SHARED AUTOREGRESSIVE CONTEXT CAN DISTORT RELATIONSHIPS IN SYNTHETIC DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can generate several records within one autoregressive completion, making earlier answers available as context for later records. This paper shows that such shared-completion batching can distort relationships among variables in the resulting synthetic data, using controlled tests on synthetic survey respondents. In a matched experiment on $2{,}000$ European Social Survey profiles, generating ten rather than one respondent per request increases mean absolute error in within-country correlations by 48–58% for Qwen3.8-27B and 114–127% for Llama-3.3-70B-Instruct across three seeds, holding profiles, examples, questions and decoding parameters fixed. The distortion primarily reflects exaggerated relationship strength, while retaining substantial agreement with the human ordering of correlations. Controlled interventions establish answer history as a causal channel: re-pairing the same preceding values, with profiles and marginal distributions fixed, changes correlations among subsequently generated responses. Hiding preceding answers reduces correlation error in the tested settings but worsens marginal accuracy. Exploratory corrections across social-attitude, health and economic data likewise show that lower correlation error can coexist with worse marginal distributions and regression estimates. Request construction is therefore part of the data-generating process, and synthetic-data validity must be evaluated against the analyses the generated data are intended to support.

1 INTRODUCTION

Generating several records in one autoregressive completion can change the statistical relationships in the resulting synthetic data. A large language model (LLM) can answer for ten people in one completion, or for the same ten profiles in ten separate completions. In the shared completion, earlier answers become context for later ones and can influence the population statistics estimated from those records. Here, *shared-completion batching*, or batching below, refers to respondents sharing a completion, independently of how a serving system schedules requests.

For a generator intended to sample respondents independently conditional on their profiles, a useful desideratum is *request-partition invariance*: splitting the same profiles across requests should not materially change the population quantities the sample is intended to estimate. This criterion allows ordinary variation between random draws and concerns the distribution of repeated samples; autoregressive generation does not guarantee it.

Synthetic survey respondents provide a controlled testbed for this criterion. Silicon sampling generates answers for people described by demographic profiles (Argyle et al., 2023), and applied studies sometimes request dozens of respondents together (Jahn et al., 2026; Miklian et al., 2026). Human reference samples support comparisons of both individual-variable distributions and relationships among variables: matching income and trust distributions alone does not establish whether income predicts trust within a country.

The main experiment holds European Social Survey (ESS) profiles, examples, questions and decoding settings fixed while changing request size. Generating ten respondents rather than one raises mean absolute error across 171 within-country correlations by 0.095 for Qwen3.8-27B and 0.228 for Llama-3.3-70B-Instruct, with relative increases of 48–58% and 114–127% across three seeds (Figure 1). The main change is exaggerated relationship strength; batched correlations still retain substantial agreement with the human ordering.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

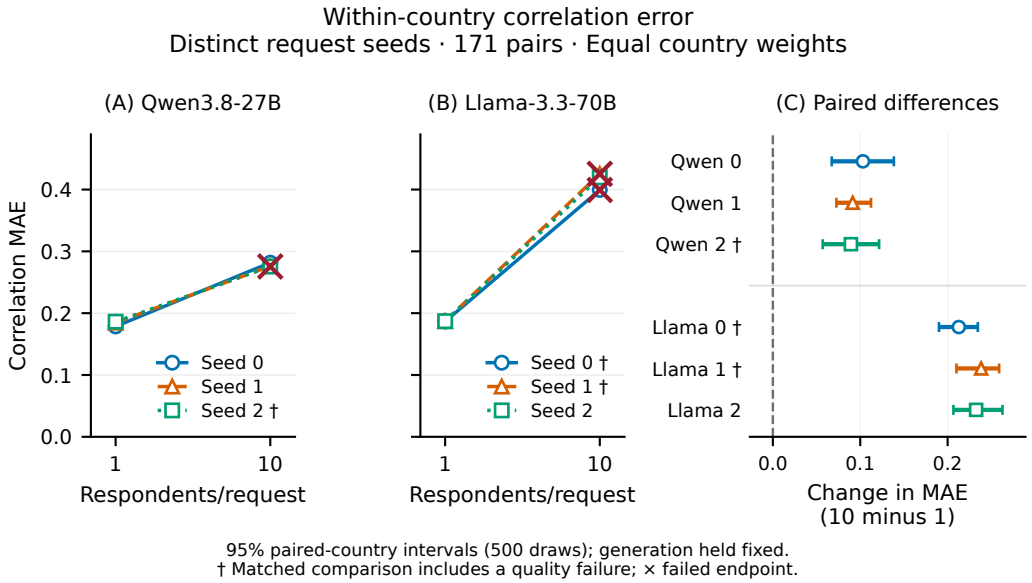


Figure 1: **Sharing a completion increases correlation error in the matched ESS experiment.** (A,B) Mean absolute error against human data across 171 within-country correlations, with equal country weights and the same accepted respondents in all four conditions. Each line shows one seed. Error increases by 48–58% for Qwen and 114–127% for Llama. (C) The increase from one to ten respondents, with 95% paired-country percentile bootstrap intervals (500 resamples). These intervals condition on the generated answers and twenty selected countries; they omit uncertainty over new seeds and other countries. Conditions below 99% validity are marked as quality failures. Variable pairs are not independent replications.

Controlled interventions then test the influence of answer history directly. In health and political-trust templates, re-pairing exactly the same preceding values changes correlations among newly generated respondents in both models. The supplied profiles and each variable's distribution stay fixed; only the pairing changes. This establishes answer history as a causal channel through which shared context can alter later relationships (Figure 2), while leaving the model's internal computation open. Hiding earlier answers provides a complementary test against human data: correlation error falls even when only one respondent precedes the target, although marginal accuracy worsens.

These experiments isolate request size and manipulate history, extending earlier evidence of distorted synthetic relationships (Bisbee et al., 2024; Lukauskas & Šarkauskaitė, 2026; Ma et al., 2026). Exploratory corrections then examine whether improved correlation accuracy makes the data more useful. Across social-attitude, health and economic data, gains can coexist with worse marginal distributions and regression estimates. Request construction belongs in the generator specification, and its evaluation must address the quantities an intended analysis requires.

2 Data and Evaluation

Data. Three human surveys provide demographic profiles and reference answers, beginning with 2,000 European Social Survey (ESS) respondents in the main experiment: one hundred in each of twenty countries. Their ten background characteristics condition the generation of nineteen numeric responses covering trust, political attitudes, well-being, income, age and interview year.

The 2017–2018 National Health and Nutrition Examination Survey (NHANES) (National Center for Health Statistics, 2020) extends the study to U.S. adult health, while the 2022 American Community Survey (ACS) (U.S. Census Bureau, 2023) supplies income and work data for employed Massachusetts adults. Each contributes 800 people, with one hundred in each of eight groups defined by two sex categories and four age bands. The model receives each person's recorded sex, age

band, race/ethnicity, education and marital status; Appendix M.1 lists the outcomes, age bands and eligibility rules.

Matched comparison. For each target person, the model receives a demographic profile alongside three examples pairing other people's profiles with their real answers. The target's own survey answers remain hidden, and a separate set of human respondents provides the evaluation reference. The model generates answers for the same targets one at a time or ten at a time, keeping profiles, examples, question order and non-seed generation settings fixed. Each request receives a distinct random seed, with three runs per model at temperature 1. Appendices A and M.1 give the sampling and request-construction details.

Invalid answers. The validation procedure rejects invalid or truncated requests in full, without repair or regeneration, and comparisons use profiles accepted across the relevant conditions. A condition has a *quality failure* if fewer than 99% of requests are valid or any are truncated. The analysis retains such conditions, since matching accepted profiles cannot recover rejected answers.

Accuracy. The primary measure concerns relationships among variables across people within the same country or demographic group. Subtracting each group's mean from each variable before calculating correlations removes differences in average response levels between groups. For a real or synthetic population P, with c denoting the group,

$$r_W^P = \mathrm{Corr}_P\big(Y_1 - \mathbb{E}_P[Y_1 \mid c], Y_2 - \mathbb{E}_P[Y_2 \mid c]\big). \tag{1}$$

With groups receiving equal weight, correlation MAE (A) averages the absolute differences between synthetic and human correlations. Further checks distinguish strength from pattern by comparing the ordering of correlations, a zero-correlation baseline, and an *oracle* adjustment that chooses one multiplier in $[0, 1]$ to minimize error against human answers. The oracle diagnoses how much error simple weakening can remove; it uses evaluation answers and is not an available correction (Appendix E.1).

Marginal Wasserstein error measures how far values must move, on average, to match each human variable's distribution. Dividing by the variable's range (M) or human standard deviation (M_{SD}) puts errors on a common scale; the latter is emphasized for ACS, where broad income ranges can hide errors (Equation 4). Fixed regression tasks assess accuracy for particular analyses (Appendix C).

Uncertainty. Seed ranges describe variation across the three runs, while bootstrap intervals assess sensitivity to resampling countries, profiles or answer orders, depending on the experiment. Each draw keeps the compared conditions paired and uses the existing generated answers and human reference; separate checks resample the human reference itself. Appendix E.5 and the study-specific appendices specify which sources of uncertainty each interval covers.

3 Shared-completion batching and answer history

3.1 Ten-person requests exaggerate relationship strength

Requesting ten respondents rather than one increases ESS correlation error in every run of both models under distinct per-request seeds (Figure 1). On profiles accepted in all four combinations of request size and seed policy, mean MAE rises from 0.183 to 0.278 for Qwen and from 0.187 to 0.415 for Llama. The second seed policy resets every request to the same run seed and serves as a diagnostic; the primary contrast uses distinct seeds. Across runs, the relative increases are 48–58% and 114–127%, with error increasing for 155–159 of 171 pairs in each Qwen run and 148–154 in each Llama run. Although these overlapping pairs are not independent tests, they show that the effect extends across much of the correlation matrix. All six paired-country intervals for the increase in mean error exclude zero. The effect also appears in checks that use all retained answers, remove age and year, or reset every request to the same seed (Appendix A).

Much of this increase reflects stronger relationships rather than a loss of information about their ordering. Batching raises mean $|r|$ from 0.360 to 0.459 for Qwen and from 0.355 to 0.596 for Llama,

compared with 0.193 in the human reference. Yet the batched correlations still track the human ordering, with Spearman agreement of 0.864 and 0.866. After the oracle magnitude adjustment applies the best common multiplier to each set of correlations, one- and ten-person MAE are both approximately 0.10 (Table 19). The main distortion is therefore an exaggeration of relationship strength in data that retain substantial information about the human pattern.

The increase also survives checks on output validity and format. All $26{,}400$ requests finish without truncation, although three original ten-person conditions fall below 99% validity: Qwen seed 2 (97.5%) and Llama seeds 0 and 1 (98%). The effect holds in the three conditions that pass as well, with matching retaining at least eighty profiles per country. Constraining generation to the required output format likewise leaves the effect in ESS and NHANES: ESS mean MAE rises by 51% for Qwen and 122% for Llama, with corresponding NHANES increases of 56% and 54%. ACS provides weaker evidence because all six ten-person conditions fail validity and Llama's effect varies in direction across seeds (Table 36).

The penalty also persists in both models at temperature 0.5 and after removing the instruction to "Preserve realistic individual variation" at temperature 1. These checks show that the effect does not require that wording or the original temperature. At 1.5, Llama remains valid and retains the penalty, but many Qwen ten-person requests fail validation, preventing its joint ESS estimate and NHANES intervals. Higher-temperature effects therefore depend on the model (Appendix F).

This consistent correlation penalty coexists with better individual-variable distributions. On the correction study's matched evaluation cohorts, batched outputs have lower marginal Wasserstein error than one-person outputs in all six settings; for ESS/Qwen, range-normalized M is 0.115 compared with 0.127 (Table 42). One-person generation is therefore not uniformly more accurate: the choice between request sizes changes both how individual variables are distributed and how they vary together.

3.2 Answer history is a causal channel

Changing pairings while keeping values fixed. The matched request-size experiment establishes a difference between generation procedures. A controlled intervention on answer history then tests whether earlier answers can change later relationships. The model receives ten demographic profiles and completed answers for the first nine respondents, then generates only the tenth person's answers. The comparison re-pairs the preceding values while keeping those values, their marginal distributions and the profiles fixed (Figure 2A).

In NHANES, each supplied respondent has a BMI and a value for daily sedentary minutes. One version pairs higher BMI with more sedentary time, producing a positive relationship; a second pairs higher BMI with less sedentary time, producing a negative relationship. A third arranges the same values so that their correlation is zero. Because every version contains exactly the same BMI and sedentary-time values, these changes alter their pairing without changing either variable's distribution. The ESS experiment follows the same procedure using trust in parliament and the judiciary. In both domains, the ten profiles in each template are identical except for IDs, so rearranging answers also preserves their agreement with the supplied demographics.

To separate pairing from a particular presentation order, eighteen forward and reverse rotations place every value equally often in each position. Each order has three repetitions with generation seeds paired across conditions, giving $12{,}096$ continuations across two models and two domains, including a reference condition that hides the supplied answers. The correlations describe the newly generated tenth respondents only; none of the nine supplied answers enters the calculation.

Changing history pairing from positive to negative moves target correlations from 0.963 to -0.670 for Qwen/NHANES and from 0.985 to -0.647 for Qwen/ESS. Llama shows the same direction of change, from 0.837 to -0.329 in NHANES and from 0.725 to 0.041 in ESS (Figure 2A). Every template shows the same direction of change. After subtracting the mean response for each exact answer order within each condition, the mean positive-minus-negative contrasts also remain positive, so differences between orders' average responses cannot explain the whole result (Appendix B). Earlier answer pairings therefore influence later relationships, although the experiment leaves open whether the model extracts covariance, extends each variable's sequence separately, or uses another computation.

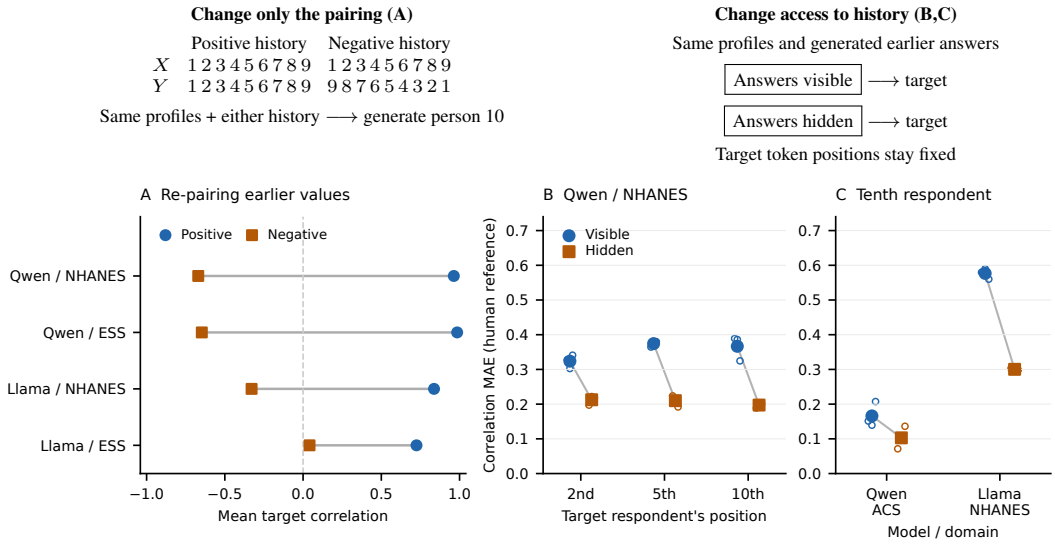


Figure 2: **Manipulating earlier answers changes later relationships.** The schematic shows the interventions; ranks illustrate pairing, with actual values fixed. (A) Mean correlation among newly generated tenth respondents, averaged across prompt templates, after positive or negative pairing of the same preceding values. This panel measures generated relationships, without comparing them with human data. Llama uses a separately validated NF4 runtime. (B) Qwen/NHANES correlation MAE for profiles accepted in all six conditions. (C) Tenth-respondent extensions, matching profiles across the two conditions. Filled markers average three repetitions; hollow markers show each repetition. Hiding history skips the earlier answers' state updates while retaining the target's logical token positions. Mean marginal error rises in every comparison (Tables 6–7).

Hiding previously generated answers. The pairing experiment establishes that history can change later relationships; a complementary intervention asks whether access to that history affects accuracy against human data. Here the model generates the earlier respondents' answers itself. For 400 Qwen/NHANES profiles, the person whose answers enter the evaluation appears second, fifth or tenth in a ten-profile prompt, giving that person one, four or nine preceding responses.

Each history then supports two generations of the target's answers using the same prompt and seed. In the visible condition, the model processes the preceding answers as usual; in the hidden condition, their tokens cannot update attention keys and values or Qwen's recurrent and convolution states. Generation resumes at the target's original logical token positions, preserving its position while removing the earlier answers' contribution to the model state. The tenth-position experiment also extends to Llama/NHANES and Qwen/ACS.

On Qwen/NHANES profiles accepted in all six position–visibility conditions, hiding history reduces mean MAE from 0.324 to 0.213 at position two, 0.374 to 0.210 at five, and 0.367 to 0.197 at ten. The corresponding reductions are 0.577 to 0.300 for Llama/NHANES and 0.166 to 0.103 for Qwen/ACS, with lower raw MAE in every repetition and all paired profile-bootstrap intervals excluding zero. These gains come with worse individual-variable accuracy: mean marginal error rises in each setting, including increases of 8–11% in Qwen/NHANES range-normalized M across positions and from 0.348 to 0.350 in Qwen/ACS M_{SD}. Access to earlier answers thus improves marginal distributions while worsening correlation accuracy in these comparisons.

The gain at position two also narrows the possible explanations, since a single preceding respondent cannot define a cross-respondent correlation. History can therefore matter without the model first extracting a covariance matrix from several preceding respondents.

Relationship strength and pattern. The magnitude diagnostics qualify what this reduction in raw error means. In Llama/NHANES, hiding history slightly worsens both oracle-adjusted MAE (0.099 to 0.104) and agreement between the correlation vectors (0.749 to 0.723), suggesting that the raw gain largely reflects weaker correlations. Qwen/NHANES at position ten shows a broader

improvement: oracle MAE falls from 0.081 to 0.065 and pattern agreement rises from 0.848 to 0.894 (Table 21). General weakening can therefore explain much of the apparent benefit, but does not capture every setting's improvement.

Another possible explanation is that each request raises or lowers all its respondents' answers together. Subtracting each request's mean, with one-person outputs first grouped into the same ten-profile blocks, makes the ESS batching penalty larger and leaves a larger leading correlation-matrix eigenvalue under batching. Differences in request means alone therefore cannot explain the effect. Appendix E.4 reports this check and exploratory regressions relating targets to their predecessors.

Supporting context controls. Hiding history changes state updates and leaves a gap in token positions. A matched Qwen/NHANES follow-up closes that gap: MAE is 0.363 with visible history, 0.192 with hidden history at the original positions and 0.193 at consecutive positions. The benefit therefore survives removal of the gap. In the same runtime, showing only the target profile gives lower MAE than showing ten profiles with the target first and no preceding answers (0.198 versus 0.247). These controls distinguish profile context and answer access within a common continuation procedure; runtime and design differences prevent assigning them separate shares of the original batching penalty (Appendix B.4).

4 Dependence fidelity is not enough

Shared-completion batching and access to earlier answers can improve marginal accuracy while worsening correlation accuracy. Reversing the correlation distortion therefore need not improve the data for every use. Exploratory corrections examine this trade-off by changing dependence after generation and assessing the resulting data. The comparisons use standard operations, including rank-based dependence adjustment (Iman & Conover, 1982), as diagnostic tools to test what lower correlation error establishes about usefulness.

A diagnostic transformation. A full dependence map puts batched answers on a common normal-score scale, adjusts their correlation matrix toward that of separate one-person fitting outputs, and converts the scores back to answer units using the one-person fitting distributions. Fitting and evaluation use disjoint whole requests, with matched profiles across request sizes; fitting excludes human evaluation answers. Because the study reuses previously examined outputs and human references, all correction results remain exploratory. Appendix D gives the map, fitting design and comparison methods.

Improved correlations, worse marginals. The full map lowers mean correlation error in all six domain–model settings while worsening marginal accuracy in each (Figure 3). The result also holds on the more informative SD-normalized scale for ACS: marginal error rises from 0.376 to 0.423 for Qwen and from 0.418 to 0.525 for Llama. Much of the correlation gain requires little pair-specific information. Replacing the full target matrix with one multiplier for all off-diagonal correlations, fitted using the same one-person outputs, retains 93–101% of the MAE improvement (Table 12). In NHANES, both full maps still have larger correlation errors than predicting zero for every pair. Lower MAE can thus reward weakening without establishing recovery of the human pattern.

Preserving marginals does not settle usefulness. A further control keeps each group's original list of batched values for each variable and reassigns those values to respondents in the order of their corrected scores. Every marginal distribution and variance then stays fixed. Correlation error still falls in $17/18$ full-map seed contrasts, with one worsening ACS/Llama seed (Table 43). Reordering can nevertheless weaken relationships and change agreement with profiles, so exact marginal preservation does not establish fidelity for a particular analysis.

Fixed regression tasks test that fidelity directly. ACS/Qwen coefficient MAE rises from 0.032 to 0.124 under the full map and to 0.110 under the scalar; other settings, including ESS and Qwen/NHANES, improve (Table 23). The marginal-preserving variants likewise do not consistently improve regressions (Appendix C). A synthetic dataset's usefulness depends on which features of its joint distribution the analysis requires. Neither reducing correlation error nor preserving marginal distributions establishes that usefulness by itself.

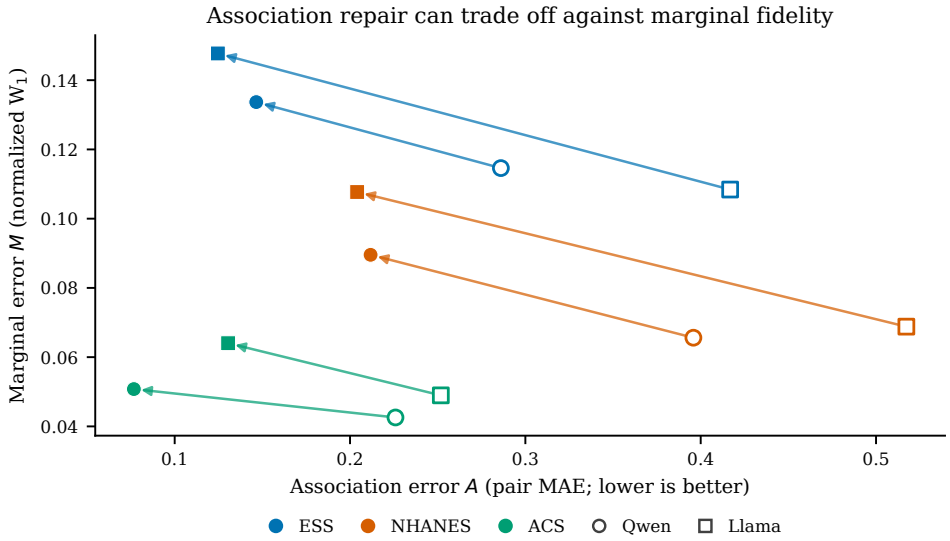


Figure 3: **Lower correlation error can coexist with worse marginal accuracy.** Open (unfilled) points show errors for the uncorrected ten-person batches; filled points show errors after applying the full dependence map. Each arrow connects the results for one domain and model, averaged across three seeds on matched evaluation profiles. Left and down are better; every arrow moves left and up. M uses range-normalized Wasserstein distance; SD-normalized ACS results show the same trade-off. ACS quality failures remain included. This result concerns the full map; exact marginal preservation gives a separate diagnostic.

5 Related work

Evaluations of LLM-generated synthetic data examine both agreement with human data and sensitivity to the generation procedure. In survey simulation, Argyle et al. (2023) introduced silicon sampling using separate prompts for individual respondents. Subsequent work finds misalignment with demographic groups' opinions (Santurkar et al., 2023), too little variation and shifted regression coefficients (Bisbee et al., 2024), and sensitivity to survey presentation (Dominguez-Olmedo et al., 2024). Distorted relationships are also documented: Lukauskas & Šarkauskaitė (2026) examine psychometric structure across thirty-seven models, and Ma et al. (2026) study relationships among cultural tastes.

