MTDT:** The method enables assessment of internal stability by constructing response pattern networks from token-level probability distributions across branching conversational paths. We quantify stability without requiring repeated administrations, a key advantage over traditional test-retest methodologies. **MTRC:** The approach addresses the fundamental challenge of evaluating free-text responses by analyzing reward model score trajectories across multi-turn interactions. It captures alignment stability through variance analysis, revealing that open-ended generation exhibits fundamentally different stability characteristics than structured responses.

RQ_2 Our cross-metric analysis reveals a nuanced answer with theoretical implications. While surface-level response patterns (Likert vs. reward-proxy) show no correlation, structural consistency measures align significantly. The significant correlation in Cronbach's α demonstrates that models achieving internal consistency in one format tend toward consistency in the other, despite mean response divergence. This validates our dual-metric approach: both MTDT and MTRC capture a shared underlying dimension of construct validity, even though they measure different behavioral aspects. Models with architectural features supporting stable representations maintain this stability across output formats, while models lacking such features exhibit instability regardless of generation constraints.

Implication Studies using LLMs as synthetic participants (Argyle et al., 2023; Larooij and Törnberg, 2025) must determine whether LLM consistency differs in degree or kind from human participants before drawing substantive conclusions. Without demonstrating acceptable internal consistency and cross-format stability (via MTDT and MTRC combined with customized classifiers as proxy measurements), findings about simulated attitudes, opinions, or behaviors lack empirical grounding.

Future Work The structural misalignment between human and LLM response patterns raises a deeper question: *why* do LLMs organize constructs so differently from humans? The answer may lie in the internal representations learned during pretraining. Future work should integrate external psychometric assessments (Ye et al., 2026; Münker, 2025a) with circuit-level analysis (Dunefsky et al., 2024), examining how models internally represent and manipulate constructs. This mechanistic approach moves from describing failures based on the output towards understanding their origins, a prerequisite for designing structures and algorithms that reduce misalignment rather than masking it.

Limitations

Model Coverage Our analysis focuses on open-weight models from 1B to 14B parameters, excluding proprietary models like GPT-4, Claude, and Gemini (Ollion et al., 2024). While this choice enables reproducibility and detailed analysis, it limits generalization to the most capable current systems. Proprietary models may exhibit different validity patterns due to advanced training procedures, larger scale, or more sophisticated alignment techniques.

Linguistic and Cultural Scope All instruments in our study are English-language implementations developed and validated primarily in US and European contexts. Construct validity patterns may differ substantially across languages and cultural contexts. Additionally, the constructs themselves (e.g., social dominance orientation, symbolic racism) reflect Western political psychology frameworks that may not translate directly to other cultural contexts.

Instrument Selection While we selected three well-established political psychology instruments covering diverse constructs, numerous other relevant instruments exist. Our findings may not generalize to instruments with different item structures, response formats, or theoretical foundations.

Response Consistency We observe both hyper-consistency (some models showing near-zero variance on specific items, potentially inflating α) and hyper-variability (negative similarity scores indicating pattern inversions). Human psychometric theory assumes moderate test-retest correlation; LLMs violate this by exhibiting extremes depending on temperature and context.

Ecological Validity Even when models achieve acceptable α values, the response patterns may not resemble human construct organization. Future work should compare LLM probability networks to human response covariance structures using techniques like Exploratory Graph Analysis (Golino and Epskamp, 2017) to assess whether the underlying factor structures align.

Nomological Networks Our focus on internal consistency does not address whether measured constructs predict theoretically expected outcomes. A model might show perfect α while failing to predict policy preferences or behavioral intentions that the construct should predict in humans. Establishing nomological validity requires additional experiments beyond the scope of this work.

Ethical Considerations

Potential for Misuse Our work demonstrates how to measure political attitudes in LLMs, which could potentially be misused to: Actors might a) select models based on political alignment with their interests, potentially creating echo chambers or reinforcing existing biases or claim model neutrality (according to our metrics) while deploying systems with harmful political biases, as the connection (Nomological Validity) between our experimental setup and real-world deployment (e.g.; as social bot) remains unclear.

Representation and Bias Our human baseline data comes primarily from US samples, which may not represent global human diversity in political attitudes. Comparing LLMs to these restricted samples risks normalizing Western political perspectives as universal standards. Future work should incorporate more diverse human reference populations and explicitly acknowledge the cultural specificity of constructs.

