Aligned Items, Broken Inferences: When to Trust Large Language Models in Social Survey Simulation

Anonymous Authors

Anonymous Affiliation

Abstract. Simulating human populations with large language models has become a common computational approach to studying social attitudes. When researchers use LLMs as synthetic survey respondents, the central question is which conclusions still hold. Recent work fine-tunes the model on demographic-conditioned real responses, which raises item-level accuracy but leaves open whether the analyses researchers run on the synthetic data remain reliable. We test this on the trust module of the seventh-wave World Values Survey, applying a fine-tuning method to three open-source models and auditing the cross-national analyses a researcher would run across its 65 countries. For country-level aggregate analyses, the cost is small and the synthetic data agrees closely with the real survey, with a country-ranking correlation of 0.94. For regression analyses, the cost is moderate, with 95 percent confidence intervals covering the real estimate at 34 to 37 percent, well below the 83 percent that two independent samples of the same survey would reach, so the analysis runs but its inferences cannot be trusted. Among these regressions, seven in ten significant findings on the synthetic data correspond to predictors that are null in the real data. For less advantaged, lower-income countries, sign agreement drops by another 9 to 11 percentage points. For cluster analyses that group countries into a trust typology, the cost is largest, with the adjusted Rand index near zero, no better than random or majority-guess baselines.The three architectures produce the same errors in the same places, with Cohen's kappa between 0.71 and 0.80. Together these costs form a decision rubric that maps the three analytical purposes we audit to their safety on synthetic survey data. The rubric helps researchers choose synthetic survey data for each analytical purpose, with the inferential cost made explicit rather than assumed.

Keywords: Large language models · Survey simulation · Computational social science · Decision rubric

1 Introduction

A growing body of work in computational social science uses large language models as synthetic survey respondents in cross-national research [2,4,6,18,3]. A researcher prompts the model with a respondent's demographic profile and reads

back a synthetic answer to a survey item [2,15,16]. Across many profiles, the same procedure produces a synthetic corpus that researchers can analyse in place of a real survey [18,10,15]. Real cross-national surveys are expensive and slow to field, and a synthetic substitute promises to relieve both constraints [18,11,8]. The question that bears most directly on whether the substitute is sound is not about item-by-item resemblance between synthetic and real responses [4,14,13]. It is about whether the conclusions a researcher would draw from analyses on the synthetic corpus match those the corresponding real survey would have supported [4,6,19,20].

When language models are queried directly with a respondent's demographic profile, their answers align imperfectly with real responses at the item level [5,16,18,1]. To close this gap, recent work fine-tunes the model on a corpus of real survey responses paired with their respondents' demographics [19,20,11]. The fine-tuned model produces synthetic answers whose item-level accuracy is substantially higher than that of the prompted model [19,11]. The validation reported for such fine-tuned synthetic corpora has so far concentrated on item-level alignment and on the agreement of country averages between synthetic and real responses [19,16,20]. Whether the analyses researchers actually run on these corpora produce the same conclusions a real survey would have supported has not been mapped at useful granularity [20,1,7]. Country-mean agreement and the agreement of fitted regression models or cluster typologies are different objects [6,14]. A researcher who wants to decide which downstream analyses to trust needs the second kind audited as well [1,7].

A second consideration concerns the generality of failures observed in audits of fine-tuned synthetic respondents [20,1]. Such a failure could stem from the fine-tuning procedure or from the language model the procedure was applied to [1,18,19]. The two can be distinguished only when the same procedure is held fixed and the architecture is varied across several models [24,1]. The same comparison shows whether failures concentrate on identifiable subgroups of countries or respondents in a manner that repeats across architectures [14,15,16]. A finding that repeats across architectures gains the status of a procedure-level fact, while one that fails to repeat reduces to a model-specific artifact [1,24].

We take the second kind of agreement as our object of audit and construct a decision rubric for the use of fine-tuned synthetic survey data in cross-national research. We instantiate the rubric on the trust module of the seventh World Values Survey wave [12], fielded in 65 countries. The rubric considers three task tiers. The first contains analyses that compare countries on average trust. The second contains regression analyses of demographic effects. The third contains multivariate clustering of countries into a trust typology. For each tier the rubric reports the inferential cost a researcher would pay if the synthetic corpus were used in place of the real survey.

The audit returns three distinct cost levels. Aggregate comparisons of countries on average trust carry a small cost, with the rank order of countries preserved almost intact. Regression analyses carry a moderate cost, where confidence intervals miscover and most synthetic-significant effects are null in the

real data, and the penalty grows for less advantaged, lower-income countries. Clustering countries into a trust typology carries the largest cost, at the level of chance. The same ordering holds across three architectures fine-tuned by one procedure, which lets us read the costs as procedure-level rather than model-specific.