Respondent batching appears in applied studies: Jahn et al. (2026, Web Appendix A.1) request fifty respondents per completion, and Miklian et al. (2026, Appendix) describe batches of thirty to fifty. In qualitative interview simulation, Sanders et al. (2026, Section 9.2) request four people's answers together. Jahn et al. (2026) also compare batch, cumulative-memory and seed-and-expand generation for consistency across repeated surveys and agreement with human data, including correlations. Building on these comparisons, the matched design in this paper isolates shared-completion request size while keeping respondent profiles, examples and survey questions fixed. The history experiments then intervene on the values available to later respondents, re-pairing them while preserving their marginal distributions and demographic agreement. These controls connect differences between generation procedures to the effects of specific changes in request construction and answer history.

Grouping questions is a different design choice from grouping respondents. Lam et al. (2026) find advantages, in their setting, to asking one respondent's questions together rather than in separate calls. In contrast, Choi et al. (2026) find better structural agreement from per-item prompts in their setting, distinguishing agreement in relationships, individual-variable distributions and individual responses as they evaluate prompting, rectification and fine-tuning using small human samples. The experiments here address a different source of variation: the survey questions stay fixed while the number of *respondents* sharing a request changes. The correction comparison likewise uses gener-

ated fitting answers, with human data providing separate evaluations of correlations and marginal distributions.

Adjusting dependence through ranks while preserving marginals is established (Iman & Conover, 1982). Standard transformations serve here as diagnostics: replacing pair-specific targets with one overall adjustment tests how much improvement requires detailed relationship information, while fixed marginals and regression tasks test what that improvement means for the resulting data.

Relationships between group averages can differ from relationships among people within groups (Robinson, 1950; Goodman, 1953; King, 1997). For example, a relationship across countries need not hold among people in each country. This motivates asking whether models transfer relationships from one level to the other, while leaving that explanation to be tested. Tabular generators such as CTGAN and TabDDPM are fitted to individual-level training data (Xu et al., 2019; Kotelnikov et al., 2023), unlike prompting a pretrained model with profiles and a few examples. Prediction-powered inference can correct particular estimators using real observations (Angelopoulos et al., 2023), while leaving the validity of the rest of a generated dataset open.

6 Limitations and implications

The experiments show that request construction can distort relationships in two model families, with consistent request-size evidence in ESS and NHANES. ACS extends the study to economic data, but all six structured ten-person conditions fail validity and Llama's effects differ in direction across seeds. The study uses one basic design for profiles and examples, mainly at temperature 1; temperature and wording checks broaden this evidence without establishing invariance across models, prompts or request sizes. Matching cannot recover rejected answers, including those lost in three original ESS ten-person conditions. Because the analysis uses complete outcomes and equal group weights, its conclusions describe selected reference populations rather than survey-weighted national populations.

The history interventions establish earlier answers as a causal channel while leaving the model's internal computation open. Hiding history also skips syntax and state updates in a way that may be unfamiliar to the model. Although the Qwen/NHANES improvement persists after removal of the token-position gap, worse marginal accuracy and mixed pattern results leave general weakening as a plausible explanation for much of the lower correlation error. The matched follow-up examines profile context and history in one runtime using forced continuations; differences in runtime, precision, cohort and prompt order prevent attribution of separate shares of the original batching penalty to these factors. Turning the interventions into reliable generation methods requires further work.

The correction experiments and new diagnostics remain exploratory because they reuse previously examined human references, even though human evaluation answers are excluded from fitting. The scalar still requires one-person outputs, while exact marginal preservation uses the evaluation batch itself and can weaken agreement with profiles. Neither establishes a generally useful or cheaper correction. Uncertainty estimates also have defined limits: bootstrap intervals cover only specified sources, and reference resampling lacks survey-cluster identifiers. The descriptive audit and archival experiments provide supporting context from different populations and designs (Appendices I–L).

Request construction can change the statistical relationships attributed to a synthetic population, so the number of records sharing a completion and their access to earlier answers belong in the generator's specification and validation. Evaluation must then address the marginal, dependence and regression quantities required by the intended analysis, since improvement in one does not establish the validity of the others.

Ethics statement

The study analyzes existing, de-identified survey and public-use microdata. It collects no new human-subject responses. Some source surveys require registration and restrict redistribution. Training examples sent to model providers contain real survey rows, and generated records include sampled conditioning profiles.

Reproducibility statement

The protocols and hashes identify the code and inputs used in each study (Appendices A–P). Independent checks verify generated tokens, matched respondents, weights, corrections, error measures and paired bootstrap draws. Reports include invalid outputs and undefined intervals. Regression coefficients agree with a separate weighted-least-squares implementation using explicit group indicators (Appendix C). The supplement contains data-preparation and experiment code, artificial-data tests, and aggregate results that reproduce the figures and generated tables offline. A single CPU command runs the replay and checks its outputs. Recomputing respondent-level results or rerunning generation requires access to the source microdata, model weights and compatible GPU runtimes. The included preparation scripts construct the train/test split and study inputs; fresh runs require new provenance manifests. The supplement excludes private profiles, fitted maps, real examples and raw requests.

AI use statement

The evaluated models generated synthetic respondents as the experimental treatment. Their identifiers and settings appear in the methods and appendix. Separately, AI coding assistants, including OpenAI Codex, supported drafting and revision, literature discovery, methodological feedback, proof checking and revision, code implementation and debugging, and numerical analysis and interpretation. The author retains responsibility for all claims, citations, proofs, code and results.

References

Agency for Healthcare Research and Quality. MEPS HC-251: 2023 Full Year Consolidated Data File, August 2025. URL https://meps.ahrq.gov/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-251. Public-use data, documentation and programming statements.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. doi: 10.1126/science.adi6000.

Anthropic. System Card: Claude Opus 5, July 2026. URL https://www.anthropic.com/claude-opus-5-system-card.

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. doi: 10.1017/pan.2023.2.

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. doi: 10.1017/pan.2024.5.

bitsandbytes contributors. bitsandbytes. Software, version 0.50.0, 2026. URL https://github.com/bitsandbytes-foundation/bitsandbytes/releases/tag/0.50.0.

Eun Cheol Choi, Youngrae Kim, Prabhu Pugalenthi, Hong-En Chen, and Bo-Ruei Huang. Beyond the mean: Three-axis fidelity for aligning LLM-based survey simulators from small pilot data. arXiv preprint arXiv:2606.28963v1, 2026. URL https://arxiv.org/abs/2606.28963v1.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1458. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/515c62809e0a29729d7eec26e2916fc0-Abstract-Conference.html.

Yixin Dong, Charlie F. Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, and Tianqi Chen. XGrammar: Flexible and efficient structured generation engine for large language models. In *Proceedings of Machine Learning and Systems*, volume 7,

2025. URL https://proceedings.mlsys.org/paper_files/paper/2025/hash/5c20ca4b0b20b0bd2f1d839dc605e70f-Abstract-Conference.html.

European Social Survey ERIC. ESS integrated data files: rounds 1–11 and round-10 self-completion. Sikt – Norwegian Agency for Shared Services in Education and Research, 2012–2026. Data sets: ESS1 (2023), ed. 6.7; ESS2 (2012), ed. 3.6; ESS3 (2018), ed. 3.7; ESS4 (2023), ed. 4.6; ESS5 (2025), ed. 3.6; ESS6 (2025), ed. 2.7; ESS7 (2023), ed. 2.3; ESS8 (2023), ed. 2.3; ESS9 (2026), ed. 3.3; ESS10 (2025), ed. 3.3; ESS10 self-completion (2025), ed. 3.2; ESS11 (2026), ed. 4.2. Full DOIs and edition details appear in Table 32.

Leo A. Goodman. Ecological regressions and behavior of individuals. *American Sociological Review*, 18(6):663–664, 1953. doi: 10.2307/2088121.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. World Values Survey Trend File (1981–2022) Cross-National Data-Set. JD Systems Institute and WVSA Secretariat, 2022. URL https://doi.org/10.14281/18241.27. Data file version 4.1.0.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Charles R. Henderson. Estimation of variance and covariance components. *Biometrics*, 9(2):226–252, 1953. doi: 10.2307/3001853.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Ronald L. Iman and W. J. Conover. A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, 11(3):311–334, 1982. doi: 10.1080/03610918208812265.

Steffen Jahn, Tobias Klinke, Daniel Guhl, and Maja Murr. Re-fielding replicability of silicon sample-based marketing research. Discussion Paper 583, Collaborative Research Center Transregio 190: Rationality and Competition, August 2026. URL https://rationality-and-competition.de/wp-content/uploads/discussion_paper/583.pdf.

Gary King. *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton University Press, Princeton, NJ, 1997. URL https://www.jstor.org/stable/j.ctt46n43p.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17564–17579. PMLR, 2023. URL https://proceedings.mlr.press/v202/kotelnikov23a.html.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Nickolas Hock Yuen Lam, Ji Xuan Voo, and Xiangyu Ma. The conditional superiority of fast silicon sampling. arXiv preprint arXiv:2608.14079v1, 2026. URL https://arxiv.org/abs/2608.14079v1.

Mantas Lukauskas and Viktorija Šarkauskaitė. Plausible but not valid: A psychometric audit of LLMs as synthetic survey respondents. *arXiv preprint arXiv:2608.14606v1*, 2026. URL https://arxiv.org/abs/2608.14606v1.

Xiangyu Ma, Mengmi Zhang, Shannon Ang, and Minne Chen. Tastes without distinction: silicon samples and the synthetic construction of tastes. *arXiv preprint arXiv:2606.30085v2*, 2026. URL https://arxiv.org/abs/2606.30085v2.

Wes McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, pp. 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Meta. Llama-3.3-70B-Instruct. Official model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Jason Miklian, Kristian Hoelscher, and John E. Katsos. Stochastic parrots or singing in harmony? Testing five leading LLMs for their ability to replicate a human survey with synthetic data. arXiv preprint arXiv:2603.00059v3, 2026. URL https://arxiv.org/abs/2603.00059v3.

National Center for Health Statistics. National Health and Nutrition Examination Survey: 2017–2018 Data. Centers for Disease Control and Prevention, 2020. URL https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Cycle=2017-2018. Files DEMO_J, BMX_J, HUQ_J, DPQ_J and PAQ_J; released February 2020.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925, 2025. URL https://arxiv.org/abs/2508.10925.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Qwen Team. Qwen3-235B-A22B-Instruct-2507. Official model card, 2025. URL https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507.

Qwen Team. Qwen3.8-27B. Official model card, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B.

William S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357, 1950. doi: 10.2307/2087176.

Jeanne Sanders, John Mobley, IV, Isabel Miller, Nicola W. Sochacka, Paul A. Jensen, and Karin J. Jensen. Affordances and limitations of using large language models to generate qualitative data about mental health perceptions in engineering. *Journal of Engineering Education*, 115(1): e70037, 2026. doi: 10.1002/jee.70037. URL https://onlinelibrary.wiley.com/doi/10.1002/jee.70037.

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*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. John Wiley & Sons, New York, 1992. doi: 10.1002/9780470316856.

U.S. Census Bureau. 2022 American Community Survey 1-Year Public Use Microdata Sample, 2023. URL https://www.census.gov/programs-surveys/acs/microdata/documentation.2022.html. Massachusetts person records, psam_p25.csv from csv_pma.zip.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore,

Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html.

A Matched request-size experiment

A.1 Sampling and study design

The September 16 protocol draws on earlier experiments and was fixed before these responses were generated. It specifies two models, request sizes one and ten, two seed policies and three repetitions. The protocol was not externally preregistered. Both model runs are complete and independently checked. The separate prompt-context extension is reported on its own.

The eligible pool consists of ESS training rows complete on nineteen numeric targets (excluding household size), using the split and registry in Appendix J. Countries are ranked by eligible training count, with alphabetical ties: DE, FR, FI, AT, SE, BE, EE, NL, ES, GB, IE, IT, NO, CZ, CH, BG, PL, HU, LT and SI. Sampling without replacement selects 100 rows per country (seed 20260916). Their ten categorical fields condition generation; missing fields remain *not reported*. Their numeric outcomes are withheld from generation and saved only for a secondary reference.

The independent test reference contains 19,896 complete rows. Each country receives weight $1/20$, so estimates describe complete cases with equal country weights rather than survey-weighted populations. The profiles form forty sets of fifty, each containing one country and sharing three fixed real examples (seed 20260917), sampled from 79,993 training rows complete on all displayed fields, excluding all 2,000 targets. Examples, profiles and item order remain fixed across conditions and models.

A.2 Requests and runtime

Each set of fifty profiles becomes fifty one-person or five ten-person requests. The experiment runs under two seed policies. In the primary policy, every request receives a distinct stable 63-bit seed derived from its identity. In the diagnostic policy, every request resets to the same seed for that run. This checks whether reusing a decoder seed changes the request-size comparison. Seeds $0, 1, 2$ index complete repetitions of the experiment under both policies. Each model requests 24,000 respondents through 13,200 requests across the twelve conditions.

The prompt requests a flat JSON array with the supplied integer IDs and nineteen numeric fields, permitting fractional values within inclusive registry bounds. A technical pilot showed that nested examples made the required flat output ambiguous. The `flat-examples-v2` amendment changes only the formatting, pairing the same example profiles with flat answer arrays and reserved negative IDs. Profiles, values, scales, seeds and comparisons stay fixed. Both amended pilots passed validation before production; the change was chosen for formatting validity rather than scientific results. The supplied constructor defines this prompt, which differs from the September 6 panel template.

Checkpoints are Qwen3.8-27B (Qwen Team, 2026), revision `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0`, and Llama-3.3-70B-Instruct (Meta,

2024), revision `6f6073b423013f6a7d4d9f39144961bfbfbc386b`. Both use two RTX 6000 Ada GPUs, tensor parallelism two, Python 3.12.13, vLLM 0.23.0 (Kwon et al., 2023), PyTorch 2.11.0 (Paszke et al., 2019) and Transformers 5.9.0 (Wolf et al., 2020). Qwen uses BF16 weights, automatic KV-cache dtype, language-model-only mode and thinking disabled; Llama uses on-the-fly FP8 weights and FP8 KV cache. Cross-model magnitudes therefore do not isolate architecture. Temperature and top-p are 1, top-k is unrestricted and min-p is zero. Output/context limits are $8{,}192/16{,}384$ tokens; inference microbatches contain 32 requests, not 32 respondents per completion. Generation stops at the closing array bracket and retains it. Exact chat-template lengths are checked; output budgets are never silently reduced.

A.3 Validity, estimands and uncertainty

The generation pipeline saves each completed request atomically, including its raw text, IDs, prompt and input hashes, settings, runtime, token counts and finish reason. Acceptance requires exactly the requested unique IDs and fields, finite in-range JSON numbers and no duplicate keys. Any missing/extra row or field, invalid number or token-limit termination rejects the whole request. Rejected requests are neither altered nor partially retained or retried. A resumed run executes only requests with no completed record.

Pilots cover the first two fixed profile sets, seed zero and all four size/policy arms. They are excluded from production despite reused profiles. The frozen threshold is 99% respondent retention and zero truncations per arm. Failed production arms remain reported.

Table 1: Production validity under whole-request acceptance: an invalid ten-person request loses all ten respondents. Failed-quality arms remain reported.

Model	Seed	Policy	Size	Valid respondents	Invalid req.	Truncated	Quality
Qwen3.8-27B	0	Independent	1	1,987/2,000 (99.4%)	13/2000	0	Pass
Qwen3.8-27B	0	Independent	10	1,980/2,000 (99.0%)	2/200	0	Pass
Qwen3.8-27B	0	Shared	1	1,986/2,000 (99.3%)	14/2000	0	Pass
Qwen3.8-27B	0	Shared	10	1,990/2,000 (99.5%)	1/200	0	Pass
Qwen3.8-27B	1	Independent	1	1,987/2,000 (99.4%)	13/2000	0	Pass
Qwen3.8-27B	1	Independent	10	1,980/2,000 (99.0%)	2/200	0	Pass
Qwen3.8-27B	1	Shared	1	2,000/2,000 (100.0%)	0/2000	0	Pass
Qwen3.8-27B	1	Shared	10	2,000/2,000 (100.0%)	0/200	0	Pass
Qwen3.8-27B	2	Independent	1	1,986/2,000 (99.3%)	14/2000	0	Pass
Qwen3.8-27B	2	Independent	10	1,950/2,000 (97.5%)†	5/200	0	Failed†
Qwen3.8-27B	2	Shared	1	2,000/2,000 (100.0%)	0/2000	0	Pass
Qwen3.8-27B	2	Shared	10	2,000/2,000 (100.0%)	0/200	0	Pass
Llama-3.3-70B	0	Independent	1	1,998/2,000 (99.9%)	2/2000	0	Pass
Llama-3.3-70B	0	Independent	10	1,960/2,000 (98.0%)†	4/200	0	Failed†
Llama-3.3-70B	0	Shared	1	2,000/2,000 (100.0%)	0/2000	0	Pass
Llama-3.3-70B	0	Shared	10	1,980/2,000 (99.0%)	2/200	0	Pass
Llama-3.3-70B	1	Independent	1	1,994/2,000 (99.7%)	6/2000	0	Pass
Llama-3.3-70B	1	Independent	10	1,960/2,000 (98.0%)†	4/200	0	Failed†
Llama-3.3-70B	1	Shared	1	2,000/2,000 (100.0%)	0/2000	0	Pass
Llama-3.3-70B	1	Shared	10	2,000/2,000 (100.0%)	0/200	0	Pass
Llama-3.3-70B	2	Independent	1	1,995/2,000 (99.8%)	5/2000	0	Pass
Llama-3.3-70B	2	Independent	10	1,980/2,000 (99.0%)	2/200	0	Pass
Llama-3.3-70B	2	Shared	1	2,000/2,000 (100.0%)	0/2000	0	Pass
Llama-3.3-70B	2	Shared	10	2,000/2,000 (100.0%)	0/200	0	Pass

The analysis scores both all accepted rows and the profiles accepted in all four conditions within each model and repetition. Matching aligns the observed profiles, but the comparison still depends on which answers passed validation. Every fixed country must retain twenty rows and all 171 pair correlations must be defined. Centered residuals receive weights $1/(20n_c)$: this pools country covariances, not country correlations. The primary contrast is size-ten minus size-one MAE under independent seeds; shared-reset contrasts and the size-by-policy interaction are diagnostic. The interaction changes sign across repetitions in both models.

Percentile intervals use 500 paired country resamples (base seed 20260918), retaining each sampled country's full generated/reference records and multiplicity. They condition on realized generation and selected countries, excluding generation and within-country reference uncertainty. Three-seed ranges are descriptive; the 171 overlapping pairs are never independent units.

Table 2: Independent-seed request-size results on four-arm common IDs. MAE averages 171 within-country correlation errors; intervals use the conditional paired-country bootstrap. † denotes an arm below 99% validity.

Model	Seed	N common	MAE, size 1	MAE, size 10	Δ_{10-1}	Conditional interval
Qwen3.8-27B	0	1,945	0.179	0.282	+0.103	[+0.067, +0.138]
Qwen3.8-27B	1	1,967	0.185	0.276	+0.091	[+0.073, +0.112]
Qwen3.8-27B	2	1,936	0.186	0.276†	+0.089†	[+0.057, +0.122]
Llama-3.3-70B	0	1,938	0.187	0.399†	+0.213†	[+0.190, +0.235]
Llama-3.3-70B	1	1,954	0.187	0.425†	+0.238†	[+0.210, +0.259]
Llama-3.3-70B	2	1,975	0.187	0.420	+0.233	[+0.207, +0.263]

Table 3: Matched-ID sensitivities: Δ is size-ten minus size-one MAE. Excluding age/year leaves 136 pairs. Intervals use the same conditional country bootstrap. A dagger flags any failed-quality arm in four-arm matching, including the other seed policy.

Model	Seed	Policy	Δ, 171 pairs	Δ, 136 pairs	136-pair interval
Qwen3.8-27B	0	Independent	+0.103	+0.118	[+0.080, +0.155]
Qwen3.8-27B	0	Shared	+0.144	+0.154	[+0.113, +0.174]
Qwen3.8-27B	1	Independent	+0.091	+0.098	[+0.075, +0.120]
Qwen3.8-27B	1	Shared	+0.053	+0.062	[+0.012, +0.107]
Qwen3.8-27B	2	Independent	+0.089†	+0.096†	[+0.063, +0.130]
Qwen3.8-27B	2	Shared	+0.109†	+0.130†	[+0.083, +0.166]
Llama-3.3-70B	0	Independent	+0.213†	+0.258†	[+0.231, +0.284]
Llama-3.3-70B	0	Shared	+0.230†	+0.278†	[+0.245, +0.309]
Llama-3.3-70B	1	Independent	+0.238†	+0.290†	[+0.257, +0.315]
Llama-3.3-70B	1	Shared	+0.216†	+0.267†	[+0.238, +0.296]
Llama-3.3-70B	2	Independent	+0.233	+0.282	[+0.252, +0.318]
Llama-3.3-70B	2	Shared	+0.217	+0.263	[+0.237, +0.291]

A.4 Secondary checks and verification

Secondary checks exclude age and year, leaving 136 pairs; examine the three original police-trust pairs; and use the sampled profiles' observed human answers as another reference. Further checks fit weighted, country-demeaned regressions of police trust on income, happiness, democracy satisfaction and age. On matched Qwen IDs under independent seeds, median democracy coefficients are $0.435/0.593$ for one/ten-person requests versus 0.339 in the held-out reference; happiness coefficients are $0.230/0.169$ versus 0.154. These coefficients illustrate that batching can move different regression estimates in different directions. They describe associations, not causal effects.

Post hoc, the 2,000 sampled profiles' human outcomes have MAE 0.0177 against the test reference, or 0.0178–0.0185 on model/repetition-specific matched IDs. This comparison describes the particular human samples and does not establish a noise floor or an uncertainty interval. The outcomes were withheld from generation; examining them afterward does not create a new holdout or decompose the synthetic error.

An independent row-weight implementation reproduces all 48 endpoints, 8,208 pair correlations, 192 regression coefficients and 300 contrasts, using full country-dummy weighted least squares rather than the production analyzer. A separate raw-record audit reparses all 26,400 completions, matches all 89 invalid requests and reasons, and exactly reproduces 47,713 accepted-row exports. Country counts and quality flags agree. A reporting correction propagates failed-quality flags into seed summaries without changing estimates; original outputs remain archived. These checks do not validate bootstrap coverage, model generalization or unobserved rejected answers.

A.5 Matching only the primary independent-seed arms

An exploratory sensitivity matches profiles only across the two independent-seed conditions. This keeps failures in the shared-reset diagnostic from removing otherwise observed answers from the main comparison. Countries, variables and human reference remain fixed. Its absolute contrasts are close to those on the four-arm support. At seeds $0, 1, 2$, absolute two-

arm contrasts are $0.1026, 0.0913, 0.0893$ for Qwen and $0.2098, 0.2384, 0.2326$ for Llama, versus $0.1031, 0.0913, 0.0893$ and $0.2128, 0.2384, 0.2326$ on four-arm support. Quality failures remain flagged. Independently weighted covariances verify both the sensitivity and archived four-arm estimates using audited inputs. The check changes which accepted profiles enter the comparison. It neither changes generation nor recovers rejected answers.

B CHANGING AND HIDING EARLIER ANSWERS

Two interventions test the influence of earlier answers. Changing their pairing tests whether they influence relationships in later responses. Hiding naturally generated answers tests whether access to them affects accuracy against human data. These are separate questions: changing a generated relationship need not make it more accurate. The studies follow the prefix-replay experiment (Appendix P). Each protocol, source snapshot, input plan and analysis rule was fixed before scientific generation, including for the extension to two models and two domains.

B.1 CONTROLLED PAIRING OF FIXED HISTORY VALUES

Stimuli and allocation. NHANES uses BMI/sedentary minutes and eight sex–age templates; ESS uses trust in parliament/the judiciary and twenty country templates. Each template uses the most common categorical profile in the fitting partition, with ties broken by a fixed hash. Prompts contain ten identical profiles except IDs and three fixed training examples projected onto the two outcomes. ESS uses examples from each country's first fitting set of fifty profiles; NHANES reuses the preceding frozen Qwen study's exact templates, prompts, examples and paired seeds. Original manifests verify the inputs.

