# Wind Tunnel: Evidence-Separated Population Modeling and AI-Assisted Questionnaire Pretesting

TACITUS

23 November 2026

**Abstract**

Generative models can produce plausible answers to survey questions without establishing what a population believes. This working paper describes Wind Tunnel, a research workspace that separates demographic descriptions, human survey evidence, statistical estimates, authored assumptions, and generated outputs. Its CiviSphere US pretest combines twelve age-by-education cells from a fitted 2020–2024 American Community Survey demographic artifact with model-generated answer distributions. Authored qualitative dossiers inform wording review but carry no population weight. Frozen study revisions connect instruments, audience definitions, source hashes, execution settings, and reports. We evaluate the released numeric recipe retrospectively against twelve institutional-confidence items in the 2024 General Social Survey, with the 2022 survey and an unstratified model call as comparators. Mean total variation is 2.48 percentage points for the earlier human survey, 7.50 for the national model call, and 8.16 for the Census-weighted model recipe. Demographic conditioning did not improve mean distributional agreement in this comparison. The contribution is an inspectable research workflow and a reproducible aggregate evaluation, rather than a validated replacement for human surveys. We specify the boundary between computational reproducibility and empirical validity, document the optional legacy Signal model separately, and identify the prospective tests needed before broader claims are justified.

**Working paper WT-2026-01 · Version 1.0.** Published by TACITUS. This report has not undergone independent peer review. Correspondence: hello@tacitus.me.

**Keywords:** computational social science; questionnaire pretesting; synthetic survey responses; population modeling; provenance; language models.

## 1 Introduction

A researcher preparing a questionnaire must decide whom a question concerns, what its words imply, which trade-offs respondents may notice, and what evidence would change the interpretation. A language model can assist with these tasks, but a fluent answer does not identify a population estimand. Likewise, fitting a demographic table does not measure institutional trust, political preferences, or likely behavior. These distinctions become consequential when synthetic outputs are displayed beside human survey results.

Wind Tunnel is a software system for population exploration, study design, exploratory scenario analysis, and the organization of human evidence. This paper documents its methodological boundaries and a released US questionnaire-pretest recipe. It distinguishes five kinds of information: source estimates; derived demographic quantities; fitted or synthetic population representations; statistical estimates from human observations; and authored or generated scenario assumptions.

The distinction is retained when information moves into a saved study and a report.

Three contributions are described. First, the system makes the relationship between a research universe, demographic controls, and generated answers explicit. Second, an immutable study revision records the materials and settings used for a particular analysis. Third, a public benchmark permits independent recomputation of distributional errors from released aggregates. None of these contributions, by itself, establishes predictive validity. The empirical question is whether conditioning a language model on a small demographic partition improves agreement with human survey distributions. The released comparison provides a negative result on its selected item battery.

The scope is deliberately bounded. This is a methods and software report, with a retrospective aggregate comparison. It does not estimate causal effects of questionnaire wording, certify the 27 legacy archetypes as population groups, or validate the full suite of operational scenario indicators. Statements about the current implementation refer to the source revision recorded in Appendix B; later product changes require a new method version.

## 2 Related work and methodological position

Research on language-model simulation has shown that conditioning can recover some patterns in particular human datasets (Argyle et al. 2023). Other evaluations identify failures in subgroup relationships and sensitivity to prompt and model choices (Bisbee et al. 2024). Survey-design experiments further show that a model's response to a wording perturbation need not reproduce the response of human participants (Tjuatja et al. 2024). These results motivate evaluation of an explicit model–instrument–population combination, rather than a general claim that synthetic answers represent people.

Demographic fitting and opinion estimation also have different roles. Adjusting a sample table to known margins has a long statistical history (Deming and Stephan 1940). Multilevel regression and poststratification combines a fitted response model with population cell counts (Gelman and Little 1997). Wind Tunnel's US numeric pretest uses a weighted average of model-supplied cell distributions. It does not fit a multilevel response model to human answers and is therefore not described here as calibrated MRP.

