RADIUS: Ranking, Distribution, and Significance—A Comprehensive Alignment Suite for Survey Simulation

Weronika Łajewska
Amazon
Luxembourg, Luxembourg
lajewska@amazon.lu

Paul Missault
Amazon
Luxembourg, Luxembourg
pmissaul@amazon.lu

George Davidson
Amazon
Luxembourg, Luxembourg
geodavix@amazon.lu

Saab Mansour
Amazon
Barcelona, Spain
saabm@amazon.es

Abstract

Simulating survey responses with LLMs offers a scalable alternative to costly human data collection, yet evaluation methods remain ad hoc and inconsistent. Current metrics, typically borrowed from adjacent domains, focus on accuracy or distributional similarity while overlooking ranking alignment—whether simulations preserve the relative preference orderings observed in human populations. A simulator may achieve low distributional divergence yet fail to identify the option humans most prefer, a critical gap for decision-making applications. We propose RADIUS, a two-dimensional alignment suite that evaluates survey simulation along complementary axes: 1) **RA**nking alignment and 2) **DI**strib**U**tion alignment, accompanied by statistical **S**ignificance testing. Experiments across multiple social surveys and persona sources demonstrate that RADIUS exposes limitations of single-metric evaluation and offers stronger discriminative power. On an independently reproduced simulator, we further show that distributional improvement often coincides with misidentifying the option humans most prefer—a decision-relevant error RADIUS surfaces but distributional metrics miss. We open-source RADIUS to support reproducible and comparable assessment of survey simulators.[1]

1 Introduction

Surveys are a cornerstone for understanding human perspectives, from large-scale social science studies to targeted e-commerce systems. However, their use is fundamentally constrained by the availability and representativeness of respondents. This limitation has motivated a growing line of work using LLMs to simulate survey responses at scale [Gao et al., 2024a, Park et al., 2024, Cao et al., 2025]. Although LLMs enable agent-based simulation with controllable heterogeneity [Gao et al., 2024a], modeling realistic user behavior remains challenging due to the complexity of human decision-making processes. A general framework for LLM-based simulation includes three stages: 1) behavioral data collection to "train" a simulator, 2) behavioral data generation using the simulator, 3) evaluation to measure divergence from human responses (see Figure 1). We focus on the simulation alignment evaluation (stage 3) to address the lack of a standardized evaluation framework. Survey

[1] https://github.com/amazon-science/survey_sim_alignment_suite

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Workshop: Grounded User Simulation for Model Evaluation and Training: Diversity, Fidelity, and Validity.

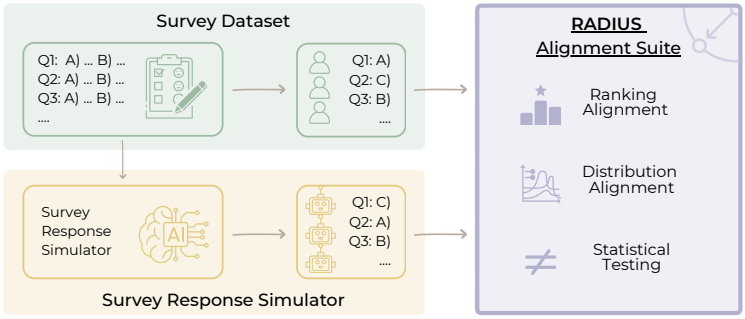


Figure 1: User simulation framework for surveys.

simulation is a population-level instance of user simulation: the simulator must reproduce how a population responds rather than a single interaction. This setting offers abundant aggregated human ground truth, making it well suited to studying how faithfully simulators match real responses—provided that simulation fidelity is measured appropriately.

Most agent-based approaches simulate user behavior by assigning personas to steer model behavior [Wang et al., 2024, Tseng et al., 2024, Hu and Collier, 2024]. This shift toward persona-based simulation changes the evaluation problem: unlike earlier approaches that focused primarily on matching aggregate response distributions, persona-based LLMs generate individual-level responses that must collectively reflect population-level behavior. Recent work has demonstrated the promise of persona-based LLMs for survey simulation [Park et al., 2024], yet evaluation methods have not kept pace with this shift. Prior studies either evaluate alignment at the individual level using accuracy [Wang et al., 2025a], or at population level using distributional metrics such as KL divergence [Namikoshi et al., 2024], Wasserstein distance [Suh et al., 2025], or Cramer's V [Argyle et al., 2023]. However, individual-level metrics require one-to-one correspondence between agents and real respondents, which is rarely available, while population-level human responses exhibit complex patterns and variability not fully captured by any single metric. As a result, evaluation practices vary widely across studies, with metrics chosen ad hoc, making systematic comparison difficult. Together, these limitations underscore the need to rethink survey simulation evaluation and motivate a comprehensive alignment framework.

We propose RADIUS, a population-level alignment suite for survey simulation that defines simulation success across two complementary dimensions: 1) *ranking alignment* measuring the preservation of relative option ordering and top-option alignment, and 2) *distribution alignment* quantifying the probability "mass" shift required to match agent and human distributions and testing for statistical differences between the two. RADIUS enables systematic quality assessment, identifies gaps in survey simulation performance in terms of different alignment dimensions, and enhances interpretability by clarifying tool capabilities for informed use.

