EXPERT-GROUNDED SYNTHETIC DATA FOR COGNITIVE BIASES: CONTROLLED GENERATION AND BLIND EVALUATION OF CONSTRUCT FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable systems for identifying cognitive-bias mechanisms require datasets whose labels are psychologically faithful, domain-grounded, and able to distinguish neighboring constructs. Large language models can scale candidate generation, but generic prompting may produce fluent items that express a different mechanism from the intended target, creating mechanism-level label noise.

We study this problem using a private, psychologist-curated Cognitive Bias Domain Atlas containing 229 cognitive biases, 30 application domains, and 2,385 curated bias–domain relationships whose records may include expert definitions, domain-specific justifications, and reference examples. We compare three conditioning packages over 31 matched relationships spanning 10 biases and five domains: *Public*, using a literature-grounded definition; *Atlas-DJ*, using an expert definition and domain-specific justification; and *Atlas-Full*, which additionally supplies an expert reference example with no-copy and overlap safeguards. Across 93 generation cells, the system produced 930 candidate survey items, of which 857 were eligible for condition- and target-blind review by a psychology expert. Primary inference is conducted at the relationship level.

Relative to *Public*, *Atlas-Full* had higher matched point estimates for Bias-ID accuracy ($+4.2$ percentage points), contamination-free items ($+8.4$ points), the strict fidelity composite ($+8.9$ points), and the soft composite ($+7.4$ points). The reported confidence intervals included zero, and none of the 12 primary comparisons survived Holm correction. Effects were heterogeneous: for Confirmation Bias, identification increased descriptively from 10.7% under *Public* to 75.9% under *Atlas-DJ* and 65.4% under *Atlas-Full*; most Public-condition misidentifications were assigned Bandwagon Effect or Authority Bias. These results support a conditional, package-level interpretation of expert grounding and an iterative workflow combining generation, blind review, and confusion analysis, while leaving downstream model utility and broader generalization for future study.

1 INTRODUCTION

Training and evaluating systems that identify cognitive-bias mechanisms requires datasets whose labels are psychologically faithful, not merely fluent. Useful examples must express the intended mechanism, remain grounded in a realistic domain, avoid confusion with neighboring biases, and support reliable model development and evaluation. Manual construction is expensive and difficult to scale, motivating broader cognitive-bias resources (Malberg et al., 2025). Large language models can increase generation volume, but a fluent and domain-plausible item may still instantiate the wrong psychological mechanism. This produces a consequential form of synthetic-data noise: *mechanism-level label noise*, in which the text appears valid while its label is psychologically imprecise.

Recent synthetic-data work increasingly uses structured and domain-aware conditioning to improve controllability and usefulness (Nadăş et al., 2025; Bauer et al., 2024). However, controlled evidence remains limited on whether expert-authored domain knowledge preserves fine-grained distinctions among neighboring psychological constructs. It is unclear how expert definitions and domain-specific justifications compare with general public definitions, or whether adding an expert reference example

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.

provides further benefit. We study these questions as a problem of labeled exemplar construction: the objective is to generate examples that instantiate a specified mechanism, not to determine whether a language model itself exhibits a cognitive bias.

We investigate this problem using a private, psychologist-curated Cognitive Bias Domain Atlas containing 229 biases, 30 application domains, and 2,385 curated bias–domain relationships. We evaluate 31 matched supported relationships spanning 10 biases and five domains under three conditioning packages: *Public*, which uses a literature-grounded public definition; *Atlas-DJ*, which uses an expert definition and domain-specific justification; and *Atlas-Full*, which adds an expert reference example together with safeguards against copying and reference overlap. The conditions share the same generation task and are compared as complete conditioning packages rather than isolated component effects. Across 93 experimental cells, the pipeline produced 930 candidates, of which 857 were eligible for condition-blind expert review. Because items are nested within bias–domain relationships, primary inference is conducted at the relationship level rather than treating reviewed items as independent observations.