The documentation approach draws on model reporting (Mitchell et al. 2019), structured descriptions of simulation assumptions (Grimm et al. 2020), and survey transparency standards (American Association for Public Opinion Research n.d.). A small provenance graph provides a useful conceptual account of entities, transformations, and responsibility, consistent with the distinctions made in PROV-O (Lebo, Sahoo, and McGuinness 2013). This conceptual comparison does not assert that every Wind Tunnel export implements the PROV-O ontology. The scientific benefit of the records is that assumptions can be located and challenged; recording a claim does not make it true.

## 3 Evidence model and research universe

### 3.1 Evidence classes

Table 1 separates the interpretation of the principal inputs and outputs. A record's type is part of its meaning. An analyst should not silently convert an authored dossier into an observed behavioral distribution, or a generated response into a completed interview.

| Evidence class | Example | What the record supports |
|---|---|---|
| Source estimate | Published ACS table with universe and period | Description within the source's definitions and uncertainty |
| Derived quantity | Adult age-band count from source cells | A documented transformation of the input estimates |
| Fitted representation | Age-by-education population masses | Compatibility with the declared donor construction and controls |
| Human evidence | Weighted GSS answer distribution | Measurement under that survey's instrument and design |
| Statistical estimate | A separately fitted response estimator | Inference conditional on that estimator and its validation |
| Authored assumption | Qualitative dossier or scenario override | A hypothesis or research choice for investigation |
| Generated output | A model's cell-level answer shares | The output of the specified computational procedure |

Table 1. Evidence classes describe inferential roles, not a ranking of document quality.

### 3.2 Population foundation

The US demographic foundation uses the 2020–2024 American Community Survey (ACS) five-year tables and Public Use Microdata Sample documentation (U.S. Census Bureau 2026). A five-year period estimate is not a count of today's population. Source periods, eligible universes, table variables, geographic definitions, and available uncertainty must accompany demographic quantities. National, state, and metropolitan products need compatible geography definitions; overlapping metropolitan and state totals cannot simply be added.

The released questionnaire-pretest recipe uses four adult age groups—18–34, 35–44, 45–64, and 65 years or older—and three education groups: high school or less; some college or an associate degree; and a bachelor's degree or higher. Their cross-product contains twelve cells. The product artifact retains both resident-adult and household-adult masses. The GSS benchmark uses the household-adult masses to better align with its household survey universe. Matching that broad universe does not remove differences in coverage, nonresponse, field dates, language, or mode.

These cell masses come from the fitted demographic artifact, rather than an assumption that age and education are independent. The source release and fitted-artifact hashes identify the chosen inputs. Agreement with selected controls is a diagnostic of the demographic construction. It does not demonstrate that every unobserved local joint distribution is recovered, and it supplies no direct observation of a person's political beliefs. In particular, adding a personality label to a demographic case would require its own evidence and validation.

### 3.3 Human evidence and qualitative dossiers

Human evidence has a separate contract: an exact item, eligible population, field period, substantive response coding, survey weight, and appropriate uncertainty method. CPS participation measures, ANES attitude measures, and GSS confidence items cannot be treated as interchangeable simply because each concerns civic life. Evidence from one geography or universe does not automatically establish an estimate for another.

The current US catalog also includes six authored qualitative dossiers. A study may attach up to two to the wording-review stage. Each dossier describes hypothesized constraints, priorities, protected values, a trade-off, information context, possible friction, a counterexample, and a fieldwork probe. Protected-values research provides a conceptual motivation (Baron and Spranca 1997). A dossier is an authored research hypothesis, not a survey respondent, validated personality type, or demographic allocation. Its population weight is zero. The separation prevents an evocative character from silently changing the numeric sample composition.

