WHAT DO LLM-SIMULATED PARTICIPANTS TRACK? MODEL BIAS, REPLICATION ESTIMATES, AND COLLAPSED VARIANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM simulations of human participants are usually validated against published effect sizes. Published effects are inflated by publication bias, and a model can also produce an effect because it shows the bias itself, so agreement with the literature cannot tell a faithful simulation from one that echoes the literature or its own habits. We propose a decomposition that separates these sources across effects and apply it at scale. For 61 psychology effects with both an original estimate and a large-sample replication estimate, we run text-adapted analogues of each experiment (four LLMs from three developers, about 43,000 simulated participants in the main protocol) and regress the simulated effect size on the model's own bias, measured by giving it the same task with no participant framing, on the replication estimate, and on the gap between original and replication. When two independent transcriptions of each experiment are averaged, simulated effects covary positively with the model's own bias and with the replication estimate in all three model families, but in one family either coefficient becomes uncertain under a single transcription. Whether simulations also inherit the gap cannot be decided at this scale. The decomposition therefore diagnoses what simulations track but cannot score one simulator. A simulator told to copy the original findings goes undetected when scored alone, while a paired comparison of two simulators on the same effects shows a signal. Simulated samples have almost no within-condition variance, so simulated effects are several times larger than human ones, and where the model shows no bias of its own, agreement with the direction of replicated effects is not distinguishable from chance. One model refuses selectively on aversive treatment arms. In an exploratory construct battery, correlations the literature makes famous are exaggerated relative to a matched human panel. We argue that simulators should be compared on shared replication-anchored protocols with a model-bias control, and that current simulations should not be used to infer effects the model does not already show.

1 INTRODUCTION

A growing line of work uses large language models as simulated participants. The model is given a persona and an experimental stimulus, and its answers are aggregated as if they came from a human sample (Argyle et al., 2023; Aher et al., 2023), and the prospect of replacing human participants is under active debate (Dillion et al., 2023; Lin, 2025). The promise is cheap piloting, power analysis before data collection, and exploration of designs costly or unethical to run on people. Datasets such as Twin-2K-500 (Toubia et al., 2025) pair thousands of real respondents with LLM "digital twins" to support exactly this use.

How is such a simulation validated? In practice, almost always by comparing simulated results to published findings, asking it to reproduce what the literature reports and be judged faithful. This standard has a structural problem that is well known in metascience but has not, to our knowledge, been examined in the simulation setting. Published original estimates are systematically inflated by

Our source code is available at `https://anonymous.4open.science/r/llm-simulation-replication-gaps-1F76/`.

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.

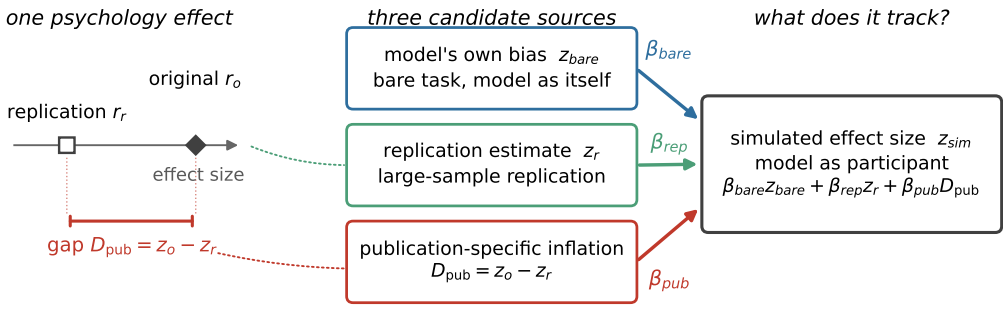


Figure 1: The three-source comparison. Each effect has an original estimate r_o and a large-sample replication estimate r_r. Their difference in Fisher-z space, $D_{\text{pub}} = z_o - z_r$, is used to probe publication-specific inflation, although it may also reflect differences in samples and implementation. For effects with a defined bare task, we measure the model both directly and under simulated-participant framing. Regressing the simulated effect size on the bare-task effect, the replication estimate, and D_{pub} estimates how strongly it tracks each quantity.

selective publication and analytic flexibility (Rosenthal, 1979; Simmons et al., 2011). Large-scale replication projects find that replication effect sizes are on average about half the original ones (Open Science Collaboration, 2015; Camerer et al., 2018), and databases of original-replication pairs, such as the Replication Database (FReD) (Röseler et al., 2024), document this gap effect by effect. A language model is trained on the same literature that contains the inflated numbers, and language models are known to memorize frequently repeated content from their training data (Carlini et al., 2023), with memorization and exploitation of that content being separable behaviors (Magar & Schwartz, 2022). If the model learned the literature's statistics, a simulation would reproduce the inflated values, pass validation against them, and fail silently exactly where it is supposed to substitute for new data.

The same gap that creates this risk also creates a way to measure it. For any effect where the original estimate and a large replication disagree, the two numbers make competing predictions about what a faithful simulation should produce. If the simulation learned from the literature, its effect size should sit near the original estimate. If the simulation reproduces the mechanism that survives replication, it should sit near the replication estimate. Publication bias, normally a confound, becomes the identifying contrast.

Running this test requires one further control that earlier validation work does not include. A language model can produce an effect without reading about it and without simulating anyone, simply because it exhibits the underlying judgment bias itself. On classic judgment tasks, models produce human-like biases in their own answers, varying by task and model generation (Binz & Schulz, 2023; Hagendorff et al., 2023). We therefore measure, wherever the task allows it, the model's *own bias*: the effect size the model produces when given the same judgment task directly, with no persona and no mention of an experiment. This separates "the simulation inherited the literature" from "the model is like this anyway."

We assemble 61 psychology effects from the Replication Database, each with a text-based paradigm, an original estimate, and a large-sample replication estimate, and run a text-adapted analogue of each experiment (four LLMs from three developers, about 43,000 simulated participants in the main protocol and 68,500 in all, where the two Claude models form the main analysis, with the decomposition estimated on the 38 effects admitting a bare-task control). Our pre-specified analysis decomposes each simulated effect size into the model's own bias, the replication estimate, and the inflation specific to the original publication.

Three findings emerge. First, simulated effect sizes covary positively with the model's bare-task effect and with the replication estimate in every model family when two independent transcriptions of each experiment are averaged, though in the Claude models either becomes uncertain under one transcription, and the gap component is inconclusive. Simulated samples have almost no within-condition variance, so these coefficients describe which effects a simulation makes large, not calibrated transfer. Second, literature influence reappears at a different level. In an exploratory construct battery, correlations that the literature treats as textbook findings are exaggerated relative to

a matched human panel while undiscussed correlations of the same size are not. Third, one model refuses selectively on the ethically aversive arm of manipulations, which deletes treatment arms instead of adding noise.

LLM-based simulation is increasingly built and benchmarked as an AI capability, from digital twin datasets (Toubia et al., 2025) to agent-based models of whole populations (Park et al., 2024), and the usual benchmark target is agreement with published effects. Our results show that this target rewards a model for its own biases and for the literature's loud claims, and that it cannot detect collapsed variance. Where the model does not show an effect when asked directly, agreement with the direction of the replicated effect is not distinguishable from chance, so there is no evidence that simulations recover effects the model does not already show. Changing this is a modeling problem, and the design shows what a decisive test of future simulators requires.

Contributions. (1) A decomposition of a simulated effect into the model's own bias, the replication estimate, and the original-replication gap, which at feasible scale diagnoses what simulations track across effects rather than scoring one simulator. (2) An application to 61 effects and four models. Simulated effects track the first two sources in every family when transcriptions are averaged, and the gap component cannot be decided at this scale. Positive controls and a Bayesian power analysis show what the design can and cannot detect and how large a decisive test must be. (3) Exploratory evidence that literature influence appears in which construct relations the model exaggerates, persisting after fame is separated from political content and item wording is controlled, though sensitive to how uncomputable correlations are treated. (4) Practical warnings for simulation users, namely variance collapse and condition-targeted refusal, with concrete reporting recommendations.

2 Related work

LLMs as simulated participants, and their validation. Persona-conditioned models reproduce survey distributions (Argyle et al., 2023), replicate classic experiments (Aher et al., 2023), act as economic agents (Horton et al., 2023), and extend to interview-grounded twins (Park et al., 2024; Toubia et al., 2025). Simulated effects are known to match direction but exceed human ones (Cui et al., 2024), and a parallel line documents distortions including compressed variance (Santurkar et al., 2023; Bisbee et al., 2024; Wang et al., 2025; Peng et al., 2025; Tjuatja et al., 2024), formalizes validation (Hullman et al., 2026; Lin, 2025), and argues that such simulations need explicit boundaries on what they can claim (Wu et al., 2026). All score simulations against a single human reference. What is new here is asking which of several competing human quantities, and how much of the model's own bias, the simulation tracks.

Models' own biases, memorization, and contamination. That LLMs show human-like judgment biases when asked directly is established (Binz & Schulz, 2023; Hagendorff et al., 2023; Coda-Forno et al., 2024; Cheung et al., 2025), and we repurpose such measurements as a competing regressor. Models memorize repeated training content (Carlini et al., 2023), memorization is separable from its use (Magar & Schwartz, 2022), and contamination threatens evaluation (Cheng et al., 2025). Published effect sizes are the social science analogue of a contaminated test set.

Publication bias and replication. Large replication projects find estimates well below originals (Open Science Collaboration, 2015; Camerer et al., 2018; Klein et al., 2018), databases aggregate the pairs (Röseler et al., 2024), and the mechanisms of inflation are well characterized (Rosenthal, 1979; Simmons et al., 2011; Ioannidis, 2005; Kvarven et al., 2020). We turn the original-replication gap into a probe of what models learned from that literature (extended review in Appendix Q).

3 A three-source decomposition of simulated effect sizes

3.1 Setup and notation

Figure 1 summarizes the design. An *effect* is a published psychology finding with a directional claim, for example "a harmful side effect is judged intentional, a helpful one is not." For each effect we have two human estimates on the correlation scale, namely the original published effect size r_o

and the effect size r_r from a large, preregistered replication of the same design. We work throughout in Fisher-z space, $z = \operatorname{arctanh}(r)$, which removes the ceiling compression near $|r| = 1$. We write z_o and z_r for the transformed estimates.

A *simulation* of an effect proceeds as in prior work on LLM participants. The model receives a system prompt stating that it powers a social science simulation, a minimal participant framing ("answer as a typical human study participant would," with no demographic persona), the experimental stimulus of one condition, and a response format. Simulated participants are independent calls. Aggregating their answers with the original study's statistical test yields a simulated effect size r_{sim}, transformed to z_{sim}.

The third measurement is the *bare task*, in the tradition of testing models directly on cognitive psychology paradigms (Binz & Schulz, 2023; Coda-Forno et al., 2024). The model receives the same judgment task with the experiment framing and the participant framing stripped away, and answers as itself. Aggregating bare-task answers with the same test yields z_{bare}, the effect the model produces with nobody to simulate. The bare task is only defined for paradigms where the dependent variable is a judgment the model can make in its own voice, for example a probability estimate or a preference. Paradigms whose dependent variable is intrinsically about a simulated person's inner state, such as recalling one's own secret, have no bare version and are excluded from the three-source estimation analysis.