The aggregate results show positive but uncertain Full–Public contrasts. Atlas-Full has higher point estimates than Public by 4.2 percentage points for Bias-ID accuracy, 8.4 points for contamination-free items, 8.9 points for the strict composite, and 7.4 points for the soft composite. The corresponding confidence intervals are wide and include zero, and none of the 12 primary comparisons survived relationship-level clustering and Holm correction. Effects are also heterogeneous across target constructs. In closed-set Bias-ID evaluation, Confirmation Bias accuracy increases from 10.7% under *Public* (3/28) to 75.9% under *Atlas-DJ* (22/29) and 65.4% under *Atlas-Full* (17/26), while public-condition misidentifications are concentrated among neighboring mechanisms such as Bandwagon Effect and Authority Bias. These results support a conditional interpretation: expert grounding may be most valuable where generic conditioning produces mechanism confusion.

This paper makes three contributions:

- We present a controlled experiment for expert-grounded synthetic generation over matched, psychologist-curated bias–domain relationships, comparing public and expert conditioning packages under a common generation task.
- We introduce a blind construct-level evaluation that measures bias identification and neighboring-construct contamination in addition to composite fidelity and usability, with relationship-level inference that accounts for dependence among generated items.
- We characterize heterogeneous, construct-specific conditioning effects and show how blind confusion analysis can support iterative dataset development, in which difficult distinctions guide targeted generation, review, and subsequent evaluation.

2 Related Work

2.1 LLM Synthetic Data and Knowledge-Grounded Generation

Large language models are increasingly used to construct synthetic datasets. Self-Instruct demonstrated that model-generated instructions and examples can reduce reliance on manual supervision, while later work explored larger-scale generation under task and attribute constraints (Wang et al., 2023; Honovich et al., 2023; Yu et al., 2023). Surveys describe this literature as a generate–curate–evaluate pipeline in which controllability, filtering, label quality, and downstream utility remain central challenges (Long et al., 2024; Nadăş et al., 2025; Bauer et al., 2024).

Recent approaches condition generation on external structure or domain knowledge. DataGen combines constraints, attribute guidance, retrieval, and automated checking; Bonito and CRAFT derive task-specific data from specialized corpora; and ClinGen and SNaRe incorporate domain information to improve task alignment and reduce domain drift (Huang et al., 2025; Nayak et al., 2024; Ziegler et al., 2025; Xu et al., 2024; Parekh et al., 2025). Other systems emphasize reproducible workflows and iterative refinement (Patel et al., 2024; Lee et al., 2024). We study a different question: whether the source and specificity of conceptual conditioning affect how faithfully an LLM instantiates a fine-grained expert-defined label. A fluent and structurally valid item may express a neighboring construct rather than the intended one, producing *construct drift* and mechanism-level

label error. This is a particularly difficult form of label noise because the mislabeled example remains linguistically plausible (Frénay & Verleysen, 2014; Northcutt et al., 2021).

2.2 Psychological Item Generation and Construct Fidelity

Automatic item generation has a substantial history in psychometrics, including recurrent and construct-specific transformer models and evaluation against psychometric criteria (von Davier, 2018; Hommel et al., 2022; Götz et al., 2024). Recent LLM-based work extends this direction through psychometric evaluation, multi-agent generation and review, personality situational-judgment items, and broader psychological-assessment applications (Bhandari et al., 2024; Lee et al., 2026; Li et al., 2026; Brickman et al., 2025).

The distinction between plausible language and valid construct representation predates LLMs. Classical measurement theory treats construct validity as evidence that observed measurements support the theoretical interpretation assigned to them, rather than merely appearing consistent with a label (Cronbach & Meehl, 1955; Messick, 1995). Recent LLM research raises analogous concerns: semantic appropriateness does not establish psychological construct validity, benchmark behavior may not support its intended latent construct, and theoretically distinct constructs may occupy overlapping representational regions (Lin, 2026; Alaa et al., 2025; Bean et al., 2025; Wulff & Mata, 2025).

We build on this literature by treating *construct discrimination* as a first-class synthetic-data outcome. In blind evaluation, the reviewer selects the construct best represented by each item from the study's closed label vocabulary without seeing the hidden target. Fidelity is assessed by comparing that judgment with the target and separately recording neighboring-construct contamination. An item may therefore be readable, domain appropriate, and usable while still failing the construct-level criterion required to justify its label.