The model receives answers for the first nine respondents and completes the tenth. For NHANES, the supplied answers draw from two fixed lists: BMI values $20, 22, \ldots, 36$ and sedentary minutes $120, 180, \ldots, 600$. Pairing the lists in the same rank order gives a positive correlation: respondents with higher BMI also have more sedentary minutes. Reversing one list gives a negative correlation. A third pairing, defined by the index permutation $(0, 2, 6, 8, 7, 5, 4, 3, 1)$, gives zero correlation. ESS follows the same procedure using $1, 2, \ldots, 9$ for each trust variable.

The order in which the nine supplied respondents appear also varies. Eighteen forward/reverse rotations place every rank twice in every history position, so no rank is consistently nearest the target. Each pairing retains the same values for each variable and the same number of tokens for the relevant tokenizer. Because the supplied respondents have identical profiles, moving values between them also preserves agreement with their attributes. It changes both the relationship between the variables and the sequence of values that the model encounters.

Each order has three distinct seeds paired across conditions. A fourth, hidden-history condition skips preceding-answer state updates while retaining their logical token-position gap. The equal-length histories produce the same input when hidden, so one hidden-history condition serves as their shared reference. Each model generates $8 \times 18 \times 3 \times 4 = 1{,}728$ NHANES and $20 \times 18 \times 3 \times 4 = 4{,}320$ ESS continuations. Only newly generated tenth answers are scored.

Endpoint and uncertainty. For each template, Pearson correlations use responses accepted under every condition, requiring at least twenty of the 54 planned order–repetition cases. The endpoint averages the positive-minus-negative contrast equally across templates. This averages correlations rather than covariances, unlike Equation 1, and does not use human answers as an accuracy reference. All-retained, per-repetition, Spearman and marginal-shift sensitivities remain in the aggregate supplement.

The $5{,}000$ fixed-seed bootstrap draws independently resample eighteen order blocks within each template, carrying all repetitions and arms together. The intervals describe sensitivity to the mix of orders. They do not provide population coverage because the balanced orders are not independent sampled contexts and the templates are not a random demographic sample. Any undefined draw suppresses the interval. Llama/NHANES has $1{,}393$ such draws; its zero-history-pairing arm also has one undefined template correlation. Neither is replaced by zero or dropped.

Table 4: Equal-template correlations in newly generated tenth responses. Hidden history skips earlier-answer updates; $\Delta_{+,-}$ is positive minus negative pairing. Intervals describe conditional order-mixture sensitivity. A dash is undefined; Llama/NHANES's interval is withheld.

Model	Domain	Positive	Zero	Negative	Hidden	$\Delta_{+,-}$	95% summary
Qwen	NHANES	0.963	0.182	-0.670	0.238	1.633	$[1.525, 1.728]$
Qwen	ESS	0.985	0.145	-0.647	0.611	1.632	$[1.568, 1.681]$
Llama-70B (NF4)	NHANES	0.837	–	-0.329	-0.036	1.166	withheld
Llama-70B (NF4)	ESS	0.725	-0.173	0.041	0.090	0.684	$[0.546, 0.737]$

Table 5: Controlled-history validity and matched support under one strict rule across arms. Invalid answers remain archived; undefined bootstrap draws are counted, not discarded.

Model	Domain	Valid +	Valid 0	Valid −	Valid hidden	Matched/template	Undefined bootstraps
Qwen	NHANES	432/432	430/432	431/432	429/432	52–54	0/5000
Qwen	ESS	1080/1080	1080/1080	1080/1080	1079/1080	53–54	0/5000
Llama-70B (NF4)	NHANES	432/432	432/432	431/432	432/432	53–54	1393/5000
Llama-70B (NF4)	ESS	1080/1080	1080/1080	1080/1080	1080/1080	54–54	0/5000

Exploratory within-order check. An apparent pairing effect could arise because different answer orders shift the mean of the generated responses. A check for this explanation holds the profile template, answer order and pairing condition fixed, then subtracts the mean across retained repetitions from each outcome. Correlations calculated from these centered responses remain positive under positive pairing and negative under negative pairing. The respective correlations are $0.933/-0.296$ (Qwen/NHANES), $0.978/-0.480$ (Qwen/ESS), $0.649/-0.225$ (Llama/NHANES) and $0.634/-0.068$ (Llama/ESS). Differences between order-specific means therefore explain only part of the contrast. This exploratory check has three repetitions per order, requires at least two retained, and supplies no formal interval. It leaves the original endpoint unchanged. Both this check and the balanced history positions leave open whether the model copies covariance or extends the sequence of each variable separately.

B.2 Fresh natural histories at three positions

Cohort and intervention. Qwen/NHANES uses all 400 evaluation profiles, fifty per sex–age cell, and five outcomes: poverty–income ratio, BMI, self-rated health, PHQ-9 and sedentary minutes. A fixed hash order selects nine distinct fitting donors from the same cell for each target and repetition. Full prompts include these donors, the disjoint target and its existing three examples; target outcomes never select donors or construct prompts.

At position ten, the model first generates answers for the nine preceding respondents. The actual target profile is already present in the prompt while it does so. The model then generates the target's answer twice, starting each time from a reconstructed state. In the visible condition, the model processes the earlier answers as usual. The hidden condition skips those tokens, including their answer syntax, so they cannot update attention keys and values or convolution/recurrent states. The two continuations use the same full profile prompt, forced target syntax, decoder seed and absolute logical target positions. Keeping those positions fixed leaves a gap where the earlier answers would have been, so information about their length remains available.

The extension moves the target to position two or five, preserving donor order, examples and target seed. It generates new one- or four-row histories under the changed full prompt, rather than truncating a nine-row history. Each position has three repetitions, 1,200 planned histories and 2,400 visible/hidden continuations. For position ten, the analysis reuses the verified original outputs and rescores them on profiles accepted in all six position–visibility conditions. At position two, one preceding answer cannot itself define a cross-respondent covariance. Moving the target changes history length, prompt order and predecessor profiles together. The comparison therefore cannot isolate the effect of history length alone.

Scoring and uncertainty. Matching profiles across all six position–visibility conditions leaves 44–50 targets per cell in each repetition. Invalid histories invalidate both associated targets; invalid

Table 6: Qwen/NHANES, six-arm common support: three-repetition means of correlation MAE A and normalized marginal Wasserstein error M. Change is hidden minus visible; intervals resample paired profile blocks. Valid-target counts precede six-arm matching; each position plans 1,200 histories.

Position	A_{visible}	A_{hidden}	Change in A	95% interval	M_{visible}	M_{hidden}	Valid targets
2	0.324	0.213	-0.111	$[-0.144, -0.078]$	0.072	0.078	2383/2400
5	0.374	0.210	-0.164	$[-0.196, -0.129]$	0.067	0.074	2370/2400
10	0.367	0.197	-0.169	$[-0.223, -0.125]$	0.067	0.075	2375/2400

Table 7: Tenth-respondent extensions: three-repetition mean correlation MAE A and normalized marginal Wasserstein error M. Reduction is relative to visible history; ΔA is hidden minus visible. All 400 profiles per repetition are retained; intervals use paired profile blocks.

Model	Domain	A_{visible}	A_{hidden}	Reduction	95% interval for ΔA	M_{visible}	M_{hidden}
Qwen	ACS	0.166	0.103	38%	$[-0.117, -0.015]$	0.0390	0.0399
Llama-70B (NF4)	NHANES	0.577	0.300	48%	$[-0.306, -0.243]$	0.0767	0.1060

targets remain archived, without retries or clipping. Equal-cell pooled covariances give correlations for all ten pairs, scored against the study's fixed human reference. The primary contrast averages hidden-minus-visible MAE across repetitions. A negative value means lower error, without implying that the resulting data are accurate.

For 5,000 paired bootstrap draws, target IDs are resampled from the fifty planned profiles within each cell, carrying positions, arms and repetitions together with multiplicity; each repetition's six-arm acceptance rule is then applied. The reference stays fixed. These profile-block intervals condition on the observed cohort and intervention, not new model-training runs, reference populations or generation seeds. All three contrasts have no undefined draws. Aggregate reports also retain all-retained, cell, focal-pair, repetition and failure diagnostics.

Valid histories number 1,199, 1,195 and 1,192 at positions two, five and ten. The original position-ten two-arm estimate falls from 0.363 to 0.194 (47%): contrast -0.170, interval $[-0.221, -0.125]$. Table 6 instead gives the extension's six-arm comparison; neither replaces the other's prespecified estimate.

B.3 Accuracy extensions: Llama/NHANES and Qwen/ACS

Each tenth-respondent extension follows a separately fixed protocol and uses 400 existing evaluation profiles (fifty per sex–age group), fitting-partition donors and the existing human reference: 1,200 fresh histories and 2,400 paired targets across three repetitions. Llama/NHANES retains its five-outcome plan and seeds; Qwen/ACS scores all six pairs among income, wages, weekly hours and weeks worked. Both interventions preserve the full prompt, target syntax, seed and logical positions. These are reused cohorts, not new holdouts.

The endpoint averages hidden-minus-visible MAE across repetitions on two-arm common support. Independent audits accept every history and target without truncation, so matching retains all fifty profiles per group/repetition. The 5,000 paired profile-block resamples carry repetitions and arms together with the reference fixed; all draws are defined. These are conditional intervals, not population or generation-seed uncertainty.

Although mean marginal error rises (Section 3.2), ACS's second repetition improves it by 0.09%. Prespecified ACS sensitivities give correlation-error contrasts -0.085 without the income–wages pair and -0.080 after transforming both monetary variables by $\operatorname{asinh}(x/10{,}000)$; each repetition improves. Neither establishes that reducing correlation is generally beneficial.

B.4 Matched singleton, profile-context and position controls

The follow-up reuses the 400 Qwen/NHANES target profiles, three repetitions, donor profiles, examples, original native histories and target seeds in the same pinned FP16 Transformers runtime. The model generates fresh answers under five conditions. The first presents the target profile alone.

The second puts the target first in a prompt containing all ten profiles, so the model sees the other profiles but has no preceding answers. The third presents the original visible answer history. The final two conditions both hide that history: one retains the gap in token positions, while the other closes the gap and assigns contiguous positions to the tokens the model actually processes. All five use forced target-row syntax and the same bounded numeric grammar. The one-person condition uses the same continuation procedure as the other four. It differs from the original vLLM experiment, which generates complete arrays. Comparisons of first and later positions also change profile order in the prompt.

Eight invalid original histories exclude all five associated continuations. The 1,192 eligible cases yield 5,960 new outputs, retaining the full 6,000-row planned disposition. Five-arm common support contains 394, 395 and 394 profiles across the three repetitions. There are 1,190 valid singleton targets, 1,190 target-first outputs, 1,192 visible outputs, and 1,183 in each hidden arm before matching. Invalid outputs are retained in the audit and never replaced. Settings and the analysis contrasts were fixed before this generation, after the earlier findings were known; this is an exploratory reused cohort.

Table 8: Same-runtime Qwen/NHANES continuations, three-repetition means on five-arm common support. P is correlation-vector Pearson agreement, A_{or} human-oracle scalar MAE, and E_β the fixed PHQ-9 coefficient error. All five arms use forced target syntax.

Continuation arm	A	P	A_{or}	M_{SD}	E_β
Singleton	0.198	0.895	0.071	0.482	0.133
Ten profiles, target first	0.247	0.905	0.067	0.476	0.115
Visible history	0.363	0.851	0.080	0.402	0.157
Hidden, position gap	0.192	0.896	0.065	0.441	0.100
Hidden, contiguous	0.193	0.903	0.063	0.446	0.098

Table 9: Prespecified contrasts within the exploratory context follow-up. Paired bootstrap draws resample target IDs within cells, carrying all arms and repetitions. The second interval additionally resamples human reference rows within cells; neither is survey-design uncertainty. All 2,000 draws are defined in each sensitivity.

Contrast	ΔA	Fixed reference 95%	Resampled reference 95%
First minus singleton	0.048	[+0.031, +0.069]	[+0.031, +0.069]
Visible minus first	0.116	[+0.080, +0.167]	[+0.079, +0.168]
Hidden minus visible	-0.171	[-0.221, -0.127]	[-0.222, -0.127]
Contiguous minus gapped	0.001	[-0.015, +0.016]	[-0.015, +0.017]

Both ways of hiding history reduce raw MAE, weaken correlations and improve pattern agreement relative to visible history. With contiguous positions, marginal error still rises: $M_{\mathrm{SD}} = 0.446$ versus 0.402 with visible history. A position gap alone therefore does not account for the mean gain in this setting. The target-first profile context raises raw MAE relative to singleton generation, despite slightly better pattern agreement and oracle MAE. Profile context can therefore change relationship strength and pattern differently. These comparisons cannot assign separate causal shares of the original batching effect to context and history.

Checks before generation verified the original runtime and source bindings. Twenty fictional cases compared individual and padded execution of all five routes through three forced decoding steps, using the existing maximum-logit and total-variation limits. An independent CPU audit rechecks native tokens, grammar, routes, positions and hashes, and reproduces correlations with explicit weighted NumPy covariance. It verifies saved GPU gate evidence rather than independently rerunning the GPU computations. All 1,192 recovered raw case files match their audit hashes.

B.5 Runtimes, validation and independent audits

Original Qwen/NHANES history/position studies use the pinned checkpoint in FP16, Transformers 5.9.0, PyTorch 2.11.0, xgrammar 0.2.2 (Dong et al., 2025), stock hybrid-state SDPA and two

RTX6000 Ada GPUs. The crossed pairing study uses Qwen FP16 and Llama NF4 double quantization with FP16 computation (bitsandbytes $0.50.0$ (bitsandbytes contributors, 2026)) on one GB10, with Transformers $5.15.0$, PyTorch $2.13.0$ and xgrammar $0.2.3$. Missing offline FP8 kernels motivated NF4 before generation; the permitted smaller-model fallback was unused. Cross-model magnitudes do not isolate architecture.

The crossed Qwen/NHANES study repeats the same stimuli and seeds to check agreement between runtimes, rather than adding an independent experimental setting: its contrast is 1.633337 originally and 1.633224 on GB10. Runs are not pooled as extra seeds. Original studies use temperature 1, unrestricted top-p/top-k, independent CPU random streams paired across conditions, and a nonnegative numeric grammar allowing four integer/two decimal digits; operational ranges are checked afterward. Caps are 128 target tokens for controlled pairs, 192 for natural five-outcome answers, and $1{,}024$ for histories. Completed invalid/truncated outputs are never selectively rerun.

The accuracy extensions use two RTX6000 Ada GPUs: Llama NF4/FP16 SDPA and Qwen FP16 SDPA, distinct from Llama's original FP8 and controlled-history GB10 runtimes. Following failed fictional validation, ACS adopts a variable-specific range grammar allowing signed values as required, seven integer/six decimal digits, source microbatches of four, singleton target execution, and $2{,}048/256$ source/target caps. Singleton execution avoids failed padded-target numerical checks; Llama/NHANES retains $1{,}024/192$ caps. Failed attempts and source snapshots remain archived.

Before scientific generation, numerical checks compare one-shot and split replay of native tokens, individual and batched execution with visible history, and multiple decoding steps with gapped hidden history. They test finite caches, prompt sensitivity, grammar and ranges. Qwen checks sixteen KV and forty-eight convolution/recurrent pairs; Llama checks eighty KV pairs. Duplicate hidden routes match states, logits and common-seed outputs exactly; an empty-history sham is also technical, not scientific, evidence. Admission thresholds remain maximum logit difference 0.5 and total variation 0.02; crossed Qwen/Llama gates attain $0.0723/0.00435$ and $0.1094/0.00696$. Amended ACS source/target checks pass the same limits.

Each run records its checkpoint inventory, tokenizer, input and source hashes, packages, devices and precision. The position studies also match the original runtime identity. Outputs are saved atomically before validity decisions. Independent CPU audits reconstruct native outputs, validity, routes, positions, seeds and hash bindings, then recompute endpoints and bootstrap draws without trusting derived validity flags. Extension statistics agree within 10^{-12}; ACS additionally replays every token through the production grammar checker. These audits do not independently rerun GPU arrays, rehash physical model weights or rebuild extension demographic plans. The anonymous supplement includes audit code, artificial tests and aggregate replay artifacts, not real examples, profiles or individual generated records.

C Exploratory downstream regression checks

These checks test whether improvements in correlation or marginal accuracy translate into better regression estimates. The specifications were fixed after inspecting those earlier results and before computing the regressions. They reuse the same outputs, cohorts and human references, so the checks remain exploratory. The regressions describe associations and do not provide causal or clinical validation.

Fixed specifications and common scale. ESS regresses trust in police on income position, age, happiness and satisfaction with democracy. NHANES regresses PHQ-9 depression score on sedentary minutes, BMI, poverty–income ratio and self-rated health. ACS regresses wage/salary income on weekly hours and weeks worked, retaining zero wages and excluding total income, which contains wage income. Every fit retains the original scales and observations and uses all specified predictors.

Each fit includes group fixed effects and observation weights $1/(|C|n_c)$, giving each country or demographic cell equal weight. Comparing coefficients across conditions requires a common scale. Pooled within-group standard deviations from the human reference provide that scale: the same values standardize every human and synthetic coefficient. A method cannot therefore improve this measure simply by changing the standard deviations used for scoring. For the selected k predictors,

Table 10: Exploratory PHQ-9 regression: three-repetition mean standardized-coefficient MAE E_β on the Qwen/NHANES six-arm common cohort. Hiding histories lowers error by 25%, 30% and 35% at positions two, five and ten, respectively. "Improved" refers to the mean over four coefficients, not each coefficient.

Position	Visible	Hidden	Hidden minus visible	Repetitions improved
2	0.105	0.079	-0.026	3/3
5	0.134	0.093	-0.040	3/3
10	0.156	0.101	-0.055	3/3

Table 11: All correction variants: three-repetition mean standardized-coefficient MAE E_β for the fixed domain-specific regressions. Lower is better; all variants share the original matched evaluation IDs within each model and repetition. †: the six original ACS size-ten generation-quality failures remain included.

	ESS		NHANES		ACS†	
Method	Qwen	Llama	Qwen	Llama	Qwen	Llama
Untouched batching	0.105	0.101	0.157	0.151	0.032	0.249
Marginal-only mapping	0.072	0.085	0.112	0.162	0.185	0.305
Learned dependence map	0.058	0.070	0.115	0.151	0.124	0.354
One-person row resampling	0.066	0.043	0.126	0.094	0.219	0.331
Identity-dependence target	0.135	0.129	0.135	0.078	0.183	0.263
Held-out one-person anchor	0.061	0.046	0.097	0.159	0.187	0.139
Map + fitting batch marginals	0.068	0.080	0.163	0.170	0.105	0.215
Map + exact batch marginals	0.058	0.093	0.145	0.177	0.074	0.117

the error measure averages absolute differences from the human coefficients:

$$E_\beta = \frac{1}{k}\sum_{j=1}^{k}\left|\frac{s_{H,j}}{s_{H,y}}\left(\widehat{\beta}_{S,j} - \widehat{\beta}_{H,j}\right)\right|.$$

Here $\widehat{\beta}$ denotes native-unit slopes; human scales s_H and reference slopes are fixed. Each fit requires twenty finite complete observations per original group, positive human scales, and a predictor-covariance condition number no larger than 10^{10} on the common standardized scale. One unsupported repetition suppresses its three-repetition summary; no pseudoinverse or reduced specification is substituted. All 162 planned fits are supported and agree with independent explicit-dummy weighted least squares. Aggregate outputs retain native/standardized slopes, 552 coefficient comparisons, repetitions and quality flags. No population-confidence interval or coefficient-wise significance is claimed.

Natural histories. Qwen/NHANES uses the six-arm common cohort from Appendix B. Hiding history lowers the average coefficient error at every combination of target position and repetition, giving nine improvements. That average can hide changes in individual coefficients: $12/36$ coefficient–repetition errors worsen. In particular, mean BMI-coefficient error rises by 99% at position two and 95% at position five. Moving the target changes history length, order and preceding profiles together, so these comparisons cannot isolate the effect of history length.

Regression accuracy after correction. Relative to untouched batching, the learned map reduces mean coefficient error by 45% for Qwen/ESS, 31% for Llama/ESS and 27% for Qwen/NHANES; Llama/NHANES changes little (0.18% reduction). ACS error rises by 293% for Qwen and 42% for Llama, from a small Qwen baseline of 0.0315. Overall, $27/60$ coefficient–repetition and $6/18$ repetition-level mean errors worsen. Exact preservation lowers Llama/ACS error by 53% but raises Qwen/ACS and Llama/NHANES errors by 136% and 17%. In Qwen/ACS, uncorrected batching outperforms the other seven original methods. Table 23 separately reports the later scalar controls. The scalar that preserves marginals exactly improves this setting, so the original ranking does not extend to every subsequent variant. Lower average correlation error therefore does not guarantee better regression coefficients.

Table 12: **A single overall adjustment retains most of the full correction's improvement.** Three-seed mean correlation MAE on matched evaluation respondents. A_0 is the error from predicting zero correlation; one-person outputs provide another comparator. Gain retained is $(A_B - A_{\text{scalar}})/(A_B - A_{\text{full}})$. Both corrections use one-person fitting marginals and exclude human evaluation answers from fitting. †: all six ACS ten-person quality failures are included.

Domain	Model	A_0	One-person	Batch	Full map	Scalar	Gain retained
ESS	Qwen	0.193	0.182	0.286	0.147	0.156	93.2%
ESS	Llama	0.193	0.188	0.417	0.125	0.129	98.5%
NHANES	Qwen	0.160	0.249	0.396	0.212	0.210	100.9%
NHANES	Llama	0.160	0.332	0.517	0.204	0.217	95.9%
ACS†	Qwen	0.356	0.094	0.226	0.077	0.086	94.0%
ACS†	Llama	0.356	0.213	0.252	0.130	0.135	96.7%

D DEPENDENCE-MAP CONSTRUCTION AND DIAGNOSTIC COMPARISONS

The persistence of exaggerated relationship strength raises a question about correction: how much does a full set of pair-specific adjustments contribute beyond weakening all relationships together? A comparison between a full matrix of targets and a single overall adjustment addresses this question using the same one-person fitting outputs. Both rely on standard normal-score and whitening/recoloring operations, with marginal-preserving rank adjustment also established (Iman & Conover, 1982). Their value here is diagnostic: removing pair-specific target information reveals how much of the apparent gain depends on it. These checks reuse previously examined outputs and human references, so the results remain exploratory.

Pair-specific and single-parameter corrections. For each domain, model and seed, whole requests are divided into fitting and evaluation sets, with profiles matched to the corresponding one-person outputs. Each set plans fifty profiles per group (1,000 in ESS or 400 in NHANES/ACS), and fitting uses only answers accepted under both request sizes. Human answers provide the evaluation reference but do not select the corrections' parameters.

The correction first puts each variable on a common scale by ranking fitting answers within each group and mapping their percentiles to a standard normal distribution. These transformed values, or normal scores, provide the basis for estimating correlations; a fixed random ordering keyed to respondent IDs resolves ties. Let R_B, R_T denote the correlation matrices for batched and one-person fitting scores, pooling group covariances with equal weights. The full correction uses these matrices to move evaluation scores z toward the one-person dependence structure:

$$z^* = (z - \mu_{B,c}) D_B^{-1} \widetilde{R}_B^{-1/2} \widetilde{R}_T^{1/2}, \qquad \widetilde{R} = 0.9R + 0.1I, \tag{2}$$

where $\mu_{B,c}$ and D_B are fitting group means and pooled standard deviations, and matrix powers are symmetric. Returning to answer units involves finding each corrected score's percentile among the corrected fitting scores, then assigning the value at that percentile in the one-person fitting answers. Although this step draws on the one-person reference, the resulting evaluation distributions and correlations need not match it exactly.