3.2 The decomposition

Our pre-specified model explains simulated effect sizes by three sources:

$$z_{\text{sim}} \;=\; \beta_0 \;+\; \beta_{\text{bare}}\, z_{\text{bare}} \;+\; \beta_{\text{rep}}\, z_r \;+\; \beta_{\text{pub}}\,(z_o - z_r) \;+\; \gamma\, m \;+\; u_{\text{effect}} \;+\; \varepsilon, \qquad (1)$$

where $m \in \{0, 1\}$ is a model indicator, u_{effect} is a random intercept per effect, and the unit of analysis is an effect-model pair. The three coefficients correspond to three hypotheses about what the simulated participant is doing:

- β_{bare} measures how much of the simulation is the model's own cognition, present with or without a participant to play. It partly reflects stimuli shared with the simulation (Appendix G).
- β_{rep} measures association with the replication estimate. It need not mean transfer, since effects that replicate may be those whose manipulation is strongly expressed in the stimulus text, to which humans and models both respond.
- β_{pub} is the quantity of interest. The difference $z_o - z_r$ is the part of the original estimate that replication does not support. If simulations inherit the literature's inflated numbers, this difference should predict z_{sim} over and above the other two sources. We call a positive β_{pub} a *publication bias fingerprint*.

Both z_o and z_r are estimates with known sampling variance $1/(n-3)$ from their study sizes. Ignoring this would bias β_{pub}, because a noisily measured z_r leaves residual variance that z_o could absorb even without any fingerprint. We therefore fit Equation 1 as a Bayesian measurement error regression (Fuller, 2009) in which the true z_o and z_r are latent variables with the observed values as noisy measurements.

3.3 Pre-specified decision rule

The decision rule was fixed before the confirmatory data were analyzed. It was not externally registered. We call the 71 units entering Equation 1 the estimation sample, and 19 effects entered through pilot waves with earlier protocol versions, so the pooled estimate is not uniformly confirmatory, and the uniform-protocol final wave is reported separately as the cleanest subset. The rule specifies an equivalence region of $|\beta_{\text{pub}}| < 0.10$ in Fisher-z slope units and a three-way rule. If the 95% credible interval lies entirely above zero, the fingerprint is supported. If the interval lies entirely inside the equivalence region, there is evidence of absence. If the interval crosses zero but extends beyond the region, the result is inconclusive. The rule also fixes the primary sample, the primary framing, and four sensitivity analyses in advance (Appendix M). No power analysis preceded data collection. At the realized sample size the equivalence region is unreachable and a large fingerprint is detected less than half the time (Appendix K).

The gap $z_o - z_r$ is not purely publication bias. It also contains sample and implementation differences between original and replication, true heterogeneity, and sampling error of the original study. A positive β_{pub} therefore means that simulations inherit the component of the original estimate that replication does not support. Attributing that component to publication bias specifically relies on the metascience consensus about what drives original-replication gaps, and we phrase all claims accordingly.

A second caveat concerns scale. The coefficients are slopes across effects in Fisher-z units. Simulated and human effect sizes share that scale only nominally, because the simulated samples have far less within-condition variance than human ones (Section 5), so the coefficients measure how simulated effects rank and covary with each source, not how much of a human effect is reproduced.

4 Experimental setup

4.1 Effects

We draw effects from the Replication Database (FReD; Röseler et al., 2024), which pairs original studies with replication attempts. From 2,164 recorded pairs we kept those where both effect sizes convert to the correlation scale and both sample sizes are known, which leaves 710 pairs, and deduplicated to one pair per original paper, keeping the best-powered replication. From this pool we selected 61 effects in two steps. Nineteen entered through two pilot waves that oversampled large original-replication gaps and well-known findings, and the remaining 42 were drawn by stratified sampling over gap size, discipline, and paradigm type, run in a final wave with one uniform protocol. In both steps the paradigm had to be runnable as a text interaction. Reaction-time and psychophysical paradigms were excluded. The 61 effects span social priming, judgment and decision biases, moral judgment, questionnaire self-report, and consumer behavior. The identifying contrast $z_o - z_r$ ranges from -0.47 to $+0.93$ in Fisher-z units across the 61 effects (median 0.26, negative for 11 effects), so the two human estimates genuinely disagree for a substantial part of the sample.

Each experiment was transcribed into a text protocol in two independent variants. In 20 of the 122 templates the text comes from the original materials, and the rest reconstruct the paradigm from its canonical description (per-template provenance in the repository), so each simulation is a text-adapted analogue of the original design, not a rerun of it.

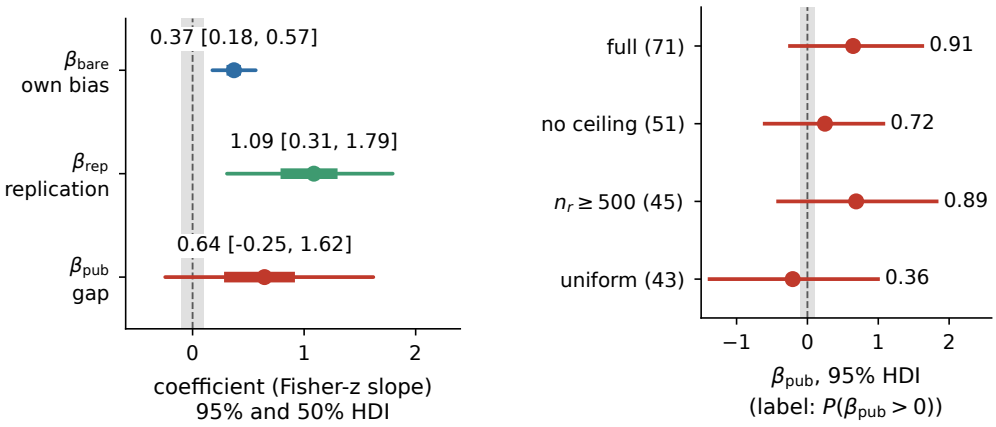
4.2 Models and simulated participants

We use two models of the same family, Claude Sonnet 5 (no temperature parameter passed, so the API's default stochastic sampling applies, not greedy decoding) and Claude Sonnet 4.5 (temperature 1.0), treated as two instruments. Between-participant variance comes from sampling across independent calls. Each cell, meaning one transcription variant of one effect on one model, was simulated with independent calls cycled across conditions (about 60 rollouts per cell, roughly 30 per condition arm) under the participant framing of Section 3. Runs with demographic personas were collected in parallel and serve as a robustness check rather than the primary condition. In total the main experiment comprises about 33,000 simulated participants across the three data collection waves (counts for every experiment in Table 1). The full protocol was also run on GPT-5.6-terra and DeepSeek-v4-flash as a cross-family robustness check (Section 5.5). Refusals were recorded verbatim (Section 5.7, Appendix O).

4.3 Bare-task battery

For the 43 effects whose dependent variable is a judgment the model can make in its own voice, we measured the model's own bias with the experiment framing stripped. To prevent a single wording from driving the measurement, each bare task was rephrased in five wording seeds, and z_{bare} aggregates over seeds. Five effects failed the bare measurement (degenerate or unparseable responses), leaving 38 effects and 71 effect-model units in the estimation sample (five single-model bare failures account for 71 rather than 76 units).

Transcription of each experiment into two independent text protocols, parsing, and scoring with the original study's statistical test are described in Appendix E.



(a) Pre-specified fingerprint test (71 units, 38 effects).

(b) Sensitivity of β_{pub}.

Figure 2: The three-source decomposition. (a) Simulated effects covary with the model's own bias and with the replication estimate, while the gap coefficient is inconclusive (shaded band, equivalence region ± 0.10). (b) The positive point estimate for β_{pub} is lower without ceiling units and in the uniform-protocol wave, but the differences between subsets are themselves uncertain.

5 Results

5.1 What the simulated effect size tracks

Figure 2a shows the three coefficients of Equation 1 on the estimation sample (71 effect-model units, 38 effects). Simulated effects covary positively with two sources. The bare-task coefficient is $\beta_{\text{bare}} = 0.37$, 95% HDI $[0.18, 0.57]$, and the replication coefficient is $\beta_{\text{rep}} = 1.09$, $[0.31, 1.79]$. The fingerprint coefficient is $\beta_{\text{pub}} = 0.64$, $[-0.25, 1.62]$. Under the pre-specified rule this is inconclusive, since the interval crosses zero and is far wider than the equivalence region.

In the Claude models both coefficients are clearly positive only when the simulated effect averages the two independent transcriptions of each experiment, which agree only moderately (Fisher-z correlation 0.68). With either transcription alone one interval reaches zero, whereas GPT and DeepSeek stay positive under either one (Table 3). Transcription variation is thus a measurement problem as serious as the collapsed variance.

5.2 Scale, ranking, and what conventional validation would see

The mean simulated effect is $z = 0.81$, against 0.38 for originals and 0.15 for replications, because the near-zero within-condition variance of simulated samples (Appendix L) turns into large standardized effects. Personas do not restore the variance, and a rank-based check points the same way but is weaker (Appendix G).

Among units with a real replicated effect, simulations match its sign in 80% of Claude units when the model shows the bias bare (partly by construction, since the bare task and the simulation share stimuli), but agreement is not distinguishable from chance when it does not (16 of 29 across the three families, 95% CI 36% to 74%, Appendix G).

A conventional validation that benchmarks simulations against original findings would see a correlation of 0.48 across units and agreement in sign for 70% of them, and would judge the simulation faithful, but against replications they are nearly the same (0.47 and 66%), so such a benchmark cannot tell which quantity the simulation tracks.

5.3 The positive point estimate is fragile

Figure 2b decomposes the inconclusive result. The full estimation sample gives $P(\beta_{\text{pub}} > 0) = 0.91$. Excluding units at the response ceiling ($|r_{\text{sim}}| > 0.95$, 20 of 71 units) lowers the estimate from 0.64 to 0.25, restricting to high-precision replications ($n_r \geq 500$,

45 units) changes little, and the final wave, which used one uniform protocol for all effects (not pre-specified), gives -0.21. The subset differences are themselves uncertain (interaction estimates in Appendix G), so the positive point estimate may come from ceiling compression and protocol differences, and a shrinking estimate in a smaller subset is not evidence of absence.

The remaining checks leave this picture unchanged, namely paradigm-class fixed effects, a rerun under a system prompt that does not mention replication, a hierarchical prior with simulation-based calibration, and censored and measurement-error likelihoods (Appendices F, M, and G). Remeasuring the 14 ceiling effects with 0–100 scales moves the gap coefficient below zero in all three families (Appendix L).

5.4 Positive controls

We built two simulators that are told, in one added prompt paragraph, either the original study's finding and effect size or the replication's. Fitted separately, neither shows a clear gap coefficient, so at 38 effects the decomposition cannot score a single simulator. The within-unit difference between the two simulators does follow the gap (slope 0.66 $[0.08, 1.24]$), so a paired comparison on the same effects shows a signal, although a stacked interaction model is less certain (Appendix I). Both also overshoot the numbers they are given (mean simulated z of 1.39 and 1.08 against 0.38 and 0.15). A model told the effect size still cannot produce it, because it does not generate the within-condition variation that the effect size depends on.

5.5 Cross-family replication

We reran the full protocol bank and bare-task battery on GPT-5.6-terra and DeepSeek-v4-flash (about 5,000 simulated participants each). In all three families both non-gap coefficients are positive and the gap coefficient is inconclusive, with a family-specific bare-task weight (0.37, 0.58, 0.50, full estimates in Appendix F).