# 4 Computational procedure

## 4.1 Frozen study revision

A study revision binds the scenario or proposal variants to the questionnaire, audience selection, demographic cells, selected dossiers, and execution recipe. It records the population release hash, fitted-artifact hash, runtime hash, model identifier, temperature, thinking setting, and output-token cap. The report records actual model use and the number of human interviews. The US generated pretest reports zero human interviews.

Figure 1 shows the logical information flow. Demographic data supply the weights; the language model supplies proposed response distributions. Human evidence is an external comparator, rather than an implicit source of extra generated respondents. Dossiers enter a separate qualitative review. The graph makes these dependencies inspectable without implying that graph connectivity proves empirical truth.

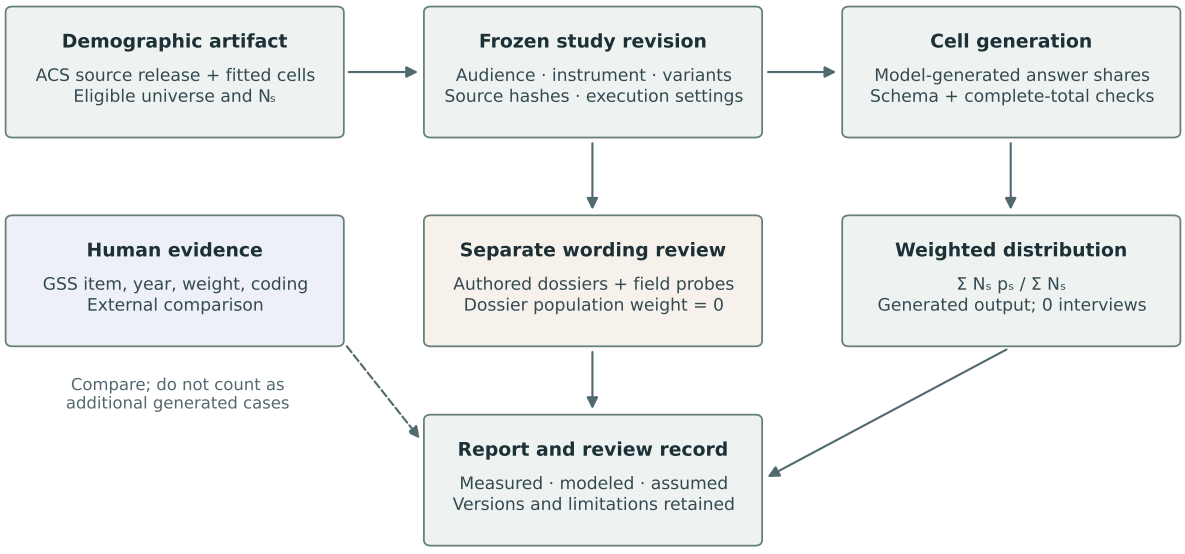


**Figure 1:** Information flow in the US pretest. Solid arrows indicate computational dependencies; the human-evidence comparison remains separate from generation. Authored dossiers contribute to wording review with zero population weight.

### 4.2 Cell-level answer generation and validation

For each proposal version, the same required multiple-choice questionnaire is supplied to the model for every eligible demographic cell. The current US recipe accepts one to four questions, each with two to eight distinct answer options. Randomized presentation, conditional questions, and an additional don't-know option are outside this particular contract. The result is an integer percentage for each option, in the instrument's question order, summing to 100 for each question. Invalid or incomplete records fail validation rather than being silently treated as complete results.

Let $s$ index the included cells, $q$ a question, and $k$ an answer category. Let $N_s$ be the selected fitted population mass, and let $p_{sqk}$ be a validated model share expressed on the unit interval. The aggregate generated distribution is

$$\hat{p}_{qk} = \frac{\sum_{s \in S} N_s p_{sqk}}{\sum_{s \in S} N_s}.$$

The weights determine how much each cell contributes. They do not establish that $p_{sqk}$ is a valid estimate of the corresponding human response probability. The twelve calls are not twelve independent survey respondents, and the population masses are not the generated sample size.