2.3 Cognitive-Bias Foundations, Benchmarks, and Positioning

The evaluated constructs are grounded in established psychological research rather than defined solely by the private Atlas. Foundational work characterizes anchoring and availability as judgment heuristics, demonstrates framing-dependent choice reversals, and examines Confirmation Bias as preferential seeking or interpretation of belief-consistent evidence (Tversky & Kahneman, 1974; 1981; Nickerson, 1998). Positive-test strategies provide a more specific account of some confirmation phenomena, while established literatures cover Status Quo Bias, the Endowment Effect, Sunk Cost effects, and Default effects (Klayman & Ha, 1987; Samuelson & Zeckhauser, 1988; Kahneman et al., 1990; Arkes & Blumer, 1985; Johnson & Goldstein, 2003).

Social-influence theory is relevant to the confusion patterns examined here. Informational influence treats others' responses as evidence about what is correct, whereas normative influence reflects pressure to conform independently of evidential value (Deutsch & Gerard, 1955; Cialdini & Goldstein, 2004). Thus, language involving confirming evidence, group consensus, and authority cues may appear similar while corresponding to different mechanisms.

A parallel literature studies whether language models exhibit cognitive biases. Prior work examines LLM failures through cognitive-bias paradigms, intuitive reasoning biases, evaluator biases, mitigation strategies, and conversational prediction of biased human decision making. Malberg et al. construct approximately 30,000 bias test cases spanning 30 biases and 200 managerial scenarios (Jones & Steinhardt, 2022; Hagendorff et al., 2023; Koo et al., 2024; Sumita et al., 2025; Pilli & Nallur, 2026; Malberg et al., 2025). Our objective is complementary but distinct: we ask whether an LLM can generate a labeled exemplar that faithfully represents a specified bias mechanism in a target domain. The primary failure mode is therefore construct-level label error in synthetic data, not biased model behavior.

Taken together, prior work establishes that LLMs can generate synthetic data, structured or expert knowledge can improve generation control, and psychological items require construct-sensitive validation. We contribute a controlled empirical study of how expert-grounded conditioning relates to construct fidelity and neighboring-label contamination in cognitive-bias dataset development. The contribution is an evaluation framework for identifying when apparently valid synthetic examples preserve the expert distinctions encoded by their labels, not a new cognitive-bias taxonomy or a claim of universal benefit from expert context.

3 EXPERT-CURATED ATLAS AND STUDY SUPPORT

The generation framework uses a private, psychologist-curated resource linking cognitive biases to applied domains, referred to as the *Atlas*. The complete Atlas contains 229 cognitive biases, 30 application domains, and 2,385 curated bias–domain relationships. Supported relationships may include an expert-authored definition, a domain-specific justification describing how the mechanism manifests in the target domain, and, where available, a reference example. Coverage is intentionally sparse: generation is restricted to relationships with curated conditioning records, while unsupported bias–domain pairs are not interpreted as psychologically invalid or inapplicable. The experiment evaluates 31 supported relationships spanning 10 biases and five domains under all three conditioning packages, yielding matched bias–domain composition across conditions. The Atlas defines where expert-grounded generation is available; it is not proposed as a new cognitive-bias taxonomy or graph-learning method.

The complete Atlas remains private, including its full inventory of 229 cognitive biases, 30 application domains, and 2,385 curated bias–domain relationships. For reproducibility, the supplementary package releases only the approved study-support subset used in this experiment: the 31 evaluated bias–domain relationships and the corresponding Atlas records required to audit the Public, Atlas-DJ, and Atlas-Full conditioning packages. Where included in the evaluated prompts, these records contain the relevant definitions, domain-specific justifications, reference examples, identifiers, and non-sensitive provenance fields. The released subset does not expose the complete Atlas, unrelated relationships, internal packet keys, private join files, or unreleased proprietary records.