Simulated effects are large where the model shows the bias when asked directly and near zero for manipulations it does not show bare, such as social priming, and participant framing can release behavior the model does not produce as itself (examples in Appendix H).

5.6 Literature influence at the level of structure

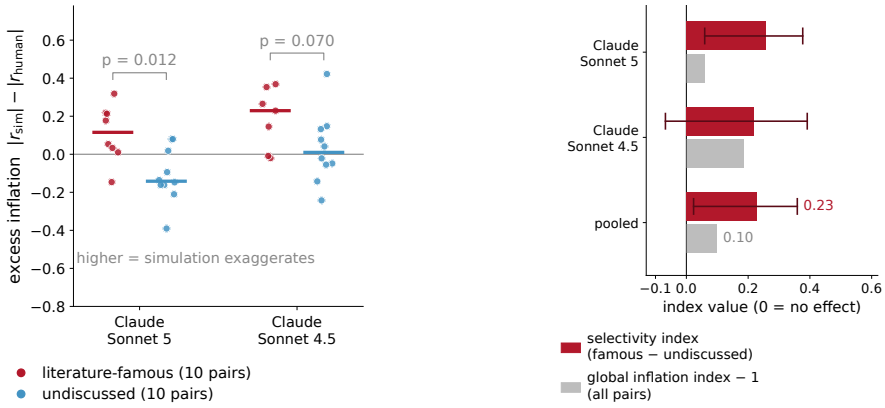
The decomposition asked whether simulations copy effect size *values* from the literature. A different form of literature influence would leave values intact and instead shape *which relations between variables* the model believes exist. We test this at the level of correlation structure.

5.6.1 Design

This experiment is exploratory (designed after the effect-size results were known) and uses the same two Claude models as the main analysis. We compare simulated and human correlations over a battery of 22 constructs that parallel measures in the Twin-2K-500 panel (Toubia et al., 2025) ($n = 2{,}058$; battery details in Appendix R). From the panel's construct pairs we preselected two groups of ten: *literature-famous* pairs whose correlation is a textbook finding (for example conscientiousness with performance, ideology with religiosity), and *undiscussed* pairs of comparable magnitude [1] in the human data that no literature we could locate discusses as a finding. If literature influence operates at the structural level, famous pairs should be exaggerated in simulation relative to the human panel, and undiscussed pairs should not. Two other features could produce the same pattern and are tested below. Many famous correlations involve politics, and LLMs hold systematic political positions (Santurkar et al., 2023), so political content could drive exaggeration by itself. Famous constructs could also simply be worded more alike, which would make a model answer them consistently without any knowledge of the literature.

Because the panel's item text is not redistributable, our battery is a parallel operationalization of the same constructs built from public domain instruments, not a verbatim rerun. Human-simulation dif-

[1] A winner's-curse simulation bounds the bias from selecting pairs on noisy panel estimates well below the observed group difference (Appendix S).



(a) Excess inflation by pair group.

(b) Selectivity versus global inflation.

Figure 3: Literature influence at the structure level. Only literature-famous construct pairs are exaggerated relative to the human panel (a). The selectivity index is distinguishable from zero while the global inflation index is small (b). Per-pair detail in Appendix R.

ferences therefore include an operationalization component, which we disclose as the main validity limit of this experiment.

5.6.2 Famous relations are selectively exaggerated

Figure 3 shows the result. Averaged over both models, famous pairs are exaggerated relative to the human panel while undiscussed pairs are not. The mean difference in excess correlation ($|r_{\text{sim}}| - |r_{\text{human}}|$, positive when the simulation exaggerates) between the two groups is $+0.18$ (Mann-Whitney $p = .002$; the median-based pooled index is $+0.23$ with bootstrap 95% CI $[0.02, 0.36]$). A global inflation account, in which simulation sharpens all correlations uniformly, predicts no famous-undiscussed difference. The observed global inflation component is small (median ratio 1.10, and below 1 after correcting the human panel for test-retest attenuation). The exaggeration is selective, not global.

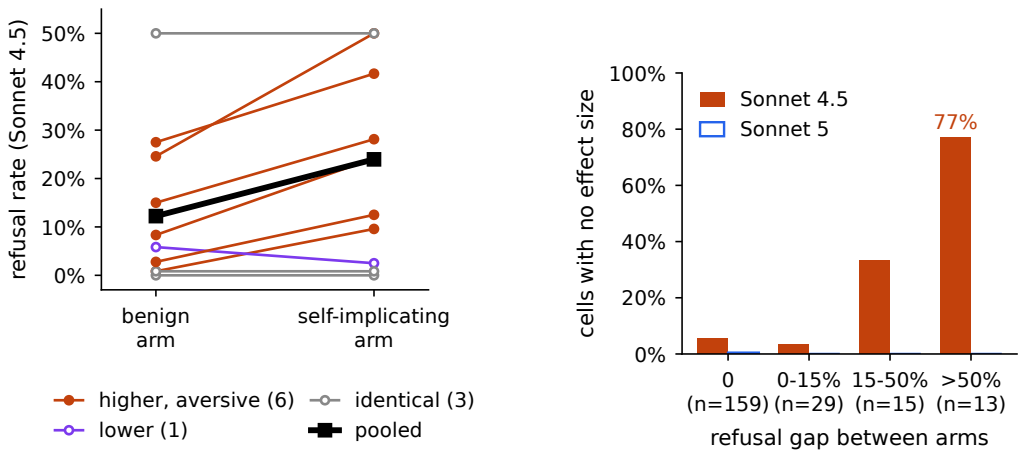
To separate fame from political content, the design crosses fame with content (famous or undiscussed by political or nonpolitical, 36 pairs, with undiscussed pairs magnitude-matched to famous ones). The famous nonpolitical and undiscussed political cells were collected in a second batch under the same protocol. With the pair as the unit, the fame main effect is $+0.16$ ($p = .007$) and the political-content main effect is $+0.005$ ($p = .92$). The effect survives neutral treatments of the uncomputable ability correlations, which fall mostly in the famous nonpolitical cell, but not a worst-case bound, and a control for item wording leaves it unchanged (Appendices S and G).

Because the battery is a parallel operationalization, part of the human-simulation gap is measurement rather than model. With this caveat, the pattern is consistent with selective exaggeration of relations the literature loudly asserts, which the effect-size analysis does not examine.

5.7 Refusal as structural missingness, a practical warning

Simulated participants sometimes refuse, which is a practical hazard for anyone running simulations. In our data refusal is essentially specific to one model. Across the 30,624 participant-framed rollouts of the two Claude models (15,312 per model), Sonnet 4.5 refused 8.4% while Sonnet 5 refused a single rollout, so the analysis below concerns Sonnet 4.5 only.

Refusal concentrates on content rather than on role-play in general (Figure 4), a pattern related to exaggerated safety behaviors (Röttger et al., 2024). Among ten effects whose arms differ only in self-implication, the aversive arm is refused twice as often (odds ratio 3.46, $[2.30, 5.20]$, Appendix O). Because refusal hits one arm of a manipulation, it deletes the contrast rather than adding noise, and most cells with between-arm refusal asymmetry above 50% yield no computable effect size. Users should therefore report refusal rates per arm, not per study.



(a) Within-effect arm comparison.

(b) Cells with no computable effect size.

Figure 4: Refusal is condition-targeted and deletes contrasts. Within the same effect, refusal is higher on the aversive arm (a), and arm-differential refusal makes entire cells uncomputable rather than adding noise (b). Refusal rates for all 25 condition types are in Appendix O.

6 Discussion

6.1 Where LLM-based social simulation should go

It has been argued that LLM-based social simulations require explicit boundaries on what they can claim (Wu et al., 2026). Our results indicate where such a boundary currently lies and what it requires of simulators and their benchmarks.

Evaluate against the right target. Benchmarking against published findings rewards a model for its own biases and for the literature's loud claims, and in our data it cannot tell originals from replications. Benchmarks should anchor on large replications, include a bare-task control, and compare simulators with each other on a shared protocol bank, since planted copying shows a signal only in such a paired comparison.

Treat variance as a first-class objective. Simulated samples are nearly deterministic within a condition, so standardized effects are inflated severalfold even when the simulator is told the target effect size (Section 5.4). This is closer to distributional alignment than to prompt design, and simulators should report within-condition variance next to any effect size.

Restrict claims until those problems are solved. Because direction agreement is not distinguishable from chance where the model shows no bias, current simulations should not be used to infer the direction of such effects, to estimate magnitudes or plan sample sizes, or to test hypotheses the literature already asserts, where they exaggerate.

6.2 Limitations

Five limits bound our claims. First, a null gap coefficient cannot separate no literature influence from influence by a corpus that includes replication outcomes (Appendix D). Second, power was assessed only after data collection, and only a large fingerprint is detectable (Appendix K). Third, the gap is not purely publication bias, and 23 of 61 effects have no bare task. Fourth, collapsed variance and transcription choice are measurement problems of equal weight. The coefficients are slopes, not calibrated transfer, and in the Claude models each is clearly positive only when two transcriptions are averaged (Appendix G). Fifth, only 20 of 122 protocols are primary-source (Appendix J), and the construct battery is a parallel operationalization.

6.3 Conclusion

Simulated effects follow the model's own bias and replications, the share of what merely got published cannot yet be pinned down, and famous relations are exaggerated (Appendix P).

Reproducibility statement

All experiment protocols, parsing code, and analysis code are included in the GitHub repository. Raw model responses are available from the authors on request. Appendix A documents effect selection and Appendix N shows the prompt formats.

AI use statement

We used generative AI tools to help develop theoretical models or conceptual frameworks, design or provide feedback on research methodology or experiments, implement code, assist with translation, clean and reformat datasets, and support qualitative and thematic data analysis. We have not used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, or interpret results. In addition, language models are the object of study in this paper: all simulated participants, bare-task responses, and part of the annotation of effects (paradigm classification and corpus-narrative labels, the latter with multi-model consensus and human adjudication) were generated by the LLMs named in the text. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pp. 337–371. PMLR, 2023.

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.

Solomon E Asch. Forming impressions of personality. *The journal of abnormal and social psychology*, 41(3):258, 1946.

František Bartoš, Maximilian Maier, David R Shanks, TD Stanley, Martina Sladekova, and Eric-Jan Wagenmakers. Meta-analyses in psychology often overestimate evidence for and size of effects. *Royal Society Open Science*, 10(7):230224, 2023.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

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.

Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature human behaviour*, 2(9):637–644, 2018.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Yuxing Cheng, Yi Chang, and Yuan Wu. A survey on data contamination for large language models. *arXiv preprint arXiv:2502.14425*, 2025.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122 (25):e2412015122, 2025.

Julian Coda-Forno, Marcel Binz, Jane X Wang, and Eric Schulz. CogBench: a large language model walks into a psychology lab. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9076–9108. PMLR, 2024.

Ziyan Cui, Ning Li, and Huaikang Zhou. Can large language models replace human subjects? a large-scale replication of scenario-based experiments in psychology and management. *arXiv preprint arXiv:2409.00128*, 2024.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in cognitive sciences*, 27(7):597–600, 2023.