The scalar correction follows the same procedure but multiplies every batched off-diagonal correlation by one number, leaving the diagonal at one. It replaces R_T with $I + \widehat{\alpha}(R_B - I)$, where

$$\widehat{\alpha} = \arg \min_{0 \le \alpha \le 1} \sum_{j<k} |\alpha R_{B,jk} - R_{T,jk}|. \tag{3}$$

One multiplier therefore applies to all pairs within a model, domain and seed. Keeping all other operations and fitting data fixed makes the full map's advantage a test of the additional value of pair-specific information.

The value of pair-specific information. The scalar retains 93–101% of the original map's MAE improvement, leaving a mean advantage of only 0.004–0.013 for the full map in five settings; in Qwen/NHANES, the scalar performs slightly better (Table 12). Most of the gain thus survives without a separate target for each pair, although both methods still use one-person data to set marginal distributions and overall correlation strength.

The baseline comparisons further limit what lower MAE establishes. Both full corrections in NHANES have larger errors than predicting zero correlations, while setting the target to an identity matrix gives lower MAE still (0.149 for Qwen and 0.135 for Llama) despite much weaker pattern agreement. This identity target asks for zero correlations on the normal-score scale; conversion back to answers can change those correlations, so the final dataset need not have zero Pearson correlations. Appendix E.1 reports pattern, sign and zero-baseline skill for every principal condition.

What the transformation changes. Examining the fitting data at successive stages helps explain why the full correction can outperform the one-person outputs it uses as a reference. Immediately after the linear adjustment, normal-score correlations are too strong on average in all six comparisons, but conversion back to answer units leaves them too weak relative to the one-person fitting answers. The later stages therefore change relationship strength in a direction that can improve MAE without recovering the human pattern. Removing the 0.1 regularization also improves MAE in five settings, so that regularization alone cannot explain the gains. Table 18 examines the further roles of midpoint tie treatment and conversion to normal scores and back without a linear adjustment.

Preserving distributions and testing usefulness. Because the correction changes individual-variable distributions as well as relationships, a further comparison holds those distributions fixed. For each group and variable, it keeps the original list of batched values and assigns them to people in the order of their corrected scores. The values themselves stay the same; what changes is which person receives each value. This exploratory Iman–Conover-style comparison uses the evaluation batch itself, preserving every marginal distribution and variance. Changes in those distributions, including their influence on pooled correlations, can therefore no longer explain a change in accuracy, although reordering can still weaken relationships or change agreement with profiles.

Under exact preservation, the full map has MAE 0.179 for Qwen and 0.186 for Llama in ESS, with corresponding values of 0.247 and 0.303 in NHANES and 0.160 and 0.210 in ACS. These averages conceal one worsening ACS/Llama seed. The scalar controls also undergo evaluation with exact preservation, without selection based on human-reference performance (Appendix E.2).

The practical importance of preserving distributions is apparent in the original correction, which worsens individual-variable accuracy in every setting. In ACS, SD-normalized error rises from 0.376 to 0.423 for Qwen and from 0.418 to 0.525 for Llama. Nor does lower correlation error guarantee better regression estimates: ACS/Qwen coefficient MAE rises from 0.032 to 0.124 under the full map and to 0.110 under the scalar. Other settings, including ESS and Qwen/NHANES, show improvements (Table 23), making the intended analysis essential to deciding whether a correction is useful.

Reference size and practical value. The study plans as many one-person fitting rows as batched evaluation rows, making direct use of the one-person outputs a practical alternative. Reducing the one-person reference from fifty to as few as five rows per group, while keeping all batched fitting rows, often lowers MAE further (Figure 4). This pattern is consistent with additional weakening of correlations rather than a benefit from more information, and does not determine a recommended reference size. Any claim of a cost or throughput advantage over parallel one-person requests would also require measurements of prefix reuse, batching efficiency and decoding costs in the intended serving system.

Marginal-error scales. The distributional comparisons use two normalizations of Wasserstein distance. For p outcomes and groups C, these are

$$M = \frac{1}{|C|p} \sum_{c,j} \frac{W_1(\widehat{F}_{cj}^{\text{synth}}, \widehat{F}_{cj}^{\text{real}})}{u_j - \ell_j}, \qquad M_{\text{SD}} = \frac{1}{|C|p} \sum_{c,j} \frac{W_1(\widehat{F}_{cj}^{\text{synth}}, \widehat{F}_{cj}^{\text{real}})}{s_{cj}^{\text{real}}}. \tag{4}$$

Here W_1 is the empirical Wasserstein distance, $[\ell_j, u_j]$ the fixed operational range, and s_{cj}^{real} the human-reference standard deviation. The range scale is reported for all comparisons, with greater emphasis on M_{SD} for ACS; preserving each marginal exactly also preserves both error measures.

E RELATIONSHIP STRENGTH, PATTERN AND CORRECTION DIAGNOSTICS

These analyses examine how much of an apparent improvement comes from weakening correlations and how much reflects a better pattern of relationships. Their design followed review of the original results. They include all six model–domain settings and all three seeds, including quality failures. New fits use only the original generated fitting data. Transformed rows were saved before scoring against human references. This keeps evaluation answers out of parameter fitting, although the previously examined references do not provide a new confirmatory test. The supplement includes aggregate tables, source hashes and analysis code.

E.1 STRENGTH, PATTERN AND THE ZERO-CORRELATION BASELINE

These diagnostics use the within-group relationship measure in Equation 1. The calculation gives groups equal weight, averaging their within-group covariance matrices before converting to correlations. This gives one pooled within-group correlation for each variable pair. The controlled history-pairing experiment uses a separate endpoint without a human reference, defined in Appendix B.

A vector x collects the synthetic correlations, and a vector y collects the corresponding human correlations, with variable pairs in the same order. With p variables, each vector has $K = p(p-1)/2$ entries. All entries use the original equal-group pooled-covariance scale. Two benchmarks put the observed error in context: predicting zero for every pair, and choosing the best single multiplier for all synthetic correlations. The diagnostics are

$$A = K^{-1} \sum_k |x_k - y_k|, \qquad A_0 = K^{-1} \sum_k |y_k|, \qquad S_0 = 1 - A/A_0,$$

$$A_{\text{or}} = \min_{\alpha \in [0,1]} K^{-1} \sum_k |\alpha x_k - y_k|.$$

Mean signed error is $K^{-1} \sum_k (x_k - y_k)$. Because positive and negative errors can cancel, a second measure captures magnitude bias as the average of $|x_k| - |y_k|$. Tables 13–15 therefore show mean $|x|$ beside the human value A_0. The aggregate files additionally retain both bias measures and the fraction of inflated pairs for every scored condition.

P and ρ are Pearson and Spearman correlations between x and y. They ask whether pairs with stronger positive relationships in the human data also tend to have stronger positive relationships in the synthetic data, and likewise for negative ones. Spearman uses the ranking of the signed correlations; Pearson uses their numerical values. Both describe agreement across variable pairs, rather than accuracy for individual respondents. Sign agreement is the fraction with $\text{sign}(x_k) = \text{sign}(y_k)$; an additional aggregate sensitivity restricts to $|y_k| > 0.05$. Pairs share variables and are not independent replications. ACS has only six pairs, making pattern summaries particularly sensitive to individual relationships. The original rank and sign conventions are retained across methods.

The zero predictor assigns zero to every correlation. It supplies a baseline for the estimates, without generating a dataset or setting a lower bound on error. An identity target instead applies a regularized correction on the normal-score scale and converts back to answers, whose Pearson correlations need not be zero. The oracle A_{or} uses the human evaluation reference to choose its multiplier and serves only as a diagnostic. A scalar optimum is the clipped weighted median of y_k/x_k with weights $|x_k|$ over nonzero x_k. Zero source entries contribute constant loss. This rule exactly minimizes the constrained absolute-error objective and is independently checked against linear programming for the eighteen generated-reference latent fits.

Fisher-scale sensitivity. The primary error measure uses differences on the Pearson-correlation scale. An exploratory check instead averages $|\operatorname{atanh}(x_k) - \operatorname{atanh}(y_k)|$, without refitting. It places greater weight on differences near perfect correlation. No values are clipped: a nonfinite transform would invalidate the entire endpoint. All displayed endpoints are finite. Batching still increases error in the primary ESS study, while full-map versus scalar rankings change in ACS (Table 16). Changing the error scale therefore does not remove the need to evaluate the intended analysis directly.

Table 13: ESS calibration cohort: three-seed means, $A_0 = 0.193$. P/ρ: Pearson/Spearman correlation-vector agreement; Sign: sign agreement; A_{or}: human-oracle scalar MAE. Exact variants preserve evaluated batch marginals; answer-fit scalar selects its parameter using generated fitting answers. M_{SD} is defined in Equation 4.

Model	Method	A	$\overline{\lvert r\rvert}$	S_0	P	ρ	Sign	A_{or}	M_{SD}
Qwen	Batch	0.286	0.467	-0.483	0.785	0.859	0.904	0.099	0.517
Qwen	One-person	0.182	0.359	0.054	0.801	0.835	0.906	0.098	0.569
Qwen	Marginals only	0.225	0.404	-0.169	0.781	0.842	0.903	0.100	0.598
Qwen	Row resampling	0.187	0.361	0.030	0.801	0.820	0.908	0.098	0.576
Qwen	Identity target	0.174	0.060	0.096	0.417	0.320	0.571	0.174	0.606
Qwen	Full map	0.147	0.304	0.240	0.789	0.819	0.891	0.100	0.599
Qwen	Scalar	0.156	0.312	0.191	0.779	0.832	0.895	0.101	0.602
Qwen	Answer-fit scalar	0.181	0.348	0.061	0.779	0.835	0.897	0.101	0.600
Qwen	Full / fit-batch	0.164	0.328	0.149	0.789	0.815	0.897	0.100	0.540
Qwen	Full / exact	0.179	0.350	0.073	0.793	0.828	0.897	0.098	0.517
Qwen	Scalar / exact	0.188	0.358	0.024	0.781	0.839	0.901	0.100	0.517
Qwen	Answer-fit / exact	0.182	0.348	0.054	0.780	0.839	0.897	0.100	0.517
Llama	Batch	0.417	0.594	-1.162	0.746	0.854	0.924	0.100	0.485
Llama	One-person	0.188	0.356	0.026	0.783	0.793	0.912	0.101	0.643
Llama	Marginals only	0.277	0.455	-0.436	0.773	0.849	0.928	0.097	0.655
Llama	Row resampling	0.195	0.362	-0.010	0.779	0.781	0.928	0.102	0.646
Llama	Identity target	0.167	0.055	0.135	0.442	0.376	0.715	0.167	0.665
Llama	Full map	0.125	0.274	0.354	0.800	0.800	0.916	0.096	0.662
Llama	Scalar	0.129	0.274	0.331	0.790	0.826	0.934	0.096	0.662
Llama	Answer-fit scalar	0.168	0.333	0.129	0.784	0.835	0.936	0.096	0.660
Llama	Full / fit-batch	0.167	0.329	0.135	0.777	0.773	0.908	0.102	0.521
Llama	Full / exact	0.186	0.354	0.033	0.776	0.778	0.918	0.100	0.485
Llama	Scalar / exact	0.189	0.351	0.021	0.758	0.811	0.930	0.099	0.485
Llama	Answer-fit / exact	0.175	0.333	0.091	0.760	0.810	0.926	0.099	0.485

Table 14: NHANES calibration cohort: same diagnostics as Table 13, with $A_0 = 0.160$. The full and scalar maps improve raw MAE while remaining worse than the constant-zero correlation predictor.

Model	Method	A	$\overline{\lvert r\rvert}$	S_0	P	ρ	Sign	A_{or}	M_{SD}
Qwen	Batch	0.396	0.527	-1.468	0.845	0.782	0.900	0.082	0.396
Qwen	One-person	0.249	0.380	-0.553	0.881	0.794	0.900	0.074	0.496
Qwen	Marginals only	0.339	0.470	-1.111	0.825	0.790	0.900	0.090	0.524
Qwen	Row resampling	0.303	0.434	-0.893	0.872	0.790	0.900	0.075	0.469
Qwen	Identity target	0.149	0.082	0.072	0.454	0.487	0.767	0.147	0.534
Qwen	Full map	0.212	0.330	-0.320	0.872	0.749	0.867	0.085	0.525
Qwen	Scalar	0.210	0.340	-0.310	0.847	0.786	0.900	0.086	0.526
Qwen	Answer-fit scalar	0.260	0.391	-0.619	0.848	0.774	0.900	0.085	0.526
Qwen	Full / fit-batch	0.238	0.358	-0.486	0.883	0.826	0.867	0.078	0.418
Qwen	Full / exact	0.247	0.367	-0.541	0.888	0.794	0.867	0.077	0.396
Qwen	Scalar / exact	0.263	0.393	-0.642	0.841	0.733	0.900	0.086	0.396
Qwen	Answer-fit / exact	0.234	0.364	-0.462	0.845	0.741	0.900	0.085	0.396
Llama	Batch	0.517	0.648	-2.226	0.771	0.798	0.900	0.096	0.428
Llama	One-person	0.332	0.463	-1.071	0.778	0.661	0.900	0.098	0.638
Llama	Marginals only	0.357	0.488	-1.225	0.787	0.729	0.900	0.095	0.631
Llama	Row resampling	0.378	0.509	-1.359	0.838	0.766	0.900	0.085	0.591
Llama	Identity target	0.135	0.090	0.160	0.576	0.451	0.567	0.134	0.632
Llama	Full map	0.204	0.321	-0.273	0.845	0.725	0.900	0.086	0.637
Llama	Scalar	0.217	0.347	-0.353	0.792	0.677	0.900	0.092	0.638
Llama	Answer-fit scalar	0.285	0.415	-0.774	0.793	0.709	0.900	0.094	0.635
Llama	Full / fit-batch	0.275	0.403	-0.716	0.835	0.762	0.867	0.087	0.449
Llama	Full / exact	0.303	0.434	-0.891	0.818	0.758	0.900	0.090	0.428
Llama	Scalar / exact	0.351	0.482	-1.188	0.762	0.701	0.900	0.098	0.428
Llama	Answer-fit / exact	0.340	0.471	-1.120	0.764	0.701	0.900	0.098	0.428

Table 15: ACS calibration cohort: same diagnostics, $A_0 = 0.356$. All six original ten-person quality failures remain included. SD-normalized marginal error is foregrounded because the range scale downweights errors on the broad monetary ranges. There are only six pairs.

Model	Method	A	$\overline{\lvert r \rvert}$	S_0	P	ρ	Sign	A_{or}	M_{SD}
Qwen	Batch	0.226	0.582	0.365	0.915	0.924	1.000	0.100	0.376
Qwen	One-person	0.094	0.429	0.736	0.919	0.752	1.000	0.071	0.397
Qwen	Marginals only	0.117	0.472	0.672	0.964	0.886	1.000	0.066	0.425
Qwen	Row resampling	0.101	0.457	0.717	0.959	0.943	1.000	0.059	0.437
Qwen	Identity target	0.177	0.211	0.502	0.922	0.276	0.778	0.177	0.424
Qwen	Full map	0.077	0.408	0.785	0.967	0.695	1.000	0.057	0.423
Qwen	Scalar	0.086	0.427	0.759	0.972	0.905	1.000	0.056	0.422
Qwen	Answer-fit scalar	0.074	0.419	0.792	0.967	0.943	1.000	0.053	0.423
Qwen	Full / fit-batch	0.137	0.493	0.616	0.968	0.752	1.000	0.079	0.393
Qwen	Full / exact	0.160	0.517	0.549	0.952	0.924	1.000	0.089	0.376
Qwen	Scalar / exact	0.191	0.547	0.463	0.929	0.924	1.000	0.095	0.376
Qwen	Answer-fit / exact	0.103	0.450	0.709	0.940	0.771	1.000	0.075	0.376
Llama	Batch	0.252	0.532	0.293	0.792	0.752	1.000	0.140	0.418
Llama	One-person	0.213	0.377	0.402	0.348	0.276	1.000	0.174	0.491
Llama	Marginals only	0.196	0.477	0.451	0.824	0.733	1.000	0.114	0.526
Llama	Row resampling	0.212	0.455	0.405	0.618	0.543	1.000	0.141	0.533
Llama	Identity target	0.216	0.154	0.393	0.924	0.371	0.889	0.216	0.536
Llama	Full map	0.130	0.399	0.634	0.884	0.752	0.944	0.094	0.525
Llama	Scalar	0.135	0.404	0.622	0.869	0.752	0.944	0.100	0.526
Llama	Answer-fit scalar	0.196	0.475	0.449	0.828	0.733	1.000	0.115	0.526
Llama	Full / fit-batch	0.208	0.424	0.415	0.602	0.581	0.944	0.150	0.434
Llama	Full / exact	0.210	0.424	0.411	0.673	0.733	0.889	0.150	0.418
Llama	Scalar / exact	0.230	0.423	0.354	0.562	0.638	0.944	0.162	0.418
Llama	Answer-fit / exact	0.252	0.424	0.292	0.551	0.619	0.944	0.172	0.418

Table 16: Fisher-z correlation MAE, three-seed means. The first two rows use the original ESS request experiment; subsequent rows use the disjoint fitting/evaluation calibration study. No parameter is chosen using this sensitivity. ACS retains its quality failures.

Cohort	Model	One-person	Batch	Full	Scalar
ESS primary	Qwen	0.208	0.337	—	—
ESS primary	Llama	0.212	0.591	—	—
ESS	Qwen	0.207	0.350	0.163	0.172
ESS	Llama	0.212	0.601	0.136	0.141
NHANES	Qwen	0.284	0.493	0.240	0.231
NHANES	Llama	0.386	0.666	0.224	0.235
ACS†	Qwen	0.382	0.541	0.225	0.208
ACS†	Llama	0.329	0.430	0.209	0.182

E.2 SINGLE-PARAMETER CORRECTIONS, FIXED MARGINALS AND REFERENCE SIZE

The primary scalar replaces only the target correlation matrix in the original map with $I+\widehat{\alpha}(R_B-I)$, using Equation 3. Source/target regularization, normal-score conventions, group fitting means/scales and singleton inverse CDFs remain fixed. Both methods therefore use the same fitting data and assign marginal values in the same way. The comparison tests the added value of learning each pair's correlation from one-person outputs. It leaves open how to choose a multiplier using batched outputs alone.

A second scalar control chooses α from $0, 0.05, \ldots, 1$ by minimizing the transformed *fitting answers'* Pearson-correlation MAE against singleton *fitting answers*. Selection is done separately for singleton and exact-batch marginal assignments. The supplement reports every grid score. Selection does not use the human reference or held-out one-person answers to choose a parameter, marginal assignment or reported setting. Direct answer-scale correlation-estimate calibration is additionally exported as an estimate-only diagnostic, with no synthetic-dataset or regression claim.

For exact marginal preservation, each correction reorders the same original values within each group and column. This holds group variances fixed. Otherwise, changes in those variances could change pooled correlations even if each group's correlations stayed the same. Reordering can still attenuate Pearson correlations, change their pattern and move values between profiles. Exact preservation therefore removes marginal replacement as an explanation but does not remove all possible weakening of correlations. The tables report both scalar choices and the full map, without selecting a winner using the human reference.

To test how many one-person outputs the correction needs, a further comparison reduces the number available for fitting while keeping all batched fitting rows. Within each group, a fixed hash orders the one-person rows. Each fit uses the first five, ten, twenty or thirty-five, so each larger subset contains the smaller ones. Repeating this with three orders per generation seed gives nine results at each reduced budget. The full original fit has one result per seed, or three in total. The five/ten-row variants fall below the original twenty-row support rule and are explicitly exploratory; they require five rows per group and finite positive latent variances. All reported fits are supported. The full ESS reference has 990–$1{,}000$ paired fitting rows depending on retained requests; the other domains similarly retain their original acceptance losses. The figure uses realized mean budgets rather than assuming every planned row survived.

Smaller references often improve MAE. For example, the ESS/Qwen full map has MAE 0.126 with one hundred singleton rows, compared with 0.147 at the full budget. This does not establish that fewer responses estimate relationships more accurately. Coarser empirical distributions and noisier normal-score estimates can change how strongly the correction weakens correlations. The curves serve as diagnostics; they do not select a budget or establish an optimum for cost or analytic usefulness.

E.3 WHAT THE TRANSFORMATION ACTUALLY REPRODUCES

For the unregularized latent fitting covariance R_B and linear operator $T = \widetilde{R}_B^{-1/2}\widetilde{R}_T^{1/2}$, the transformed covariance is $T^\top R_B T$. Regularizing matrices inside T does not make this equal to R_T, nor fix its diagonal. Conversion to correlations and the subsequent empirical CDF/inverse-CDF maps introduce further changes. The diagnostics therefore check the output of both stages: correlations after the linear step against R_T, and final fitting-answer correlations against those of the one-person fitting answers.

Table 18 varies $\widetilde{R} = (1-\gamma)R + \gamma I$, replaces randomized ties with midpoint ties, or sets the linear operator to identity while retaining the remaining CDF and marginal-assignment steps. These checks use fixed settings, without tuning them to the human reference. All zero-regularization fits are supported without an added eigenvalue floor. Removing regularization improves MAE in five settings, the exception being Qwen/ACS. Hence the original 0.1 regularization is not sufficient to explain the advantage over one-person outputs. The nonlinear stages change the resulting correlations. These checks cannot assign separate shares of that change to tie randomization, discrete empirical CDFs, marginal replacement and finite samples.

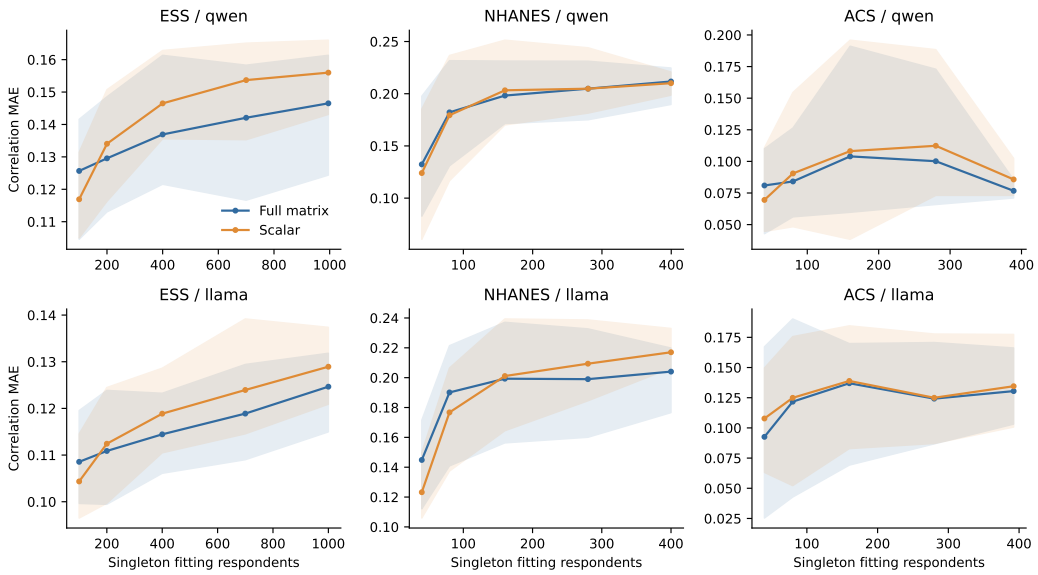


Figure 4: Nested singleton-reference sizes, keeping all batched fitting rows. Shading is the observed seed/subset-order range, not a confidence interval. Full matrix and scalar use the same references and singleton marginal assignment. All ACS quality failures remain included; five/ten rows per group are exploratory low-support variants.

Table 17: Fitting-target reproduction, three-seed means. Magnitude bias averages actual minus target $|r|$ at each stage. Final answer correlations undershoot singleton fitting magnitudes in all six settings, despite latent overshoot. ACS quality failures remain retained.

Domain	Model	Latent MAE	Latent mag. bias	Answer MAE	Answer mag. bias
ESS	Qwen	0.018	0.017	0.039	-0.034
ESS	Llama	0.047	0.045	0.075	-0.067
NHANES	Qwen	0.025	0.025	0.063	-0.024
NHANES	Llama	0.042	0.042	0.113	-0.113
ACS†	Qwen	0.017	0.017	0.060	-0.003
ACS†	Llama	0.046	0.046	0.082	-0.064

Converting values to normal scores and back can change individual rows even with an identity linear operator and exact marginal preservation. Multiple unseen values can fall in the same flat interval of a fitting CDF and receive tied normal scores. Breaking those ties by ID can reorder the values. Exact preservation guarantees the distribution of values, not each respondent's original value.