This paper makes four contributions to the use of fine-tuned synthetic survey data in social computing research. First, we audit such data at the layer of fitted regression and multivariate analyses, rather than at the layer of item-level or country-mean alignment. Second, we measure properties of the fine-tuning procedure we apply, observed across three architectures of comparable scale, rather than of any particular model. Third, we decompose the regression cost by economic advantage and quantify the additional inferential penalty researchers pay for less advantaged, lower-income countries. Fourth, we collect the costs into a decision rubric that maps these three analytical purposes to their safety on synthetic survey data. The rubric helps a social science researcher choose the data source for each purpose with the inferential cost in view.

2 Related Work

2.1 Large Language Models as Synthetic Survey Respondents

Large language models prompted with respondent demographics have been used to generate synthetic survey responses across social science domains [2,15,16,19]. Early reports observed that the marginal distributions of synthetic responses can track those of real respondents in some populations [2,10,19]. Subsequent work found that the alignment is uneven across items and subpopulations, and that aggregate-level appearances of fit do not always survive finer-grained inspection [5,16,18,14]. Practitioners who treat the synthetic corpus as a stand-in for a real survey therefore need stronger evidence than item-level agreement before they generalise conclusions [4,18,1,22].

A second line fine-tunes language models on real survey responses paired with respondent demographics to raise item-level alignment beyond what prompting achieves [19,20,11]. The reported gains are substantial, with reductions in the synthetic-to-human gap of up to forty-six percent in one large study [19]. Some of this work asks whether fine-tuned models also reproduce regression coefficients and finds the agreement only partial [20]. The evaluation has stayed at the level of single coefficients or marginal distributions, leaving the validity of the regression inference and the multivariate analyses that anchor cross-national social science unaudited [19,20].

2.2 Audits of Synthetic Survey Data

A growing body of methodological work audits when synthetic survey data can be trusted for downstream use [4,14,13,1]. One line compares synthetic and real responses at the item or distribution level and reports the conditions under

which alignment holds [18,17,16,14]. A second line tests whether published social-science effects replicate on synthetic respondents, with one study of 156 studies recovering around eighty percent of main effects [6]. Coefficient-level audits on national survey data instead report divergence in sign or magnitude that persists even after fine-tuning on a relevant human sample [4,20]. A third line proposes principles or protocols for evaluating synthetic data [1,7,24].

Two gaps in this body of work motivate the present paper. The first concerns granularity, since the validity of interval and significance inference and the multivariate cluster structure are properties that coefficient-level agreement does not constrain, so a corpus can reproduce single effects yet still fail the fitted analyses a comparative researcher relies on [19,20]. Recent methodological commentary notes general shortcomings of language model social simulation but does not map each to a specific analytical purpose with a corresponding cost [13,1]. The second concerns architectural generality, since an audit of a single model cannot separate failures of the fine-tuning procedure from failures of the particular model it was applied to [1,24]. Our paper takes both gaps as its starting point.

3 Methods

3.1 Audit Framework

Our audit operates at the level of fitted analyses rather than at the level of the data on which they are fitted. For each of three task tiers, we compute the inferential output a researcher would produce on the synthetic corpus and compare it to the corresponding output on the real survey. The synthetic and real corpora are aligned by demographic profile, so each synthetic respondent corresponds to a real respondent with the same profile. The cost of the synthetic corpus for a tier is the discrepancy between the two outputs under a tier-appropriate metric.

The synthetic corpus is produced by a supervised fine-tuning procedure that trains a base language model on real survey responses paired with their respondents' demographics and then queries the model with a demographic profile to obtain a synthetic response. Repeating the query across all profiles yields a synthetic corpus comparable in shape to the real survey. We treat the procedure as a black box and audit only its outputs at the analysis layer, without modifying the procedure or the underlying model.

3.2 Data and Models

Our corpus is the trust module of the seventh World Values Survey wave, comprising 244,219 respondents drawn from 65 countries. The module contains thirteen items. Seven concern trust in other people, ranging from close relations to strangers, namely generalised interpersonal trust and trust in one's family, personally known acquaintances, the neighbourhood, strangers met for the first time, people of another religion, and people of another nationality. The remaining six concern confidence in public institutions, namely churches, the press,

Table 1. Decision rubric for fine-tuned synthetic survey data. For each analytical purpose it gives the headline evidence on the synthetic corpus, a colour-coded verdict, and the recommended action. Estimates are procedure-level across the three architectures.