Wayne A Fuller. *Measurement error models*. John Wiley & Sons, 2009.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838, 2023.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Jessica Hullman, David Broska, Huaman Sun, and Aaron Shaw. This human study did not involve human subjects: Validating llm simulations as behavioral evidence. *arXiv preprint arXiv:2602.15785*, 2026.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Richard A Klein, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks, Claudia Chloe Brumbaugh, et al. Investigating variation in replicability. *Social psychology*, 2014.

Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.

Joshua Knobe. Intentional action and side effects in ordinary language. *Analysis*, 63(3):190–194, 2003.

Amanda Kvarven, Eirik Strømland, and Magnus Johannesson. Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, 4(4):423–434, 2020.

Zhicheng Lin. Six fallacies in substituting large language models for human participants. *Advances in Methods and Practices in Psychological Science*, 8(3):25152459251357566, 2025.

Inbal Magar and Roy Schwartz. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 157–165, 2022.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015.

Daniel M Oppenheimer and Benoît Monin. The retrospective gamblers fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, 4(5):326–334, 2009.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Joon Sung Park, Carolyn Q Zou, Jonne Kamphorst, Niles Egan, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Percy Liang, Robb Willer, et al. Llm agents grounded in self-reports enable general-purpose simulation of individuals. *arXiv preprint arXiv:2411.10109*, 2024.

Tianyi Peng, George Gui, Melanie Brucks, Daniel J Merlau, Grace Jiarui Fan, Malek Ben Sliman, Eric J Johnson, Abdullah Althenayyan, Silvia Bellezza, Dante Donati, et al. Digital twins as funhouse mirrors: Five key distortions. *arXiv preprint arXiv:2509.19088*, 2025.

Lukas Röseler, Leonard Kaiser, Christopher Doetsch, Noah Klett, Christian Seida, Astrid Schütz, Balazs Aczel, Nadia Adelina, Valeria Agostini, Samuel Alarie, et al. The replication database: Documenting the replicability of psychological science. *Journal of Open Psychology Data*, 12 (1):8–8, 2024.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86 (3):638, 1979.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5377–5400, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024.

Olivier Toubia, George Z Gui, Tianyi Peng, Daniel J Merlau, Ang Li, and Haozhe Chen. Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455, 2025.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232, 1973.

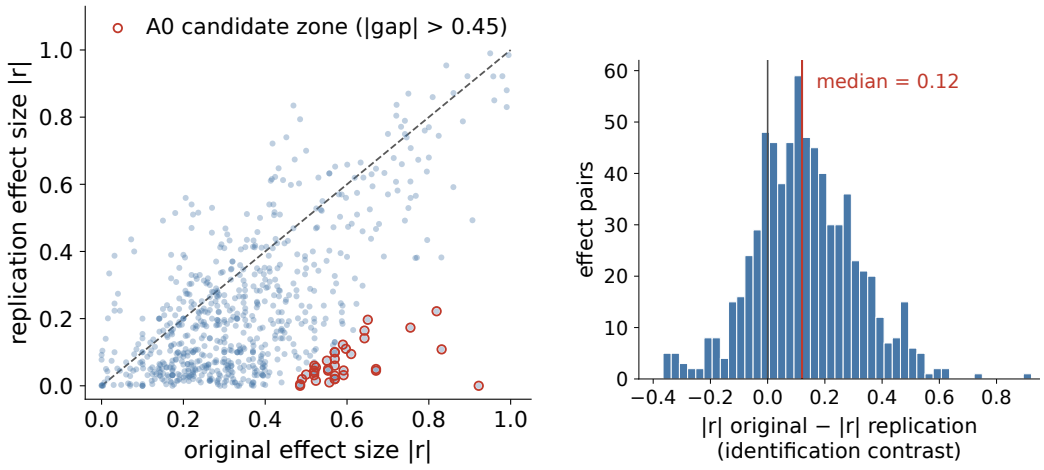
Felipe Valencia-Clavijo. Anchors in the machine: Behavioral and attributional evidence of anchoring bias in llms. *arXiv preprint arXiv:2511.05766*, 2025.

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

Zengqing Wu, Run Peng, Takayuki Ito, Makoto Onizuka, and Chuan Xiao. Llm-based social simulations require a boundary. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

A Effect selection and audit

Figure 5 summarizes the original–replication pairs used to construct the stratified sample. The stratified sample of 61 effects, with original and replication estimates, sample sizes, paradigm class, estimation-sample membership, and transcription provenance per variant, is provided as a table in the GitHub repository (`protocols/effect_summary_61.csv`), and the original references and DOIs are in `protocols/effect_DOI_list_61.csv`.



(a) Original versus replication effect sizes.

(b) Distribution of the identifying contrast.

Figure 5: Audit of the Replication Database (FReD) pairs used for effect selection. Most replications land below the original (a), and 249 pairs have contrast above 0.2 (b).

B PILOT WAVES

Data collection proceeded in three waves. Wave A0 (10 effects with the largest original-replication gaps) and wave A1 (9 effects in a 2×2 of literature fame by replication outcome, gap-matched between the critical cells) were design pilots whose protocols differ slightly from the final uniform protocol, and the final wave covers the remaining 42 effects under one protocol. All waves enter the descriptive analyses. The sensitivity analysis in Figure 2b isolates the uniform-protocol wave. Figure 6 shows the results of the A1 pilot.

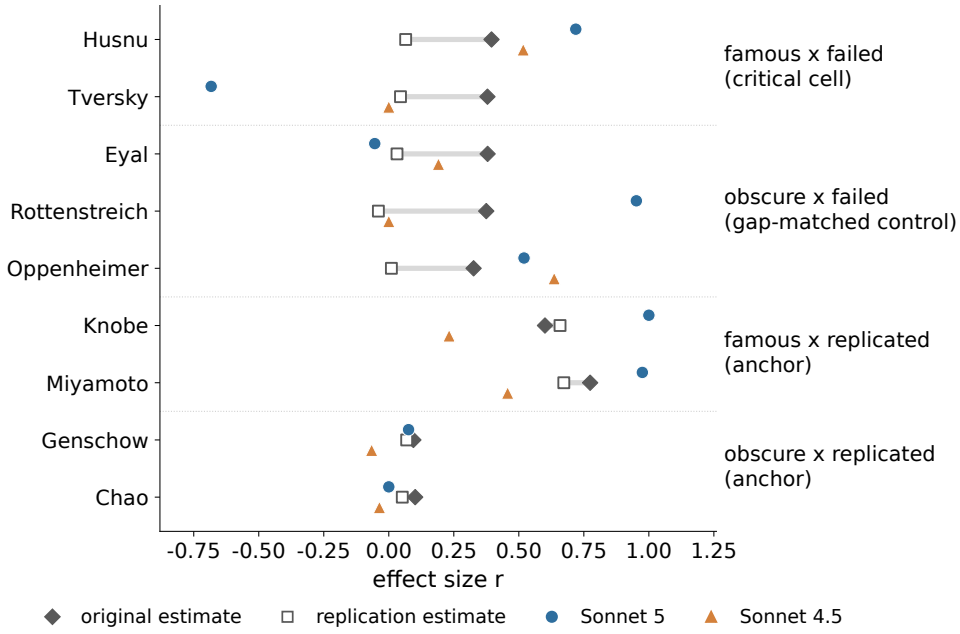


Figure 6: Wave A1: nine new effects in a fame-by-replication-outcome design with gap matching. Simulated effects do not sort by fame. They sort by whether the model itself exhibits the bias, the pattern the bare-task control in the main analysis quantifies directly.

C BARE VERSUS SIMULATED EFFECTS

The full bare-task battery, with all five wording seeds per task, is in the GitHub repository (`protocols/bare_task_bank.json`), and per-effect bare effect sizes are tabulated there as well (`analysis_tables/`). The three regressors of Equation 1 are separable. Across the 71 estimation units the variance inflation factors are 1.19 for z_{bare}, 1.39 for z_r, and 1.19 for D_{pub}, and at the effect level z_r and D_{pub} correlate at -0.38 while z_o retains 45% of its variance after controlling for z_r. On the outcome side, z_{bare} and z_{sim} correlate marginally at 0.51, and the correlation is 0.41 after partialling out z_r, D_{pub}, and the model indicator. The coefficient $\beta_{\text{bare}} = 0.37$ is this partial relation, the contribution of the bare measurement holding the human anchors fixed. The marginal and partial values differ because the bare measurement also covaries with the human anchors (0.39 with z_r).

D NARRATIVE LABELS

This appendix probes an identification threat. Because replication outcomes are themselves in the training corpus, $\beta_{\text{pub}} \approx 0$ cannot by itself distinguish a model that ignores the literature's numbers from one that tracks a literature which includes the replication outcome. If literature influence drives simulation, it should be strongest for effects whose original finding still dominates what the corpus says.

We labeled each effect by the dominant narrative about it in the broad scientific and popular corpus (original finding still told as true, replication failure better known, contested, or rarely discussed). Labels are the consensus of four independent annotation rounds by three model families (two Claude Sonnet 5 rounds, one Claude Opus 5 round, one GPT-5.6-terra round run by the authors), requiring agreement of at least three rounds, and effects without such agreement are labeled uncertain. All round-level labels and the consensus are in the GitHub repository (`analysis_tables/narrative_final_fourround.csv`, mapping in `analysis_tables/narrative_map.json`).

Stratifying the fingerprint slope by this label (an original-still-dominates indicator interacting with D_{pub} in the main specification) gives no robust evidence that literature influence concentrates where the original finding dominates. On the observed values the fourteen units whose original finding still dominates the corpus show $\beta_{\text{pub}} = 1.69$, 95% credible interval $[0.03, 3.12]$, $P(\beta > 0) = 0.98$, against 0.52, $[-0.39, 1.43]$ for the rest (interaction 1.17, $[-0.19, 2.59]$). This stratum also concentrates the response ceiling (36% of its units at $|r_{\text{sim}}| > .95$ versus 26% elsewhere), and on the ceiling-remediated measurements of Appendix L the original-dominates estimate is -0.20, $[-1.36, 0.99]$, $P(\beta > 0) = 0.36$, with an interaction of 0.04, $[-1.03, 1.18]$.

Two structural limits make this a probe rather than a test. The sharpest contrast, effects whose replication failure is better known than the original finding, is unavailable. All four such effects fall outside the estimation sample because their paradigms admit no bare task. And fourteen units is far too few for the interaction to be decisive. The analysis therefore cannot distinguish no literature influence from literature influence that includes replication outcomes.

E ADDITIONAL SETUP DETAILS

Cross-family scoring adaptations. Parsers for 2 of the 42 uniform-protocol effects needed GPT-specific label fallbacks, the parsers for the 19 pilot-wave effects were rewritten with a more tolerant label search, and the scoring keys that map answers to the claimed direction are derived deterministically from the protocol bank. All adaptation code is in the repository (`code/`).

E.1 TRANSCRIPTION OF PARADIGMS

Each effect was transcribed into a prompt template with the original's conditions, stimuli, and response scale, working from the original method section or the replication report where those were available (20 of 122 templates) and otherwise from the paradigm's canonical description in the literature, and reconstructions include cases of simplified or invented stimulus items, recorded per

template. Because wording choices could drive results, each effect was transcribed twice by independent passes, and the analysis treats the two transcriptions as parallel forms. Their consistency is moderate. Across the 72 effect-model pairs of the uniform-protocol wave with both variants measured (minimal framing), the two variants' simulated effects correlate at 0.68 on the Fisher-z scale, agree in sign in 74% of pairs, and differ by a median $|\Delta r| = 0.13$ (IQR 0.01–0.45). The unit-level analysis aggregates the two variants. Transcription fidelity is documented per effect in the protocol bank (`protocols/P2_master_protocol_bank.json`, fields `evidence_tier` and `notes`), including which effects were transcribed from primary sources and which from replication reports.