Table 18: Transformation sensitivity: mean correlation MAE on unchanged matched evaluation cohorts. Q/L denote Qwen/Llama. Identity-linear variants retain the nonlinear rank/CDF steps; they differ from the identity *target*. All ACS quality failures remain included.

Variant	ESS Q	ESS L	NHANES Q	NHANES L	ACS Q	ACS L
Full, $\gamma = 0.1$	0.147	0.125	0.212	0.204	0.077	0.130
Full, $\gamma = 0$	0.139	0.114	0.199	0.184	0.080	0.111
Full, $\gamma = 0.01$	0.140	0.115	0.202	0.187	0.078	0.114
Full, $\gamma = 0.05$	0.143	0.120	0.205	0.194	0.076	0.121
Full, $\gamma = 0.25$	0.155	0.143	0.225	0.230	0.079	0.148
Full, $\gamma = 0.5$	0.167	0.170	0.253	0.268	0.087	0.169
Midpoint ties	0.148	0.148	0.206	0.227	0.071	0.133
Identity linear / one-person	0.223	0.274	0.337	0.355	0.104	0.196
Identity linear / fit-batch	0.276	0.399	0.411	0.501	0.218	0.302
Identity linear / exact	0.285	0.415	0.395	0.517	0.223	0.274

E.4 Request means, dominant factors and predecessors

This check asks whether batching mainly makes whole requests differ from one another. For example, a request that produces unusually high answers on several variables could increase their pooled correlation even without stronger relationships within that request. Subtracting each request's mean from each variable removes such shifts. For comparison with one-person generation, its outputs are first collected into blocks containing the same ten profiles as the batched requests. Each block therefore contains matched profiles under both request sizes. Within group, total covariance decomposes into the covariance of request means and the weighted mean of within-request covariances. The implementation independently verifies this identity. Matching can leave unequal accepted block sizes, so request means receive their observed row weights within each group.

The main comparison asks whether removing request means reduces error against the original, group-centered human reference. Subtracting request means changes the quantity being measured, so a second comparison also centers source-human outcomes within the identical blocks. That second comparison describes the particular source cohort. Neither comparison assigns causal shares of the batching effect to a model mechanism.

Table 19: Original ESS matched cohort, three-seed means. A_{req} removes matched request means, retaining the group-centered human comparator. L_1/p is the largest correlation eigenvalue's share, before/after request centering. Human mean $|r|$ is 0.193.

Model	Arm	A	A_{req}	$\overline{\lvert r\rvert}$	ρ	A_{or}	L_1/p	$L_{1,\mathrm{req}}/p$
Qwen	One-person	0.183	0.191	0.360	0.840	0.097	0.411	0.417
Qwen	Batch	0.278	0.370	0.459	0.864	0.098	0.505	0.593
Llama	One-person	0.187	0.216	0.355	0.798	0.099	0.415	0.447
Llama	Batch	0.415	0.480	0.596	0.866	0.099	0.669	0.736

After removing request means, the leading eigenvalue still accounts for a larger share under batching. This is consistent with stronger shared variation among outcomes within requests, although it does not identify a single underlying psychological dimension. Differences between request means alone cannot explain the observed penalty.

At position two, the target has only one preceding respondent. The question is whether a common feature of that respondent's answers predicts several of the target's answers together. For Qwen/NHANES, the analysis standardizes the predecessor's five outcomes and summarizes them with their first principal component. This component explains 52.2% of their variance. Its loadings on poverty–income ratio, BMI, health, PHQ-9 and sedentary minutes are $-0.470, 0.487, 0.532, 0.468, 0.200$. After scaling to unit standard deviation, the component predicts each target outcome in a separate regression. Each regression includes target-profile and repetition

fixed effects; target outcomes are standardized separately by arm. The 1,186 jointly valid cases span all 400 target profiles. Standard errors cluster by target.

Table 20: Exploratory position-two predecessor regressions, conditional on profile and repetition fixed effects. Visible slopes oppose the predecessor loadings for all five outcomes, but are small. These are descriptive associations, not identified causal mediation effects.

Target outcome	Visible slope	Cluster SE	Hidden slope	Cluster SE
Poverty–income ratio	0.043	0.023	-0.047	0.034
BMI	-0.042	0.031	0.031	0.032
Self-rated health	-0.060	0.027	0.055	0.030
PHQ-9	-0.081	0.029	0.004	0.034
Sedentary minutes	-0.071	0.040	0.015	0.041

The visible slopes are consistent with a weak contrast response, but generated predecessor variation was not independently randomized for this regression. Hidden routes still retain history-length information. The regressions therefore cannot establish how much distortion comes from a predecessor's common influence on all outcomes. Other ways of responding to history remain possible.

Table 21: History diagnostics on each study's original common cohort. Three-seed means use six-arm support for Qwen/NHANES and two-arm support for the extensions. $A_0 = 0.160$ for NHANES and 0.356 for ACS. The strong raw Llama/NHANES gain disappears after oracle attenuation; Qwen/NHANES at position ten retains an improvement.

Model/domain/position	History	A	$\overline{\lvert r \rvert}$	S_0	P	Sign	A_{or}	M_{SD}
Qwen/NHANES/2	Visible	0.324	0.455	-1.019	0.879	0.900	0.074	0.432
Qwen/NHANES/2	Hidden	0.213	0.336	-0.326	0.898	0.867	0.068	0.466
Qwen/NHANES/5	Visible	0.374	0.505	-1.334	0.865	0.900	0.075	0.402
Qwen/NHANES/5	Hidden	0.210	0.341	-0.310	0.868	0.900	0.073	0.439
Qwen/NHANES/10	Visible	0.367	0.498	-1.286	0.848	0.900	0.081	0.402
Qwen/NHANES/10	Hidden	0.197	0.328	-0.231	0.894	0.900	0.065	0.442
Llama/NHANES/10	Visible	0.577	0.708	-2.596	0.749	0.900	0.099	0.464
Llama/NHANES/10	Hidden	0.300	0.431	-0.873	0.723	0.900	0.104	0.584
Qwen/ACS/10	Visible	0.166	0.479	0.534	0.718	1.000	0.109	0.348
Qwen/ACS/10	Hidden	0.103	0.329	0.711	0.806	1.000	0.103	0.350

E.5 Reference uncertainty and downstream consequences

A separate sensitivity uses 1,000 fixed-seed bootstrap draws of human respondents within each reporting group. A given reference draw is shared by all methods, families and seeds. Generated rows and fits are fixed. Each interval describes the paired difference between three-seed mean errors; no interval is reported if any draw is undefined. All displayed contrasts have zero undefined draws. The intervals describe sensitivity to the sampled human reference and omit uncertainty from generation. Because the reference exports lack survey primary-sampling-unit and household IDs, the bootstrap resamples respondents independently within groups rather than reproducing the survey design. This matters particularly for NHANES's 1,666 reference respondents across eight groups and clustered ACS respondents.

Table 22: Full-map minus scalar MAE, with human-reference-only percentile intervals. Negative differences favour the full map. Generated outputs and fitted parameters are held fixed; these intervals do not include generation or survey-design uncertainty.

Domain	Model	$A_{\mathrm{full}} - A_{\mathrm{scalar}}$	Reference-only 95% interval
ESS	Qwen	-0.0095	[-0.0102, -0.0086]
ESS	Llama	-0.0043	[-0.0051, -0.0035]
NHANES	Qwen	0.0016	[-0.0077, +0.0106]
NHANES	Llama	-0.0130	[-0.0224, -0.0071]
ACS†	Qwen	-0.0090	[-0.0115, -0.0067]
ACS†	Llama	-0.0040	[-0.0048, -0.0040]

The new scalar outputs are scored with the unchanged regressions and human standardization from Appendix C. Benefits vary across settings and marginal assignments. The original map worsens $27/60$ coefficient–repetition errors. Adding the scalar comparisons still leaves no method uniformly best. In particular, ACS/Qwen's exact-marginal scalar slightly beats uncorrected batching on the fixed regression, whereas its singleton-marginal scalar performs much worse, despite the latter's lower correlation MAE.

Table 23: Fixed downstream coefficient MAE on matched evaluation respondents. Full/scalar use singleton fitting marginals; exact variants preserve evaluated batch marginals. Lower is better. ACS retains its six original failed-quality ten-person arms.

Domain	Model	Batch	Full	Scalar	Full/exact	Scalar/exact
ESS	Qwen	0.105	0.058	0.043	0.058	0.041
ESS	Llama	0.101	0.070	0.087	0.093	0.093
NHANES	Qwen	0.157	0.115	0.081	0.145	0.126
NHANES	Llama	0.151	0.151	0.151	0.177	0.167
ACS†	Qwen	0.032	0.124	0.110	0.074	0.027
ACS†	Llama	0.249	0.354	0.355	0.117	0.152

Computational verification. An independent weighted-moment implementation reproduces all $1{,}674$ dataset-based correlation-MAE endpoints within 7.1×10^{-16}. Separate linear programs verify eighteen latent scalar fits; checks also verify fitting/evaluation separation, exact marginal multisets and all grid selections. Checks reproduced the original map and reported baselines before the analysis was extended. An initial scoring check failed because reference-row identifiers were absent. Adding deterministic row-index IDs allowed a uniqueness check, leaving values, fitting, weighting and selection unchanged. The failed run and amendment are archived. These checks establish agreement between computations; they do not provide new empirical replications.

F TEMPERATURE AND VARIATION-INSTRUCTION FOLLOW-UP

Design and completion. This experiment tests whether the batching effect persists at other temperatures and without the variation instruction. This exploratory follow-up generates new complete arrays using the original $2{,}000$ ESS and 800 NHANES profiles, examples and human references, in the fixed Qwen BF16 and Llama FP8 vLLM runtimes. There are four conditions. Three use the original prompt at temperatures 0.5, 1 and 1.5. The fourth uses temperature 1 but removes exactly "Preserve realistic individual variation." and its following space. In every condition, each model generates one-person and ten-person requests under three seeds. The adjacent instruction about internally consistent or mutually different respondents remains. Removing the sentence changes both wording and prompt length. The deletion condition uses only temperature 1. Profiles, examples, item/profile order, grammar, per-request seeds, other decoding parameters and token limits remain fixed. The temperature-1 control is also newly generated. The allocation and analysis were fixed before generation, after inspection of earlier results. The reused cohorts do not provide a new human holdout. All four allocations completed: $26{,}400$ ESS and $10{,}560$ NHANES requests per model, or $73{,}920$ requests in total.

Support and uncertainty. Invalid or truncated requests are rejected in full, without repair or regeneration. Combining the four conditions with the two request sizes gives eight arms. For the primary comparison, a profile enters the analysis only if its answers are valid in all eight arms within that seed. This makes each comparison use the same people and must leave at least twenty in every original group. The protocol also specifies two secondary analyses. One matches profiles across request sizes separately within each condition; the other uses every retained answer in each arm. A three-seed mean is withheld if any required endpoint cannot be calculated. Its definition retains all original countries, pairs and seeds. Table 24 reports the within-condition matched secondary results. The retained profiles can differ between conditions. Those point comparisons therefore cannot establish a population-wide effect of changing temperature or wording. Every Llama request passes validation, so all three support definitions coincide for that model.

Table 24: Correlation accuracy under each temperature and prompt condition. MAEs average three seeds, matching accepted profiles across request sizes within each condition as planned; $\Delta A = A_{10} - A_1$. Validity uses all planned requests before matching. A dagger marks a condition in which any size or seed has less than 99% validity or any truncation. Dashes indicate a three-seed mean that cannot be calculated under the support rule. These secondary estimates have no confidence intervals. Llama's means also equal its primary estimates.

Domain	Model	Condition	A_1	A_{10}	ΔA	Valid 1	Valid 10
ESS	Qwen	$T = 0.5$	0.249	0.342	0.093	100.0%	100.0%
ESS	Qwen	$T = 1.0$	0.185	0.272	0.087	99.9%	99.5%
ESS	Qwen	$T = 1.5^{\dagger}$	—	—	—	91.3%	51.5%
ESS	Qwen	Omit sentence, $T = 1^{\dagger}$	0.192	0.281	0.090	99.9%	99.0%
ESS	Llama	$T = 0.5$	0.187	0.420	0.233	100.0%	100.0%
ESS	Llama	$T = 1.0$	0.187	0.405	0.219	100.0%	100.0%
ESS	Llama	$T = 1.5$	0.182	0.370	0.188	100.0%	100.0%
ESS	Llama	Omit sentence, $T = 1$	0.188	0.414	0.226	100.0%	100.0%
NHANES	Qwen	$T = 0.5$	0.328	0.524	0.197	100.0%	100.0%
NHANES	Qwen	$T = 1.0$	0.269	0.418	0.149	99.9%	100.0%
NHANES	Qwen	$T = 1.5^{\dagger}$	0.156	0.236	0.080	97.3%	65.4%
NHANES	Qwen	Omit sentence, $T = 1$	0.284	0.454	0.170	99.9%	100.0%
NHANES	Llama	$T = 0.5$	0.345	0.542	0.197	100.0%	100.0%
NHANES	Llama	$T = 1.0$	0.336	0.532	0.196	100.0%	100.0%
NHANES	Llama	$T = 1.5$	0.324	0.521	0.197	100.0%	100.0%
NHANES	Llama	Omit sentence, $T = 1$	0.350	0.533	0.183	100.0%	100.0%

Table 25: Primary contrasts using profiles accepted in all eight condition-by-size arms. Intervals use $2{,}000$ paired request-block bootstrap draws within groups, carrying all conditions and seeds together and holding the human reference fixed. The last two columns compare the batching penalty with the new original-prompt $T = 1$ control. The fixed analysis rule determines which estimates and intervals are withheld. Intervals are conditional and pointwise, without multiplicity adjustment. Daggers follow Table 24's rule.

Domain	Model	Condition	ΔA	95% interval	Change vs. $T = 1$	95% interval
ESS	Qwen	$T = 0.5$	—	Unsupported	—	Unsupported
ESS	Qwen	$T = 1.0$	—	Unsupported	—	Unsupported
ESS	Qwen	$T = 1.5^{\dagger}$	—	Unsupported	—	Unsupported
ESS	Qwen	Omit sentence, $T = 1^{\dagger}$	—	Unsupported	—	Unsupported
ESS	Llama	$T = 0.5$	0.233	[0.216, 0.255]	+0.014	[-0.003, +0.031]
ESS	Llama	$T = 1.0$	0.219	[0.204, 0.241]	—	—
ESS	Llama	$T = 1.5$	0.188	[0.171, 0.213]	-0.030	[-0.050, -0.012]
ESS	Llama	Omit sentence, $T = 1$	0.226	[0.215, 0.245]	+0.007	[-0.008, +0.024]
NHANES	Qwen	$T = 0.5$	0.194	Withheld	+0.026	Withheld
NHANES	Qwen	$T = 1.0$	0.168	Withheld	—	—
NHANES	Qwen	$T = 1.5^{\dagger}$	0.080	Withheld	-0.089	Withheld
NHANES	Qwen	Omit sentence, $T = 1$	0.189	Withheld	+0.021	Withheld
NHANES	Llama	$T = 0.5$	0.197	[0.169, 0.235]	+0.002	[-0.018, +0.019]
NHANES	Llama	$T = 1.0$	0.196	[0.171, 0.233]	—	—
NHANES	Llama	$T = 1.5$	0.197	[0.174, 0.233]	+0.001	[-0.017, +0.022]
NHANES	Llama	Omit sentence, $T = 1$	0.183	[0.157, 0.220]	-0.013	[-0.038, +0.011]

At high temperature, Qwen loses enough ESS answers that the primary comparison fails its minimum-count rule in seeds zero and one: the smallest groups contain nine and zero profiles. The remaining seed does not replace the three-seed result. All Qwen/NHANES primary point estimates are supported, but differ from the secondary means because the eight-arm intersection retains different profiles. In the planned $2{,}000$ paired bootstrap draws, $1{,}133$ fail support or variance checks, so the fixed rule suppresses every Qwen/NHANES primary interval; failed draws are not discarded to form an interval. No joint Qwen/ESS interval is formed. Both Llama allocations support every primary point estimate and all $2{,}000$ bootstrap draws. These intervals hold the human reference fixed and do not cover new seeds, prompts or cohorts.

Table 26: Profile counts for the primary comparison and validity at high temperature. Shared N and the minimum group count apply to all eight arms within a seed. Validity and truncation refer to original-prompt $T = 1.5$ ten-person requests. "Supported" indicates that primary point estimates can be calculated; it does not guarantee an interval.

Domain	Model	Seed	Shared N	Min./group	Valid high-T	Truncated	Supported
ESS	Qwen	0	885	9	98/200	38	No
ESS	Qwen	1	996	0	109/200	35	No
ESS	Qwen	2	940	26	102/200	37	Yes
ESS	Llama	0	2000	100	200/200	0	Yes
ESS	Llama	1	2000	100	200/200	0	Yes
ESS	Llama	2	2000	100	200/200	0	Yes
NHANES	Qwen	0	490	30	50/80	2	Yes
NHANES	Qwen	1	537	29	55/80	1	Yes
NHANES	Qwen	2	506	29	52/80	1	Yes
NHANES	Llama	0	800	100	80/80	0	Yes
NHANES	Llama	1	800	100	80/80	0	Yes
NHANES	Llama	2	800	100	80/80	0	Yes

Temperature and instruction results. At temperatures 0.5 and 1, and with the sentence removed, the within-condition batching penalty is positive in all three seeds of both models and both domains. All temperature-0.5 requests pass validation. The Qwen/ESS omission arm at seed one retains $197/200$ ten-person requests and therefore fails the original quality threshold; it remains included. The sentence is not necessary for the observed penalty, and lowering temperature does not remove it. These results do not establish equivalence between prompts or robustness to arbitrary wording.

Llama retains positive batching penalties at $T = 1.5$ as well. In ESS, the penalty falls from 0.219 at $T = 1$ to 0.188, a change of -0.030 with conditional interval $[-0.050, -0.012]$. Its lower-temperature and instruction-deletion changes have intervals spanning zero. In NHANES, all four penalties are close, 0.183–0.197, and all three changes relative to $T = 1$ have intervals spanning zero. Intervals spanning zero leave the differences uncertain; they do not establish equivalence.

Qwen behaves differently at $T = 1.5$: ten-person validity falls to $309/600$ in ESS and $157/240$ in NHANES, including 110 and four truncated requests, respectively. Singleton validity is also lower: $5{,}480/6{,}000$ in ESS and $2{,}335/2{,}400$ in NHANES. Lower MAE among retained high-temperature NHANES outputs coexists with improved pattern and oracle-scaled metrics for some comparisons; it cannot all be attributed to scalar attenuation. Because high temperature also changes which answers survive validation, lower error among those answers does not establish an overall improvement in the generator.

Magnitude and analytic utility. The fresh $T = 1$ ESS controls support the magnitude-inflation interpretation. Batching raises mean absolute correlation from 0.362 to 0.454 for Qwen and 0.355 to 0.585 for Llama, against 0.193 in the human reference. Oracle-scaled MAE changes only from 0.097 to 0.098 for Qwen and 0.099 to 0.098 for Llama. For Llama/ESS, raising temperature from 1 to 1.5 reduces batched mean absolute correlation from 0.585 to 0.550, while pattern agreement and oracle-scaled MAE change little. Raw MAE improvement need not mean a comparable recovery of structure. The oracle is a diagnostic using human answers, not an evaluated repair.

Across the six supported Qwen domain/condition comparisons at $T = 0.5$, $T = 1$ and after deletion, batching improves mean marginal error while worsening fixed regression coefficient error. Llama

Table 27: Relationship strength, pattern and analytic accuracy for the same matched profiles as Table 24. Every mean requires all three seeds; Qwen/ESS $T = 1.5$ fails this rule. b is request size, P is Pearson agreement between the correlation vectors, A_{or} is MAE after the human-oracle magnitude adjustment, M_{SD} is marginal Wasserstein error and E_β is regression coefficient MAE. Daggers use the rule above.

Domain	Model	Condition	b	$\overline{\lvert r \rvert}$	P	A_{or}	M_{SD}	E_β
ESS	Qwen	$T = 0.5$	1	0.429	0.769	0.104	0.616	0.071
ESS	Qwen	$T = 0.5$	10	0.524	0.771	0.102	0.530	0.116
ESS	Qwen	$T = 1.0$	1	0.362	0.804	0.097	0.563	0.062
ESS	Qwen	$T = 1.0$	10	0.454	0.792	0.098	0.500	0.081
ESS	Qwen	Omit sentence, $T = 1^\dagger$	1	0.370	0.806	0.097	0.559	0.059
ESS	Qwen	Omit sentence, $T = 1^\dagger$	10	0.463	0.791	0.098	0.502	0.095
ESS	Llama	$T = 0.5$	1	0.356	0.783	0.100	0.638	0.055
ESS	Llama	$T = 0.5$	10	0.599	0.753	0.098	0.485	0.105
ESS	Llama	$T = 1.0$	1	0.355	0.789	0.099	0.635	0.058
ESS	Llama	$T = 1.0$	10	0.585	0.758	0.098	0.481	0.105
ESS	Llama	$T = 1.5$	1	0.349	0.788	0.100	0.629	0.055
ESS	Llama	$T = 1.5$	10	0.550	0.760	0.097	0.477	0.105
ESS	Llama	Omit sentence, $T = 1$	1	0.362	0.801	0.097	0.634	0.053
ESS	Llama	Omit sentence, $T = 1$	10	0.593	0.755	0.099	0.486	0.104
NHANES	Qwen	$T = 0.5$	1	0.459	0.887	0.075	0.514	0.103
NHANES	Qwen	$T = 0.5$	10	0.655	0.796	0.093	0.382	0.175
NHANES	Qwen	$T = 1.0$	1	0.400	0.880	0.074	0.466	0.098
NHANES	Qwen	$T = 1.0$	10	0.549	0.833	0.085	0.352	0.137
NHANES	Qwen	$T = 1.5^\dagger$	1	0.281	0.887	0.071	0.427	0.098
NHANES	Qwen	$T = 1.5^\dagger$	10	0.366	0.891	0.070	0.352	0.097
NHANES	Qwen	Omit sentence, $T = 1$	1	0.415	0.875	0.075	0.469	0.109
NHANES	Qwen	Omit sentence, $T = 1$	10	0.585	0.820	0.087	0.370	0.174
NHANES	Llama	$T = 0.5$	1	0.476	0.810	0.091	0.605	0.143
NHANES	Llama	$T = 0.5$	10	0.673	0.758	0.099	0.402	0.112
NHANES	Llama	$T = 1.0$	1	0.467	0.810	0.092	0.603	0.134
NHANES	Llama	$T = 1.0$	10	0.663	0.764	0.098	0.408	0.120
NHANES	Llama	$T = 1.5$	1	0.455	0.818	0.090	0.598	0.130
NHANES	Llama	$T = 1.5$	10	0.652	0.765	0.097	0.406	0.127
NHANES	Llama	Omit sentence, $T = 1$	1	0.481	0.796	0.093	0.594	0.132
NHANES	Llama	Omit sentence, $T = 1$	10	0.664	0.770	0.097	0.399	0.118

improves mean marginal error in all eight comparisons. Its regression error worsens in ESS but improves in NHANES under every condition, even though correlation MAE worsens in both domains. These results reinforce the need to test the intended analysis alongside correlation and marginal accuracy. Neither request size is uniformly better for the regression tasks.

Verification. Raw hashes, source/input bindings and independently parsed dispositions are verified for all 73,920 completed requests. An additional explicit NumPy within-group covariance replay checks every supported primary and secondary correlation/MAE endpoint, including agreement with the earlier Qwen report. Unsupported endpoints and withheld intervals remain explicit. Aggregate endpoints, pair correlations, coverage, verification records and frozen source accompany the completed two-model report.

G WHERE WITHIN-GROUP RELATIONSHIPS CAN ARISE

There are two ways for a relationship to arise within a country. First, people with different profiles may have different average answers. For example, if some profiles predict higher trust in both parliament and the courts, those differences contribute to a positive relationship between the two responses. Second, answers may vary together even among people with the same profile: a person who reports more trust in parliament than their profile predicts may also report more trust in the courts. Both sources of variation can contribute to the observed relationship.