We validate RADIUS on over 300 survey questions spanning diverse topics, including politics, family, food preferences, and religion. We find that survey simulation alignment is multi-dimensional: ranking alignment is generally easier to achieve than distribution alignment, while distribution homogeneity is the strictest criterion, exposing even subtle mismatches. Our ranking metrics show higher discriminative power than commonly used alternatives, separating persona-based simulations from non-parametric baselines across most benchmarks—differences that distributional metrics alone often fail to detect. Qualitative analysis confirms that ranking and distribution metrics expose complementary failure modes, motivating an evaluation suite that jointly considers both perspectives rather than relying on any single metric. Reproducing an independent simulator at scale, we further show that improving a distributional metric frequently coincides with misidentifying the human-preferred option—a verifiable, decision-relevant error that ranking evaluation catches and a distributional score alone cannot.

Overall, this paper introduces RADIUS, a multi-dimensional evaluation framework for LLM-based survey simulation. Our contribution is methodological: we synthesize complementary metrics—including Top Rank Match (TRM), a new tie-aware top-choice agreement metric—into a unified suite for measuring population-level simulation fidelity, demonstrate its discriminative power over single-score evaluation across domains and persona sources, and provide an open-source implementation for reproducible assessment.

2 Related Work

User simulation uses intelligent agents to mimic human interactions with systems [Balog and Zhai, 2024], supporting applications such as automatic evaluation [Jung et al., 2009, Abbasiantaeb et al., 2024], behavioral analysis [Hazrati and Ricci, 2024, Park et al., 2024], experiment piloting [Filippas et al., 2024], social scenario simulation [Gao et al., 2024b, Park et al., 2022], and synthetic data generation [Balog et al., 2025]. Effective simulation requires validity, cognitive plausibility, and behavioral variation [Balog and Zhai, 2024, Bignold et al., 2021]. Recent advances in LLMs have significantly improved the realism of user simulators [Wang et al., 2025b, Argyle et al., 2023, Aher et al., 2023], enabling agents to assume specific roles through prompting [Wang et al., 2024, Tseng et al., 2024, Hu and Collier, 2024, Park et al., 2024]. These roles are typically defined by user profiles that include demographic attributes, personality traits, and social context tailored to the target application [Chen et al., 2024, Wang et al., 2024].

User simulation is particularly valuable in social science research, where collecting human responses is costly. Park et al. [2024] showed that generative agents can reproduce survey responses with 85% accuracy, comparable to participants' own two-week consistency. Existing approaches include persona-based response generation [Park et al., 2024] and LLMs fine-tuned for multiple-choice questions [Cao et al., 2025]. Simulators may operate at the individual level to replicate specific user behaviors [Park et al., 2024, Wang et al., 2025a], or at the population level to generate response distributions for targeted sub-populations [Argyle et al., 2023, Cao et al., 2025, Suh et al., 2025]. While surveys span diverse formats, including Likert-scale, multi-select, ranking, and open-ended questions, this work focuses on multiple-choice single-select questions to enable more precise modeling and evaluation, leaving open-ended survey simulation to future work [Ma et al., 2025, 2026].

Despite promising results, replicating the complexity of real human opinions with LLM-based agents remains challenging [Kaiser et al., 2025], and simulated responses often show less variation than real survey data [Bisbee et al., 2024]. A central unresolved question is how to evaluate the alignment between simulated and human responses. Survey simulation alignment is influenced by multiple factors, including the question domain [Meister et al., 2025], persona design [Lutz et al., 2025], steering methods [Meister et al., 2025, Bui et al., 2025], and response representation (e.g., single choice, distributions, log-probabilities) [Meister et al., 2025]. Given the many factors affecting simulation alignment, developing consistent evaluation methods is challenging. Existing approaches fall into two categories. Individual-level alignment relies on metrics such as accuracy [Park et al., 2024, Wang et al., 2025a] and RMSE [Namikoshi et al., 2024]. Population-level alignment uses distributional measures, including Total Variation Distance (TVD) [Hu et al., 2026], SimBench score (SBS) [Hu et al., 2026] (TVD normalized against a uniform baseline), Kullback–Leibler divergence (KLD) [Namikoshi et al., 2024], Jensen–Shannon divergence (JSD) [Cao et al., 2025], Wasserstein distance (WD) [Cao et al., 2025, Suh et al., 2025], Cramer's V correlation (CV) [Argyle et al., 2023], ANOVA [Park et al., 2024, Kaiser et al., 2025], and model log-probabilities [Meister et al., 2025]. The prior work closest to our top-choice metric reports accuracy against the single human majority option [Cao et al., 2025], but without accounting for statistically tied top options; separately, distribution-level tests assess whether human and simulated distributions differ as a whole [Sun et al., 2024]. Our TRM is tie-aware, crediting any human top option statistically indistinguishable from the most-voted one via bootstrap grouping. Each of these metrics captures only partial aspects of alignment, and the lack of standardization hinders comparison across studies. We address this gap by proposing a dedicated alignment suite for evaluating survey simulations.

3 Survey Simulation

Since our goal is establishing a reusable evaluation framework rather than improving simulation itself, we employ a standard simulation environment with persona-based prompting, leaving more sophisticated approaches for future work [Tseng et al., 2024, Chen et al., 2024, Hu and Collier, 2024]. The core component is a population of agents generated from behavioral data. To assess how persona source affects alignment, we construct personas from three distinct domains: (1) social science survey responses (Soc) [Kolluri et al., 2025], providing attitudinal and demographic grounding; (2) Rotten

Tomatoes movie critic reviews (RT), [2] capturing cultural preferences and judgment patterns; and (3) e-commerce purchase histories (E-c) [Berke et al., 2024], reflecting consumer behavior. Each agent's persona encompasses demographics, interests, and behavioral patterns [Lutz et al., 2025], with the specific composition determined by the source domain. Personas derived from behavioral data provide context for an LLM to role-play an individual with specific traits [Chen et al., 2024], selecting the response that best aligns with portrayed character.[3] We use Claude Sonnet 4.5 [Anthropic, 2025] as the simulator across all persona sources and benchmarks.[4] Since the alignment suite is model-agnostic, it applies unchanged to any simulator that produces response distributions.