4 EXPERIMENTAL DESIGN

4.1 RESEARCH QUESTIONS AND EXPERIMENTAL SUBSET

The experiment addresses three research questions:

- **RQ1: How does expert definition-and-domain-justification conditioning compare with Public conditioning in construct fidelity?** We compare Public with Atlas-DJ over the same supported bias–domain relationships.
- **RQ2: How does the Atlas-Full conditioning package compare with Atlas-DJ?** Atlas-Full adds an expert reference example, a no-copy instruction, and reference-specific overlap validation. This is a conditioning-package comparison rather than an isolated reference-example effect.
- **RQ3: Are conditioning effects heterogeneous across cognitive biases?** Bias-specific analyses examine whether condition differences are concentrated among constructs with greater baseline confusion and are treated as exploratory.

The study evaluates 10 cognitive biases across five application domains. Of the resulting 50 possible bias–domain combinations, 31 have supported Atlas relationships. The remaining pairs are excluded because no curated Atlas record is available and are not interpreted as negative or psychologically invalid relationships. Each supported relationship is evaluated under all three conditions, yielding identical bias–domain composition across conditions. The subset was selected to provide a tractable range of constructs and domains rather than as a random or statistically representative sample of the complete Atlas.

4.2 GENERATION SETUP AND STUDY FLOW

The 31 supported relationships crossed with three conditioning packages yielded 93 experimental cells. Each cell was generated once with `gpt-4o-2024-08-06`, using a fixed generation configuration and requesting ten candidate items, for 930 raw candidates in total. No candidate was regenerated, repaired, or removed after unblinding or on the basis of psychological quality.

Predefined automated validation produced 803 accepted records, 54 warning-status records, and 73 rejected records. Accepted and warning-status candidates were eligible for blinded review, yielding $N = 857$: 288 Public, 282 Atlas-DJ, and 287 Atlas-Full items. The unequal review-set sizes

reflect pre-review validation rather than post hoc sampling. Atlas-Full additionally underwent the reference-overlap safeguard associated with its private reference example, so pre-review filtering was not perfectly symmetric across conditions. Validation outcomes are therefore treated as part of the generation procedure, not as measures of psychological fidelity. Exact decoding parameters, validation rules, and candidate-flow details are provided in the appendix.

4.3 Evaluation Outcomes

Human evaluation assessed construct fidelity, neighboring-bias contamination, domain relevance, survey usability, and editorial suitability using a fixed multidimensional rubric. Full rubric definitions and supplementary naturalness and originality fields are provided in the appendix.

Bias identification. For each blinded item, the reviewer selected the best-represented bias from the study's closed 10-label vocabulary without access to the hidden target or generation condition. We then compared the assigned label $\hat{b}$ with the hidden generation target b:

$$\mathrm{BiasMatch} = \mathbf{1}\{\hat{b} = b\}. \tag{1}$$

Bias-ID accuracy is the proportion of items with $\mathrm{BiasMatch} = 1$. Uniform selection from the closed label set gives a nominal expected accuracy of 10%.

Neighboring-bias contamination. The reviewer rated competing or additional mechanisms as none, possible, or likely and recorded an alternate bias where applicable. The primary outcome *Contamination none* is the proportion of items receiving no contamination rating; likely contamination provides the strongest indication of a different or mixed mechanism.

Domain relevance, usability, and editorial decision. Domain relevance was rated as strong, acceptable, weak, or incorrect, with the proportion rated strong treated as a secondary outcome. Survey usability was rated as usable or usable with revision; no unusable ratings were recorded in the retained review set. Editorial decisions were retain, revise, or reject and measure practical dataset utility rather than construct fidelity.