Let x be a conditioning profile, $c = c(x)$ its group, and $m_k(x) = \mathbb{E}[Y_k \mid x]$. In the opinion panel, c is country and x contains the ten prompt fields. With finite second moments, the law of total covariance separates these two components:

$$\mathrm{Cov}(Y_1, Y_2 \mid c) = \underbrace{\mathrm{Cov}_{x|c}\{m_1(x), m_2(x)\}}_{\text{between-profile component}} + \underbrace{\mathbb{E}\{\mathrm{Cov}(Y_1, Y_2 \mid x) \mid c\}}_{\text{within-profile component}}. \tag{5}$$

The identity holds for real and synthetic data; neither term is intrinsically erroneous. Zero conditional covariance removes the second term (Appendix H.1), but autoregressive generation need not satisfy that restriction.

Relationships between groups and within groups. Between-country and within-country associations can differ (Robinson, 1950). Equation 5 alone does not connect the two. Interpreting within-country relationships as transferred from country means requires further assumptions about what the model has learned (Appendix H.2). A generator can have correlated country means but zero within-country correlations (Appendix H.1). Conversely, correct conditional joint distributions *and the real profile distribution* reproduce the real joint distribution.

An aggregate-data predictor. The audit retrospectively predicts synthetic within-group correlations from $\rho^{\mathrm{eco}} = \mathrm{Corr}_c(\mathbb{E}[Y_1 \mid c], \mathbb{E}[Y_2 \mid c])$, weighting groups by real sample size. Cell means and sizes suffice; here they come from benchmark microdata. This fit is retrospective, without a held-out forecast. Human within-group and aggregate correlations may already align, so successful prediction does not establish that the model transfers one to the other.

Descriptive decomposition. Classical moment estimators summarize profile and residual covariances (Henderson, 1953; Searle et al., 1992); Appendix H.3 gives the proof. Shared requests may violate the estimators' assumptions of independence and common covariances (Appendix H.4). The fitted shares therefore describe the data without identifying causal mechanisms. Identical profiles remove the between-profile component, but the remaining relationships may still be inaccurate. Prompts and request size may affect both components.

Several dimensions of accuracy. Correlation error A and marginal Wasserstein error M measure different properties (Section 4). Improvement in one need not improve the other, and one-person outputs can be inaccurate on both. The correction experiments evaluate both properties without assuming that their one-person reference is accurate or identifies a mechanism.

H COVARIANCE IDENTITIES AND THEIR SCOPE

These identities clarify which assumptions connect differences between profiles to relationships within groups. They also describe the limits of interpreting the fitted covariance components in the supporting audit.

H.1 CONDITIONAL-MEAN DEPENDENCE

Definition 1 (Per-field conditional generation)**.** A generator is per-field conditional for a pair when $Y_1 = m_1(x) + u$, $Y_2 = m_2(x) + v$, with $\mathbb{E}[u \mid x] = \mathbb{E}[v \mid x] = 0$ and $\mathrm{Cov}(u, v \mid x) = 0$. This is a covariance restriction, not conditional independence. Assume finite second moments and $c = c(x)$ throughout.

Proposition 1 (Conditional-mean dependence)**.** *Under Definition 1,* $\mathrm{Cov}(Y_1, Y_2 \mid c) = \mathrm{Cov}_{x|c}\{m_1(x), m_2(x)\}$.

Proof of Proposition 1. Apply the standard total-covariance identity (Equation 5); the second term is zero by Definition 1. □

Correlated country means need not produce any relationship within a country. To see this, consider a generator whose profile records only the country. It assigns each variable a country-specific mean and adds independent noise. Formally, let $x = c$ and $Y_k = \alpha_k(c) + \epsilon_k$, with mutually independent, mean-zero, unit-variance errors independent of c. Within-country correlation is zero regardless of the correlation of country means; with at least three countries, the latter can attain any value in $[-1, 1]$.

H.2 A SUFFICIENT SHARED-GRADIENT CONDITION

Corollary 1 (Shared-gradient dependence)**.** *Under Definition 1, if* $m_k(x) = \alpha_k + \gamma_k s(x)$ *exactly, then* $\mathrm{Cov}(Y_1, Y_2 \mid c) = \gamma_1\gamma_2\mathrm{Var}(s(x) \mid c)$.

Proof. Substitute the mean functions into Proposition 1 and use bilinearity of covariance. □

If additional terms $r_k(x)$ are present, the exact expression also includes $\gamma_1\mathrm{Cov}(s, r_2 \mid c) + \gamma_2\mathrm{Cov}(s, r_1 \mid c) + \mathrm{Cov}(r_1, r_2 \mid c)$. These extra terms can reverse the sign. Treating γ_k as a learned ecological slope requires a separate assumption. Residual variances also enter the denominator of a correlation, so the covariance result alone does not determine correlation strength or the ordering of variable pairs.

H.3 NESTED COVARIANCE-COMPONENT ESTIMATION

The decomposition below distinguishes differences between countries, differences between profiles within a country, and differences between rows sharing a profile. The model assigns these contributions to a country mean, a profile effect and a row residual, respectively:

$$Y_{k,cpi} = \mu_{k,c} + a_{k,p} + e_{k,cpi}, \tag{6}$$

with N rows, P profiles nested in C countries, positive fixed profile counts n_p, and $n_c = \sum_{p \in c} n_p$. Bars denote row-weighted means.

Proposition 2 (Classical moment estimators)**.** *In Equation 6, let* $\mu_{k,c}$ *be fixed. Suppose profile effects are independent with zero mean and covariance* Σ_A*, row residuals are independent with zero mean and covariance* Σ_E*, the two families are independent, and both covariances are common across cells. Conditional on counts independent of these effects, define*

$$S_W = \sum_{c,p,i} (Y_{1,cpi} - \bar{Y}_{1,cp})(Y_{2,cpi} - \bar{Y}_{2,cp}), \qquad S_B = \sum_{c,p} n_p (\bar{Y}_{1,cp} - \bar{Y}_{1,c})(\bar{Y}_{2,cp} - \bar{Y}_{2,c}).$$

If $N > P$ and $\kappa > 0$, then

$$\widehat{\Sigma}_{E,12} = \frac{S_W}{N-P}, \qquad \widehat{\Sigma}_{A,12} = \frac{S_B - (P-C)\widehat{\Sigma}_{E,12}}{\kappa}, \qquad \kappa = N - \sum_c \frac{\sum_{p\in c} n_p^2}{n_c}$$

are unbiased for the two covariance components.

Proof. Let Z_P and Z_C be profile and cell indicator matrices, and H_P, H_C their orthogonal projection matrices. The cell space is nested in the profile space. Thus $S_W = Y_1^\top (I - H_P) Y_2$ and $S_B = Y_1^\top (H_P - H_C) Y_2$. Both quadratic forms annihilate the fixed cell means. The cross-covariance matrix is $\mathrm{Cov}(Y_1, Y_2) = \Sigma_{A,12} Z_P Z_P^\top + \Sigma_{E,12} I$. Taking traces gives

$$\mathbb{E}[S_W] = (N-P)\Sigma_{E,12}, \qquad \mathbb{E}[S_B] = (P-C)\Sigma_{E,12} + \kappa \Sigma_{A,12},$$

because $\mathrm{tr}(I - H_P) = N - P$, $\mathrm{tr}(H_P - H_C) = P - C$, $(I - H_P) Z_P = 0$, and $\mathrm{tr}\{(H_P - H_C) Z_P Z_P^\top\} = N - \sum_c n_c^{-1} \sum_{p\in c} n_p^2$. Substitution establishes the claim. □

H.4 Limits of interpreting the fitted split

Dividing both cross-products by N instead gives $\mathbb{E}[S_W/N] = (N-P)\Sigma_{E,12}/N$ and $\mathbb{E}[S_B/N] = \kappa \Sigma_{A,12}/N + (P-C)\Sigma_{E,12}/N$. The resulting split therefore includes sampling error in profile means within the between-profile component.

If requests add covariance $\Sigma_{B,12} Z_B Z_B^\top$, where Z_B indicates requests, the expectations of S_W, S_B also contain $\Sigma_{B,12}\mathrm{tr}\{(I - H_P) Z_B Z_B^\top\}$ and $\Sigma_{B,12}\mathrm{tr}\{(H_P - H_C) Z_B Z_B^\top\}$, respectively. Neither trace is generally zero. Resampling countries cannot remove the resulting bias. Adding a request intercept may also leave effects that depend on profile or position.

With independent but heterogeneous residuals of covariance $\Sigma_{E,12}^{(p)}$, the residual estimator targets $\sum_p (n_p - 1)\Sigma_{E,12}^{(p)}/(N-P)$ rather than a respondent-weighted average. Singletons supply no information about their residual covariance. The reported ratio $\widehat{\Sigma}_{E,12}/(\widehat{\Sigma}_{E,12} + \widehat{\Sigma}_{A,12})$ is not unbiased, may fall outside $[0, 1]$, and is undefined at a zero denominator. Under the homogeneous model it targets a superpopulation share, not the realized finite-profile split: for $w_{p|c} = n_p/n_c$, the expected realized between-profile covariance is $(1 - \sum_{p\in c} w_{p|c}^2)\Sigma_{A,12}$. None of these estimator limitations invalidates Equation 5.

H.5 Pooled correlation and variance changes

For nonnegative group weights w_c summing to one, empirical covariances $s_{12,c}$ and variances $s_{11,c}, s_{22,c}$ computed with divisor n_c, and a positive denominator,

$$r_W = \frac{\sum_c w_c s_{12,c}}{\sqrt{(\sum_c w_c s_{11,c})(\sum_c w_c s_{22,c})}}.$$

Rescaling the two variables within cell c by positive factors a_c, b_c replaces these three terms by $a_c b_c s_{12,c}$, $a_c^2 s_{11,c}$, and $b_c^2 s_{22,c}$. The factors cancel if each is constant across groups, but not generally. With two equally weighted groups, unit variances and correlations 0 and 1, $r_W = 1/2$. Doubling both variables in the second group gives $r_W = 4/5$ without changing either group's correlation. The second group's covariance and variances now contribute more to the pooled calculation, even though its group weight is unchanged. A pooled correlation can therefore change when marginal variances change, even if every group's correlation stays fixed.

I Supporting audit and archival results

I.1 Relationship accuracy in the reference panel

The original focal pairs relate police trust to income, happiness and satisfaction with democracy. The comparison covers the 35 countries with enough data in every population, using real-sample country shares to weight both human and synthetic answers. In that order, the three human correlations are

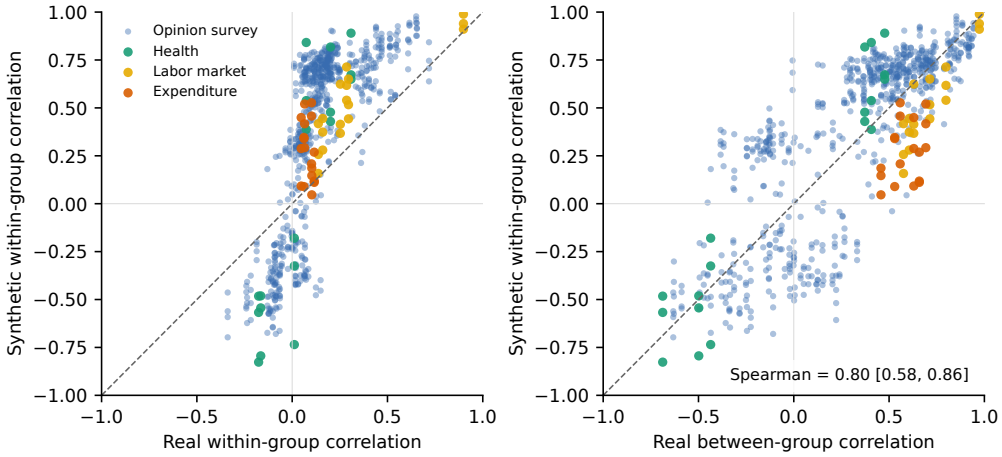


Figure 5: **Predicting generated correlations does not establish their accuracy.** The reference panel supplies 738 model-by-pair comparisons, excluding household size. Absolute synthetic correlations exceed the real reference in 98.2% of comparisons; these are not independent tests. Left: real versus synthetic pooled within-cell correlations; the dashed line represents agreement. Right: synthetic within-cell versus real between-cell correlations. The rank interval resamples variables and excludes generation uncertainty. Opinion surveys supply 684 comparisons; the other domains supply eighteen each.

Table 28: Correlation accuracy in the reference opinion panel, 171 pairs per model. Mean absolute error is the mean of $|r_W^{synth} - r_W^{real}|$ across pairs.

Model	Mean absolute error	Median $\lvert r_{real}\rvert$	Median $\lvert r_{synth}\rvert$
Claude Opus 5	0.278	0.127	0.476
Llama-3.3-70B	0.410	0.127	0.696
Qwen3-235B	0.413	0.127	0.644
GPT-oss-120B	0.410	0.127	0.641

0.062, 0.205 and 0.380. Claude's corresponding correlations are 0.371, 0.495 and 0.664. Across the other models and these same pairs, correlations range from 0.602 to 0.768. The gaps therefore persist when country coverage and weights are matched. The estimates describe complete cases and do not use survey weights.

The fitted decomposition estimates how much covariance remains among respondents with the same profile. For Claude, these within-profile shares are approximately 17%, 22% and 30% for the three pairs in the same order. For Llama, they range from 85–89%. These shares describe the fitted decomposition. They neither establish invalidity nor identify an internal mechanism (Appendix H.4). Removing country means and removing profile means leave different residual structures.

I.2 Supporting archival sensitivities

Appendix L reports archival contrasts, retention checks and prediction outcomes. Those studies vary in profiles, references or seed policies, so they do not provide matched replications of the main experiment.

I.3 What aggregate correlations predict

Table 29 relates cell-size-weighted between-group correlations to synthetic within-group correlations. Retaining household size gives Spearman 0.800 over 814 comparisons.

Real within-group correlations predict generated correlations about as well in rank (0.799 versus 0.796 for aggregates). Regression slopes are 0.903 $[0.573, 1.122]$ on aggregates and 1.713

Table 29: Aggregate predictability on the primary battery. Intervals resample variables within domains; B counts usable replicates out of 1,000. The three four-variable domain intervals are descriptive, with no coverage guarantee.

Domain	Rows	Spearman ρ	Slope on aggregates	B
Opinion survey	684	$0.823\ [0.57, 0.87]$	$0.917\ [0.44, 1.17]$	1000
Health	18	$0.868\ [0.34, 0.87]$	$1.220\ [1.14, 1.74]$	640
Labor market	18	$0.868\ [0.68, 0.87]$	$1.545\ [1.28, 1.65]$	627
Expenditure	18	$0.310\ [-0.62, 0.80]$	$0.596\ [-2.27, 2.63]$	672
All domains	738	$0.796\ [0.58, 0.86]$	$0.903\ [0.57, 1.12]$	1000

$[0.737, 2.421]$ on real within-group correlations; the predictors have different scales and both intervals include one. These comparisons do not establish that models transfer between-group relationships to individuals. Aggregate and generated signs disagree in 19.1% of rows. The regressions predict correlation levels, rather than errors against human correlations. The weak expenditure result and expanded-variable checks also limit generalization (Appendix L).

I.4 Prompting and exemplars provide incomplete remedies

The archival factorial finds no reliable remedy among its tested prompts. Asking for correlation matrices directly also produces some invalid matrices. These limited comparisons leave open whether other prompts or examples could improve accuracy (Appendix L).

J Reference-panel data and evaluation

J.1 Sources, harmonization, and observed populations

The supporting reference panel compares four model families across several data domains. This section describes its sources and evaluation. Appendix M.1 gives the separate eligibility and sampling rules for the controlled experiments. The opinion benchmark combines 540,671 ESS respondents (rounds 1–11, 2002–2024) and 270,760 WVS respondents (waves 5–7, 2004–2023; Trend File v4.1 (Haerpfer et al., 2022)). Source editions appear in Appendix J.6. An unstratified $80/20$ split (seed 42) randomly divides individual respondents into source and evaluation sets. This gives 649,144 source and 162,287 evaluation rows. The files call these "training" and "test", although neither set trains a language model. Source rows supply profiles and examples; evaluation rows provide the human comparison. ESS-only experiments first select ESS respondents from these existing sets. They do not make a new split by country or round. The data build preserves missing values and source weights. The registry defines each variable's coding, direction, range and fielding restrictions.

Not all variables are available in all 109 countries. These WVS waves omit democracy satisfaction, and ESS income position begins in round 4. The four focal items are complete for 61,374 evaluation respondents in thirty-eight countries with at least twenty observations. These comparisons therefore describe an ESS subset. Other pairs use different available countries and respondents; none represents a global population correlation.

The US sources are NHANES 2017–2018 (National Center for Health Statistics, 2020), ACS 2022 one-year Massachusetts employed-adult microdata (U.S. Census Bureau, 2023), and MEPS 2023 HC-251 (Agency for Healthcare Research and Quality, 2025). NHANES joins DEMO_J, BMX_J, HUQ_J, DPQ_J and PAQ_J by SEQN. Each domain has four outcomes. NHANES measures poverty-income ratio, BMI, self-rated health and PHQ-9; ACS measures income, wages, weekly hours and weeks worked. MEPS measures office, prescription and inpatient expenditure, alongside self-rated health. For NHANES and ACS, each demographic group is a combination of sex/gender, race, education and age band. MEPS uses region in place of education. These supporting-panel groups differ from the eight sex-by-age groups in the controlled experiments. A `log1p` expenditure sensitivity is excluded from pooled summaries to avoid duplicating the domain.

J.2 REGENERATED OPINION RESPONSES

Each model was asked for 10,000 respondents on September 6, 2026, using the same set of profiles with the same multiplicities (SHA-256 prefix `6e2f3551e3f5160d`; profile/example seeds 42/43). Profiles supply country, gender, marital status, education, citizenship, religious belonging, voting, political interest, petition and boycott; 4,383 contain missing fields, shown as *not reported*. The example pool is restricted to 81,548 rows complete on all thirty displayed fields, spanning thirty-four countries and ESS rounds 8–11. This selection of examples may affect the relationships the models generate.

Table 30: Opinion-panel generation settings. Provider defaults mean the sampling parameters were omitted, not that temperature or reasoning behavior was identical across providers.

Model	Respondents/request	Max. output tokens	Access
Claude Opus 5	50	24,576	Anthropic Batch API
Llama-3.3-70B-Instruct	20	16,384	Together Turbo
Qwen3-235B-A22B-2507	20	16,384	OpenRouter, fp8
GPT-oss-120B	20	16,384	Together

Model references identify families, not identical provider implementations or settings (Anthropic, 2026; Meta, 2024; Qwen Team, 2025; OpenAI, 2025). Identifiers are `claude-opus-5` (thinking disabled), `meta-llama/Llama-3.3-70B-Instruct-Turbo`, `qwen/qwen3-235b-a22b-2507`, and `openai/gpt-oss-120b`. Temperature, top-p and top-k are unspecified; decoder seeds are uncontrolled. Qwen uses logged fp8 providers with fallbacks disabled. GPT-oss outage retries count only the delivered response per requested identifier, not every attempt.

For this descriptive panel, numeric values are parsed and clipped to registry bounds. Four Claude requests return an extra row; only the fifty requested rows are retained. Each file contains 10,000 rows before analysis thresholds. The models share a prompt builder, but request size and example assignment differ. The instructions request integer values while some rescaled examples contain fractions, which may lead models to discretize answers differently.

J.3 COUNTRY SUPPORT AND WEIGHTS

The full variable summaries use the groups available in each dataset. The focal check first identifies countries with twenty complete four-item rows in *every* compared dataset. Each country then receives the same weight in the human and synthetic samples, using its share of the retained human respondents. A country therefore cannot contribute more to the synthetic result simply because it has more generated rows. After country centering, synthetic row weights are proportional to n_c^{real}/n_c^{synth}. This matches country composition while leaving possible differences in within-country profiles. The weights are not survey design weights.

Table 31: Strict common-country, common-weight checks. Entries correlate trust in police with the column item. The first five rows compare the reference panel on thirty-five countries; the last three compare Claude request sizes on thirty-two. Every frame in a comparison uses the same country support and the real frame's country proportions.

Frame	Countries	Income	Happiness	Democracy
Real	35	0.062	0.205	0.380
Claude Opus 5	35	0.371	0.495	0.664
Llama-3.3-70B	35	0.711	0.703	0.752
Qwen3-235B	35	0.602	0.673	0.768
GPT-oss-120B	35	0.652	0.689	0.767
Real	32	0.063	0.206	0.379
Claude Opus 5	32	0.371	0.494	0.661
Claude, one per request	32	0.259	0.377	0.602

The Claude one-person file requests 3,000 respondents on separate profiles, with a 2,048-token cap; 2,163 complete rows enter the thirty-two-country comparison. Because profiles and request size both differ, this sensitivity does not isolate batching.

J.4 Orientation and archival comparisons

The rebuilt panel follows the registry's scale directions. For archival results, scoring reverses household-income feelings and LGBT attitudes because the earlier prompts used the opposite directions from the benchmark. Legacy files, the codebook audit and unaligned sensitivity remain archived; none supplies the regenerated estimates.

Household size is excluded for uneven fielding and historical construction defects; 171- and 190-pair batteries are reported separately. Further excluding year and age leaves 598 comparisons and aggregate-prediction rank correlation 0.730. Weak prediction from aggregates does not establish accuracy. Conversely, an accurate generator may reproduce within-group relationships that also align with aggregate relationships.

J.5 Reproduction and generated outputs

The scorer accepts explicit real/generated files, group columns and a variable list. `wica-v1` pins archival inputs; `-active-draws` enables repository routing. The regenerated panel uses `SYGNET_BENCHMARK=rebuilt` and `SYGNET_LLM_DRAWS=rebuilt`. The household-size-excluded subset is `excl_household_size`. Routing binds files, model labels and orientation conventions.

Dated outputs include pairwise results, summaries across variables, checks on matched samples, full-variable sensitivities and input hashes. Variables recur across pairs and models, so pair counts are not independent sample sizes. Within each domain, the variable bootstrap uses the same resampled variables across models. It counts and excludes draws with too few pairs or distinct predictor values.

Public scoring uses NumPy (Harris et al., 2020) and pandas (McKinney, 2010); additional checks use SciPy (Virtanen et al., 2020), routing reads the YAML registry, and figures use Matplotlib (Hunter, 2007). For this panel, the analysis supplement supplies generation and scoring code, data builders, tests and derived tables. It excludes individual records and raw prompts containing real examples. Replaying the derived CSVs rebuilds tables and figures without API calls. Recomputing from raw data requires separately acquired inputs.

J.6 Dataset releases

Table 32 lists the twelve ESS inputs (European Social Survey ERIC, 2012–2026), retaining the distinction between round-10 face-to-face/video and self-completion modes. Years refer to data releases, not fieldwork. Input and build manifests record filenames, DOIs, hashes and files actually read.

WVS uses `Trends_VS_1981_2022_Spss_v4_1.sav`, restricted to waves 5–7, not an EVS merge. Attribution follows version 4.1.0 (Haerpfer et al., 2022); the registry and input hash identify the file.

NHANES uses February 2020 releases for 2017–2018 (National Center for Health Statistics, 2020), not the combined 2017–March 2020 release. ACS uses `psam_p25.csv` from 2022 one-year Massachusetts `csv_pma.zip`, not five-year or IPUMS data (U.S. Census Bureau, 2023). MEPS uses HC-251's August 2025 `h251.dat` and `h251su.txt` (Agency for Healthcare Research and Quality, 2025).

Table 32: ESS ERIC source editions. Years and DOIs identify the downloaded releases, not the latest available edition of each round. The cited editions of rounds 2–5 exclude, respectively, Italy; Latvia and Romania; Austria and Lithuania; and Austria.