Tier	Purpose	Evidence	Verdict	What to do
Aggregate	Country rankings	Spearman **0.94**	✓ **Safe**	Use as a substitute
Regression	Demographic effects	CI coverage **34**–**37**% 7 of 10 sig. null	! **Caution**	Trust direction only
	↳ lower-income	**−9** to **−11** pp	! **Caution**	Avoid for these
Cluster	Country typologies	ARI **≈ 0**	✗ **Avoid**	Do not use

government, political parties, major companies, and the United Nations. Each respondent carries a demographic profile of eight background variables, all of which the generation prompt encodes. The regression analyses below use five of these as predictors, following standard practice in the survey codebook. We split the respondents within each country into a training set and a held-out test set in equal proportion under a fixed random seed, map each response to an ordinal score, and discard model outputs that do not map to a valid response option before any metric is computed.

We fine-tune three open-source language models of comparable size, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3, holding one fine-tuning procedure fixed across all three. Each model receives low-rank supervised fine-tuning on the demographic-conditioned training set, with a LoRA adapter on all linear modules at rank 8 and $\alpha = 16$. Training uses AdamW under a cosine schedule at a peak learning rate of 2×10^{-5} in bf16, with effective batch size 128, one epoch, and a cutoff of 2048 tokens. The prompt encodes the demographic profile in a fixed natural-language template and queries one trust item per call.

The country covariate for our subgroup decomposition is economic advantage, operationalised by the World Bank income classification [23], since better-resourced countries tend to be better represented in the data the models are trained on. We group the high-income countries in our sample as the more advantaged set and the remaining countries as the comparison group. We construct a country-level binary indicator from this grouping and hold it fixed across architectures and tiers.

3.3 Inferential Cost Metrics

The first tier audits analyses that compare countries on their average response to an item. We compute the country mean of each item under the synthetic and the real corpus and report the cross-country Spearman correlation between the two, averaged across the trust items. A correlation close to one lets a researcher

Table 2. Full audit results for the trust block of the World Values Survey wave seven. Columns compare the prompted general baseline (Gen) and the fine-tuned variant (SFT) of three architectures (Q = Qwen2.5-7B-Instruct, L = Llama-3.1-8B-Instruct, M = Mistral-7B-Instruct-v0.3). Real is the reference value computed on the real survey. Higher is better for all metrics except false discovery rate, where lower is better. Cluster baselines at k=5 are a majority-prediction and a random-prediction model clustered the same way, both falling within ±0.05 of zero, the same band as every model variant.

Metric	Real	Q-Gen	Q-SFT	L-Gen	L-SFT	M-Gen	M-SFT
Aggregate tier							
Item-level accuracy	1.00	0.29	0.52	0.10	0.52	0.35	0.53
Cross-country Spearman	1.00	−0.09	0.94	−0.12	0.94	−0.09	0.94
Regression tier							
95% CI coverage	0.95	0.41	0.36	0.49	0.37	0.52	0.34
Sign agreement (aggregate)	1.00	0.60	0.70	0.63	0.70	0.56	0.71
...higher-income countries	1.00	0.69	0.77	0.71	0.77	0.72	0.78
...other countries	1.00	0.58	0.68	0.61	0.66	0.61	0.68
False discovery rate	0.00	0.75	0.72	0.74	0.71	0.75	0.70
Cluster tier							
Adjusted Rand index (k=5)	1.00	0.02	−0.03	0.00	−0.02	−0.02	−0.03

use the synthetic corpus for descriptive country-rank claims without inferential loss.

The second tier audits regression analyses that estimate the effect of respondent demographics on a target item. We fit per-country ordinary least squares regressions of each item on five demographic predictors. For these regressions we report the coverage of nominal 95 percent confidence intervals, the false discovery rate of synthetic-significant effects whose real-data counterpart is null, taken as a real p above 0.10, and the sign agreement rate among synthetic-significant effects. This rate measures discordance against the real survey within each cell and is not a multiplicity-controlled false discovery rate in the testing sense. We treat an effect as significant at the conventional 0.05 threshold and compute each rate per country-by-variable-by-predictor cell without multiple-comparison correction, so the rates describe the corpus as a practitioner would use it.