One additional structured review is requested per proposal version. It receives the stimulus, questionnaire, and selected dossiers, and returns potential wording issues and field checks using supplied identifiers. Dossiers are excluded from the numeric weighting step. This provides a traceable division between distribution generation and qualitative hypothesis generation.

Comparing proposal versions yields differences in model-supplied shares under a fixed audience and instrument. These differences can suggest an issue to test. They are not a randomized experiment with people and do not identify a human treatment effect. The paper does not claim that a particular suggested wording change has been shown to improve an actual research outcome.

### 4.3 Provenance, replay, and the small graph

The relevant graph is a bounded record of the analysis: source release, demographic artifact, audience selection, instrument, proposal variant, execution, finding, and report. Relations such as "derived from," "used in," and "compared with" distinguish computational dependence from evidentiary support. An amended assumption creates another revision, so a colleague can recover which materials accompanied a conclusion.

There are three distinct reproducibility levels. **Score replay** recomputes metrics from the published answer distributions. **Input reconstruction** requires the exact source extracts, transformations, cell masses, instrument, and prompts. **Generation replay** additionally depends on provider model availability and execution behavior. A hash can detect a changed byte sequence, but it cannot recover an unavailable input or force a hosted model to return the same answer. The public package in this report establishes score replay; it does not claim independent end-to-end replication of the proprietary generation pipeline.

# 5 Retrospective GSS comparison

## 5.1 Design and comparators

The public benchmark compares the 2022 and 2024 GSS cross-sections on twelve recurring institutional-confidence items (NORC at the University of Chicago 2026). It retains substantive response codes 1–3 and applies the released `WTSSNRPS` survey weights within each year's eligible, answered item set. The 2022 release has 3,544 eligible cases and the 2024 release 3,309; item-specific substantive answers and weighted denominators are retained in the aggregate file. These case counts are not claims of an equal effective sample size across items.

Three methods are evaluated against the 2024 weighted distribution: the corresponding 2022 human distribution; a single national model call for each item; and the twelve-cell, Census-weighted model recipe. The generation receipt records `gemini-3.5-flash-lite` on Vertex AI, completed on 23 November 2026. This identifier belongs to the benchmark receipt and should not be inferred from whichever model the live product currently offers. The model task used the substantive item text and the 2024 context. `CONLABOR` is excluded because of split wording in the target cross-section. Each retained item receives equal weight in the final average.

The comparison is retrospective. The target answers were public before generation, and training exposure is possible. The released protocol identifiers document the run; they do not establish prospective preregistration or concealment of the target data. Forms, presentation, interview mode, field dates, and language coverage can differ between the two human waves. No claim of exact instrument equivalence beyond the documented item matching is made.

## 5.2 Metrics

For a forecast distribution $p$ and the target weighted distribution $y$, with three substantive categories, total variation is

$$\mathrm{TV}(p, y) = \frac{1}{2} \sum_{k=1}^{3} |p_k - y_k|.$$

We report 100 TV in percentage points. It measures the amount of answer mass that must move to match the target. It is not the maximum error of any specific answer category.

The scorer also reports the half-scaled multiclass Brier score, averaged over the target category shares:

$$\mathrm{B}(p, y) = \frac{1}{2} \left( 1 + \sum_{k=1}^{3} p_k^2 - 2 \sum_{k=1}^{3} p_k y_k \right).$$

This is the expectation of a proper squared-error scoring rule under the observed target distribution (Brier 1950). Even a forecast equal to a nondegenerate target distribution has nonzero Brier loss. Consequently, the reported value is not the squared distance between the two aggregate distributions, and it does not establish calibration for future events. Lower scores are preferable for both metrics in this fixed comparison.

## 5.3 Results

| Method | Mean TV (pp) | Mean Brier |
|---|---|---|
| 2022 human survey | 2.48 | 0.29274 |
| National model call | 7.50 | 0.29771 |
| Census-weighted model | 8.16 | 0.29892 |

Table 2. Equal-item mean errors across twelve institutional-confidence questions. Values are recomputed from the public aggregate file; lower is better. TV is expressed in percentage points. Brier uses the half-scaled multiclass convention above.