Additionally, low construct validity on instruments measuring prejudice (e.g., SR2000) might be interpreted positively (safety alignment preventing harmful outputs) or negatively (inability to represent the full spectrum of human attitudes). This ambiguity highlights tensions between safety goals and faithful representation of human psychological diversity.

Research Ethics While our work does not directly involve human participants in the experimental phase, we use human-generated validation data from published sources. We have: (1) Properly cited all original data sources and respected original study permissions, (2) Used only aggregate statistics and de-identified data in our analyses, (3) Acknowledged limitations in demographic representation of baseline samples, (4) Made efforts to obtain diverse model perspectives rather than focusing on single systems.

Application and Downstream-Task

To address the practical utility of our proposed framework and the interpretation of its metrics, it is essential to distinguish between the two dimensions of model stability being measured. While both MTDT and MTRC quantify consistency in sequential interactions, their values signal fundamentally different aspects of LLM behavior.

MTDT, specifically its internal consistency measured via Cronbach's α, reflects the robustness of an LLM's underlying probability distributions when subjected to accumulating context. A high MTDT value indicates that a model maintains a consistent behavioral pattern across branching conversational paths, and this structural stability remains resilient even under varying sampling temperatures. Conversely, a low MTDT value signals that the model's representations are highly entangled with the context window, meaning its simulated attitudes shift erratically depending on prior conversational trajectories rather than representing stable tendencies. Consequently, MTDT is best utilized during model auditing to evaluate foundational alignment and structural robustness independent of specific sampling constraints

In contrast, the MTRC metric measures the stability of a model's reasoning and alignment in open-ended generation, utilizing reward model score trajectories as a proxy for generative coherence. High MTRC values suggest that a model consistently generates high-quality text across multiple turns without degrading or contradicting its prior reasoning. A low MTRC value indicates a vulnerability to context-driven instability, where generative quality collapses as the interaction lengthens or as sampling temperature increases. In practice, MTRC is particularly valuable for application testing, such as evaluating conversational agents where users interact via free text.

Acknowledgments

We thank Veronika Solopova, Kai Kugler, Nils Schwager and Christoph Hau for our constructive discussions. This study was conducted with a financial contribution from the EUs Horizon Europe Framework (HORIZON-CL2-2022-DEMOCRACY-01-07) under grant agreement number 101095095. This work is part of the EVAS research cooperative `https://evas-research.github.io`, a community that challenges assumptions and advances methods for the evaluation of language models as reliable human proxies.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The illusion of artificial inclusion. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating Large Language Models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Jean Baudrillard. 1994. *Simulacra and Simulation*. University of Michigan Press.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. Measuring what matters: Construct validity in Large Language Model benchmarks. *Preprint*, arXiv:2511.04703.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Analysis*, 32(4):401–416.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing ai-assisted group decision making through LLM-powered devil's advocate. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, IUI '24, pages 103–119, New York, NY, USA. Association for Computing Machinery.

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

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J. Gross, and James W. Pennebaker. 2023. Using Large Language Models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. RAFT: reward ranked finetuning for generative foundation model alignment. *Preprint*, arXiv:2304.06767.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable LLM feature circuits. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Elle. 2025. Reward model perspectives: Whose opinions do reward models reward? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14920–14944. Association for Computational Linguistics.

David Exler, Mark Schutera, Markus Reischl, and Luca Rettenberger. 2025. Large means left: Political bias in Large Language Models increases with their number of parameters. *Preprint*, arXiv:2505.04393.

Mats Faulborn, Indira Sen, Max Pellert, Andreas Spitz, and David Garcia. 2025. Only a little to the left: A theory-grounded measure of political bias in Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31684–31704, Vienna, Austria. Association for Computational Linguistics.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Hudson F. Golino and Sacha Epskamp. 2017. Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLOS ONE*, 12(6):1–26.

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.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, and 7 others. 2020. Array programming with NumPy. *nature*, 585(7825):357–362.

Patrick J Henry and David O Sears. 2002. The symbolic racism 2000 scale. *Political psychology*, 23(2):253–283.