The third tier audits the recovery of a multivariate country typology. For both the synthetic and the real data we form each country's mean profile across the thirteen trust items, perform Ward hierarchical clustering on these country profiles, and report the adjusted Rand index between the two sets of cluster labels at k clusters. To anchor the index, we cluster the same way a random-prediction model and a majority-prediction model that returns each item's modal response. An index in the range spanned by these baselines signals that the synthetic cluster labels carry no more information about the typology than chance.

We report each tier's cost separately for each of the three architectures, treating a cost level that holds across all three as a procedure-level finding and one that varies substantially as model-specific. To characterise where the errors

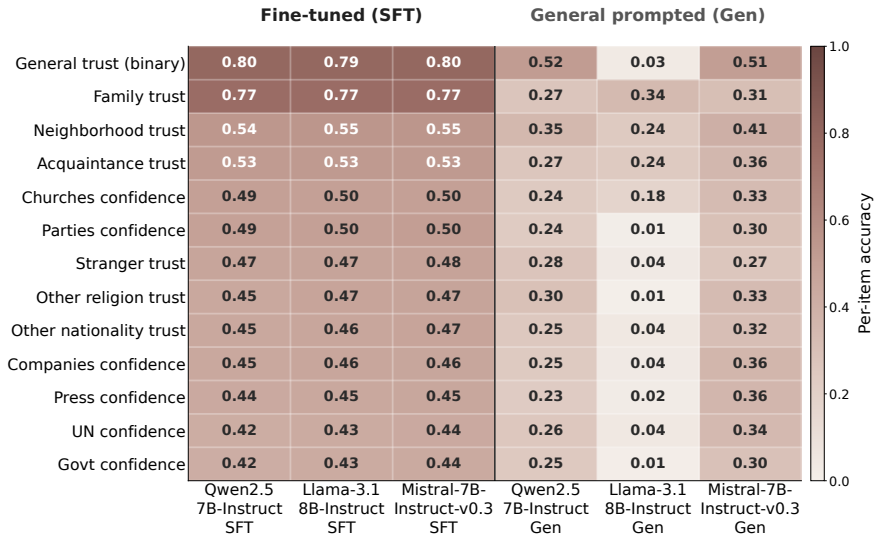

Fig. 1. Per-item prediction accuracy on the thirteen trust items across the three architectures, each as a prompted general baseline and a fine-tuned variant. Cell labels and shading encode accuracy in the held-out test set. Fine-tuning roughly doubles per-item accuracy over prompting, and the three fine-tuned columns agree to within a few percentage points on almost every item.

fall, for each pair of architectures we compute Cohen's kappa over the held-out items, treating each item as a correct or incorrect prediction, so a high kappa indicates the architectures err on the same items. For the regression tier we additionally test whether the cost differs between the more advantaged, higher-income countries and the comparison group with an independent Welch's two-sample t-test reported separately for each architecture.

4 Results and Discussion

Table 1 presents the decision rubric our audit produces, which the rest of this section establishes tier by tier. The rubric maps each analytical purpose audited in this paper to the inferential cost a researcher should expect to pay for that purpose on a fine-tuned synthetic survey corpus. The cost levels are reported as procedure-level estimates, with the architecture-specific values available in Figures 2 through 5. The rubric also records the additional penalty that applies to regression analyses on less advantaged, lower-income countries. The cost levels apply to the trust construct on this survey and should be re-evaluated before transfer to other constructs or instruments.

Table 2 reports every metric in our audit side by side for the prompted general baseline and the fine-tuned variant of each architecture, with the real-data reference in the leftmost numerical column. Within each tier the three