The earlier human survey has the lowest mean error on both metrics. The weighted cell recipe has a mean TV 0.66 percentage points higher than the national model call. The earlier human survey's advantage over the weighted recipe is 5.68 points. This comparison therefore provides no evidence that demographic conditioning improved average distributional agreement on the selected battery. No confidence interval or significance claim is attached to these differences: they summarize a fixed set of related items and a single released generation run.

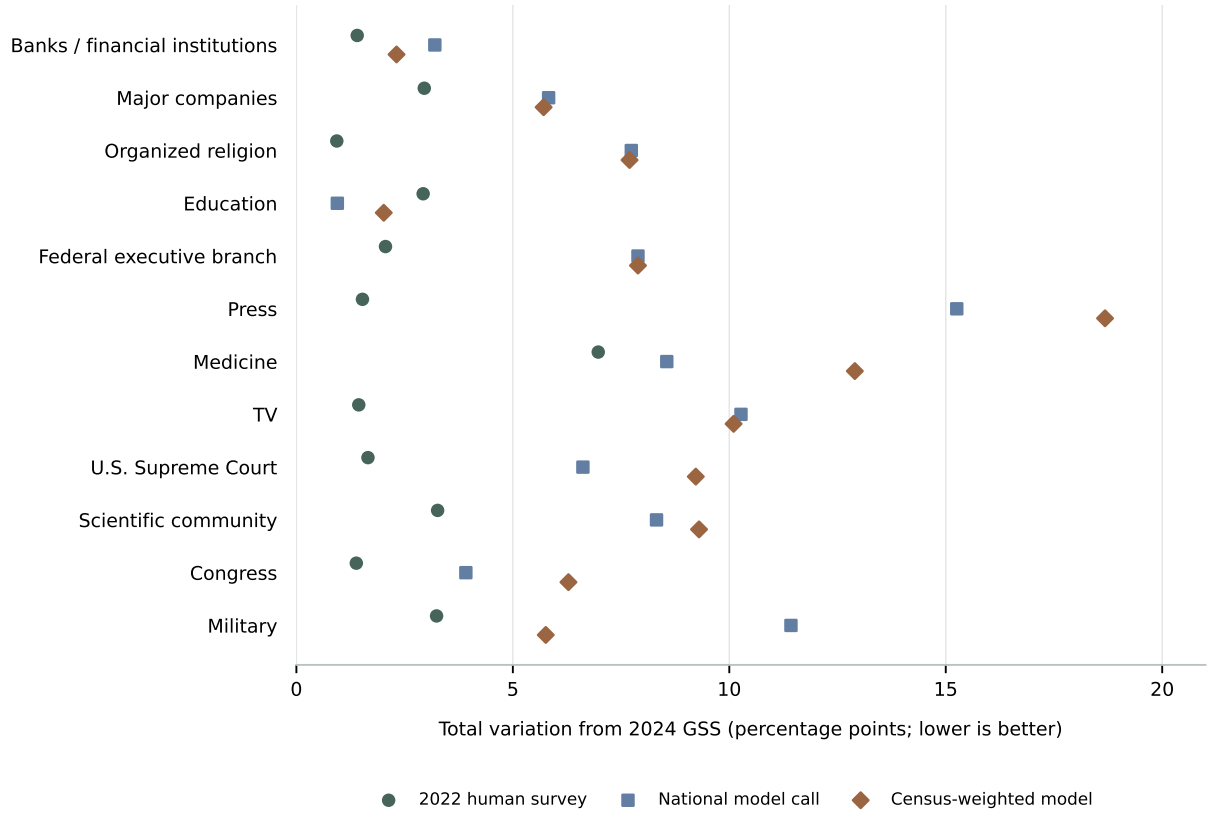


**Figure 2:** Item-level total variation from the weighted 2024 GSS distribution. Points show the two model methods and the earlier human-survey comparator. The battery is a retrospective, correlated set of items; the plot is descriptive and includes no inferential error bars.