4 RADIUS Alignment Suite

Our evaluation approach goes beyond identifying the "winner" option to assess whether simulated surveys capture the patterns and variability of human responses. The RADIUS alignment suite operates at multiple levels: question-level metrics compare response rankings and distributions, survey-level aggregations synthesize overall alignment, and statistical tests quantify the significance of differences. The ultimate goal is to ensure that simulated data reflects both the general human patterns and natural individual variability. Alignment is defined via two complementary criteria:

- *Ranking alignment* measuring how well the relative ordering of options is preserved and verifying the top-choice alignment.
- *Distribution alignment* quantifying the probability mass needed to transform agent distributions into human distributions and testing for the statistically significant differences between them.

4.1 Ranking Alignment

We evaluate ranking alignment at two levels: 1) top-choice agreement between humans and agents, and 2) agreement in the relative ordering of all answer options.

Top Rank Match (TRM) TRM quantifies whether agents correctly identify the top-ranked human preferences, accounting for statistical significance in human votes. It is binary, with 1 indicating perfect alignment with human top choice within statistical bounds. The metric recognizes that when humans cannot distinguish between their first and second choices, agents should not be penalized for selecting the second-ranked option. Conversely, when humans show a significant preference for their top choice over their second choice, agents are penalized for predicting the second-ranked option. To take confidence of human top choice into account, we create option groups using confidence intervals identified with bootstrap significance testing [Efron and Tibshirani, 1994], which is a model-free approach that extends to all types of questions and setups. For each question's human vote distribution, we perform bootstrap resampling ($n = 1000$) to determine confidence intervals and identify statistically tied options. Next, we identify the top option group by clustering options whose confidence intervals overlap with the highest-voted option. Finally, we evaluate binary top rank alignment where the simulator prediction is considered correct if its top choice falls within the human top choice option group. This metric is intuitive and easily interpretable, handles ties in human preferences via bootstrap confidence intervals, supports per-question analysis with option groups, and is robust to small sample sizes through bootstrapping. TRM generalizes the majority-choice accuracy used in prior work [Cao et al., 2025]: rather than binary top-1 agreement with the single most-voted human option, it credits any statistically indistinguishable top option via bootstrap grouping, avoiding unfair penalties when human first and second choices are tied.

Rank Correlation (RC) RC quantifies the alignment between the option ordering induced by agent responses and that induced by human responses. We implement this metric using the Spearman correlation coefficient, which measures the strength and direction of monotonic relationships between

[2] `https://www.kaggle.com/datasets/stefanoleone992/rotten-tomatoes-movies-and-critic-reviews-dataset`

[3] The simulation setup serves as a testbed for the alignment suite rather than a contribution in itself; our focus is on evaluation methodology, not simulator optimization.

[4] `anthropic.claude-sonnet-4-5-20250929` via Amazon Bedrock, decoding temperature 0 for reproducible outputs; each of the 1,000 persona-based agents per source answers independently, invalid answers are discarded, and per-question response distributions are aggregated across the persona population.

two ranked variables. The coefficient is calculated as: $RC = 1 - \frac{6\sum d^2}{n(n^2-1)}$, where d represents the absolute difference between ranks for each pair of data points and n is the total number of response options. The Spearman correlation coefficient analyzes the ranks of data points rather than raw values, making it particularly suitable for non-normally distributed data, ordinal data, or datasets containing outliers. To facilitate the interpretation of the results, we normalize the coefficient ($RC_{norm} = \frac{RC+1}{2}$), resulting in a scale where 1 indicates perfect positive correlation and 0 indicates no correlation between human and agent rankings. Compared to TRM, RC provides a softer measure of rank distance, with larger penalties for greater rank discrepancies from human responses.

4.2 Distribution Alignment

While ranking metrics capture the correct ordering of options, they are insensitive to prevalence: the distributions [0.51, 0.49] and [0.99, 0.01] have perfect ranking scores despite representing very different response frequencies. Distribution alignment metrics fill this gap.

Total Variation Distance (TVD) TVD quantifies the minimum percentage of probability mass that needs to be redistributed to make two response distributions identical, calculated as: TVD $= \frac{1}{2}\sum_{i=1}^{n} |H_i - A_i|$, where H represents the human response distribution, A represents the simulated response distribution, and n is the number of response options [Hu et al., 2026]. In principle, TVD represents the maximum possible difference between the probability of any event occurring under two different distributions. It ranges from 0 to 1, with lower values indicating better simulation fidelity. Unlike alternatives such as KL divergence, TVD remains well-defined when models assign zero probability to human responses, while also offering symmetry and boundedness for improved interpretability across datasets [Hu et al., 2026]. TVD is highly sensitive to truly categorical questions and large shifts in popular options, but less sensitive to ordered responses or small changes spread across many options.

Distribution Homogeneity (DH) The chi-square test of homogeneity assesses whether categorical response distributions are the same across different populations (agents vs humans) [Sun et al., 2024]. The null hypothesis assumes the same distributions across populations, while the alternative hypothesis states they differ. Chi-squared test p-values below 0.05 indicate a statistically significant difference between the distributions. We report DH as a binary value, with 0 indicating significant distributional mismatch for a given question. DH is complementary to TVD, as it tests whether simulated and human response distributions are statistically distinguishable rather than measuring the magnitude of their difference.