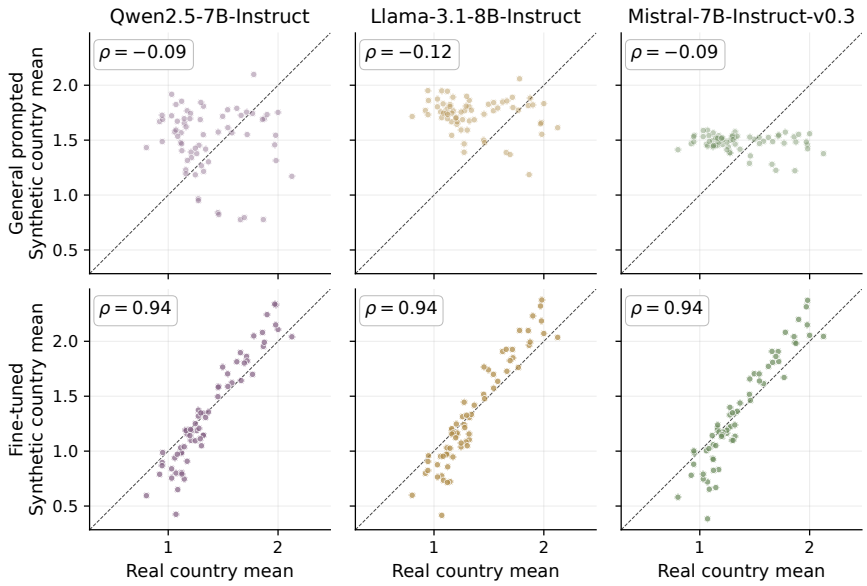

Fig. 2. Per-country composite trust mean on the synthetic corpus against the real corpus, across 65 countries. The top row reports the prompted general baselines and the bottom row reports the fine-tuned variants. Each panel reports the Spearman rank correlation with the real country means in the upper left corner. The dashed diagonal marks exact agreement with the real corpus.

architectures return near-identical values, so the table reads the same way across models.

At the item level the fine-tuned corpus looks well aligned, which is the surface agreement that has encouraged researchers to use such corpora. Figure 1 shows that fine-tuning roughly doubles per-item prediction accuracy over the prompted baseline on every architecture, lifting it to about 0.52 on average from a prompted range of 0.10 to 0.35. The three fine-tuned architectures reach almost the same accuracy on each item, and their item-level errors agree pairwise at Cohen's kappa of 0.71, 0.77, and 0.80. We therefore read the tier costs that follow as properties of this fine-tuning procedure across architectures of comparable scale rather than of any single model. The remaining sections show that this item-level alignment does not carry through to the regression and multivariate analyses a researcher would fit on the corpus.

4.1 Aggregate Tier, Small Cost

Aggregate analyses that compare countries on their average response carry a small cost on the synthetic corpus. Each fine-tuned model achieves a cross-country Spearman correlation of approximately 0.94 with the real country means, averaged across the trust items. The prompted versions of the same models reach

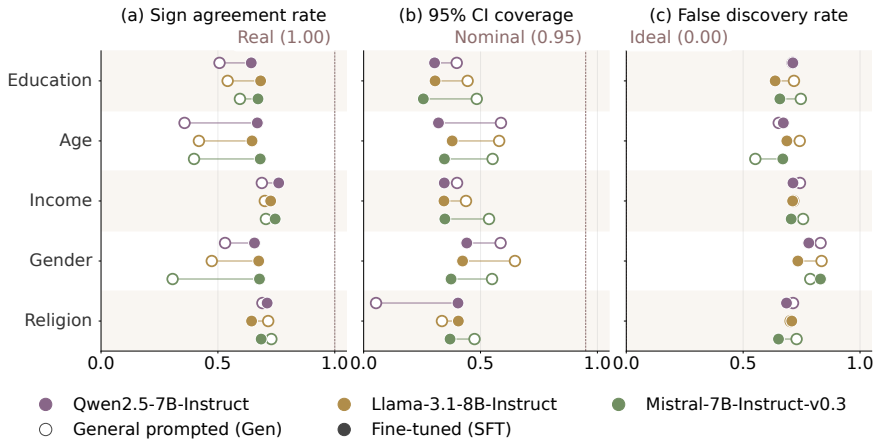

Fig. 3. Per-predictor regression costs across architectures. Rows correspond to the five demographic predictors and columns to three cost metrics. Within each row, each architecture contributes one open circle for the prompted general baseline and one filled circle for the fine-tuned variant, connected by a thin segment. The horizontal axis on each panel is the metric value, averaged across the thirteen trust items. Vertical dashed lines mark the reference value of each metric.

Spearman correlations near zero or negative on the same task, which indicates that fine-tuning makes a substantial difference at this tier. Figure 2 shows the per-architecture values. A researcher whose analysis depends only on the rank order of countries on the trust items has little to lose from the synthetic corpus at this level of aggregation.

4.2 Regression Tier, Moderate Cost

Regression analyses that estimate the effect of respondent demographics on a target item incur a moderate cost. The damage is to the inference a researcher draws, not to the ability to produce an estimate. Confidence intervals fitted on the synthetic corpus are too narrow, so they cover the corresponding real point estimate at only 34 to 37 percent against the nominal 95, and the same shortfall appears in all three fine-tuned architectures. A researcher reading these intervals would treat them as far more precise than the real survey warrants.

Figure 3 reports per-predictor sign agreement, confidence interval coverage, and the share of significant synthetic effects that correspond to a predictor null in the real data, and the same row-level pattern holds across the three fine-tuned architectures. Sign agreement averages about 0.70 across predictors, with age higher and gender lower. About seven in ten effects flagged as significant on the synthetic data correspond to predictors that are null in the real data, and the aggregate sign-agreement rate of about 0.70 masks a subgroup pattern we decompose next.