The largest weighted-recipe discrepancy occurs for confidence in the press, at 18.68 TV points; the national model call is also substantially different from the target on that item. Individual items can favor a model method even when the mean favors the human baseline. Table A1 reports every scored item so readers can inspect heterogeneity rather than relying on a headline average.

# 6 Optional legacy scenario model

The existing methodology page also documents a legacy Signal engine and 27 authored behavioral archetypes. These form an optional exploratory model, separate from the source-backed demographic releases and the GSS pretest comparison. Archetype labels, traits, and weights are authored modeling choices. They do not partition the US population empirically, and the paper assigns them no measured population totals.

The deterministic engine maps scenario text to ten signal dimensions and then to archetype reactions and operational indicators. Its outputs can be written abstractly as $z = F_\theta(x, a)$, where $x$ is the scenario, $a$ the selected archetype assumptions, and $\theta$ the versioned scoring rules. Determinism supports software replay. It does not validate the psychological interpretation of $F_\theta$. Moral Foundations research motivates some constructs, but the theory and its cross-cultural measurement do not validate the engine's authored coefficients (Atari et al. 2023).

The legacy reference lists 34 operational indicators. Historical names such as "Backlash Probability" refer to model scales and must not be read as calibrated event probabilities. Related indicators share inputs and are not independent confirmations of a result. The legacy 13-event historical showcase corpus is a curated retrospective diagnostic; reuse during heuristic development can bias apparent performance. Its outcomes and polarization labels are not interchangeable with the GSS item distributions reported above.

A seeded 50-run perturbation ensemble explores declared variation in the legacy assumptions and summarizes model-output percentiles. Those ranges describe the chosen perturbations, not survey sampling error or validated forecast intervals. The paid saved-Study modes add structured model interpretation to canonical Signal results; generated prose does not create an independent numeric estimate. A circuit breaker limits repeated provider failures but cannot detect all factual or substantive errors.

The full operational catalog remains a technical reference. It is not reproduced as a list of academically validated measures, and the GSS comparison must not be presented as validation of those indicators.

# 7 Validity, limitations, and responsible use

Several limitations constrain the conclusions. First, the twelve confidence items are a narrow, related battery. They are not a random sample of all attitudes, policy questions, products, places, or historical periods. Second, model training contamination is unresolved because the target wave was public. Third, one recorded run cannot characterize model-version variation, provider changes, stochastic variability, or prompt sensitivity.

Fourth, demographic conditioning covers age and education in this recipe, not all determinants of response. A model can stereotype a cell or miss disagreement within it. Fifth, household coverage, item nonresponse, survey modes, and period effects complicate comparisons across GSS waves and between the survey and Census frame. Sixth, released aggregate scores do not include a design-based variance analysis with the full survey design; no synthetic respondent count repairs that omission.

Seventh, the public package allows score verification but not full reconstruction of the private workbook mapping, prompt templates, and generation pipeline. Hashes document the receipt without making those materials open. Eighth, qualitative usefulness is not established by numerical distribution matching. Whether the wording review helps researchers find important defects remains a

separate empirical question.

The intended use is preparation for research: examining audience definitions, identifying unclear wording, surfacing assumptions, and planning checks with people. The output should identify which statements are measured, modeled, or hypothesized. It should not be described as representative public opinion, a completed poll, an election forecast, or evidence that an individual holds a particular belief. Researchers remain responsible for consent, source terms, privacy, and the communication of uncertainty. This report follows disclosure principles rather than claiming certification or endorsement by AAPOR or any cited author.

## 8 Evaluation agenda

A stronger validation program should test claims independently. A prospective study would freeze the model version, input sources, item-selection rule, prompts, analysis code, and primary metrics before collecting or exposing the target responses. A contemporaneous human baseline and a simple historical or marginal baseline should be included. The comparison should cover new items, relevant subgroups, unsupported or missing cells, and changes in geography and time.