Arnold K Ho, Jim Sidanius, Nour Kteily, Jennifer Sheehy-Skeffington, Felicia Pratto, Kristin E Henkel, Rob Foels, and Andrew L Stewart. 2015. The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new sdo scale. *Journal of personality and social psychology*, 109(6):1003.

D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge. 1993. Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, and 196 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for Large Language Model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pages 611–626, New York, NY, USA. Association for Computing Machinery.

Maik Larooij and Petter Törnberg. 2025. Do Large Language Models solve the problems of agent-based modeling? a critical review of generative social simulations. *Preprint*, arXiv:2504.03274.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8412–8452, Albuquerque, New Mexico. Association for Computational Linguistics.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2026. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *Preprint*, arXiv:2507.01352.

Yunting Liu, Shreya Bhandari, and Zachary A. Pardos. 2025. Leveraging LLM respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.

Bolei Ma, Xinpeng Wang, Tiancheng Hu, Anna-Carolina Haensch, Michael A. Hedderich, Barbara Plank, and Frauke Kreuter. 2024. The potential and

challenges of evaluating attitudes, opinions, and values in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8783–8805, Miami, Florida, USA. Association for Computational Linguistics.

Wes McKinney. 2011. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions. *Preprint*, huggingface.co/datasets:nvidia/Nemotron-Personas-USA.

Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. 2024. More human than human: Measuring ChatGPT political bias. *Public Choice*, 198(1):3–23.

Mervin E. Muller. 1959. A note on a method for generating points uniformly on n-dimensional spheres. *Commun. ACM*, 2(4):19–20.

Simon Münker. 2025a. Fingerprinting LLMs through survey item factor correlation: A case study on humor style questionnaire. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 245–258, Suzhou, China. Association for Computational Linguistics.

Simon Münker. 2025b. Political Bias in LLMs: Unaligned Moral Values in Agent-centric Simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138.

Simon Münker, Nils Schwager, and Achim Rettinger. 2026. Don't trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1141–1151, Rabat, Morocco. Association for Computational Linguistics.

Jum C. Nunnally. 1975. Psychometric theory. 25 years ago and now. *Educational Researcher*, 4(10):7–21.

Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2024. The dangers of using proprietary LLMs for research. *Nature Machine Intelligence*, 6(1):4–5.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of LLMs to simulate human psychological behaviours: a psychometric analysis. *Preprint*, arXiv:2405.07248.

Tenko Raykov and George A. Marcoulides. 2011. *Introduction to Psychometric Theory*. Routledge, New York.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Kirk, Hinrich Schuetze, and Dirk Hovy. 2024. Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15295–15311, Bangkok, Thailand. Association for Computational Linguistics.

David Rozado. 2023. The Political Biases of ChatGPT. *Social Sciences*, 12(3):148.

Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. 2025. CAPE: Context-aware personality evaluation framework for Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10648–10662, Suzhou, China. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matari. 2025. Personality traits in Large Language Models. *Preprint*, arXiv:2307.00184.

Murray Shanahan. 2026. Simulacra as conscious exotica. *Inquiry*, 69(6):2982–3010.

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.

Lorenz Sparrenberg, Tobias Schneider, Tobias DeuSSer, Markus Koppenborg, and Rafet Sifa. 2024. Correcting systematic bias in LLM-generated dialogues using big five personality traits. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3061–3069.

Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic biases in LLM simulations of debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 251–267, Miami, Florida, USA. Association for Computational Linguistics.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large Language Models (LLM) in computational social science: Prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):4.

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.

Yuji Ueda and Hirohiko Suwa. 2025. Bias Mitigation and Shifting Ideologies of LLM by Prompting. In *HCI International 2025 Posters*, pages 424–435, Cham. Springer Nature Switzerland.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 93 others. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272.

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2025. Large Language Models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411.

Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for LLMs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Haoran Ye, Jing Jin, Yuhang Xie, Xin Zhang, and Guojie Song. 2026. Large Language Model psychometrics: A systematic review of evaluation, validation, and enhancement. *Preprint*, arXiv:2505.08245.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2025. A survey on recent advances in LLM-based multi-turn dialogue systems. *ACM Comput. Surv.*, 58(6).