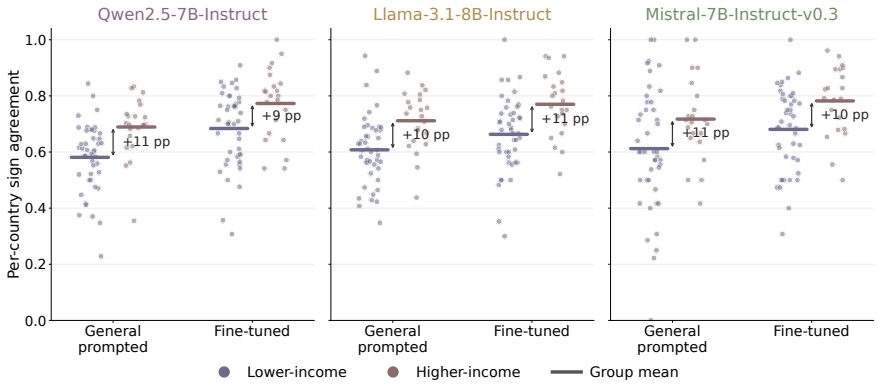

Fig. 4. Per-country sign-agreement rate by income group, under the prompted general baseline and the fine-tuned variant, separately for each architecture. Each dot is a country, colored by membership in the higher-income, more advantaged group defined in Section 3.2. The thick horizontal bars mark group means and the arrow gives the gap between them in percentage points. Within both conditions the higher-income countries sit above the rest, and fine-tuning preserves rather than closes the gap, which is significant at Welch's two-sample t-test p below 0.05 on the fine-tuned variant in every architecture.

The regression cost is further unequal across the countries in our sample. Within the more advantaged, higher-income countries defined in Section 3.2, the rate at which a significant synthetic effect matches the sign of the real estimate is approximately 78 percent. Outside this group, the same rate falls by 9 to 11 percentage points across the three architectures, a difference significant at Welch's t-test p below 0.05 in each. The same gap is already present in the prompted general baselines at 10 to 11 percentage points, so the fine-tuning procedure preserves rather than removes a divide that is present before fine-tuning. Figure 4 shows the by-architecture breakdown. A researcher whose analysis concerns less advantaged populations should add this additional cost to the regression-tier values reported above.

These regression-level signals point the same way rather than acting as independent failures. The synthetic corpus produces outputs that look workable in isolation but diverge from the real survey once the two are compared, so the cost is not a missing output but an output that cannot be trusted. A researcher who fits a regression on the corpus still obtains coefficients, confidence intervals, and significance calls, yet these do not reproduce the conclusions the real survey would have supported. We treat this tier as a moderate cost because the analysis runs to completion while its inferential content is unreliable.

4.3 Cluster Tier, Largest Cost

Clustering carries a weight in social inquiry that a single coefficient does not. In the Weberian tradition, social analysis seeks not only average tendencies but

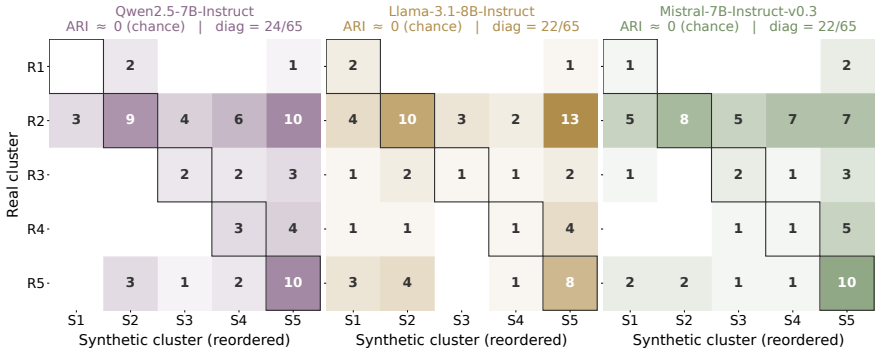

Fig. 5. Cluster-tier results at $k = 5$. Each panel shows the confusion matrix between Ward cluster labels on the real survey (rows R_1 to R_5) and on the fine-tuned synthetic corpus from one architecture (columns S_1 to S_5, reordered to maximise the diagonal). Cell entries count countries. Across all three architectures several real clusters map onto one or two dominant synthetic clusters, and even under the best alignment the diagonal recovers only about a third of the 65 countries. The adjusted Rand index stays within ± 0.05 of zero for every architecture, the same range as random-prediction and majority-prediction baselines clustered the same way, so typology recovery is at chance.