E.2 Parsing and effect size computation

Each effect has a parser for its response format and is scored with the original study's statistical test, yielding a simulated effect size on the correlation scale with a bootstrap confidence interval. The overall parse rate is 91–95% per wave. Cells with parse rates below 60% are flagged and excluded in a sensitivity analysis. All parsers and scoring code are in the GitHub repository (`code/`). Raw model responses are available from the authors on request.

Simulated participants per experiment. Table 1 counts one model call as one simulated participant. The main protocol (the two Claude models plus the cross-family runs) totals about 43,000, and all experiments together total about 68,500.

Table 1: Simulated participants (model calls) per experiment.

Experiment	Models	Calls
Main protocol, participant framing (three waves)	Sonnet 5, Sonnet 4.5	30,624
Bare-task battery	Sonnet 5, Sonnet 4.5	2,580
Cross-family protocol and bare battery	GPT-5.6-terra	4,950
Cross-family protocol and bare battery	DeepSeek-v4-flash	4,950
Main protocol subtotal		43,104
Ceiling rerun with 0–100 scales	two Claude models	4,860
Ceiling rerun with 0–100 scales	GPT, DeepSeek	4,860
Positive controls	two Claude models	9,120
Neutral system prompt	two Claude models	6,080
Construct battery, two batches	two Claude models	480
Total		68,504

F Robustness details

Likelihood treatments of ceiling and transcription noise. Replacing the normal likelihood with a censored likelihood for the 20 ceiling units (the observation is then only that z_{sim} exceeds $\operatorname{arctanh}(0.95)$ in the observed direction) gives $\beta_{\text{pub}} = 0.70$, $[-0.40, 1.73]$, essentially the main estimate, since knowing only that these units are large is compatible with the positive trend, whereas remeasuring them with headroom scales (Appendix L) removes it. Giving z_{sim} a per-unit measurement variance $(z_A - z_B)^2/4$ from the two transcription variants gives $\beta_{\text{bare}} = 0.45$, $\beta_{\text{rep}} = 1.01$, $\beta_{\text{pub}} = 0.59$, $[-0.39, 1.56]$, with the unit-level residual dropping from 0.64 to 0.55 as the transcription noise moves into the measurement term. Fitting code and posterior summaries are in the repository (`code/fit_model_treatments.py`, `analysis_tables/model_treatments_posteriors.csv`).

Cross-family estimates. We reran the full protocol bank (61 effects, both transcriptions, 30 simulated participants per transcription, cycled across conditions) and the bare-task battery on GPT-5.6-terra and DeepSeek-v4-flash, about 5,000 additional simulated participants each, with provider default sampling and a 2,000-token completion cap (scoring details in Appendix E). Fitting the same model to each family gives, for GPT, $\beta_{\text{bare}} = 0.58$ $[0.36, 0.79]$, $\beta_{\text{rep}} = 0.93$ $[0.27, 1.55]$, and $\beta_{\text{pub}} = 0.46$ $[-0.33, 1.22]$, and for DeepSeek 0.50 $[0.33, 0.68]$, 0.89 $[0.35, 1.40]$, and 0.16 $[-0.46, 0.76]$. In all three families both non-gap coefficients are positive and the gap coefficient is

inconclusive. The bare-task weight differs across families (0.37, 0.58, 0.50), consistent with each family having its own bias profile rather than sharing one literature signal.

Summary of further checks. Fixed effects for paradigm class give $\beta_{\text{pub}} = 0.31$ $[-0.66, 1.30]$. A rerun of the estimation sample under a system prompt that does not mention replication gives 0.51 $[-0.43, 1.45]$ against 0.65 for the original prompt scored the same way, with the two sets of simulated effects correlating at 0.87. A hierarchical prior on the latent anchors gives 0.73 $[-0.42, 1.85]$, and simulation-based calibration of the model passes. Censoring the ceiling units and adding transcription measurement variance leave all three coefficients essentially unchanged, while remeasuring the 14 ceiling effects with 0–100 scales moves the gap coefficient below zero in all three model families (Appendix L). Dropping the bare-task control raises the replication coefficient from 1.09 to 1.46 on the same units, the expected omitted-variable bias when the model's own bias correlates with the human effect, with overlapping intervals.

All subset fits in this section are reproduced by `code/fit_sensitivity.py` (`analysis_tables/sensitivity_posteriors.csv`). Excluding units at the response ceiling changes all three coefficients, not only the fingerprint: $\beta_{\text{bare}} = 0.16$ $[-0.03, 0.33]$, $\beta_{\text{rep}} = 0.70$ $[0.21, 1.21]$, and $\beta_{\text{pub}} = 0.25$ $[-0.60, 1.07]$ on the 51 non-ceiling units. The two non-fingerprint coefficients keep their sign and remain the dominant sources, and their attenuation reflects that ceiling units are concentrated among the strongest model-bias effects.

The two remaining pre-specified analyses give the same picture. Using the persona-inclusive outcome instead of the minimal-framing outcome leaves all three coefficients essentially unchanged (0.38, 1.00, 0.57). Excluding cells with parse rates below 60% likewise changes nothing. Dropping the bare-task control from the main model, on the same 71 units, inflates the human-effect coefficient from 1.09 to 1.46, 95% interval $[0.67, 2.19]$ (`analysis_tables/nobare_posteriors.csv`). This isolates the effect of removing the control, since sample and specification are otherwise identical. A secondary analysis without the bare-task control on all 60 effects with a computable simulated effect (118 units) gives 1.54, folding in the sample change as well. Without the bare-task control, the part of the model's own bias that points the same way as the human effect is credited to the human effect.

G Scale, transcription, prompt, and similarity checks

This section reports additional analyses of the estimation sample. All are reproduced by `code/robustness_checks.py`, with results in `analysis_tables/robustness_checks.json`.

Level calibration and ranking. Table 2 collects all level statistics of the estimation sample in one place. The intercept of the main model is 0.33 $[-0.08, 0.74]$, and the simulated effect exceeds the replication estimate by 0.66 on average ($[0.41, 0.91]$). The human studies report only standardized effects, so all comparisons use the standardized scale.

Table 2: Level of simulated and human effects in the estimation sample (71 units).

	Simulated	Original	Replication
Mean z	0.81	0.38	0.15
Mean $\lvert z \rvert$	1.03	0.47	0.27
Median $\lvert r \rvert$	0.53	0.38	0.11
Share of units with $\lvert z_{\text{sim}} \rvert$ larger	–	61%	65%

At the effect level (38 effects, averaging the two models), the marginal Spearman correlations of the simulated effect are 0.45 with the replication estimate, 0.51 with the original estimate, and 0.50 with the bare-task effect. Partial Spearman correlations, each controlling for the other two sources, are 0.35 $[-0.04, 0.66]$ for the replication estimate, 0.35 $[-0.03, 0.67]$ for the bare-task effect, and 0.24 $[-0.12, 0.55]$ for the gap (bootstrap intervals). All three partial intervals include zero, so the rank-based evidence for each source is weak on its own.

Personas and variance. Comparing the 145 cells simulated both with minimal framing and with demographic personas, the median absolute simulated effect is 0.67 without personas and 0.57 with

them, and the share of cells at the ceiling is 34% and 27% (Wilcoxon $p = .42$). Personas lower effect sizes slightly but do not restore human-like within-condition variance.

Cross-paired transcriptions. The bare-task bank was built from the first transcription of each effect, so the first transcription shares wording with the bare task and the second does not. Table 3 fits the main model with the simulated effect taken from each transcription alone and from their average, on the units where both transcriptions yield a simulated effect. For the Claude models, the two transcriptions agree only moderately (Fisher-z correlation 0.68), and each coefficient is clearly positive only when the transcriptions are averaged. With the shared transcription the replication interval includes zero, and with the independent one the lower end of the bare-task interval touches zero. On the same 66 units the bare-task coefficient is 0.75 under the shared transcription, 0.49 for averaged cell values, and 0.44 for the pooled unit values of the main analysis, so the difference comes from the transcription (the main estimate of 0.37 uses all 71 units). For GPT and DeepSeek the transcriptions agree more closely (0.82 and 0.81), and both coefficients stay positive under either transcription. The gap coefficient is unaffected throughout.

Table 3: Coefficients by transcription (posterior mean and 95% interval). First = shared with the bare task, second = independent.

Family (units)	Simulation from	β_{bare}	β_{rep}	β_{pub}
Claude (66)	average	$0.49\ [0.27, 0.71]$	$0.99\ [0.22, 1.72]$	$0.46\ [-0.50, 1.43]$
	first	$0.75\ [0.50, 0.98]$	$0.60\ [-0.15, 1.34]$	$0.45\ [-0.52, 1.36]$
	second	$0.28\ [-0.01, 0.58]$	$1.24\ [0.36, 2.10]$	$0.43\ [-0.66, 1.52]$
GPT (39)	average	$0.54\ [0.33, 0.77]$	$0.96\ [0.28, 1.63]$	$0.49\ [-0.30, 1.30]$
	first	$0.62\ [0.42, 0.82]$	$0.98\ [0.35, 1.59]$	$0.50\ [-0.27, 1.26]$
	second	$0.47\ [0.19, 0.74]$	$0.91\ [0.08, 1.72]$	$0.44\ [-0.52, 1.44]$
DeepSeek (41)	average	$0.53\ [0.34, 0.71]$	$0.88\ [0.31, 1.46]$	$0.27\ [-0.43, 0.94]$
	first	$0.55\ [0.39, 0.76]$	$0.82\ [0.29, 1.42]$	$0.48\ [-0.23, 1.14]$
	second	$0.48\ [0.25, 0.70]$	$0.87\ [0.19, 1.57]$	$0.09\ [-0.72, 0.91]$

Direction where the model shows no bias. A replicated human effect is taken as real when the replication estimate differs from zero at the 5% level, and the model is taken to show no bias when $|r_{\text{bare}}| < 0.10$. Table 4 counts units whose simulated effect has the same sign as the replication estimate. Where the model shows no bias, agreement is 16 of 29 units across the three families (55%, 95% CI 36% to 74%), not distinguishable from the one half expected by chance (binomial $p = .63$ for the 17 Claude units). The interval is wide, so moderate agreement is not excluded either. Where it does, agreement is 80% to 100%, which is partly by construction because the bare task and the simulation share stimuli. A threshold of 0.20 gives the same picture (`analysis_tables/direction_subset.csv`).

Table 4: Sign agreement between simulated and replication effects among units with a real replicated effect.

Family	Model shows no bias	Model shows the bias
Claude	10/17	24/30
GPT	3/5	16/19
DeepSeek	3/7	19/19

Paradigm class. Paradigm types could differ in both gap size and simulability. The clearest case is social priming, which has large gaps and near-zero simulated effects. Social priming has no bare task and is therefore outside the estimation sample. Within it, median gaps range from -0.01 to 0.44 across classes, and no class combines large gaps with near-zero simulated effects (Table 5). Adding class fixed effects gives $\beta_{\text{bare}} = 0.40\ [0.19, 0.62]$, $\beta_{\text{rep}} = 0.85\ [0.09, 1.59]$, and $\beta_{\text{pub}} = 0.31\ [-0.66, 1.30]$. Restricting to the largest class, judgment and decision biases (27 units, 15 effects), gives $\beta_{\text{pub}} = -0.24\ [-1.68, 1.25]$.