Composite outcomes. The strict composite requires retention without substantive revision, correct bias identification, no neighboring-bias contamination, and strict usability:

$$\mathrm{Strict} = \mathrm{Retain} \wedge \mathrm{BiasMatch} \wedge \mathrm{ContamNone} \wedge \mathrm{Usable}. \tag{2}$$

The soft composite permits revision, possible but not likely contamination, and usability either directly or after limited revision:

$$\mathrm{Soft} = (\mathrm{Retain} \vee \mathrm{Revise}) \wedge \mathrm{BiasMatch} \wedge \neg\mathrm{LikelyContam} \wedge (\mathrm{Usable} \vee \mathrm{UsableWithRevision}). \tag{3}$$

The four primary fidelity outcomes are Bias-ID accuracy, *Contamination none*, the strict composite, and the soft composite. Domain relevance, usability, and editorial decisions are secondary or descriptive outcomes. These measures assess item-level construct fidelity and dataset utility; they do not independently validate respondent behavior or the underlying Atlas taxonomy.

4.4 Statistical Analysis

The 857 reviewed items are nested within 31 matched bias–domain relationships and are not treated as independent inferential units. For each binary outcome, we computed the success proportion within each relationship–condition cell and formed within-relationship pairwise differences. Averaging these differences across the 31 relationships gives each relationship equal weight and defines the primary percentage-point contrasts. Item-weighted condition rates are reported descriptively.

Uncertainty was quantified using a nonparametric cluster bootstrap that resampled complete bias–domain relationships, with 10,000 replicates and percentile two-sided 95% confidence intervals. Matched permutation tests exchanged condition labels within relationships under the null of no condition difference, using 100,000 Monte Carlo draws. A cluster-aware Friedman test and logistic GEE models with exchangeable working correlation were used as robustness analyses.

For each primary outcome, we evaluated three prespecified pairwise contrasts: Atlas-DJ versus Public, Atlas-Full versus Public, and Atlas-Full versus Atlas-DJ. Raw and Holm-adjusted permutation p-values were reported across the resulting 12 primary comparisons. Secondary outcomes were excluded from this multiplicity family, and interpretation emphasizes effect magnitudes and confidence intervals rather than thresholded significance alone.

Bias-specific analyses, including the Confirmation Bias case study, neighboring-bias confusion patterns, and associations between public-condition difficulty and expert-conditioning effects, were treated as exploratory analyses of heterogeneity. Leave-one-bias-out analyses assessed whether aggregate contrasts were disproportionately driven by individual target biases. Unless otherwise stated, all intervals are two-sided 95% confidence intervals.

5 Blind Evaluation

All $N = 857$ review-eligible items were evaluated by a single expert reviewer with graduate training in psychology through the doctoral level and university-level research and teaching experience. The reviewer had contributed to Atlas curation and applied the fixed multidimensional rubric described in Section 4.3. This shared provenance between the reviewer and conditioning resource is addressed as a construct-alignment limitation in Section 8.

Evaluation was condition- and target-blind. Reviewer-facing records contained only the respondent-facing item, its application domain, an opaque identifier, and the fields required for scoring. The reviewer did not see the hidden target bias, generation condition, provenance metadata, or private Atlas conditioning content. The three conditions were presented in separate packets with randomized opaque condition identities and randomized item order; packet-construction details are provided in the appendix.

For each item, the reviewer selected the best-represented bias from the study's closed 10-label vocabulary and independently rated neighboring-bias contamination, domain relevance, survey usability, and the retain–revise–reject editorial decision. The same rubric and interface were used across conditions. Ratings served only as evaluation labels and did not trigger item repair, regeneration, or removal. The condition and target-label mapping was revealed only after all ratings had been finalized, after which opaque identifiers were rejoined with generation metadata for analysis. No item was rescored, reclassified, or removed after unblinding.

As an intra-rater consistency check, a stratified sample of 30 items underwent a subsequent blind re-review. The second ratings agreed with the original ratings for all 30 selected bias labels, 29 of 30 auxiliary target-presence judgments, and 27 of 30 editorial decisions. The target-presence field was not used to compute Bias-ID, which was defined exclusively by exact equality between the selected canonical label and the hidden target label. All three editorial disagreements were retain-versus-revise rather than rejection. This check supports within-reviewer consistency but does not establish inter-rater reliability or remove the limitation of the single-reviewer design.