coherent ideal types, the recurring configurations of orientation that make a worldview intelligible [21]. Multivariate clustering operationalises this aim by recovering the joint distribution of responses across items and identifying which attitudinal profiles recur across cases. Trust is a domain where this matters, because cross-national work distinguishes particularistic trust, bounded by kinship and familiarity, from generalised trust that extends to strangers and institutions [9]. These are qualitatively distinct orientations rather than points on one scale, and telling them apart requires the configuration of responses across several items at once. A corpus that fixes the marginal distributions yet loses the contingent associations among items removes the basis for any typological claim about who trusts whom, which leaves downstream inferences about cohesion and legitimacy uninterpretable.

Cluster analyses that group countries into a multivariate trust typology incur the largest cost. The adjusted Rand index between the synthetic and real cluster labels sits near zero across all three fine-tuned architectures, between -0.03 and -0.02, as shown in Figure 5. A random-prediction and a majority-prediction baseline clustered the same way fall in the same ± 0.05 band, so the synthetic clusterings recover the real typology no better than chance. Even under the column permutation that maximises the diagonal, only about a third of the countries fall on it. The same near-zero index holds at $k = 8$, and the verdict rests on the off-diagonal spread visible in the confusion matrix rather than on the precise index value. In substantive terms, the distinct typology the real survey supports is not reproduced on the synthetic corpus, so any analysis that

depends on recovering it cannot be run on this data. This is the strongest signal in our audit that the synthetic corpus does not preserve the multivariate cluster structure of the trust items in our sample.

The cluster cost is not an artifact of how the corpus is generated. Because each response is drawn by sampling rather than greedy decoding, respondents who share a demographic response receive different answers at close to the rate seen in the real survey, so clustering the synthetic corpus is not the same as clustering the demographic profiles. The aggregate tier already shows that each item's country means are preserved almost intact, with a cross-country correlation of 0.94. The typology is nevertheless not recovered, so the failure is multivariate rather than a matter of any single item. A researcher who needs a country typology therefore cannot rely on this corpus, even though its per-item country means are accurate.

Taken together, the three tiers trace a single arc. The same fine-tuned corpus that aligns well at the item level returns analyses whose reliability falls as the analysis grows more structured, from a small cost on country rankings to a moderate cost on regression and the largest cost on multivariate typology. Item-level alignment does not certify inferential validity, and the gap between the two widens with the complexity of the quantity a researcher estimates. A country mean averages over respondents and is forgiving, a regression coefficient depends on the joint distribution of a predictor and an outcome within countries, and a cluster typology depends on the joint pattern of all items at once, so each tier asks more of the synthetic corpus than the last. Because the same ordering appears under one fine-tuning procedure applied to three architectures of comparable scale, we read it as a property of the procedure rather than of any single model, and the decision rubric in Table 1 records the ordering so that a researcher can match each analytical purpose to the cost it carries before committing to the synthetic corpus.

5 Conclusion

A researcher who plans to replace a real cross-national survey with a fine-tuned synthetic corpus can now tell in advance which conclusions will survive the substitution. Country-level aggregate comparisons of trust carry a small inferential cost on the synthetic corpus. Regression analyses that estimate demographic effects carry a moderate cost, with confidence intervals that miscover and significance tests that over-report. Multivariate cluster recovery carries the largest cost, with typology recovery at chance, no better than random or majority baselines. The cost on regression is further unequal, and less advantaged, lower-income countries fall 9 to 11 percentage points behind on sign agreement. The same cost levels appear under the same procedure applied to three architectures, which lets us read them as procedure-level expectations. A researcher can now plan a study on this kind of synthetic survey data with the cost of each analytical purpose visible in advance. More broadly, as social computing turns to language models to simulate social agents and their collective attitudes, our audit shows

that fidelity must be certified at the level of the analysis rather than the item, and offers one template for doing so.

Our audit has limits that bound these claims. The cost levels are measured for one construct, interpersonal and institutional trust, on one survey. The procedure-level reading rests on three architectures of comparable scale under one fine-tuning recipe, which separates the costs from any single model but not from the broader family of similarly sized instruction-tuned models or from the recipe itself. We leave the extension of the audit to further constructs, surveys, model scales, and fine-tuning procedures to future work.