4.3 Survey-Level Aggregation and Significance Testing

Survey-level alignment is computed by averaging question-level scores for each metric. To assess differences between simulator runs, we apply paired t-tests to question-level scores, which naturally pair Simulator A and Simulator B responses for each question, testing whether the mean difference is significantly different from zero.

5 Results

To validate the sensitivity of our metrics, we evaluate simulated survey responses using RADIUS alongside metrics from prior work (Table 1) across surveys spanning diverse domains and question complexity. Our test set includes the General Social Survey (GSS; political views/religion/science/crime) and two waves of OpinionQA [Santurkar et al., 2023]: W50 (relationships/family) and W92 (political views).[5]

5.1 Relative Strictness of Alignment Criteria

We observe a clear hierarchy in alignment difficulty: ranking alignment is generally easier to achieve than distribution alignment, while DH is the most stringent criterion, requiring near-perfect agreement with no statistical differences. In contrast, TVD captures the magnitude of disagreement between

[5] Additional results on W34 (biomedical/food) are consistent with reported findings.

Table 1: Survey-level simulation alignment across persona sources (Soc, RT, E-c), test survey benchmarks (GSS, W50-92), and non-parametric baseline (Nor). Best per benchmark; * = significantly different from best ($p < 0.05$, paired t-test).

	GSS				W50				W92			
Metric	Soc	RT	E-c	Nor	Soc	RT	E-c	Nor	Soc	RT	E-c	Nor
RADIUS (proposed alignment suite)												
TRM ↑	0.70	0.67	0.77	0.45*	0.35*	0.27*	0.48	0.39	0.65	0.49*	0.58	0.36*
RC ↑	0.82	0.76	0.83	0.61*	0.53*	0.49*	0.62*	0.71	0.87	0.68*	0.76*	0.68*
TVD ↓	0.31	0.38*	0.34*	0.25	0.57*	0.66*	0.49*	0.38	0.28	0.46*	0.45*	0.31
DH ↑	0.03	0.00	0.02	0.03	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Prior work (for comparison)												
CV ↓	0.32	0.41*	0.37*	0.27	0.59*	0.67*	0.52*	0.42	0.21	0.37*	0.34*	0.35*
JSD ↓	0.09*	0.13*	0.12*	0.05	0.31*	0.37*	0.24*	0.13	0.07	0.19*	0.18*	0.09*
WD ↓	0.38	0.55*	0.49*	0.38	1.10*	1.33*	0.91*	0.69	0.35	0.70*	0.64*	0.57*
SBS ↑	-64.7	-119.2	-98.5	-0.1	-55.2	-76.2	-30.4	7.4	10.4	-49.2	-45.7	7.8

Table 2: Example questions illustrating complementary failure modes, with human and agent response counts and computed metrics. *Src* denotes the persona source (Soc: social science, RT: Rotten Tomatoes, E-c: e-commerce) and survey benchmark (GSS, W34/W50/W92 OpinionQA waves).

			RADIUS				Prior work		
Question	Src	Human/Agent Counts	TRM	RC	TVD	DH	WD	CV	JSD
1. Please tell us whether you are satisfied or dissatisfied with your social life.	Soc/W50	[3011, 4590, 1770, 447, 16] [1, 919, 22, 0, 57]	1.00	0.65	0.51	0.00	0.52	0.34	0.21
2. How often do you choose foods to eat because they are healthy and nutritious	RT/W34	[310, 1024, 784, 368, 46, 5] [5, 354, 453, 188, 0, 0]	0.00	0.96	0.19	0.00	0.33	0.23	0.05
3. Are we spending too much, too little, or about the right amount on solving the problems of big cities?	RT/GSS	[260, 410, 921] [5, 11, 984]	1.00	1.00	0.41	0.00	0.56	0.45	0.14
4. What about sexual relations between two adults of the same sex–do you think it is...	E-c/GSS	[741, 84, 137, 1163] [2, 0, 0, 998]	1.00	0.97	0.45	0.00	1.18	0.46	0.18
5. Do you restrict or limit eating sugar on a regular basis?	E-c/W34	[1455, 1082] [342, 658]	0.00	0.00	0.23	0.00	0.23	0.21	0.03
6. Which statement comes closer to your view?	Soc/W50	[2603, 2256, 48] [449, 537, 13]	0.00	0.75	0.08	0.00	0.08	0.06	0.00

distributions, providing a graded assessment even when differences are statistically significant. Ranking-based metrics are less strict because they depend only on the relative ordering of response options rather than their exact probability mass. As a result, substantial distributional mismatches can coexist with preserved rankings between human and agent responses. Among ranking metrics, RC is easier to satisfy than TRM: high correlation can be achieved even when the most preferred option differs, whereas top-rank agreement requires exact matching of the highest-ranked choice. Together, these results highlight the complementary roles and differing strictness of alignment metrics, motivating the need for a multi-dimensional evaluation framework.