6 Results

We evaluate the three conditioning packages on $N = 857$ blindly reviewed items drawn from 31 matched bias–domain relationships: 288 Public, 282 Atlas-DJ, and 287 Atlas-Full. Condition rates are item-weighted descriptive summaries, whereas inferential contrasts use the relationship-weighted matched analysis defined in Section 4.4.

6.1 Primary Outcomes

Table 1 reports the four primary fidelity outcomes. Atlas-Full has the highest item-weighted descriptive rate for each outcome.

Relationship-level contrasts showed the same directional pattern but remained imprecisely estimated. Relative to Public, Atlas-Full increased contamination-free generation by 8.4 percentage points, the strict composite by 8.9 points, and the soft composite by 7.4 points. The matched Bias-ID difference was $+4.2$ points. None of the 12 primary pairwise comparisons remained statistically significant after Holm correction, and all corresponding primary confidence intervals included zero. The observed 7–9

Table 1: Primary fidelity outcomes. Condition rates are item-weighted descriptive percentages; Full–Public contrasts are relationship-weighted matched percentage-point differences with 95% cluster-bootstrap confidence intervals.

Metric	Public	Atlas-DJ	Atlas-Full	Full–Public Δ [95% CI]
Bias-ID accuracy	86.8%	90.4%	91.6%	$+4.2$ $[-3.2, 13.1]$
Contamination none	63.2%	65.6%	71.8%	$+8.4$ $[-1.4, 18.6]$
Strict composite	43.1%	46.8%	51.9%	$+8.9$ $[-0.3, 18.7]$
Soft composite	79.2%	81.9%	87.1%	$+7.4$ $[-1.7, 17.9]$

percentage-point differences are therefore interpreted as potentially practically relevant but uncertain condition contrasts, not as established population-level effects.

Secondary outcomes did not show identical condition ordering. Retention remained approximately constant across conditions (70.1% Public, 71.3% Atlas-DJ, and 70.4% Atlas-Full), while strong domain-relevance ratings were highest under Atlas-DJ (89.2%, 95.0%, and 91.6%, respectively). In a secondary cluster-aware GEE analysis, Atlas-DJ was associated with higher odds of a strong domain rating than Public (OR $= 2.23$, 95% CI $[1.39,\ 3.58]$, $p = 0.0008$). This result is interpreted separately from the primary multiplicity family.

6.2 Bias-Level Heterogeneity and the Confirmation Bias Case

Condition differences varied substantially across target biases. Authority Bias, Default Effect, Framing Effect, and Sunk Cost Fallacy, for example, were identified with high accuracy under Public conditioning and showed limited room for improvement, whereas other constructs exhibited markedly different condition profiles. Complete bias-level results are provided in the appendix.

The clearest heterogeneity occurred for *Confirmation Bias*. Under Public conditioning, 3 of 28 items were identified correctly (10.7%; Wilson 95% CI $[3.7,\ 27.2]$), approximately the 10% nominal accuracy expected under uniform selection from the study's closed 10-label vocabulary. Accuracy increased descriptively to 75.9% under Atlas-DJ (22/29; 95% CI $[57.9,\ 87.8]$) and 65.4% under Atlas-Full (17/26; 95% CI $[46.2,\ 80.6]$). These items came from three verified Confirmation Bias–domain relationships.

Public-condition misidentifications were concentrated among neighboring constructs: 13 of 28 items were identified as Bandwagon Effect, 10 as Authority Bias, two as Availability Heuristic, and three as Confirmation Bias. The dominant failure mode was therefore confusion between adjacent mechanisms rather than absence of a recognizable bias-related signal. Exploratory item-level Fisher exact tests reflected the separation between Public and Atlas-DJ ($p = 6.8 \times 10^{-7}$) and between Public and Atlas-Full ($p = 4.8 \times 10^{-5}$). These tests are not treated as independent confirmatory evidence because the items are nested within only three Confirmation Bias–domain relationships.