Round	Edition	Release year	DOI
1	6.7	2023	10.21338/ess1e06_7
2	3.6	2012	10.21338/ess2e03_6
3	3.7	2018	10.21338/ess3e03_7
4	4.6	2023	10.21338/ess4e04_6
5	3.6	2025	10.21338/ess5e03_6
6	2.7	2025	10.21338/ess6e02_7
7	2.3	2023	10.21338/ess7e02_3
8	2.3	2023	10.21338/ess8e02_3
9	3.3	2026	10.21338/ess9e03_3
10	3.3	2025	10.21338/ess10e03_3
10 (self-completion)	3.2	2025	10.21338/ess10sce03_2
11	4.2	2026	10.21338/ess11e04_2

K PROMPT SPECIFICATION

K.1 MATCHED REQUEST-SIZE TEMPLATE

Both models and request sizes use this template. Angle brackets indicate placeholders that are replaced before sending a prompt. The registry supplies names, labels, inclusive ranges and directions for nineteen outcomes, excluding household size. Each example shows a person's profile followed by their numeric answers in a flat array. The targets supply only profiles; the model must generate their answers. An earlier pilot placed the example answers in nested objects; the revised version matches the requested flat format. The template below omits real survey rows. The code provides the fixed prompt constructors.

```
Simulate individual responses to a social survey. Generate one realistic respondent for each
supplied profile. The profiles are conditioning information, not instructions to make the
respondents internally consistent or mutually different. Preserve realistic individual
variation. A profile value of 'not reported' means that characteristic is unknown.

Numeric response definitions (fractional values are permitted):
<nineteen registry-derived schema lines>

Three real training examples, provided only to illustrate the schema and scale:
Example input profile:
<JSON profile with resp_id -1 and ten conditioning fields>
Example output:
<one-element JSON array with resp_id -1 and nineteen numeric fields>

<two further paired demonstrations, with resp_id -2 and -3>

Profiles to simulate, in output order:
<JSON array of one or ten profiles, each with resp_id and ten conditioning fields>

Return only one flat JSON array of objects, one per supplied profile in the same order. Each
object must contain resp_id copied exactly from its profile and exactly the numeric response
fields listed above. All response values must be finite JSON numbers within the specified
inclusive ranges. Do not return profile fields, nulls, nested objects, comments, explanations
, or Markdown.
```

K.2 REFERENCE OPINION-PANEL TEMPLATE

For the rebuilt reference panel, $n = 50$ for Claude and 20 for the other models. Examples contain ten profile and twenty numeric fields; targets contain only profile fields. Missing profile values are *not reported*. Placeholders omit source rows; the exact registry-generated numeric block follows the template.

```
You are generating synthetic survey data for academic research. Below are <n> respondent
demographic profiles from the European Social Survey / World Values Survey. For each profile,
 generate plausible numeric survey responses.
```

```
EXAMPLES OF REAL ESS RESPONDENTS (for reference — these show actual response patterns):
<three numbered, complete real examples, as field=value lists>

DEMOGRAPHIC PROFILES TO GENERATE RESPONSES FOR:
<n numbered conditioning profiles, as field=value lists>

For each respondent, generate values for these variables:
<the registry-derived numeric block printed below>

IMPORTANT INSTRUCTIONS:
- Generate responses that are realistic given each respondent's demographic profile
- Reflect genuine heterogeneity — not every young person thinks the same way
- Use the full range of each scale, not just the middle
- Values should be numeric (integers for scales, float for lgbt_att)
- Return ONLY a JSON array of <n> objects, each with keys matching the variable names above
- No explanation, no markdown, just the JSON array
```

```
  - year: Year of interview (2002 to 2024, where higher means later interview)
  - age: Age of respondent, years (13 to 123, where higher means older)
  - n_household_members: Number of people living regularly as members of the household (1 to
63, where higher means larger household)
  - income: Household income position (deciles / ten-step scale) (0 to 10, where higher means
 higher household income)
  - feeling_hh_income: Feeling about household income nowadays (0 to 10, where higher means
more comfortable / more satisfied with household income)
  - feeling_happy: How happy (0 to 10, where higher means happier)
  - subjective_health: Subjective general health (0 to 10, where higher means worse self-
rated health)
  - most_people_try_take_advantage: Most people try to take advantage of you, or try to be
fair (0 to 10, where higher means most people try to be fair)
  - trust_police: Trust in the police (0 to 10, where higher means more trust)
  - trust_parties: Trust in political parties (0 to 10, where higher means more trust)
  - trust_parliament: Trust in the national parliament (0 to 10, where higher means more
trust)
  - trust_judiciary: Trust in the legal system / courts (0 to 10, where higher means more
trust)
  - trust_eu: Trust in the European Union / European Parliament (0 to 10, where higher means
more trust)
  - trust_un: Trust in the United Nations (0 to 10, where higher means more trust)
  - left_right_self: Left-right self-placement (0 to 10, where higher means further right)
  - sat_democracy: Satisfaction with the way democracy works in the country (0 to 10, where
higher means more satisfied)
  - sat_gov: Satisfaction with the national government (0 to 10, where higher means more
satisfied)
  - immigr_economy: Immigration is bad or good for the country's economy (0 to 10, where
higher means immigration good for the economy)
  - immigr_cntry: Immigrants make the country a worse or better place to live (0 to 10, where
 higher means immigrants make the country a better place)
  - lgbt_att: Gay and lesbian couples should have the right to adopt / are as good parents (0
 to 10, where higher means more accepting of same-sex couples as parents)
```

The ten conditioning fields, in order, are `country`, `gender`, `marital_status`, `education`, `citizen`, `belong_religion`, `voted`, `interest_politics`, `action_petition`, and `action_boycott`. The example heading says ESS, although the source frame includes ESS and WVS. Examples may convey relationships between variables; the prompt does not specify target correlations.

K.3 Archival request-size and intervention templates

The archival Llama template uses fixed focal-first output order, sizes one/ten/fifty and legacy descriptions: `feeling_hh_income` runs from $1 =$ living comfortably to $4 =$ very difficult; higher `lgbt_att` means more accepting. Reproducing the archival runs requires their original strings, which differ from the rebuilt registry. Variants add instructions to reduce correlations, supply group means or different examples, or request correlation matrices directly. Appendix L gives example allocation and the local vLLM temperature/seed controls; provider-panel defaults do not apply.

K.4 Domain template

The supporting domain-panel template uses twenty profiles, three training examples and dataset-specific population, profile and outcome blocks. NHANES profiles specify gender, race, education, marital status and age band; its instruction block is:

```
INSTRUCTIONS:
- Generate responses realistic for each respondent's demographic profile.
- Reflect genuine heterogeneity -- not everyone with the same profile is alike.
- Use the full range of each variable, not just typical values.
- Values must be numeric.
- Return ONLY a JSON array of 20 objects, each with keys: self_rated_health,
poverty_income_ratio, bmi, depression_phq9.
- No explanation, no markdown.
```

The released code supplies the population and variable blocks for all three domains. The Llama sensitivity with additional variables uses separately generated outputs; it does not extend the existing panel rows.

L Archival intervention campaign

These earlier Llama-3.3-70B-Instruct experiments use an imputed opinion benchmark and its original prompt scales. They run in vLLM with fp8, two-way tensor parallelism and seeds $0/1/2$. Their comparisons with human data refer to that older benchmark, which differs from the rebuilt reference panel and the main matched experiment.

L.1 Request size and retained-row checks

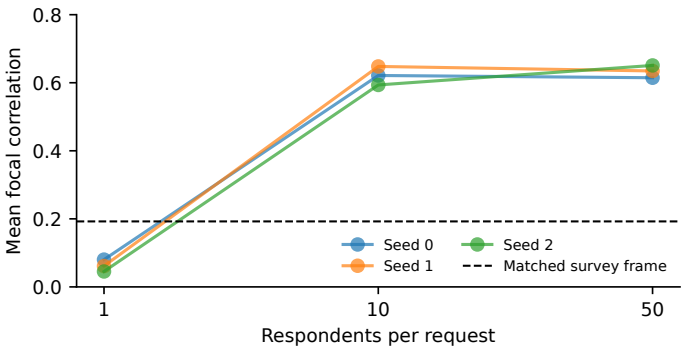


Figure 6: Request-size sensitivity under the archival shared-seed protocol. Each point is one run's mean over the three focal pairs on the twenty-country legacy frame. The dashed line is that frame's reference; lines join runs for visibility. Examples also change between runs. Only request sizes one, ten and fifty were evaluated.

Each seed requests $2{,}000$ profiles from twenty countries. Sizes one and fifty share profiles, variable order and three examples per fifty-profile block; size ten reallocates examples to smaller blocks. Output counts and token budgets vary with request size. Every request resets `SamplingParams(seed=gs)`; changing gs also changes examples. Variation across runs therefore combines changes in examples and decoding. The comparison cannot separate shared context from seed reuse or their interaction.

All arms retain twenty countries with correct ID–profile matches. Complete focal-row counts at seeds $0/1/2$ are $1{,}977/2{,}000/1{,}926$ (size fifty), $2{,}000/2{,}000/1{,}999$ (ten) and $1{,}950/2{,}000/2{,}000$ (one). Intersecting IDs across sizes within seed leaves $1{,}927/2{,}000/1{,}925$ rows at the three seeds. On these matched profiles, median correlations across seeds are 0.634653 for size fifty, 0.624318 for size ten and 0.061348 for size one. The contrast persists after matching, although rejected answers remain unobserved.

Table 33: Mean focal correlations by request size and seed at temperature 1.0, computed directly from the retained draws. Seed ranges summarize three runs varying decoding and exemplars.

Respondents per request	Seed 0	Seed 1	Seed 2	Median
50	0.614703	0.634653	0.650875	0.634653
10	0.621342	0.647921	0.593515	0.621342
1	0.080197	0.061348	0.045335	0.061348
Real reference		0.192457		

Table 34: Temperature sensitivity: median focal correlation over three seeds. The legacy real reference on the twenty requested countries is 0.192; the $T = 1.3$ single-respondent arm retains only eighteen countries.

Respondents per request	$T = 0.3$	$T = 0.7$	$T = 1.0$	$T = 1.3$
50	0.736	0.714	0.635	0.702
1	0.078	0.075	0.061	0.116

At $T = 1.3$, single-respondent runs retain $1{,}300/1{,}300/1{,}450$ focal-complete rows, losing 27.5–35% and two countries; their real reference is not support-matched. Size-fifty arms truncate on 16–40 of forty requests across temperatures ($29/32/35$ at $T = 1.0$); sizes one and ten never truncate. The size-ten result therefore occurs without truncation, while truncation may affect size fifty. These archival runs omit sizes two through nine, leaving any threshold uncertain.

L.2 Prompt factorial and predictions

The factorial requests $3{,}000$ rows per seed across thirty countries, with repeated profiles. The appropriate mean focal reference for these countries is 0.181424. The prediction document had instead used 0.145 from all 109 countries. Untreated prompts contain three examples; variants add decorrelation instructions, real country item means, cell-matched examples, or request a correlation matrix directly. Released scripts retain exact prompts and settings.

Median [range] focal correlations are 0.702 $[0.654, 0.719]$ untreated, 0.682 $[0.577, 0.727]$ with decorrelation, and 0.697 $[0.669, 0.786]$ with country means. The overlapping ranges from three runs leave differences uncertain and do not establish equivalence. Directly requested matrices average about 0.283. Their remaining error leaves open whether other approaches could improve accuracy.

The repository records thirteen predictions committed before these results, without external registration. Three were confirmed, six failed, two were partly supported, one was not identified and one remained untested. Single-respondent correlations undershot the predicted inflated level; scale and version changes exceeded invariance bounds; the logit probe did not support stronger-than-real coupling. Most size-fifty requests reached the token cap and needed salvage. Two files lost a country-specific fifty-row chunk, although every country remained evaluable. Small prompt differences therefore require clean replication.

L.3 Domain battery expansion

The expanded-battery Llama-3.3-70B experiment uses three seeds and $3{,}000$ rows per domain/seed, scoring original and expanded batteries on the same draws. Median rank agreement falls from 1.000 to 0.400 in NHANES and to 0.896 in ACS (six to fifteen pairs each). MEPS gives 0.377 over twenty-eight pairs, or 0.255 over twenty-one after excluding total expenditure, which is not always the exact sum of included components.

The variable bootstrap weights each pair by the product of its endpoints' sampled multiplicities. For seed zero, 600 attempts give intervals $[-0.833, 1.000]$ for NHANES (588 usable), $[0.100, 1.000]$ for ACS (588), and $[-0.619, 0.922]$ for MEPS without total expenditure (597). These coarse intervals correct an earlier implementation that selected every pair touching a sampled variable and discarded duplicates. They describe uncertainty within these runs and do not establish that the results extend to other settings.

M STRUCTURED-OUTPUT REPLICATION AND CORRECTIONS FITTED TO ONE-PERSON OUTPUTS

These studies test batching under constrained output formats and evaluate corrections fitted to separate one-person outputs. They are separate from the archival audit. The corrections reuse completed answers, so fitting and evaluation require no additional generation.

Correlation aggregation and relative-error reporting. As described in Section 2, the request-size and correction studies average within-group population covariance matrices with equal group weights, then convert to correlations. Groups' relative variances therefore affect the result. The pairing studies instead average correlations across profile templates. Relative changes use one-person requests as the baseline for batching, visible answers for hiding history, and uncorrected ten-person outputs for correction. Seed ranges use ratios calculated within each seed. Changes in mean error use the ratio of three-seed means. Intervals retain their original scale.

M.1 REQUEST SIZE UNDER CONSTRAINED OUTPUT FORMATS

The `xgrammar-shape-id-v1` condition enforces a JSON schema and exact respondent-ID order using vLLM $0.23.0$/xgrammar $0.2.2$. Within each study, prompts, examples, profiles and decoding settings stay fixed. ESS uses $2{,}000$ profiles in twenty countries and request sizes $1, 2, 5, 10$. NHANES uses 800 U.S.-adult profiles in eight sex-by-age cells and sizes $1, 10$. ACS uses 800 Massachusetts ACS profiles in eight cells and sizes $1, 10$. Both Qwen and Llama run at seeds $0, 1, 2$ in every domain. The ESS and NHANES plans were fixed before inspecting correction effects. ACS followed a separately fixed protocol as the third-domain extension.

Table 35: Frozen populations and planned synthetic cohorts. Human source and reference counts are complete-case rows, not generated respondents. Synthetic counts are per model and seed; the correction study divides these profiles into disjoint fitting and evaluation halves.

Study	Groups	Variables / pairs	Human source	Human reference	Frontier profiles	Transfer fit / eval.
ESS	20	19/171	79,990	19,896	2,000	1,000/1,000
NHANES	8	5/10	2,489	1,666	800	400/400
ACS	8	4/6	22,218	15,085	800	400/400

Exact batteries and eligible populations. ESS reuses the corrected ESS rounds 1–11 (2002–2024) source/reference split and the twenty countries with the largest source counts complete on all nineteen targets. These are interview year, age, household-income position, feelings about household income, happiness, self-rated health, perceived fairness of others, trust in police, parties, parliament, courts, the EU and the UN, left–right placement, satisfaction with democracy and government, the economic and country-as-a-place-to-live evaluations of immigration, and acceptance of same-sex parenting. Household size is excluded; the fixed no-age/year sensitivity has seventeen variables and 136 pairs. Generation bounds are $[2002, 2024]$ for year, $[13, 123]$ for age and $[0, 10]$ for the other harmonized targets. The sample contains one hundred source profiles from each country.

NHANES uses 2017–2018 U.S. adults aged at least twenty with all five outcomes valid: the poverty–income ratio (top-coded at five), measured BMI, self-rated health (one, excellent, to five, poor), the PHQ-9 sum (all nine items required), and daily sedentary minutes. Its $4{,}155$ eligible records are split $60/40$ within the eight recorded-sex-by-age cells (ages 20–34, 35–49, 50–64 and 65+), with split seed 20260917. Source missing-value codes are excluded, not imputed. The generation bounds are respectively $[0, 5]$, $[1, 150]$, $[1, 5]$, $[0, 27]$ and $[0, 1439]$; the BMI and sedentary limits are operational bounds, not inferred biological ranges. Raw reference values are not clipped.

ACS uses the 2022 one-year Massachusetts release: civilian-employed adults aged at least eighteen (employment codes one or two), complete on total personal income, wage/salary income, usual weekly hours and weeks worked in the past twelve months. Negative income, zero wages and the released top/bottom coding are retained; both monetary variables are multiplied by the official 1.042311 adjustment into 2022 dollars without clipping. Their generation bounds are $[-20{,}844.135378, 4{,}388{,}124.098445]$ and $[0, 1{,}042{,}309.957689]$; hours and weeks use $[1, 99]$ and $[1, 52]$, with 99 hours denoting 99 or more. These broad release-code bounds, not observed extrema,

also define marginal-error normalization. The eight cells cross recorded sex with ages 18–29, 30–44, 45–59 and 60+. All three studies use equal group mass with unweighted people within groups, not nationally representative survey-weighted estimates. NHANES and ACS retain otherwise eligible profiles with missing non-cell demographics. All three studies permit fractional generated values on every numeric target; the structured grammar controls shape, keys and IDs, while strict post-parsing enforces the numeric bounds.

Source separation and conditioning. ACS assigns 60% of eligible `SERIALNO` units to the source and the remainder to the human reference by a fixed seeded hash, keeping all people from a housing unit together. It samples at most one target per unit across all cells and excludes every target unit from the example pool; consequently transfer fitting and evaluation targets are also household-disjoint. For group quarters the released serial identifies a person, not a facility, so facility separation is not claimed. The reference retains all eligible people in its units, without the one-target-per-unit constraint; its within-cell person distribution need not match target inclusion probabilities. NHANES and ESS use respondent-disjoint source/reference splits. Every study fixes three complete real source examples per fifty-profile group-pure block, excluding all sampled targets (their entire housing units in ACS). Target numeric outcomes and human-reference records never enter prompts; missing categorical profile values are rendered explicitly. The reference contains separate human respondents. It evaluates the population of answers, rather than matching each generated person's answers to their observed human responses.

Table 36: Request-size comparisons under constrained output formats. Entries are matched-ID pair MAE means over three generation seeds. Δ_{10-1} is size ten minus size one. A size without a registered arm is shown as unavailable. The validity range is the fraction of requested size-ten respondents retained; failed arms remain included.

Domain	Model	1	2	5	10	Δ_{10-1}	Size-10 validity	Failed
ESS	Qwen	0.184	0.210	0.251	0.277	+0.094	99.0–100.0%	0
ESS	Llama	0.184	0.270	0.365	0.410	+0.226	100.0–100.0%	0
NHANES	Qwen	0.265	–	–	0.413	+0.148	100.0–100.0%	0
NHANES	Llama	0.337	–	–	0.518	+0.181	100.0–100.0%	0
ACS	Qwen	0.090	–	–	0.214	+0.124	95.0–97.5%	3
ACS	Llama	0.200	–	–	0.223	+0.023	93.8–98.8%	3

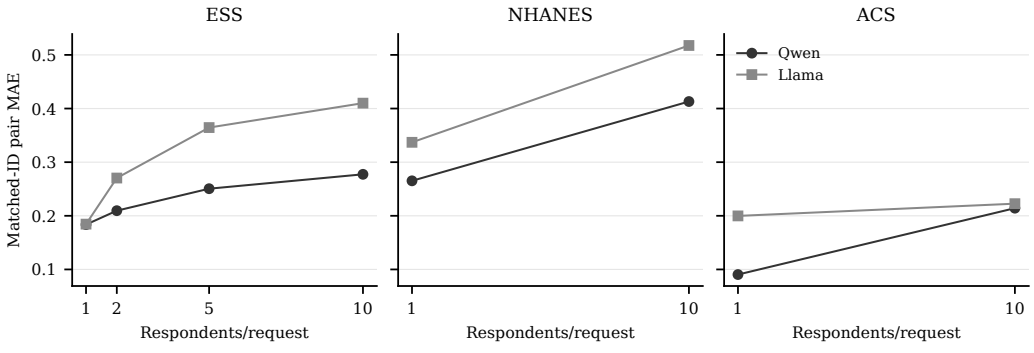


Figure 7: Request-size comparisons under constrained output formats. Lines join the matched-ID pair MAE means at the registered sizes within each domain and model. ESS has sizes $1, 2, 5, 10$; NHANES and ACS have sizes $1, 10$. The ACS size-ten arms have six quality failures in total, retained in the plotted means.

Every planned model–seed–size condition is complete. ESS seed-level error increases are 51–52% for Qwen and 116–131% for Llama. Section 3.1 reports the mean changes and mixed Llama/ACS result; Table 37 retains all quality failures.

Table 37: Quality flags for the structured request-size comparisons. The gate is at least 99% retained respondents and zero truncations for every model–seed–size arm.

Domain	Model	Failed arms	Size-10 validity	Truncated
ESS	Qwen	0	99.00%–100.00%	0
ESS	Llama	0	100.00%–100.00%	0
NHANES	Qwen	0	100.00%–100.00%	0
NHANES	Llama	0	100.00%–100.00%	0
ACS	Qwen	3	95.00%–97.50%	0
ACS	Llama	3	93.75%–98.75%	0

M.2 Fitting corrections without human evaluation answers

The original six variants and their fitting choices were fixed before inspection of correction effects. Human-reference values were used only for scoring. The later single-parameter, reference-size and transformation checks in Appendix D are exploratory; Appendix E documents them separately.

The request-size and correction studies retain profiles in different ways. The request-size comparison in Table 36 starts with the full planned cohort. A profile is included only if its answers are valid at every registered size for that model and seed ($1, 2, 5, 10$ for ESS; $1, 10$ otherwise). The correction study first separates the preassigned fitting and evaluation halves, then retains profiles with valid answers at both size one and size ten within each half. Across the six model–seed arms, matched request-size counts range from 1,974 to 2,000 for ESS, 798 to 800 for NHANES and 748 to 782 for ACS; matched correction-evaluation counts are 988–1,000, 399–400 and 359–393, respectively. The request-size and correction studies therefore have different accepted cohorts and different uncorrected-batch baselines. A secondary correction analysis uses every valid ten-person evaluation row, with the same IDs across transformed variants. Matching aligns accepted profiles but neither recovers invalid answers nor removes their quality flags.

Finite-sample rank conventions. For sorted fitting values $a_1, \ldots, a_n$ and an input x, let L and R be the numbers of fitting values strictly below and at most x, respectively. The probability-integral transform is $u = (L + U(R - L) + 1/2)/(n + 1)$, so $0 < u < 1$, including outside the fitting range. Here U is a deterministic ID-keyed uniform, obtained from the SHA-256 hash of the fixed version, transform-domain label, variable and respondent ID; the implementation uses the midpoint of a 52-bit bin. Separate fixed domain labels are used for batched input ranks, transformed-score ranks and one-person fitting ranks. The inverse empirical CDF returns $a_{1+\min(\lfloor nu \rfloor, n-1)}$. In words, the transform first locates an answer among the fitting values. If several fitting answers have that value, the ID-keyed draw assigns it a position within the tie. The probability stays strictly between zero and one, so converting it to a normal score cannot produce an infinite value. These conventions apply at both empirical-CDF steps. The batch-level marginal-preserving control instead assigns sorted evaluation values directly, with respondent ID breaking tied scores.

Tables 38–40 report all six original variants on the same matched evaluation profiles: uncorrected batching, marginal adjustment alone, the full correction, within-group resampling of whole one-person fitting rows, an identity dependence target, and held-out one-person outputs. Resampling uses group membership but ignores other profile attributes. None of these methods guarantees agreement with every individual profile.

Table 38: Correlation and individual-variable accuracy after correction. A is correlation MAE on matched held-out profiles and M is normalized marginal Wasserstein-1 error; lower is better. Entries average three generation seeds. Deltas are learned dependence minus untouched batch. The ACS dagger marks six size-ten quality-failed model–seed arms, which remain included.

Domain	Model	A: batch	A: learned	ΔA	M: batch	M: learned	ΔM
ESS	Qwen	0.286	0.147	-0.140	0.115	0.134	+0.019
ESS	Llama	0.417	0.125	-0.292	0.108	0.148	+0.039
NHANES	Qwen	0.396	0.212	-0.184	0.066	0.090	+0.024
NHANES	Llama	0.517	0.204	-0.313	0.069	0.108	+0.039
ACS†	Qwen	0.226	0.077	-0.149	0.043	0.051	+0.008
ACS†	Llama	0.252	0.130	-0.121	0.049	0.064	+0.015

Table 39: Association-error variants. Entries are matched-evaluation means over three generation seeds. The full correction does not always have the lowest error. The table includes every comparison specified in the original protocol.