5.2 Discriminative Power of Metrics

Alignment varies substantially by topic, question complexity, and persona source. On W92, all RADIUS metrics identify Soc as the best configuration (or tied on DH). In contrast, on GSS and W50 no single configuration is best across all metrics: persona sources achieve the best top-rank alignment (TRM) while the normal baseline achieves the best distributional alignment. A single-score metric would collapse this tradeoff into a misleading verdict: the SimBench score [Hu et al., 2026], which measures distributional improvement over a uniform baseline, is negative for 8 of 9 persona configurations, meaning distributions are further from human responses than uniform guessing, despite strong ranking alignment (TRM up to 0.77, RC up to 0.87). RADIUS exposes this

tradeoff, enabling informed choices based on application requirements—a distinction that single-score evaluations would obscure.

5.3 Alignment Metrics Robustness

To assess metric robustness, we compare persona-based simulations against a normal distribution baseline (Nor), which we found consistently outperforms random (Dirichlet-Multinomial) and uniform alternatives (full results for all three baselines in Appendix B). Alignment scores from this baseline are compared to persona-based LLM simulations, testing whether the non-parametric responses are statistically distinguishable from those conditioned in personas. Overall, our ranking metrics (TRM, RC) assign significantly worse scores to baselines than to persona-based simulations on GSS and W92, while commonly used distribution metrics (JSD, CV) often show no significant advantage for persona sources or even favor the baselines. For surveys where LLM simulations are misaligned (W50), the normal baseline sometimes scores better, highlighting cases where LLMs do not outperform trivial baselines. Distribution alignment metrics often favor non-parametric baselines: on GSS and W50, normal distributions achieve lower TVD and JSD than all persona sources, indicating better distributional shape despite the baseline being content-agnostic. This pattern reflects the *alignment-simulation tradeoff* identified by Hu et al. [2026]: instruction-tuned models used for persona role-playing produce confident, decisive outputs, causing mode collapse—concentrating responses on one or two options rather than reproducing the full diversity of human opinion. Together, these results confirm that RADIUS provides more robust discrimination between meaningful simulation and statistical noise than commonly used distributional measures alone.

5.4 Metric Complementarity

We qualitatively examine question scores to assess whether ranking- and distribution-based metrics capture complementary failure modes in survey simulation (see Table 2; examples span all evaluated benchmarks including W34). Across questions, we observe that strong performance under one metric can mask substantial misalignment under another, underscoring the need for multi-dimensional evaluation. In some cases, agreement on the most frequent response masks severe distributional mismatch. For Q1, the simulator achieves perfect TRM (1.00), but moderate RC (0.65) and poor distribution alignment (TVD = 0.51). While both humans and agents most often select the 2nd option, the agent concentrates probability on a narrow subset of options versus the much broader human distribution—a manifestation of mode collapse. Conversely, low top-rank agreement does not necessarily imply poor overall alignment. For Q2, TRM is 0.00, while RC is high (0.96) and TVD is low (0.19), reflecting only a minor swap among the top options rather than significant disagreement in preferences. On the other hand, in Q3, both TRM and RC are perfect (1.00), yet distribution alignment remains poor (TVD = 0.41), indicating that identical orderings can coexist with distribution mismatch. Standard distributional metrics can also give misleading signals in both directions. For Q4, ranking alignment is near-perfect (TRM = 1.00, RC = 0.97), yet WD is high (1.18) as the agent correctly identifies the top option but concentrates nearly all probability mass on it. In the opposite direction, standard distributional metrics can miss ranking errors. For Q5, the agent completely flips the human preference ordering (TRM = 0), yet JSD remains very low (0.03), failing to flag a clear ranking disagreement. Similarly, in Q6, CV is low (0.06) despite the agent selecting the wrong top option (TRM = 0). Notably, DH is 0.00 for all examples, including Q6 where TVD and JSD indicate near-perfect distributional match, demonstrating that DH is the strictest criterion, detecting statistically significant differences even when the magnitude of mismatch is negligible. Altogether, these examples show that ranking and distribution metrics surface distinct and complementary aspects of survey alignment, and that relying on any single metric can lead to misleading conclusions.

5.5 RADIUS Beyond Our Experimental Setup

The analyses above establish metric complementarity within our own experimental setup (Section 5.4). To test whether it holds at scale and beyond our own design choices, we reuse the large-scale World Values Survey data and reproduce a fine-tuned simulator of Cao et al. [2025]—an independently trained model, dataset, and method not designed by us—as a second testbed, and evaluate it with RADIUS.[6] Figure 2 plots, per question, the change in distribution alignment (ΔTVD) against the

[6]Reproduction details, including our fidelity check against their published scores, are in Appendix C.

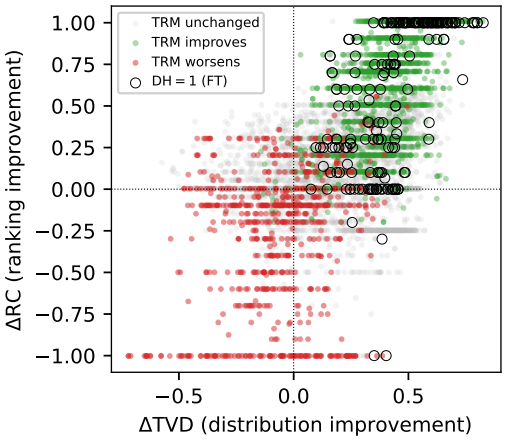


Figure 2: Per-question change after fine-tuning on the Cao et al. [2025] simulator, evaluated with RADIUS. Each point is one question: x = distribution improvement (ΔTVD), y = ranking improvement (ΔRC); colour marks the change in top-rank agreement (TRM), and thin circles mark questions reaching distributional homogeneity (DH=1) after fine-tuning.

change in ranking alignment (ΔRC) after fine-tuning, with the change in TRM shown by colour and questions reaching distributional homogeneity (DH=1) marked. The distribution and ranking axes are correlated but far from equivalent: fine-tuning improves TVD on 82.5% of questions, yet 67.7% of those show no improvement in TRM and 4.9% show TRM *worsening*, while DH is attained almost only where both axes improve.