A Reproducibility and Code Availability

All experimental procedures, statistical analyses, and visualization tools are implemented using Python 3.13+ with standard scientific computing libraries (NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), pandas (McKinney, 2011), scikit-learn (Pedregosa et al., 2011), seaborn (Waskom, 2021), Plotly). LLM inference uses the vLLM (Kwon et al., 2023) library with Hugging Face Transformers (Wolf et al., 2020) for model loading. Complete code implementations, experimental configurations, processed datasets, and analysis notebooks are available through the following GitHub repository to allow replication and extension of our work: `https://github.com/simon-muenker/LLM-Centered-Constructs`.

B Psychological Instruments

We select two established instruments from political psychology, chosen for their theoretical importance, psychometric validation in human populations, and coverage of diverse political constructs. Table 2 in the Appendix presents detailed psychometric properties from validation studies with human participants.

B.1 Social Dominance Orientation Scale (SDO-7)

The SDO-7 scale (Ho et al., 2015) measures individual differences in the preference for inequality between social groups. This 16-item instrument captures two theoretically distinct dimensions: SDO-Dominance (preference for systems of group-based dominance and hierarchies maintained through overt oppression) and SDO-Egalitarianism (opposition to equality between groups, often expressed through opposition to policies promoting equality). The scale uses a 7-point response format (strongly oppose to strongly favor) and demonstrates strong psychometric properties across diverse populations.

Instrument	Construct	Population	μ	σ	α
Social Dominance Orientation Scale SDO-7 (Ho et al., 2015)	SDO-Dominance	White (US)	2.83	1.24	0.88
		Black (US)	2.52	1.18	0.83
	SDO-Egalitarianism	White (US)	2.63	1.32	0.91
		Black (US)	2.38	1.21	0.82
Symbolic Racism 2000 Scale SR2000 (Henry and Sears, 2002)	Symbolic Racism	White	0.49	0.20	0.77
		Black	0.36	0.16	0.55
		Asian	0.53	0.17	0.74
		Latino	0.36	0.16	0.42

Table 2: Human baseline psychometric properties across all psychological instruments used in the study.

SDO-7 - Dominance *Instruction:* Show how much you favor or oppose each idea below by selecting a number from 1 to 7 on the scale below. You can work quickly; your first feeling is generally best. *Response Scale:* 1 = Strongly Oppose, 2 = Somewhat Oppose, 3 = Slightly Oppose, 4 = Neutral, 5 = Slightly Favor, 6 = Somewhat Favor, 7 = Strongly Favor

1. Some groups of people must be kept in their place. [+]
2. It's probably a good thing that certain groups are at the top and other groups are at the bottom. [+]
3. An ideal society requires some groups to be on top and others to be on the bottom. [+]
4. Some groups of people are simply inferior to other groups. [+]
5. Groups at the bottom are just as deserving as groups at the top. [−]
6. No one group should dominate in society. [−]
7. Groups at the bottom should not have to stay in their place. [−]
8. Group dominance is a poor principle. [−]

SDO-7 - Anti-egalitarianism *Instruction:* Show how much you favor or oppose each idea below by selecting a number from 1 to 7 on the scale below. You can work quickly; your first feeling is generally best. *Response Scale:* 1 = Strongly Oppose, 2 = Somewhat Oppose, 3 = Slightly Oppose, 4 = Neutral, 5 = Slightly Favor, 6 = Somewhat Favor, 7 = Strongly Favor

1. We should not push for group equality. [+]
2. We shouldn't try to guarantee that every group has the same quality of life. [+]
3. It is unjust to try to make groups equal. [+]
4. Group equality should not be our primary goal. [+]
5. We should work to give all groups an equal chance to succeed. [−]
6. We should do what we can to equalize conditions for different groups. [−]
7. No matter how much effort it takes, we ought to strive to ensure that all groups have the same chance in life. [−]
8. Group equality should be our ideal. [−]

Note: Negative weights indicate reversed items.

B.2 Symbolic Racism 2000 Scale (SR2000)

The SR2000 scale (Henry and Sears, 2002) assesses contemporary forms of racial prejudice expressed through seemingly race-neutral political and social attitudes. This instrument measures symbolic racism through four key themes: work ethic (beliefs about Black Americans' work ethic), excessive demands (perceptions that Black Americans demand too much), undeserved advantage (beliefs that Black Americans receive unfair benefits), and denial of continuing discrimination.