Differences between sensitivity subsets. Adding an indicator for ceiling units and its interaction with the gap gives an interaction of $0.19\ [-0.77, 1.23]$. The corresponding interaction for the uniform-protocol wave is $-0.90\ [-2.24, 0.48]$, with posterior probability 0.10 of being positive. Neither subset difference is established.

Table 5: Gap D_{pub} and mean simulated effect by paradigm class in the estimation sample.

Class	Effects	Median gap	Gap range	Mean z_{sim}
Judgment and decision biases	15	0.24	−0.17 to 0.43	0.41
Social judgment	5	0.20	−0.16 to 0.76	0.73
Trait inference	4	0.23	0.17 to 0.58	1.53
Moral judgment	4	0.44	−0.10 to 0.53	1.01
Questionnaire	4	−0.01	−0.17 to 0.38	−0.05
Linguistic intuition	2	0.23	−0.47 to 0.93	2.65
Social imagery	2	0.34	0.33 to 0.35	1.69
Incentive	2	0.29	0.05 to 0.53	0.66

Neutral system prompt. The system prompt of the main runs mentions replication (Appendix N). We reran the estimation sample on both Claude models with a prompt that mentions neither replication nor metascience (6,080 rollouts, 40 simulated participants per transcription, same participant framing), scored with the released pipeline. Scoring both runs from the cell tables in the same way, the neutral prompt gives $\beta_{\mathrm{pub}} = 0.51$ $[-0.43, 1.45]$ against 0.65 $[-0.38, 1.65]$ for the original prompt on the same 70 units, with $\beta_{\mathrm{bare}} = 0.68$ and $\beta_{\mathrm{rep}} = 0.86$. The two sets of simulated effects correlate at 0.87, their mean differs by 0.01, and their mean distance to the replication estimate is 1.04 for the original prompt and 1.02 for the neutral one. If the word had pushed simulations toward replications, the neutral prompt would have raised the gap coefficient and moved simulated effects away from the replication anchor, and it did neither.

Conventional validation. Regressing the simulated effect on the original estimate alone gives a slope of 1.51 $[0.84, 2.17]$, and on the replication estimate alone 1.43 $[0.86, 1.99]$ (cluster-robust). Correlations are 0.48 and 0.47, and sign agreement is 70% and 66%. By these standard criteria the simulation would pass against either reference.

Item similarity in the construct battery. For each construct we embedded every item with two sentence-embedding models (all-mpnet-base-v2 and all-MiniLM-L6-v2), averaged the item vectors, and took the cosine similarity of the two constructs in each pair, with a TF-IDF cosine as a lexical alternative. Similarity is 0.26 for famous and 0.16 for undiscussed pairs (Mann-Whitney $p = .25$). In the pair-level regression the fame effect is $+0.156$ without and $+0.164$, $+0.164$, and $+0.167$ with the three similarity covariates (all $p \leq .008$), and no similarity coefficient approaches significance. Embedding similarity measures how alike the items are worded, not how the model represents the association between the two constructs, so this check rules out only a wording account. A commonsense account, in which famous correlations are simply the intuitively plausible ones, would need lay predictions of each correlation as a covariate, which we do not have. One construct, gender, is a persona attribute rather than an item and is represented by its persona sentence.

H What drives simulated effects

Simulated effect sizes are large where the model shows the bias when asked directly (marginal and partial associations are compared in Appendix C). Judgment biases such as the side-effect effect (Knobe, 2003) and the retrospective gambler's fallacy (Oppenheimer & Monin, 2009) produce large simulated effects, and the model produces them bare. Social priming and minimal-incentive manipulations, which the model does not show bare, produce simulated effects near zero whatever the literature reports. For paradigms with a verifiable answer the model often gives the true answer rather than the biased one, for example judging letter position frequencies correctly where humans show the availability bias (Tversky & Kahneman, 1973). Framing is not a neutral wrapper either. The model shows no warm-cold trait inference (Asch, 1946) bare ($r \approx 0$) but a ceiling-level effect under participant framing, so participant framing can release behavior the model does not produce as itself.

I Positive controls

Each control adds one paragraph before the simulation prompt of every estimation-sample protocol (both transcriptions, 30 simulated participants per transcription, both Claude models, the main

system prompt). For the original-copying simulator the paragraph reads, for the warm-cold effect, "Background for this simulation. The original published study (Asch, 1946; n = 166) found the following: People described with a "warm" personality trait were rated as more generous than people described with a "cold" personality trait. The published effect size was r = 0.65. Simulate the participants so that the results reproduce this published finding and its effect size." The replication-copying paragraph gives the replication's sample size, direction (or "little or no effect"), and effect size in the same format. The one-sentence finding for each effect was generated from the protocol bank's scoring notes and checked by hand. Of 9,120 rollouts, 8,974 returned a response, and responses were scored with the released pipeline.

Table 6 reports the main model for each simulator on the same 71 units. Both simulators still produce effects far larger than the numbers they are given (mean simulated z of 1.39 and 1.08, against 0.38 and 0.15 for the original and replication estimates), so even an explicit instruction does not overcome the collapsed variance. A stacked model with simulator interactions gives a gap-by-simulator interaction of 0.58 $[-0.10, 1.25]$ ($P > 0 = 0.95$). Regressing the within-unit difference between the two simulators on the gap, with standard errors clustered by effect, gives a slope of 0.66 $[0.08, 1.24]$ ($p = .027$), so a paired design shows a signal of the planted difference where separate fits do not. This rests on one control pair and an interval whose lower end is close to zero.

Table 6: Main model fitted to the two positive-control simulators.

Simulator	β_{bare}	β_{rep}	β_{pub}
Original-copying	0.39 $[0.19, 0.59]$	1.39 $[0.56, 2.19]$	0.79 $[-0.25, 1.84]$
Replication-copying	0.35 $[0.18, 0.52]$	1.53 $[0.80, 2.25]$	0.22 $[-0.73, 1.19]$

J Effects with at least one primary-source transcription, descriptively

Eight of the 38 estimation-sample effects have at least one prompt template transcribed from primary-source materials (16 effect-model units). For three of the eight (Eyal, Genschow, Oppenheimer) only one of the two transcription variants is primary-source, and the simulated values below aggregate both variants, so this is not a pure primary-source result. The subset is too small to identify the three-coefficient model, so we report it descriptively and draw no parametric inference from it. Table 7 lists the four quantities per effect. In Fisher-z distance, the simulated effect is closer to the original estimate in 8 of the 16 units and closer to the replication estimate in the other 8, matching the inconclusive full-sample verdict, and the simulated direction agrees with the replication direction in 12 of 16 units. The subset shows no systematic preference for either anchor, but it is too small to establish robustness to transcription fidelity.

Table 7: The eight estimation-sample effects with primary-source templates, namely human anchors, bare-task effect, and simulated effect (mean over the two models where both are measured).

Effect	r (orig.)	r (repl.)	r (bare)	r (sim.)
Aryani2020 word arousal	0.42	0.72	1.00	1.00
Cooney2016 fairness	0.57	0.12	0.68	0.44
Eyal2008 moral distance	0.38	0.03	−0.03	0.07
Genschow2017 free will attribution	0.09	0.07	0.21	0.00
Glikson2017 smiley competence	0.34	0.11	−0.64	0.10
Klink2000 sound symbolism	0.82	0.22	1.00	0.99
Oppenheimer2009 retrospective gambler	0.33	0.01	−0.09	0.58
Tamir2012 self-disclosure	0.49	0.00	−0.99	0.82

K Design precision and the effects without a bare task

Power simulation. To quantify what the estimation sample can and cannot detect, we simulated data from the fitted model at the observed design (71 units, the observed z_{bare}, z_r, and D_{pub} values, effect-level intercept spread and residual scale at their posterior means, $\beta_{\text{bare}} = 0.37$, $\beta_{\text{rep}} = 1$) while varying the true β_{pub}, and refit each simulated dataset with a mixed model (500 simulations per value). The mixed-model approximation is slightly conservative. On the real data its interval for

β_{pub} is somewhat wider than the Bayesian measurement-error model's. The probability that the 95% interval excludes zero is 0.07 at a true coefficient of 0 (approximately consistent with the nominal 0.05 rate given 500 simulations), 0.06 at 0.10, 0.16 at 0.50, 0.31 at 0.75, and 0.46 at 1.00. The probability of an interval entirely inside the equivalence region $(-0.10, 0.10)$ is 0.00 at every value tested. The same simulation with the actual Bayesian measurement-error model as the fitting model (100 simulated datasets per value, `code/calibration_and_power.py`) gives detection rates of 0.05, 0.08, and 0.38 at true values 0, 0.5, and 1.0, and an equivalence rate of 0.00 throughout, in line with the approximation. Two consequences follow. Even a gap coefficient of one would be detected less than half the time, and the pre-specified evidence-of-absence verdict was effectively unreachable at this design size. The inconclusive verdict on β_{pub} therefore reflects limited precision rather than evidence against a fingerprint, and the informative results of the design are the two coefficients it does estimate clearly. Precision is limited jointly by the number of bare-measurable effects (38) and by the spread of D_{pub} relative to the effect-level residual scale.

What a decisive design would take. Rerunning the simulation on scaled versions of the same design (resampling the observed covariate distribution, two model rows per effect) gives the design size needed for a decisive test. Detection power for a full-transfer fingerprint ($\beta_{\text{pub}} = 1$) reaches 0.68 at 75 bare-measurable effects and 0.93 at 150. For a moderate fingerprint (0.5) it is 0.46 at 150 and 0.71 at 300. The equivalence verdict remains out of reach even at 300 effects, at every true value tested, because the ± 0.10 region is small relative to the effect-level residual scale (0.76): concluding absence requires designs that shrink the residual, for example response scales with headroom and trial-level outcome models, not only more effects. The practical summary is that the intersection of current replication databases with text-simulable paradigms supports detecting only a large fingerprint, and a decisive test needs roughly 150 or more bare-measurable effects for detection, and lower residual noise before evidence of absence is attainable at all. Simulation code and both result tables are in the repository (`code/power_simulation.py`).

The effects without a bare task. The 47 effect-model units without a bare measurement cover 26 effects: 22 with no bare-measurable unit at all (18 whose paradigm admits no bare task plus four where the bare measurement failed for both models) and four with a bare measurement for only one model. These units could in principle hide a fingerprint. Two descriptive checks give no indication of that. Their publication gaps are distributed like the estimation sample's (median D_{pub} 0.27 versus 0.24, Mann-Whitney $p = 0.22$), so the estimation sample is not selected toward small gaps. And among these units the simulated effect correlates with the replication estimate ($r = 0.32$) but not with the publication gap ($r = 0.01$), the same qualitative pattern the estimation sample shows. These are marginal associations without the bare-task control and support no causal reading. They show only that nothing in the excluded effects visibly contradicts the main-sample picture.

L Exploratory rerun of the ceiling effects with headroom scales

The 14 effects with at least one unit at $|r_{\text{sim}}| > .95$ were rerun with the response format changed to a continuous 0–100 scale and everything else in the protocols unchanged (both transcription variants, 60 simulated participants per variant cycled across conditions, and a per-condition bare battery with five wording seeds, on both Claude models. Scoring is a generic numeric parser with the per-effect group contrasts fixed in advance in the protocol bank). This rerun is exploratory, because the response format changes the measurement, so the new values are not a retest of the same estimand.