Crucially, this divergence corresponds to real, verifiable errors rather than mere metric disagreement. Among the questions whose TVD improves after fine-tuning, 45.9% still assign the top rank to an option other than the human population's most-preferred one (TRM=0, verified against the ground-truth response distribution), and 21.4% of all questions do so despite appearing well-aligned distributionally (TVD < 0.3). A practitioner relying on distributional improvement alone would accept these simulators as improved, yet they misidentify the option humans most prefer—precisely the decision-relevant error RADIUS's ranking axis surfaces and a single distributional score cannot. This confirms, on data and a simulator not designed by us, that RADIUS's dimensions capture distinct and decision-relevant information.

6 Conclusions

We introduced RADIUS, a population-level evaluation framework for survey simulation that jointly captures ranking alignment, distribution alignment, and statistical significance through complementary, interpretable metrics supporting discriminative comparison between approaches. Our analyses show that alignment is inherently multi-faceted: ranking and distribution metrics capture related but distinct information, and relying on either alone yields a misleading verdict—persona-based simulations can preserve human preference structure while exhibiting mode collapse in their distributions, or match a distribution's shape while misidentifying the most-preferred option. We observe this complementarity both across our own persona sources and benchmarks and, independently, when reproducing a fine-tuned simulator not designed by us, where improving distributional alignment frequently leaves the human-preferred option misidentified—an error verifiable against real responses that ranking evaluation catches. By framing survey alignment as a multi-dimensional construct, RADIUS provides a foundation for systematically assessing and comparing the population-level fidelity of survey simulation methods. Future work could investigate the relationship between individual- and group-level alignment and develop simulation approaches that mitigate mode collapse while preserving preference structure.

Limitations

Our evaluation focuses on multiple-choice single-select questions—the format with the broadest public ground-truth coverage (GSS, OpinionQA). This reflects data availability rather than a restriction of the metrics: TRM's bootstrap option grouping is model-free and applies to any categorical response options set (including multi-select), and RC uses Spearman rank correlation, which is defined on ordinal data and already rewards "close" answers on Likert scales, penalizing adjacent transpositions less than distant ones. Distribution metrics that treat options as unordered (TVD, DH) can be extended to ordinal scales via distance-weighted variants such as the WD we already report, while open-ended responses would require a prior discretization or embedding step. TRM's tie rule based on overlapping confidence intervals is conservative (overlap does not imply a non-significant difference), and tie groups depend on human sample size; testing the top-option difference directly is a natural refinement. We also do not study prompt variants or repeated runs, which may affect absolute scores. Finally, our primary experiments use a single prompt-based simulator with personas derived from specific behavioral datasets; although we additionally evaluate an independently fine-tuned simulator (Section 5.5), alignment scores in general depend on both persona design and the underlying model, and absolute values may shift under alternative simulation setups. As our contribution is the evaluation framework rather than the simulator, we expect the relative, multi-dimensional comparisons RADIUS enables to remain informative across such choices.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. Let the LLMs talk: Simulating human-to-human conversational QA via zero-shot LLM-to-LLM interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, pages 8–17, 2024.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, ICML '23, pages 337–371, 2023.

Anthropic. Introducing claude sonnet 4.5. `https://www.anthropic.com/news/claude-sonnet-4-5`, 2025.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Krisztian Balog and ChengXiang Zhai. User simulation for evaluating information access systems. *Foundations and Trends in Information Retrieval*, 18(1-2):1–261, 2024.

Krisztian Balog, Nolwenn Bernard, Saber Zerhoudi, and ChengXiang Zhai. Theory and toolkits for user simulation in the era of generative AI: User modeling, synthetic data generation, and system evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pages 4138–4141, 2025.

Alex Berke, Dan Calacci, Robert Mahari, Takahiro Yabe, Kent Larson, and Sandy Pentland. Open e-commerce 1.0: Five years of crowdsourced U.S. Amazon purchase histories with user demographics. *Scientific Data*, 11:491, 2024.

Adam Bignold, Francisco Cruz, Richard Dazeley, Peter Vamplew, and Cameron Foale. An evaluation methodology for interactive reinforcement learning with simulated users. *Biomimetics*, 6(1), 2021.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024.

Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. Mixture-of-personas language models for population simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, ACL '25, pages 24761–24778, 2025.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL-HLT '25, pages 3141–3154, 2025.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*, 2024.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from homo silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pages 614–615, 2024.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1259), 2024a.

Chen Gao, Fengli Xu, Xu Chen, Xiang Wang, Xiangnan He, and Yong Li. Simulating human society with large language model agents: City, social media, and economic system. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, pages 1290–1293, 2024b.

Naieme Hazrati and Francesco Ricci. Choice models and recommender systems effects on users' choices. *User Modeling and User-Adapted Interaction*, 34(1):109–145, 2024.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '24, pages 10289–10307, 2024.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *The Fourteenth International Conference on Learning Representations*, ICLR '26, 2026.

Sangkeun Jung, Cheongjae Lee, Kyungduk Kim, Minwoo Jeong, and Gary Geunbae Lee. Data-driven user simulation for automated evaluation of spoken dialog systems. *Computer Speech & Language*, 23(4):479–509, 2009.

Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, pages 82–86, 2025.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, EMNLP '25, pages 30096–30111, 2025.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23212–23237, 2025.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Aßenmacher. Algorithmic fidelity of large language models in generating synthetic German public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '25, pages 1785–1809, 2025.

Bolei Ma, Yong Cao, Indira Sen, Anna-Carolina Haensch, Frauke Kreuter, Barbara Plank, and Daniel Hershcovich. Too open for opinion? embracing open-endedness in large language models for social simulation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, EACL '26, pages 1998–2016, 2026.

Nicole Meister, Carlos Guestrin, and Tatsunori Hashimoto. Benchmarking distributional alignment of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL-HLT '25, pages 24–49, 2025.

Keiichi Namikoshi, Alexandre Filipowicz, David Ayman Shamma, Rumen I Iliev, Candice L Hogan, and Nikos Arechiga. Using LLMs to model the beliefs and preferences of targeted populations. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, 2022.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. Generative agent simulations of 1,000 people. *arXiv*, cs.AI/2411.10109v1, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '25, pages 21147–21170, 2025.

Seungjong Sun, Eungu Lee, Dongyan Nan, Xiangying Zhao, Wonbyung Lee, Bernard J. Jansen, and Jang Hyun Kim. Random silicon sampling: Simulating human sub-population opinion using a large language model based on group-level demographic information. *arXiv*, cs.AI/2402.18144, 2024.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, EMNLP '24, pages 16612–16631, 2024.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. SocioBench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26257–26289, 2025a.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Front. Comput. Sci.*, 18(6), 2024.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.*, 43(2), 2025b.

A Results on OpinionQA W34

For space, the main text reports GSS, W50, and W92 (Table 1). Table 3 reports the additional OpinionQA W34 wave (biomedical and food-habit topics), across all three persona sources (Soc, RT, E-c) and all three non-parametric baselines (Nor, Rand, Uni). Results are consistent with the reported findings: persona-based simulators achieve the strongest ranking alignment (E-c is best on TRM, RC), while the tradeoff with distributional metrics observed on other benchmarks persists.

Table 3: Simulation alignment on OpinionQA W34 across persona sources (Soc, RT, E-c) and non-parametric baselines (Nor, Rand, Uni). Underline marks the best value; (*) indicates a statistically significant difference from the best (paired t-test, $p < 0.05$).

Metric	Soc	RT	E-c	Nor	Rand	Uni
RADIUS (proposed alignment suite)						
TRM ↑	0.74	0.71	0.77	0.36*	0.33*	0.32*
RC ↑	0.82*	0.81*	0.89	0.64*	0.47*	0.33*
TVD ↓	0.37*	0.38	0.33	0.36	0.47*	0.37
DH ↑	0.00	0.00	0.00	0.03	0.03	0.05
Prior work (for comparison)						
CV ↓	0.39*	0.40*	0.35	0.42*	0.52*	0.45*
JSD ↓	0.14	0.15	0.11	0.11	0.19*	0.12
WD ↓	0.48	0.52*	0.42	0.66*	0.80*	0.69*
SBS ↑	-37.9	-38.8	-9.0	2.8	-50.1	0.0

Table 4: Non-parametric baseline alignment (Nor, Rand, Uni) for GSS, W50, and W92. Persona-source results (Soc, RT, E-c), and statistical comparisons against them, are reported in Table 1; W34 baselines appear in Table 3.

Survey	Baseline	TRM ↑	RC ↑	TVD ↓	DH ↑	CV ↓	JSD ↓	WD ↓	SBS ↑
GSS	Nor	0.450	0.613	0.252	0.033	0.270	0.054	0.379	-0.1
	Rand	0.383	0.479	0.386	0.000	0.413	0.121	0.600	-105.8
	Uni	0.100	0.172	0.254	0.017	0.279	0.051	0.394	0.0
W50	Nor	0.391	0.710	0.380	0.000	0.424	0.126	0.688	7.4
	Rand	0.312	0.500	0.492	0.000	0.527	0.206	0.861	-25.1
	Uni	0.078	0.173	0.408	0.000	0.485	0.144	0.766	0.0
W92	Nor	0.364	0.675	0.306	0.000	0.348	0.094	0.568	7.8
	Rand	0.286	0.464	0.454	0.000	0.471	0.185	0.768	-43.1
	Uni	0.299	0.464	0.331	0.000	0.424	0.108	0.612	0.1

B Non-Parametric Baseline Results

For space, the main-text results (Table 1) report only the normal (Nor) non-parametric baseline. Table 4 reports the complete set of non-parametric baselines—normal (Nor), random (Rand, Dirichlet-Multinomial), and uniform (Uni)—for GSS, W50, and W92; the corresponding W34 baselines appear alongside its persona-source results in Table 3. Persona-source results for these benchmarks are given in Table 1. Here we isolate the baselines to characterize the non-parametric reference points against which persona-based simulation is compared. The normal baseline consistently dominates random and uniform on distributional alignment (TVD, JSD, WD), motivating its use as the single representative baseline in the main text.

C Reproducibility of Cao et al. [2025]

Section 5.5 uses the fine-tuned survey simulator of Cao et al. [2025] as an independent, large-scale testbed for metric complementarity. Here we detail that reproduction and verify its fidelity against their published results. They fine-tune LLMs to align first-token option probabilities with World Values Survey (WVS) response distributions via a KL-divergence objective, evaluating with a Jensen–Shannon-based similarity score (S_{JS}) and Earth Mover's Distance (EMD). We use these two metrics only for the fidelity check below (Section C.2) and evaluate complementarity in the main text with RADIUS.