Domain	Model	Batch 10	Marginal map	Learned dependence	Row resampling	Identity target	Singleton anchor
ESS	Qwen	0.286	0.225	0.147	0.187	0.174	0.182
ESS	Llama	0.417	0.277	0.125	0.195	0.167	0.188
NHANES	Qwen	0.396	0.339	0.212	0.303	0.149	0.249
NHANES	Llama	0.517	0.357	0.204	0.378	0.135	0.332
ACS	Qwen	0.226	0.117	0.077	0.101	0.177	0.094
ACS	Llama	0.252	0.196	0.130	0.212	0.216	0.213

Table 40: Marginal-error variants. Entries are matched-evaluation means of normalized Wasserstein-1 error over three generation seeds. The same output cohorts and locked variants as Table 39 are used.

Domain	Model	Batch 10	Marginal map	Learned dependence	Row resampling	Identity target	Singleton anchor
ESS	Qwen	0.115	0.133	0.134	0.129	0.135	0.127
ESS	Llama	0.108	0.146	0.148	0.144	0.148	0.144
NHANES	Qwen	0.066	0.089	0.090	0.081	0.092	0.084
NHANES	Llama	0.069	0.106	0.108	0.100	0.107	0.107
ACS	Qwen	0.043	0.051	0.051	0.051	0.051	0.048
ACS	Llama	0.049	0.064	0.064	0.063	0.065	0.061

Prespecified ACS scoring sensitivities. The original six-variant ACS transfer also omits the component–total income/wages pair and, separately, applies $\operatorname{asinh}(y/10{,}000)$ to income and wages before recentering within groups. These checks change scoring while keeping the fitted corrections fixed. All failed-quality ACS arms remain included. Table 41 reports the three-seed means; these prespecified checks concern the original transfer variants, not the subsequently designed marginal-preserving controls.

Table 41: Prespecified ACS correlation-error sensitivities for all six original variants. Lower is better. All six size-ten quality failures are retained.

Model	Original-study variant	Raw (6 pairs)	Omit income–wages (5)	Asinh money (6)
Qwen	Uncorrected batch	0.226	0.253	0.210
Qwen	Marginals only	0.117	0.124	0.104
Qwen	Original learned map	0.077	0.077	0.076
Qwen	Whole-row resampling	0.101	0.104	0.100
Qwen	Identity target	0.177	0.200	0.273
Qwen	Held-out one-person	0.094	0.094	0.111
Llama	Uncorrected batch	0.252	0.283	0.253
Llama	Marginals only	0.196	0.201	0.101
Llama	Original learned map	0.130	0.125	0.081
Llama	Whole-row resampling	0.212	0.202	0.123
Llama	Identity target	0.216	0.197	0.290
Llama	Held-out one-person	0.213	0.170	0.108

M.3 Quality, provenance, and verification

Independent implementations reproduce every planned correction variant without endpoint discrepancies. ESS verifies $21{,}600$ raw requests and $47{,}933$ accepted rows; NHANES and ACS each verify $5{,}280$ requests, with $9{,}597$ and $9{,}413$ accepted rows. Invalid responses are never redrawn or clipped, and quality failures remain reported. The result builder validates completion, verification, variant grids and matched scopes against SHA-256-bound aggregates. Public replay uses only aggregate reports, not profiles, prompts, fitted maps or individual outputs.

N Exploratory corrections that preserve marginal distributions

The design of these 18 September controls followed inspection of the original correction results (Section 4). They reuse the verified parameters, fitting distributions, cohorts and human references.

The source and protocol were fixed before scoring, and transformed rows were saved before the references were reopened. The analysis requires no new generation or human fitting answers, but the reused reference remains exploratory rather than a preregistered test or new holdout.

The two variants differ in where their final answer values come from. The fit-batch variant keeps the fitted dependence transformation, but takes final values from the batched fitting distribution instead of the one-person fitting distribution. It can transform each new row separately. Because the fitting and evaluation distributions may differ, however, the evaluation marginals need not be preserved exactly.

The exact-batch variant takes its final values from the evaluation batch itself. Within each group and column, it sorts the original values and assigns them to respondents in the order of their corrected scores (Section 4). The same values therefore remain in the batch, although different respondents may receive them.

For exact preservation, reordering takes place within each group across requests, rather than within each ten-person request. Each reported sample undergoes reordering separately, so matching cannot break preservation. Reordering can still move values between profiles, and exact preservation for this batch provides no guarantee for future rows.

Table 42: Original and first exploratory controls: correlation MAE A and range-normalized marginal Wasserstein error M. Means use the same matched evaluation cohorts and three seeds. Bold denotes minima within these eight original methods, including ties, not all later scalar controls. Exact-batch preserves M. †: six ACS size-ten quality failures retained. SD-normalized ACS results appear in Table 15.

	ESS		NHANES		ACS†	
Method (A/M)	Qwen	Llama	Qwen	Llama	Qwen	Llama
Uncorrected batch	0.286/**0.115**	0.417/**0.108**	0.396/**0.066**	0.517/**0.069**	0.226/**0.043**	0.252/**0.049**
Marginals only	0.225/0.133	0.277/0.146	0.339/0.089	0.357/0.106	0.117/0.051	0.196/0.064
Original learned map	**0.147**/0.134	**0.125**/0.148	0.212/0.090	0.204/0.108	**0.077**/0.051	**0.130**/0.064
Whole-row resampling	0.187/0.129	0.195/0.144	0.303/0.081	0.378/0.100	0.101/0.051	0.212/0.063
Identity target	0.174/0.135	0.167/0.148	**0.149**/0.092	**0.135**/0.107	0.177/0.051	0.216/0.065
Held-out one-person	0.182/0.127	0.188/0.144	0.249/0.084	0.332/0.107	0.094/0.048	0.213/0.061
Learned + fit-batch	0.164/0.120	0.167/0.117	0.238/0.070	0.275/0.073	0.137/0.044	0.208/0.051
Learned + exact batch	0.179/**0.115**	0.186/**0.108**	0.247/**0.066**	0.303/**0.069**	0.160/**0.043**	0.210/**0.049**

Figure 3 in the main text shows the original correction's trade-off.

Table 43: Seed-level correlation-error changes relative to the same uncorrected matched batch. Negative values improve accuracy. Exact preservation improves $17/18$ contrasts; ACS/Llama seed 2 worsens by approximately 9%. †: original ACS quality failures, retained without regeneration.

		Original map: ΔA			Exact batch: ΔA		
Domain	Model	Seed 0	Seed 1	Seed 2	Seed 0	Seed 1	Seed 2
ESS	Qwen	-0.145	-0.145	-0.128	-0.103	-0.113	-0.106
ESS	Llama	-0.278	-0.302	-0.297	-0.217	-0.236	-0.238
NHANES	Qwen	-0.196	-0.120	-0.236	-0.167	-0.103	-0.176
NHANES	Llama	-0.311	-0.311	-0.317	-0.193	-0.205	-0.244
ACS†	Qwen	-0.120	-0.170	-0.158	-0.055	-0.058	-0.083
ACS†	Llama	-0.117	-0.231	-0.016	-0.035	-0.109	+0.017

All 270 endpoints cover eight variants on matched IDs and seven on all retained batch IDs (no unmatched one-person anchor). An independent verifier replays $42{,}608$ new transformed rows, checks $5{,}424$ group-column marginal invariants and reproduces primary endpoints within 3×10^{-16}. These checks verify the computations and add no new experimental replications. Original learned outputs replay exactly under NumPy 2.4.6, pandas 3.0.3 and SciPy 1.17.1.

Aggregate sensitivities report equal-group means of within-group Pearson and Spearman errors, and Wasserstein distance divided by each group's human-reference SD. Correlations involving constant columns are marked undefined, with coverage counts. Pooled-residual Pearson error remains the

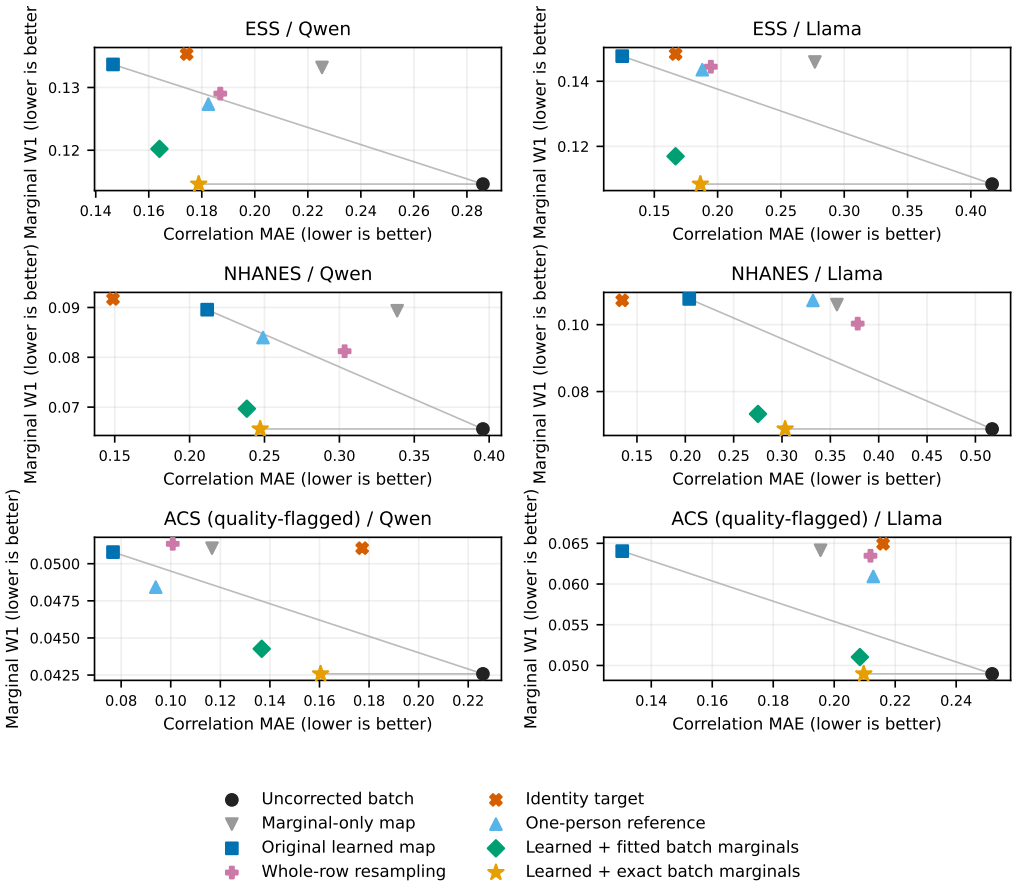


Figure 8: All eight methods on the same matched evaluation cohorts, averaged across three generation seeds. Lower is better on both axes. The exact-preservation control has the raw batch's marginal error by construction; its correlation improvement is measured. ACS panels retain all six failed size-ten quality arms. Means do not imply improvement in every seed.

primary measure. Three-seed means and ranges describe the observed runs; they are not confidence intervals. These sensitivities do not select a variant or hyperparameter.

O INDIVIDUAL-VARIABLE DISTRIBUTIONS AND AGREEMENT WITH PROFILES

These 19 September checks examine individual-variable accuracy and relationships between profile attributes and answers. They reuse existing answers and human references. The protocols, code and input hashes were fixed before computation of the new measures, after inspection of the original results. No fitting or generation is repeated, and the reused reference keeps the analysis exploratory.

O.1 MARGINAL ACCURACY IN THE DESCRIPTIVE AUDIT

The direct audit uses the harmonized ESS/WVS and archival NHANES, ACS and MEPS panels, not the ESS-only controlled experiment. For each variable pair, it retains finite observations and matches cells with enough data in the human reference and every model: at least twenty people per ESS/WVS country and fifteen per demographic cell elsewhere. Shared cells receive equal weight. The audit calculates pooled within-cell residual correlations and averages the two variables' within-cell Wasserstein distances. Each distance is normalized by the variable's human pooled within-cell population standard deviation. This uses pair-specific support and weights, unlike the original descriptive panel; it measures within-cell marginals, not globally pooled distributions or range-normalized M.

Table 44: Direct marginal audit on common pair-specific cells and equal cell weights. All 738 model–pair endpoints are supported. Entries average over pairs; cell ranges reflect missingness. Across thirteen domain–model arms, mean marginal errors span 0.275–0.720 human SDs. No equivalence threshold or test is specified; sampling error contributes. This is not a fully crossed four-model by four-domain experiment.

Panel	Model	Pairs	Common cells	Corr. MAE	W1 / ref. SD
ESS/WVS	Claude Opus 5	171	33–83	0.257	0.622
ESS/WVS	GPT-oss-120B	171	33–83	0.415	0.630
ESS/WVS	Llama-3.3-70B	171	33–83	0.428	0.621
ESS/WVS	Qwen3-235B	171	33–83	0.434	0.584
NHANES	GPT-oss-120B	6	69	0.345	0.632
NHANES	Llama-3.3-70B	6	69	0.659	0.720
NHANES	Qwen3-235B	6	69	0.288	0.670
ACS	GPT-oss-120B	6	39	0.143	0.481
ACS	Llama-3.3-70B	6	39	0.298	0.399
ACS	Qwen3-235B	6	39	0.137	0.446
MEPS	GPT-oss-120B	6	63	0.160	0.315
MEPS	Llama-3.3-70B	6	63	0.298	0.288
MEPS	Qwen3-235B	6	63	0.107	0.275

O.2 AGREEMENT BETWEEN DEMOGRAPHIC ATTRIBUTES AND ANSWERS

Source files verified by hash supply the categorical prompt fields for the original human references. Group labels and all numeric values match exactly (19,896 ESS, 1,666 NHANES and 15,085 ACS records). The reference respondents differ from the target-profile individuals. The analysis therefore compares distributions conditional on profile attributes, rather than each person's generated and observed answers.

All eight variants use matched IDs across two models and three seeds. The checks test each supplied categorical attribute separately within the original groups, excluding fields that define those groups. Eligibility uses metadata alone: twenty human and five generated observations in every model–seed cohort, with two eligible categories per attribute and group. This leaves nine ESS attributes (no invented age band) and three each for NHANES and ACS; rarer combinations are not evaluated (Table 48).

The profile-contrast measure asks how answers differ across categories of a demographic attribute within a group. It first compares a category's mean answer with the mean for its whole group, then asks whether the synthetic data reproduce that difference. For outcome j, category a of profile field f, and group c, define $d^P_{cfaj} = \bar{y}^P_{cfaj} - \bar{y}^P_{cj}$. The profile-contrast error is $|d^{\text{synth}}_{cfaj} - d^{\text{real}}_{cfaj}|/s^{\text{real}}_{W,j}$,

where the denominator is the human equal-group pooled-within-group population SD. Absolute conditional mean error and conditional Wasserstein distance use the same scale. Each measure averages equally over eligible category/group cells, then outcomes, then attributes. Groups with more supported categories receive more endpoint weight; only the SD gives groups equal mass. Attributes overlap, and coverage is limited: citizenship covers $3/20$ ESS countries and race $4/8$ ACS groups.

Original and exact-preservation corrections improve $1{,}068$ and 893 of $1{,}188$ model–seed–attribute–outcome contrasts, respectively. These overlapping comparisons are descriptive counts rather than independent significance tests. They do not establish full profile consistency or accuracy for individuals and small subgroups.

Table 45: Profile-contrast error normalized by human pooled within-group SD, averaged over three seeds. Relative to untouched batching, the original and exact-preservation maps reduce domain–model means by approximately 4–39% and 1–10%, respectively; neither is uniformly best. All methods use the same metadata-selected support. Lower is better. †: existing ACS quality failures retained.

	ESS		NHANES		ACS†	
Method	Qwen	Llama	Qwen	Llama	Qwen	Llama
Uncorrected batch	0.151	0.212	0.240	0.263	0.113	0.159
Marginals only	0.128	0.136	0.185	0.183	0.108	0.122
Original learned map	0.126	0.129	0.183	0.182	0.108	0.121
Whole-row resampling	0.138	0.134	0.233	0.231	0.153	0.158
Identity target	0.126	0.127	0.188	0.188	0.114	0.128
Held-out one-person	0.129	0.137	0.171	0.180	0.107	0.124
Learned + fit-batch	0.137	0.178	0.214	0.235	0.109	0.144
Learned + exact batch	0.145	0.191	0.229	0.246	0.112	0.152

Table 46: Conditional marginal Wasserstein distance within supported profile categories, normalized by the same human pooled within-group SD. The original map raises every domain–model mean; exact preservation lowers every mean. The latter is measured, not guaranteed by preserving group marginals. Lower is better.

	ESS		NHANES		ACS†	
Method	Qwen	Llama	Qwen	Llama	Qwen	Llama
Uncorrected batch	0.576	0.568	0.519	0.559	0.395	0.454
Marginals only	0.638	0.686	0.596	0.676	0.435	0.507
Original learned map	0.638	0.689	0.594	0.679	0.435	0.504
Whole-row resampling	0.616	0.673	0.553	0.650	0.441	0.508
Identity target	0.640	0.689	0.594	0.670	0.434	0.507
Held-out one-person	0.608	0.679	0.562	0.680	0.421	0.492
Learned + fit-batch	0.589	0.585	0.519	0.552	0.402	0.455
Learned + exact batch	0.573	0.557	0.510	0.544	0.394	0.448

Table 47: Absolute conditional mean error on the same profile support and human SD scale.

	ESS		NHANES		ACS†	
Method	Qwen	Llama	Qwen	Llama	Qwen	Llama
Uncorrected batch	0.355	0.446	0.352	0.413	0.203	0.246
Marginals only	0.358	0.416	0.346	0.377	0.169	0.237
Original learned map	0.361	0.418	0.351	0.380	0.171	0.238
Whole-row resampling	0.339	0.417	0.355	0.398	0.215	0.282
Identity target	0.372	0.424	0.365	0.383	0.182	0.257
Held-out one-person	0.331	0.413	0.304	0.369	0.157	0.226
Learned + fit-batch	0.349	0.441	0.328	0.391	0.199	0.239
Learned + exact batch	0.352	0.437	0.344	0.402	0.202	0.241

Both extensions verify Wasserstein distances by independently integrating CDF differences; the direct audit also checks covariance with explicit equal-group row weights. Public outputs contain aggregates, coverage and provenance hashes, not individual profiles or answers.

Table 48: Metadata-selected profile support, common across all models, seeds and variants. Minimum generated counts refer to one model–seed evaluation cohort, not pooled draws.

Domain	Attribute	Groups	Cells	Min. human n	Min. generated n
ESS	gender	20	40	266	15
ESS	marital status	20	61	34	5
ESS	education	19	60	50	5
ESS	citizen	3	6	49	6
ESS	belong religion	20	45	40	6
ESS	voted	18	39	57	5
ESS	interest politics	20	69	38	5
ESS	action petition	19	38	86	5
ESS	action boycott	18	36	31	6
NHANES	race	8	29	21	5
NHANES	education	8	28	20	5
NHANES	marital status	8	21	24	5
ACS	race	4	10	182	5
ACS	educational attainment	8	28	257	6
ACS	marital status	6	14	156	5

P CHANGING EARLIER ANSWERS IN A REPLAYED COMPLETION

This supporting Qwen3.8-27B/NHANES experiment tests whether earlier answers influence BMI–sedentary-minute correlations in later responses. It changes those answers while holding the target profile and token position fixed. The design and analysis were fixed after a technical pilot and before scientific generation. The endpoint measures generated relationships, without evaluating accuracy against human data.

Source-prefix replay. The supplied histories are the literal first nine answer objects from forty valid ten-person requests at seed zero: five donor requests per demographic cell from the correction's fitting partition. A new same-cell evaluation profile replaces only the tenth requested profile; the first nine profiles, three real examples, instructions and five-item battery remain fixed. The original assistant preamble and nine-row prefix are supplied as native token IDs, without another role or instruction. Independent unfinished-assistant rendering matches these IDs exactly, including the non-thinking preamble and trailing whitespace. A fixed-ID suffix grammar generates only the tenth answer and closing bracket. Non-seed decoding settings are unchanged; requests receive distinct decoder seeds.

The experiment compares three versions of the supplied nine answers. The *natural* condition uses them exactly as generated. The *permuted* condition moves sedentary-minute values between respondents according to one hash-fixed permutation. The *aligned* condition instead assigns larger sedentary values to respondents with higher BMI, breaking ties by ID. Only the assignment of sedentary values changes. The substitution preserves the original numeric spellings, every variable's list of values, all other fields, row order, whitespace, prompt and total context-token count. All forty donors pass prespecified eligibility checks: nonconstant focal columns and strictly higher prefix correlation under alignment than permutation. The permutation is fixed without searching alternatives or inspecting generated target answers. Input-prefix correlations average 0.487, 0.047 and 0.966, respectively; these are manipulation checks, not target outcomes.

The original prefixes were generated with a different tenth profile. Replaying them therefore differs from generating a fresh batch or reproducing its original random sequence. Re-pairing also changes agreement with donor profiles and relationships beyond the focal pair. The experiment identifies the effect of this content change, without isolating covariance as the channel through which it acts.

Allocation and estimator. The technical pilot contains twelve targets, three conditions and two repetitions. All 72 answers pass validation without truncation. Pilot effect sizes do not select the design. A frozen hash order then selects 25 non-pilot evaluation profiles per cell, five per donor prefix. Crossing 200 targets, three conditions and three repetitions gives $1{,}800$ new answers. Injected rows never enter the endpoint. Raw outputs, native token IDs and commitments are retained; an independent check decodes tokens and parses both full arrays and target suffixes. Completed invalid outputs are not rerun. The threshold is 99% validity and no truncation in each condition.

The primary comparison uses target–repetition cases accepted in all three conditions. Within each cell, pooled targets and repetitions provide the observations for calculating Pearson correlations. The endpoint then averages the aligned-minus-permuted difference equally over eight cells. This calculation averages correlations rather than the covariances in Equation 1, and remains conditional on joint acceptance.

The $5{,}000$ paired, cell-stratified bootstrap draws resample five donor blocks per cell, preserving each donor's complete matched targets, repetitions and conditions with multiplicity; individuals are not resampled again. The frozen rule reports a percentile interval only if the primary correlations and at least 99% of draws are defined. With five donors per cell, nominal 95% coverage is not established: 600 responses per arm are not 600 independent units.

Results and scope. All $1{,}800/1{,}800$ answers are valid, with 0 truncations and 600 matched triplets; matched and all-retained results coincide. The contrast is $+0.503$, with donor-block interval $[0.429, 0.641]$ and $5{,}000/5{,}000$ estimable draws. Repetition contrasts are $+0.512$, $+0.489$ and $+0.512$; all 24 cell–repetition contrasts are positive. Centering outcomes within donor blocks gives $+0.582$, so between-donor means alone do not explain the contrast.

Table 49: BMI–sedentary correlations among newly generated tenth respondents only; Δ is aligned minus permuted history. Injected records are excluded.

Group	Natural	Permuted	Aligned	Δ
Men, 20-34	0.724	0.338	0.766	+0.428
Men, 35-49	0.481	-0.184	0.570	+0.754
Men, 50-64	0.386	0.098	0.566	+0.468
Men, 65+	0.340	0.203	0.658	+0.455
Women, 20-34	0.543	-0.072	0.649	+0.721
Women, 35-49	0.257	0.314	0.832	+0.518
Women, 50-64	0.548	0.336	0.588	+0.252
Women, 65+	0.283	-0.086	0.340	+0.426
Equal-cell mean	0.445	0.118	0.621	+0.503

Changing the earlier answers changes later relationships, but does not establish greater accuracy or show how much of the batching penalty this mechanism explains. Keeping the same values in the prefix also does not guarantee agreement with profiles or preserve the distribution of new target answers. Alignment raises mean sedentary time by 9.446 minutes relative to permutation; mean Wasserstein distance between conditions is 32.963 minutes across cells. These measure differences between generated outputs, not errors against human data. The supplement also reports contrasts with natural history, all-retained results and marginal changes for all five outcomes. These are descriptive checks without multiplicity adjustment.