1. It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites. *Scale:* 1 = strongly agree, 2 = somewhat agree, 3 = somewhat disagree, 4 = strongly disagree
2. Irish, Italian, Jewish, and many other minorities overcame prejudice and worked their way up. Blacks should do the same. *Scale:* 1 = strongly agree, 2 = somewhat agree, 3 = somewhat disagree, 4 = strongly disagree
3. Some say that black leaders have been trying to push too fast. Others feel that they haven't pushed fast enough. What do you

think? *Scale:* 1 = trying to push too fast, 2 = going too slowly, 3 = moving at about the right speed

4. How much of the racial tension that exists in the United States today do you think blacks are responsible for creating? *Scale:* 1 = all of it, 2 = most, 3 = some, 4 = not much at all
5. How much discrimination against blacks do you feel there is in the United States today, limiting their chances to get ahead? *Scale:* 1 = a lot, 2 = some, 3 = just a little, 4 = none at all
6. Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class. *Scale:* 1 = strongly agree, 2 = somewhat agree, 3 = somewhat disagree, 4 = strongly disagree
7. Over the past few years, blacks have gotten less than they deserve. *Scale:* 1 = strongly agree, 2 = somewhat agree, 3 = somewhat disagree, 4 = strongly disagree
8. Over the past few years, blacks have gotten more economically than they deserve. *Scale:* 1 = strongly agree, 2 = somewhat agree, 3 = somewhat disagree, 4 = strongly disagree

C Reward Models

Reward models are trained to predict human preferences over generated responses. The training data consists of pairwise comparisons (x, y_p, y_l), where x is a prompt and y_p and y_l are the preferred and less preferred responses, annotated by humans. The models are trained using the Bradley–Terry formulation of pairwise preference learning, which calculates a relative preference probability for each pair of answers [$p = \sigma(r(x, y_1) - r(x, y_2))$]. Thus, for each prompt-response pair, reward models output a scalar score indicating how well the response fits the prompt.

SkyworkRewardV2 The models were trained on SynPref40M, a largescale preference dataset containing approximately 40 million preference pairs with a broad range of prompt categories, including Information seeking, Coding & Debugging and Advice seeking (Liu et al., 2026).

GRM-Gemma2-2B It was trained using the HH-RLHF dataset by Ganguli et al. (2022), which contains human preference comparisons for helpful and harmless LLM responses (Dong et al., 2023).

RM-Mistral-7B The model uses the Unified-Feedback dataset, a large-scale collection of diverse human preference pairs covering conversational, reasoning, and safety-oriented prompts (Yang et al., 2024).

D Temperature Ablation Study

A core assumption underlying both MTDT and MTRC is that the metrics capture structural properties of the model, what we term *architectural stability* (see Sec. 4), rather than artifacts of the sampling procedure. This distinction matters: if observed consistency scores were primarily governed by the entropy of sampling rather than by the model's underlying representation structure, the cross-metric α correlation reported in Sec. 4 could be attributed to a shared sensitivity to temperature rather than to a shared architectural property. Falsifying this alternative explanation is the primary purpose of this ablation.

We vary the sampling temperature $T \in \{0.0, 0.2, \ldots, 2.0\}$ and examine how Cronbach's α changes for both metrics. Temperature controls the entropy of the token-level probability distribution: at $T = 0$ inference is deterministic (greedy decoding), while increasing T introduces progressively more stochastic variation. We conduct the ablation using Qwen 3 14B across all three psychological instruments, as this model achieves the highest MTDT α in the main analysis and therefore provides the strongest test case. For MTRC, we apply all four reward models independently and report per-scorer α values. For MTDT, α is computed on the aggregated response profile across the full decision tree (see Sec. 3.4).

Findings

Figure 3 presents Cronbach's α as a function of temperature for both metrics and all instruments. The results reveal a asymmetry between the two metrics that bears directly on the validity argument in the main paper.