References

1. Abdurahman, S., Salkhordeh Ziabari, A., Moore, A.K., Bartels, D.M., Dehghani, M.: A primer for evaluating large language models in social-science research. Advances in Methods and Practices in Psychological Science (2025)
2. Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. Political Analysis **31**(3), 337–351 (2023)
3. Bail, C.A.: Can generative AI improve social science? Proceedings of the National Academy of Sciences **121**(21), e2314021121 (2024). https://doi.org/10.1073/pnas.2314021121
4. Bisbee, J., Clinton, J.D., Dorff, C., Kenkel, B., Larson, J.M.: Synthetic replacements for human survey data? the perils of large language models. Political Analysis **32**(4) (2024)
5. Cheung, V., Maier, M., Lieder, F.: Large language models show amplified cognitive biases in moral decision-making. Proceedings of the National Academy of Sciences **122**(25), e2412015122 (2025). https://doi.org/10.1073/pnas.2412015122
6. Cui, Z., Li, N., Zhou, H.: A large-scale replication of scenario-based experiments in psychology and management using large language models. Nature Computational Science **5**(8), 627–634 (2025)
7. Cummins, J.: The threat of analytic flexibility in using large language models to simulate human data: A call to attention. arXiv preprint arXiv:2509.13397 (2025)
8. Davidson, T., Karell, D.: Integrating generative artificial intelligence into social science research: Measurement, prompting, and simulation. Sociological Methods & Research **54**(3), 775–793 (2025). https://doi.org/10.1177/00491241251339184
9. Delhey, J., Newton, K.: Predicting cross-national levels of social trust. European Sociological Review **21**(4), 311–327 (2005)
10. González-Bustamante, B., Verelst, N., Cisternas, C.: Emulating public opinion: A proof-of-concept of ai-generated synthetic survey responses for the chilean case. arXiv preprint arXiv:2509.09871 (2025)
11. Holtdirk, T., Assenmacher, D., Bleier, A., Wagner, C.: Learning from convenience samples: A case study on fine-tuning LLMs for survey non-response in the german longitudinal election study. arXiv preprint arXiv:2509.25063 (2025)
12. Inglehart, R., Basanez, M., Diez-Medrano, J., Halman, L., Luijkx, R.: World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. Ann Arbor-Michigan, Institute for Social Research, ICPSR version (2000)
13. Kozlowski, A.C., Evans, J.: Simulating subjects: The promise and peril of artificial intelligence stand-ins for social agents and interactions. Sociological Methods & Research (2025). https://doi.org/10.1177/00491241251337316

14. Li, D., Li, L., Qiu, H.S.: ChatGPT is not a man but das man: Representativeness and structural consistency of silicon samples generated by large language models. arXiv preprint arXiv:2507.02919 (2025)
15. Lu, J.G., Song, L.L., Zhang, L.D.: Cultural tendencies in generative AI. Nature Human Behaviour **9**, 2360–2369 (2025). https://doi.org/10.1038/s41562-025-02242-1
16. Ma, B., Yoztyurk, B., Haensch, A.C., Wang, X., Herklotz, M., Kreuter, F., Plank, B., Assenmacher, M.: 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 (2025)
17. Salecha, A., Ireland, M.E., Subrahmanya, S., Sedoc, J., Ungar, L.H., Eichstaedt, J.C.: Large language models display human-like social desirability biases in Big Five personality surveys. PNAS Nexus **3**(12), pgae533 (2024). https://doi.org/10.1093/pnasnexus/pgae533
18. Sarstedt, M., Adler, S.J., Rau, L., Schmitt, B.: Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. Psychology & Marketing **41**(6) (2024)
19. Suh, J., Jahanparast, E., Moon, S., Kang, M., Chang, K.W.: 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. pp. 21147–21170 (2025)
20. Wang, S., Hunt, K., Tang, S., Joseph, K.: Can finetuning LLMs on small human samples increase heterogeneity, alignment, and belief-action coherence? arXiv preprint arXiv:2511.21218 (2025)
21. Weber, M.: The Methodology of the Social Sciences. Free Press (1949), translated by E. A. Shils and H. A. Finch
22. Westwood, S.J.: The potential existential threat of large language models to online survey research. Proceedings of the National Academy of Sciences **122** (2025). https://doi.org/10.1073/pnas.2518075122
23. World Bank: World bank country and lending groups. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups (2024), accessed 1 June 2026
24. Zhou, J., Huang, J.t., Zhou, X., Lam, M.H., Wang, X., Zhu, H., Wang, W., Sap, M.: The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. arXiv preprint arXiv:2509.18052 (2025)