Two results. First, headroom alone does not remove the ceiling. Only 9 of the 20 originally ceiling unit values leave the ceiling under the 0–100 scale. The reason is measurable and diagnostic. Within an experimental arm the ratings are nearly deterministic, with a median within-arm standard deviation of 1.8 points on the 100-point scale (80% of arms below 3, pooled over the two Claude models), so any between-arm separation saturates the effect size. The ceiling is a property of collapsed response variance in the simulated participants, not of the response scale. Second, replacing the 27 available effect-model units' simulated and bare values with the headroom measurements and refitting the main model moves the fingerprint coefficient to $\beta_{\text{pub}} = -0.24\ [-0.91, 0.44]$, $P(\beta_{\text{pub}} > 0) = 0.23$, while the other two sources sharpen ($\beta_{\text{bare}} = 0.48\ [0.33, 0.64]$, $\beta_{\text{rep}} = 1.09\ [0.61, 1.62]$) and both residual scales shrink by about a third (effect-level $0.76 \rightarrow 0.48$, unit-level

$0.59 \to 0.42$). This is consistent with the sensitivity analyses in the main text. The positive trend in the fingerprint coefficient is carried by ceiling-compressed units, and reducing the compression and replacing the ceiling-selected measurements removes the positive trend. It also bears on the design size of Appendix K: response-scale headroom delivers a real residual reduction, but the floor is set by the collapsed between-participant variance of simulated respondents, so decisive designs need variance-generating protocols, not only wider scales.

Both findings replicate across families. The same rerun on GPT-5.6-terra and DeepSeek-v4-flash shows the same variance collapse in graded form, with a median within-arm standard deviation is 2.3 points for GPT (68% of arms below 3, and only 1 of its 7 ceiling effects escapes the ceiling) and 4.1 points for DeepSeek (28% below 3), which also has the fewest cells at ceiling. Refitting each family's model with the remediated values moves the fingerprint coefficient in the same direction in all three families: -0.24 $[-0.91, 0.44]$ for the Claude models, -0.54 $[-1.43, 0.26]$ for GPT, and -0.31 $[-0.86, 0.24]$ for DeepSeek, with $P(\beta_{\text{pub}} > 0)$ of 0.23, 0.11, and 0.13, while the own-bias coefficient stays positive and stable (0.48, 0.49, 0.46). Across families, the positive point estimate of the fingerprint coefficient is not seen after ceiling remediation, although the intervals remain wide. The rerun kit, the per-cell and hybrid unit tables, the refit script, and the full posteriors are in the repository (`ceiling_kit/`, `analysis_tables/CEIL_*`).

M Model specification, priors, and diagnostics

The pre-specified decision rule fixes the interpretation of each coefficient pattern in advance, the primary sample (effects with a defined bare task), the primary framing condition (minimal participant framing without demographic personas), and four sensitivity analyses (excluding response-ceiling units, excluding low-parse cells, adding the persona units, and restricting to replications with $n_r \geq 500$).

The main model is a Bayesian measurement-error regression fit with PyMC. The two human anchors enter as latent true values with known sampling variances, $z_o^{\text{true}} \sim \mathcal{N}(z_o, \sigma_o)$ and $z_r^{\text{true}} \sim \mathcal{N}(z_r, \sigma_r)$, where $\sigma^2 = 1/(n-3)$ from the recorded sample sizes, while z_{sim} and z_{bare} enter as point values. Point values for z_{bare} could attenuate β_{bare} and shift part of the model's own bias into β_{rep}, since z_{bare} and z_r correlate at 0.39. Giving z_{bare} a latent true value with a population prior and sampling variance $1/(n_{\text{bare}} - 3)$ (median 30 scored responses), and inflating that variance up to fourfold for wording-seed variation, gives β_{bare} from 0.39 to 0.43 and β_{rep} from 1.12 to 1.16, so β_{rep} does not absorb bare-task error (`code/bare_measurement_error.py`). A measurement-error treatment of z_{sim} also leaves the results unchanged (Appendix F). (Parameterizing the latent value by the observed estimate is shorthand for the posterior of the true value under a flat prior given the observed estimate and its known sampling variance.) There is one latent anchor pair per effect (38 pairs), shared by that effect's two model rows. The outcome model is

$$z_{\text{sim}} \sim \mathcal{N}\big(\beta_0 + \beta_{\text{bare}} z_{\text{bare}} + \beta_{\text{rep}} z_r^{\text{true}} + \beta_{\text{pub}}(z_o^{\text{true}} - z_r^{\text{true}}) + \beta_m M + u_e,\ \sigma\big)\,,$$

with M an indicator for the model (Sonnet 4.5 versus Sonnet 5) and u_e a random intercept per effect. Priors are $\mathcal{N}(0,1)$ on all regression coefficients, $u_e \sim \mathcal{N}(0, \text{sd}_{\text{eff}})$ with $\text{sd}_{\text{eff}} \sim \text{HalfNormal}(0.5)$, and $\sigma \sim \text{HalfNormal}(1)$. Sampling used NUTS with 4 chains, 1,000 tuning and 1,500 kept draws per chain, and target acceptance 0.95. Diagnostics are clean, with zero divergent transitions, maximum $\widehat{R} = 1.004$ over all parameters, and minimum bulk effective sample size 2,384. The fitting script and the posterior draws are in the repository (`code/fit_main_model.py`, `analysis_tables/P2_main_model_idata.nc`).

Two known approximations. The $1/(n-3)$ variance is exact for Pearson correlations and approximate for effect sizes converted from other statistics. The outcome z_{sim} and the covariate z_{bare} enter as point estimates, and their sampling error is absorbed by the residual rather than modeled, which is conservative for the coefficient on z_{bare} (attenuation) and adds noise to the outcome.

Structural prior and calibration. A hierarchical variant, in which the true replication effects and true gaps are drawn from normal population distributions with estimated means and spreads and the observed estimates enter as measurements, gives $\beta_{\text{bare}} = 0.38$ $[0.18, 0.57]$, $\beta_{\text{rep}} = 1.05$ $[0.26, 1.80]$, and $\beta_{\text{pub}} = 0.73$ $[-0.42, 1.85]$ (estimated mean gap 0.24, spread 0.25). For simulation-based calibration we drew 200 parameter sets from the prior, simulated outcomes at the observed design,

refit each, and recorded the rank of the true value among 199 posterior draws. Rank histograms are consistent with uniformity for all three coefficients (χ^2 test $p = 0.66$, 0.42, and 0.23 for β_{bare}, β_{rep}, and β_{pub}), and nominal 95% intervals cover the truth in 94.5%, 96%, and 97% of fits. Both analyses use a NumPyro implementation of the same model (`code/nm_models.py`), which reproduces the PyMC posterior within Monte Carlo error.

N PROMPT FORMATS

All prompts are built from per-effect templates in the released protocol bank, and the wrapper components are constant across effects. The system prompt for all simulation rollouts is:

```
You are powering a social-science simulation that generates
synthetic survey-participant responses for metascience
research on replication.  Respond as the described
participant would, in their voice, following the requested
output format exactly.
```

This prompt mentions replication. A rerun of the estimation sample under a neutral system prompt that does not mention it leaves the gap coefficient essentially unchanged (Appendix G).

The minimal participant framing prepended to each simulation prompt is:

```
For the purposes of this research simulation, answer as
a typical human study participant would -- a plausible,
ordinary adult.  Stay in that role for every question.
```

An example effect template (moral brightness, one of two conditions):

```
Please recall a time when you behaved unethically.  Describe
it in two sentences.  Then rate how bright the room you
are in right now seems to you, from 1 (very dim) to 7 (very
bright).  Reply in the format BRIGHTNESS: <number>.
```

Bare-task prompts strip the system prompt and participant framing, remove laboratory scene-setting sentences from the template, and prepend "Answer the following questions directly and honestly, as yourself." Each bare task was rephrased in five wording seeds, and all seed rephrasings are in the GitHub repository (`protocols/bare_task_bank.json`) and all 122 simulation templates in `protocols/P2_master_protocol_bank.json`.

O REFUSAL CODING AND CONDITION-TYPE RATES

Refusals are data, not noise. In the main estimates refused rollouts are missing responses. Cells are scored from answered rollouts, and heavily refused cells fail the computability rules and drop out as units. Refusal is rare in the estimation sample itself (mean 1.9% per unit, maximum 19%).

Figure 7 breaks down refusal rates by condition type. Among ten effects whose arms differ only in self-implication, the aversive arm is refused at 24.0% versus 12.2% (odds ratio 3.46, 95% CI $[2.30, 5.20]$, effect fixed effects, cluster-robust). Against the no-manipulation baseline of 3.8%, the largest odds ratios are unethical self-recall 36.8, ethical self-recall 11.7, and secret recall 9.8. Refusal was coded from raw responses by fixed rules (hard API refusals, plus soft refusals that decline in text without answering), with the full condition-type mapping (`analysis_tables/S1_condition_map.csv`) and per-cell refusal rates (`analysis_tables/S1_refusal_units.csv`) in the GitHub repository. The logistic models use cluster-robust standard errors by effect. Five condition types were dropped from the condition-type model for complete separation (zero refusals).

P A TWO-LEVEL PICTURE OF LITERATURE INFLUENCE

Simulated effect sizes covary with the model's bare-task effects and replication estimates, while at the level of construct structure famous relations are selectively exaggerated. The two are consistent if

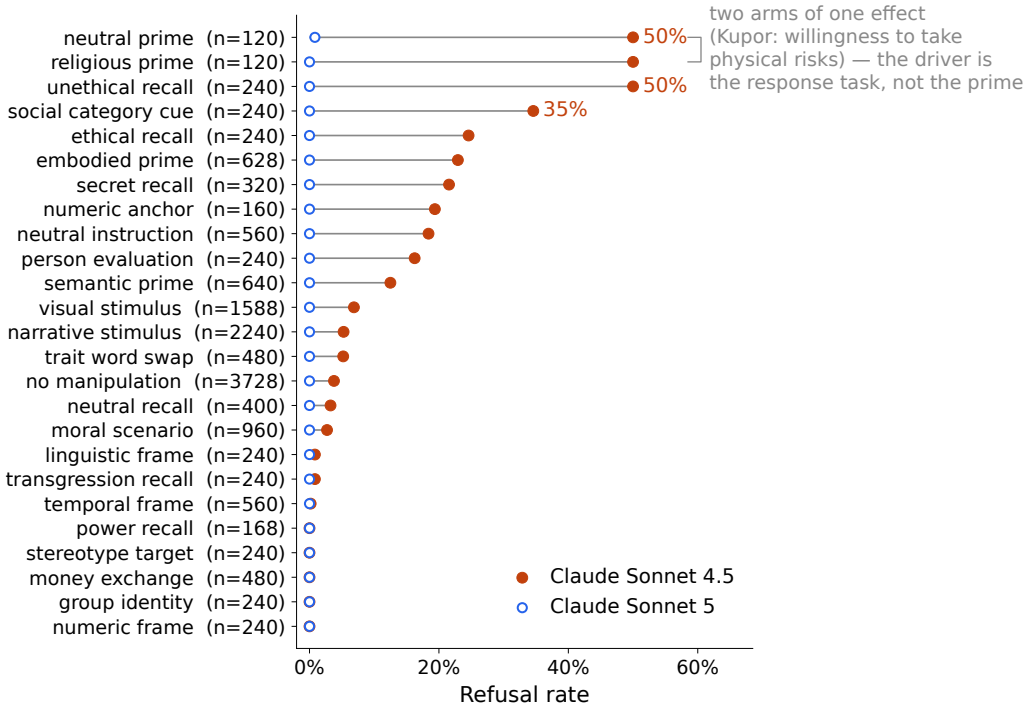


Figure 7: Sonnet 4.5 refusal rate by condition type, across all 25 condition types in the protocol bank. Refusal is highest where the response task asks for a transgression, a belief report, or a judgment of a named person.

what the model absorbs from the literature is not statistics but assertions, which relations exist and in which direction, with magnitude reconstructed at generation time. This explains why exaggeration appears where the literature is loud rather than merely inflated, and predicts that literature influence scales with assertion frequency, not published magnitude, a corpus-level test we leave open.