An ablation study should distinguish demographic weighting from demographic prompting, dossier review, and model choice. Multiple runs and model versions should assess stability. Human survey uncertainty should be estimated with the appropriate design information, while model variability should be reported separately. Results should include distributional error, within-group variation, and systematic failures, rather than relying on aggregate correlation alone.

The wording-review feature needs a separate blinded evaluation. Researchers could compare AI-flagged and independently identified questionnaire defects, record false positives, and test whether revisions improve comprehension or measurement in cognitive interviews and a human pilot. Prospective usefulness would then be a measured outcome, rather than an inference from plausible generated criticism. These are proposed evaluations; they have not been completed by the benchmark in this paper.

## 9 Conclusion

Wind Tunnel organizes the context of research so that demographic inputs, human measurements, authored assumptions, and generated outputs can be examined separately. The US pretest supplies a bounded numeric procedure and a distinct qualitative review, with versioned records connecting inputs to reports. Its public GSS comparison is reproducible at the aggregate-scoring level and shows that the earlier human survey outperforms both released model methods on the selected battery. The defensible contribution is inspectability and support for research preparation. Claims about representative opinion, human wording effects, or predictive accuracy require additional, independent evidence.

## 10 Availability, authorship, and disclosures

**Publication and citation.** TACITUS. 2026. *Wind Tunnel: Evidence-Separated Population Modeling and AI-Assisted Questionnaire Pretesting.* Working paper WT-2026-01, version 1.0. The canonical release, PDF, manuscript source, bibliography, figures, aggregate benchmark, and scoring code are linked from the public research repository. No journal acceptance, external review, DOI registration, or institutional endorsement is claimed.

**Authorship and interests.** TACITUS is the institutional author and publisher and develops the software described here. This commercial interest should be considered when interpreting its claims. AI assistance was used in preparing the manuscript and checking sources and code; it is not an independent scientific reviewer. The manuscript distinguishes directly recomputed results from implementation descriptions and proposed research.

**Data and materials.** The benchmark release contains aggregate response distributions, item identifiers, source references, run hashes, and a Node.js scorer. It contains no individual GSS respondent records. The GSS, ACS, and cited instruments retain their original source terms. This manuscript and its original figures follow the repository's license. The private application and generation templates are not represented as open source.

## Appendix A. Item-level benchmark

| GSS item / institution | 2022 human | National model | Weighted model |
|---|---|---|---|
| `confinan` · Banks and financial institutions | 1.40 | 3.20 | 2.31 |
| `conbus` · Major companies | 2.95 | 5.83 | 5.71 |
| `conclerg` · Organized religion | 0.93 | 7.74 | 7.70 |
| `coneduc` · Education | 2.93 | 0.95 | 2.02 |
| `confed` · Executive branch of the federal government | 2.06 | 7.89 | 7.89 |
| `conpress` · Press | 1.53 | 15.25 | 18.68 |
| `conmedic` · Medicine | 6.97 | 8.56 | 12.90 |
| `contv` · TV | 1.44 | 10.27 | 10.10 |
| `conjudge` · U.S. Supreme Court | 1.65 | 6.62 | 9.23 |
| `consci` · Scientific community | 3.26 | 8.32 | 9.30 |
| `conlegis` · Congress | 1.38 | 3.91 | 6.28 |
| `conarmy` · Military | 3.24 | 11.42 | 5.76 |

Table A1. Total variation in percentage points by retained GSS item. The item identifiers and all three-category distributions are in the aggregate benchmark. Each item contributes equally to Table 2.