C.1 Setup

We fine-tune `DeepSeek-R1-Distill-Qwen-7B`, a state-of-the-art open model with strong performance across diverse benchmarks, using their released code, data splits (train: $C_1 \times Q_1$, 6,841

Table 5: Our reproduction vs. Cao et al. [2025]'s published DeepSeek-R1-Distill-Qwen-7B results (matching test subset, $n = 6{,}673$). S_{JS} is their similarity metric (Eq. (1)), distinct from our JSD.

	$S_{JS} \uparrow$		EMD $\downarrow$	
	Paper	Ours	Paper	Ours
Zero-shot (ZS)	0.648	0.593	0.109	0.122
Fine-tuned (FT)	0.801	0.797	0.081	0.080

Table 6: RADIUS metrics on full test set ($n = 7{,}329$) and matching test subset ($n = 6{,}673$).

	Full test set		Matching test subset	
Metric	Zero Shot	Fine-Tuned	Zero Shot	Fine-Tuned
TRM $\uparrow$	0.374	0.540	0.379	0.566
RC $\uparrow$	0.482	0.698	0.467	0.707
TVD $\downarrow$	0.498	0.256	0.498	0.239
DH $\uparrow$	0.006	0.029	0.004	0.031

samples; validation: $C_1 \times Q_2$, 1,586 samples; test: 7,329 samples across seven country-question combinations), and hyperparameters (LoRA rank 8, α=32, dropout 0.05, target modules {q_proj, v_proj}; AdamW, lr $1\mathrm{e}{-4}$). We train for up to 5 epochs.[7] Following their unmodified training script's default behavior, which checkpoints only on validation-loss improvement, without per-epoch saves, the retained checkpoint corresponds to the epoch with lowest validation loss (epoch 2 of 5 in our run), not necessarily the final epoch. We replicate their prompt construction exactly, including their custom chat-style template (a non-standard reuse of the model's `<|begin_of_sentence|>` token as a role separator, rather than DeepSeek-R1's native template) and their exact option-token extraction logic, as the fine-tuned checkpoint was trained on this specific format.

Two deviations from their setup remain: (1) due to GPU memory constraints (single A10G, 24GB, vs. their A100, 80GB) we use a training batch size of 4 with 4-bit quantization instead of their batch size of 16 at full precision; (2) their main results table averages over five of the seven test country-question conditions (6,673 samples), omitting two ($C_2 \times Q_2$, $C_3 \times Q_2$). We report both the full test set and this matching 6,673-sample subset. Cao et al. [2025] further report that their fine-tuning yields similar performance regardless of the base model, so this choice is not critical to the comparison.

C.2 Evaluation Metrics

We follow the metric definitions and implementation from Cao et al. [2025]'s released repository, alongside RADIUS. Cao et al. [2025] define their similarity metric as $1 - d_{JS}(P, Q)$, where d_{JS} is the Jensen–Shannon distance between the human and predicted probability vectors P, Q (their code passes unnormalized percentages for P, which `scipy.spatial.distance.jensenshannon` normalizes internally):

$$S_{JS} = 1 - d_{JS}(P, Q), \quad d_{JS}(P, Q) = \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P \| M) + \tfrac{1}{2} D_{\mathrm{KL}}(Q \| M)}, \; M = \tfrac{1}{2}(P + Q). \quad (1)$$

Since this differs from the standard Jensen–Shannon divergence used elsewhere in this paper, we denote their metric S_{JS} to avoid confusion.[8] Their EMD is computed via `scipy.stats.wasserstein_distance` applied directly to P and Q as unweighted sample values, rather than as probability masses over a shared ordinal support. We replicate this exactly to enable direct comparison with their published results. These two metrics are used only to verify reproduction fidelity (Table 5). The complementarity analysis (Section 5.5) uses RADIUS's own metrics.

[7] All fine-tuning and inference ran on a single AWS EC2 `g5.xlarge` instance (1× A10G, 24GB). Fine-tuning took ∼7 hours, each inference pass ∼20 minutes.

[8] Our JSD is the divergence itself, without the square root: $\mathrm{JSD}(P, Q) = \tfrac{1}{2} D_{\mathrm{KL}}(P \| M) + \tfrac{1}{2} D_{\mathrm{KL}}(Q \| M) = d_{JS}(P, Q)^2$. Thus $S_{JS} \neq 1 - \mathrm{JSD}$.

C.3 Results

Table 5 compares our reproduction (matching test subset, n=6,673) to their published values. Fine-tuned scores closely match (S_{JS}: 0.797 vs. 0.801; EMD: 0.080 vs. 0.081), with a small residual gap in the zero-shot condition. The close match on the fine-tuned model, despite using their default best-validation checkpoint and our hardware deviations, confirms the fidelity of our reproduction.

Table 6 reports RADIUS on the same predictions. On average, fine-tuning improves every metric: both their distributional scores and all four RADIUS metrics (TRM: 0.37→0.54, RC: 0.48→0.70, TVD: 0.50→0.26, DH: 0.006→0.029 on the full set). Taken alone, these aggregates agree with Cao et al. [2025]'s own metrics. The per-question analysis in the main text (Figure 2) shows that this aggregate agreement nonetheless masks substantial complementarity: distribution and ranking improvements are correlated but far from equivalent, and among questions whose TVD improves, 45.9% still misidentify the human top choice (TRM=0), confirming that RADIUS surfaces decision-relevant errors invisible to distributional metrics.