Q Extended Related Work

LLMs as simulated participants. Simulation of human respondents with language models was established by work showing that conditioning on demographic personas reproduces the response distributions of survey subpopulations (Argyle et al., 2023), that classic experimental results can be replicated with simulated subject pools (Aher et al., 2023), and that models can stand in for economic agents (Horton et al., 2023). The largest replication effort to date reports that scenario-based experiments reproduce direction well while effect sizes come out systematically larger than in humans (Cui et al., 2024). Persona-based simulation has since scaled to interview-grounded agents (Park et al., 2023; 2024) and to panel-scale digital twins (Toubia et al., 2025; Peng et al., 2025). Our design differs from all of these in what it asks: not whether the simulation matches a human result, but which of several competing human quantities it matches when they disagree.

Skepticism and validation methodology. A parallel line documents failure modes, such as opinion distributions that track some demographics and miss others (Santurkar et al., 2023; Bisbee et al., 2024), flattened portrayals of identity groups (Wang et al., 2025), survey response patterns unlike human ones (Tjuatja et al., 2024), and conceptual fallacies in substitution claims (Lin, 2025). Recent methodological work formalizes what validating a simulation should mean (Hullman et al., 2026). Within that framing, our contribution is a decomposition that places the simulated quantity against two competing human benchmarks plus the model's own baseline, rather than scoring agreement with a single reference. At feasible scale it serves as a diagnostic across effects rather than as a score for one simulator.

Models' own cognitive biases. That language models exhibit human-like judgment biases when tested directly is established for several paradigms (Binz & Schulz, 2023; Hagendorff et al., 2023), has been systematized into benchmarks (Coda-Forno et al., 2024), and extends to anchoring (Valencia-Clavijo, 2025) and moral judgment, where model biases can exceed human ones (Cheung et al., 2025). These measurements are the precedent for our bare-task battery. What is new here is using them as a regressor that competes with literature-derived quantities to explain simulated effect sizes.

Training data memorization and contamination. Memorization of repeated training content scales with model size and duplication (Carlini et al., 2023), and memorization is separable from exploitation of the memorized content (Magar & Schwartz, 2022), and contamination of evaluation data is a recognized threat to benchmark validity (Cheng et al., 2025). Our question is the social science analogue of benchmark contamination. Published effect sizes are the test set a simulated participant could have memorized. The effect-size results cannot establish whether memorized estimates are exploited, while the structure-level results suggest what is retained is closer to asserted relations than to numbers, echoing the memorization-exploitation distinction. We cannot measure exposure directly, since membership-inference and pretraining-data detection methods require token-level probabilities or model weights, which closed-weight models do not expose, so exposure to the original papers is assumed probable for these decades-old, heavily cited findings rather than verified per paper.

Publication bias and replication. The empirical basis for our identifying contrast comes from large replication projects (Open Science Collaboration, 2015; Camerer et al., 2018; Klein et al., 2014; 2018), the databases that aggregate them (Röseler et al., 2024), and analyses of why published and replicated estimates diverge (Ioannidis, 2005; Simmons et al., 2011; Rosenthal, 1979; Kvarven et al., 2020; Bartoš et al., 2023). We take this literature's central quantity, the original-replication gap, and repurpose it as a probe of what language models learned from the literature that reports it.

R STRUCTURE EXPERIMENT: PER-PAIR DETAIL

Battery. The 22 constructs cover personality traits, cognitive ability, economic preferences, and consumer attitudes from the panel's psychometric waves, plus political ideology and religiosity from its demographic items. Both models answered the full 99-item battery as 120 synthetic adult respondents each.

Figure 8 shows the first batch of the structure experiment at the level of individual construct pairs. For each of its 20 pairs (40 pair-model values, with the 16 second-batch pairs tabulated in Appendix S) it plots the human panel correlation next to the two simulated correlations. Two patterns are visible that the aggregate indices in Section 5.6 summarize. On literature-famous pairs the simulated correlations tend to sit farther from zero than the human value, in the direction the literature reports. On undiscussed pairs the simulated values scatter around the human value without a systematic outward shift, including several sign disagreements. Pairs marked n.d. yielded no computable simulated correlation for one model because of refusals. Construct score reliability in the simulated samples is essentially equal across groups (mean intraclass correlation 0.66 for famous, 0.68 for undiscussed constructs), so the group difference is not carried by differential measurement reliability.

S THE FAME-BY-CONTENT DESIGN

The design crosses fame (famous or undiscussed) with content (political or nonpolitical). The first batch of 20 pairs (Appendix R) supplies the famous political and undiscussed nonpolitical cells, plus one famous nonpolitical pair. The second batch supplies eight famous nonpolitical and eight undiscussed political pairs, selected from the same human panel and magnitude-matched pair by pair (ranked by $|r_{\text{human}}|$; mean $|\Delta r| = 0.011$, max 0.050). Famous pairs are review-level or textbook associations with no political content (numeracy with financial literacy, age with neuroticism, conscientiousness with educational attainment, neuroticism with maximizing tendency, income with financial literacy, openness with fluid ability, agreeableness with

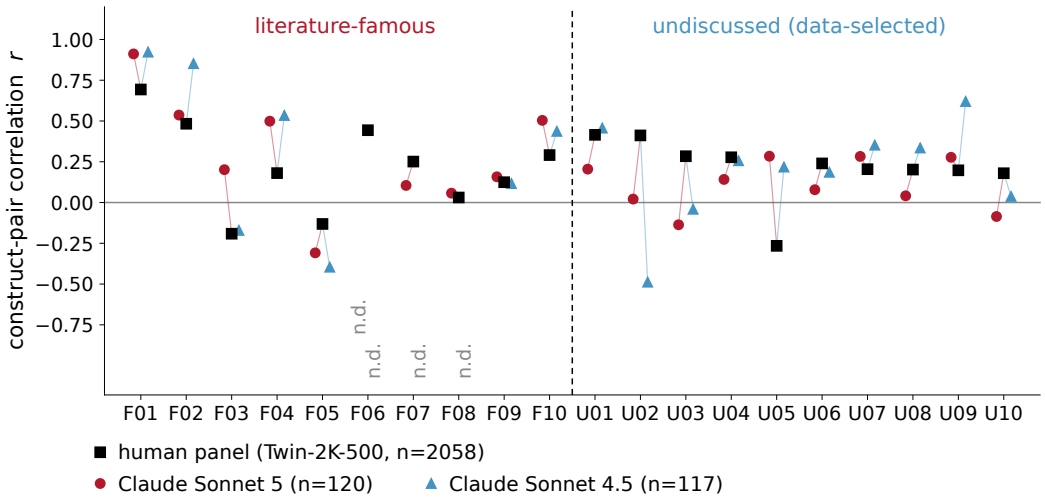


Figure 8: The 20 first-batch construct pairs: human panel correlation and the two simulated correlations, for literature-famous and undiscussed pairs.

dictator-game giving, cognitive reflection with conditional reasoning), each backed by a named review, meta-analysis, or monograph source, and each undiscussed pair by a documented null search (`protocols/ceil2x2_fame_evidence.md`). Undiscussed political pairs put one political construct (policy support, self-rated ideology, religious attendance, religious fundamentalism) against a judgment-and-decision task that no literature we could locate discusses in relation to it, and no construct appears more than twice per cell. Items for the fourteen constructs not in the first batch were written as parallel forms of standard instruments, and fresh samples of 120 synthetic participants per model answered the full extended battery under the same protocol (97% of construct scores computable, response rate 233/240).

The outcome is the excess correlation $|r_{\text{sim}}| - |r_{\text{human}}|$. Famous nonpolitical pairs are exaggerated (mean excess $+0.19$) while undiscussed political pairs are not ($+0.03$). The pair is the statistical unit. Per-pair values average the two models, and the two main effects are the coefficients of a pair-level regression of excess on a fame and a political indicator (33 computable pairs). Cell means at the observation level are famous political $+0.14$, famous nonpolitical $+0.20$, undiscussed nonpolitical -0.04, undiscussed political $+0.04$. The fame main effect is $+0.16$ (SE 0.05, $p = .007$; one-sided pair-level Mann-Whitney $p = .004$ over all pairs, $p = .07$ on the second-batch cells alone with 6 and 8 computable pairs), and the political-content main effect is $+0.005$ ($p = .92$; equal-weight cell-mean contrast $+0.01$). All numbers in this appendix are generated by `code/analyze_ceil2x2.py`. Selective exaggeration follows what the literature discusses, not political content.

Because undiscussed pairs are selected on their observed panel magnitude, a winner's curse could inflate their human correlations. A simulation at the panel's precision (n = 2,058, standard error about 0.022) bounds this bias below 0.001 for magnitude-matched selection, the procedure used, and below 0.05 even for worst-case largest-first selection from the entire panel pool, several times smaller than the observed famous–undiscussed difference.

Pairs built on ability tests (numeracy, financial literacy, fluid ability, cognitive reflection, conditional reasoning) are partly uncomputable, because the simulated samples answer ability items at or near 100% correct with zero between-participant variance, so the corresponding correlations have no denominator. This is the same variance collapse documented in Appendix L, here in its starkest form. Simulated participants do not reproduce human ability variation at all. The cell means above use the computable values. Pair-level results are in the repository (`analysis_tables/ceil2x2_results.csv`, with the design table and item bank alongside).

Missing values. In the second-batch cells, 10 of 16 pair-model values are computable for famous nonpolitical pairs and 14 of 16 for undiscussed political pairs. The first batch contributes one further famous nonpolitical pair (a cognitive-reflection score with a probability-maximizing

choice) that is uncomputable for both models, so the full famous nonpolitical cell has 10 of 18 computable values, against 16 of 18 and 20 of 20 in the two first-batch cells. Three famous nonpolitical pairs in all are uncomputable for both models. Because this missingness is concentrated in one cell, we bound its influence on the pair-level regression. Setting all missing values to zero gives a fame effect of $+0.12$ ($p = .016$), and restricting to the 27 pairs computable in both models gives $+0.18$ ($p = .007$). A worst-case bound that assigns every missing famous value the lowest excess observed anywhere (-0.39) removes the effect ($+0.02$, $p = .83$). The fame effect is therefore robust to neutral treatments of the missing values but not to the least favorable one (`analysis_tables/c2x2_missingness_bounds.json`).