## Appendix B. Reproduction record

The aggregate benchmark and original scorer are pinned to public repository commit `2b51dfee375b1ee0e5fd1ee7a99edd1623d8cac4`, directory `research/wind-tunnel-us/v1`. The application implementation reviewed for this report is version 3.12.1 at commit `546b24d1f47671ff917ac1a09eb3bc5d9ab54d15`. The paper's generation script recomputes the tables and figure from the released distributions; it does not call an AI model.

From the public repository root, aggregate verification requires Node.js 20 or later:

```
node research/wind-tunnel-us/v1/score.mjs verify \
  research/wind-tunnel-us/v1/benchmark.json
```

Rebuilding the publication requires Python 3 with Matplotlib, Pandoc with citeproc, and XeLaTeX. The publication builder records the benchmark SHA-256, publication version, and output hashes in `manifest.json`. Changing the source creates a new publication artifact; the versioned release should remain unchanged.

The benchmark receipt retains protocol hash `f4752712c623d1fec445d913ebd50645df9620b2e6feb9f693b2d01eb14b941a`, input hash `4beacae21cafa9bc39b9a87d4b29fafee3657f81ea1c15d63f270d1757b6a9f9`, and output hash `7f1fbffe45f03aa2254710c816b0eb60f93780aca98b35ea2e544d2db6de6a04`. These identify the recorded run; they are not proof of a concealed target or independent replication.

## References


American Association for Public Opinion Research. n.d. "Transparency Initiative: Disclosure Standards." Accessed November 23, 2026. https://aapor.org/standards-and-ethics/transparency-initiative/.

Argyle, Lisa P., 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): 337–51. https://doi.org/10.1017/pan.2023.2.

Atari, Mohammad, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T. Stevens, and Morteza Dehghani. 2023. "Morality Beyond the WEIRD: How the Nomological Network of Morality Varies Across Cultures." *Journal of Personality and Social Psychology* 125 (5): 1157–88. https://doi.org/10.1037/pspp0000470.

Baron, Jonathan, and Mark Spranca. 1997. "Protected Values." *Organizational Behavior and Human Decision Processes* 70 (1): 1–16. https://doi.org/10.1006/obhd.1997.2690.

Bisbee, James, 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): 401–16. https://doi.org/10.1017/pan.2024.5.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3. https://doi.org/10.1175/1520-0493(1950)078%3C0001:vofeit%3E2.0.co;2.

Deming, W. Edwards, and Frederick F. Stephan. 1940. "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known." *The Annals of Mathematical Statistics* 11 (4): 427–44. https://doi.org/10.1214/aoms/1177731829.

Gelman, Andrew, and Thomas C. Little. 1997. "Poststratification into Many Categories Using Hierarchical Logistic Regression." *Survey Methodology* 23 (2): 127–35. https://www150.statcan.gc.ca/n1/pub/12-001-x/1997002/article/3616-eng.pdf.

Grimm, Volker, Steven F. Railsback, Christian E. Vincenot, Uta Berger, Cara Gallagher, Donald L. DeAngelis, Bruce Edmonds, et al. 2020. "The ODD Protocol for Describing Agent-Based and Other Simulation Models: A Second Update to Improve Clarity, Replication, and Structural Realism." *Journal of Artificial Societies and Social Simulation* 23 (2). https://doi.org/10.18564/jasss.4259.

Lebo, Timothy, Satya Sahoo, and Deborah McGuinness. 2013. "PROV-O: The PROV Ontology." W3C Recommendation. World Wide Web Consortium. https://www.w3.org/TR/prov-o/.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. "Model Cards for Model Reporting." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–29. https://doi.org/10.1145/3287560.3287596.

NORC at the University of Chicago. 2026. "General Social Survey 1972–2024: Cross-Section Documentation and Public Data." https://gss.norc.org/get-documentation.html.

Tjuatja, Lindia, 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: 1011–26. https://doi.org/10.1162/tacl_a_00685.

U.S. Census Bureau. 2026. "2020–2024 American Community Survey: Five-Year Estimates and Public Use Microdata Sample Documentation." https://www.census.gov/programs-surveys/acs/microdata/documentation.html.