MTDT α reflects model structure. Across all three instruments, Likert-based α values remain stable as temperature increases from 0.0 to 2.0. For SDO-Dominance, α fluctuates narrowly between 0.734 and 0.754 across the full temperature range, a spread of less than 0.02. For SDO-Anti-egalitarianism, the range is similarly narrow (0.591–0.638). The SR2000 instrument shows uniformly low α (0.066–0.101) at all tempera-

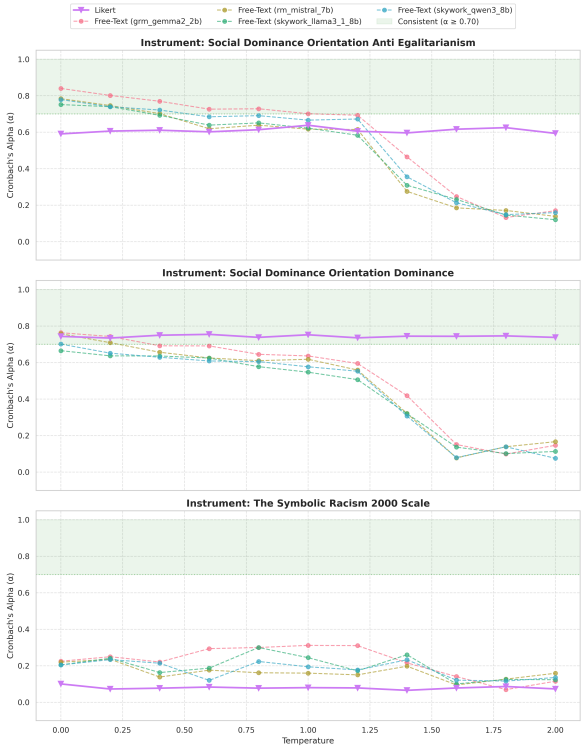

Figure 3: Cronbach's α as a function of sampling temperature for Qwen 3 14B across three psychological instruments. The shaded region indicates the conventional acceptability threshold ($\alpha \geq 0.70$).

tures. This stability has an explanation rooted in MTDT's design: the method constructs decision-tree branches from logits extracted before temperature rescaling is applied. Temperature modulates which tokens are ultimately sampled but does not alter the relative ordering of the probability mass across response options.

MTRC α tracks real-world generative coherence. In contrast, free-text α values show a pronounced and consistent decline as temperature rises, with the relationship becoming markedly non-linear above $T = 1.2$. For SDO-Anti-egalitarianism, α for the GRM-Gemma2-2B scorer drops from 0.840 at $T = 0.0$ to 0.170 at $T = 2.0$; the RM-Mistral-7B scorer declines from 0.784 to 0.139 over the same range. The same monotonic degradation holds for SDO-Dominance (GRM-Gemma2-2B: 0.763 to 0.146) and SR2000, though the SR2000 baseline is already low given the instrument's own psychometric properties. All four reward models exhibit this qualitative pattern, ruling out scorer-specific sensitivity as an explanation and reinforcing the ensemble rationale described in Sec. 3.5.

This sensitivity is not a methodological weakness. MTRC measures the coherence of generated text as evaluated by reward models; the quality of that text is directly affected by how much entropy is injected at the sampling stage. As temperature rises, token-level decisions depart further from the distribution that sustained the model's reasoning trajectory at prior turns, producing free-text responses whose surface form becomes progressively less coherent. The reward model registers this as reduced response quality, leading to higher inter-item variance and lower α.

Crucially, this temperature sensitivity means that MTRC makes deployment-relevant instabilities measurable. A practitioner using an LLM as a synthetic survey respondent at $T = 1.4$ will observe degraded consistency in open-ended responses even if the model's probability distributions are structurally stable. The fact that the two metrics diverge under temperature perturbation is therefore a feature of the dual-metric design, not a confound: they are sensitive to different failure modes, and together they triangulate architectural stability from two independent directions. Their convergence in α at low temperature (the main analysis setting) is the construct-validity signal; their divergence under temperature stress confirms that the convergence is not artifactual.

Practical temperature range for psychometric applications. From a practical standpoint, temperatures in the range $T \in [0.0, 1.2]$ preserve acceptable MTRC α for the SDO instruments in the present study, while values above $T = 1.4$ produce a rapid collapse in free-text coherence. We recommend this range as a reference for future work employing LLMs in psychometric contexts with free-text elicitation. For researchers requiring temperature-invariant measurements of response structure, for instance, when comparing models across configurations, MTDT provides robust estimates throughout the full temperature range. Taken together, the two metrics offer complementary operating regimes: MTDT for structural comparisons insulated from sampling variance, and MTRC for deployment-realistic assessments that are explicitly sensitive to the entropy of the generation process.