Confirmation Bias accounts for most of the item-weighted Bias-ID advantage of Atlas-Full over Public: the aggregate difference was $+4.8$ percentage points across all biases but approximately -0.8 points after its exclusion. The overall Bias-ID contrast therefore reflects heterogeneous construct-specific differences rather than broad improvement across all ten biases. The other fidelity outcomes remained positive after excluding Confirmation Bias, with Full–Public differences of $+5.9$ points for contamination-free generation, $+6.0$ points for the strict composite, and $+2.7$ points for the soft composite.

6.3 Robustness and Sensitivity Analysis

The directional contamination and composite-fidelity patterns were stable across weighting and modeling choices. Item-weighted and relationship-weighted contrasts were closely aligned: the contamination-free difference was $+8.6$ points under item weighting and $+8.4$ points under relationship weighting, while the strict-composite difference was $+8.9$ points under both approaches. Cluster-aware GEE models produced the same directional ordering as the matched analysis, with Atlas-Full odds ratios above one relative to Public for all four primary outcomes, although all confidence intervals included the null. Leave-one-bias-out analyses showed that the Bias-ID contrast was

highly sensitive to Confirmation Bias, whereas the contamination and composite-fidelity contrasts remained positive after excluding any individual bias. Complete robustness outputs are provided in the appendix.

7 Discussion

7.1 Interpretation of the Main Findings

The results distinguish surface acceptability from construct fidelity. Retention remained approximately constant across conditions, while the Atlas-Full conditioning package was associated with higher point estimates for contamination-free generation and the strict and soft fidelity composites. The observed differences therefore concerned the proportion of items satisfying mechanism-level criteria rather than the overall amount of acceptable text. These contrasts remain uncertain: confidence intervals include zero and none of the 12 primary comparisons survived Holm correction. They should be interpreted as potentially practically relevant condition contrasts requiring broader confirmation, not as evidence of uniform superiority.

Effects were heterogeneous across target constructs. Several biases were already identified reliably under Public conditioning, whereas Confirmation Bias showed substantial public-condition confusion concentrated among Bandwagon Effect and Authority Bias. The Atlas-DJ package was associated with greater descriptive discriminability within this confusion set, supporting construct-specific evaluation beyond fluency, domain relevance, or generation yield. The results do not support a monotonic "more context is better" interpretation: Atlas-Full did not consistently outperform Atlas-DJ, and Atlas-DJ produced the highest descriptive Confirmation Bias identification rate. The Atlas-Full package may provide useful grounding while also introducing exemplar-specific tendencies or constraints (Yu et al., 2023). Because it adds a reference example, a no-copy instruction, and reference-overlap validation, this comparison does not isolate the effect of reference examples. Similarly, the Atlas-DJ result does not isolate the contributions of expert definitions and domain justifications. Adaptive allocation of richer conditioning to relationships with low baseline fidelity or high neighboring-label confusion is therefore a future-work hypothesis, not a tested result.

7.2 Implications for Expert-Grounded Synthetic Data

When labels represent fine-grained expert constructs, generic definitions may produce fluent examples without preserving the distinctions that define the target label. Expert-grounded generation incorporates conceptual and domain knowledge before generation, while expert review remains the operational criterion for construct fidelity. The results further suggest that the type of context may matter more than its quantity, since Atlas-Full did not consistently produce the best outcomes across constructs.

Private expert knowledge creates an auditability challenge. Our framework separates conditioning content from research artifacts while retaining versioned identifiers, configuration fingerprints, and prompt-level hashes that link generated records to their retained context. These mechanisms support provenance and traceability without permitting reconstruction of the private source. More broadly, expert-grounded generation may enable controlled construction of candidate artifacts from nondistributable resources, provided that source leakage, provenance, and human validation are addressed.

7.3 Implications for Cognitive-Bias Dataset and Benchmark Development

The practical value of expert-grounded generation is the development of resources that preserve distinctions among neighboring cognitive-bias mechanisms. Expert-validated candidates could expand coverage for sparsely represented bias–domain relationships, create targeted challenge cases, and support controlled evaluation of successive model versions. The Confirmation Bias analysis illustrates how observed generation failures can reveal structured confusion sets: repeated Public-condition assignments to Bandwagon Effect or Authority Bias could be recorded as hard negatives or targeted challenge examples.

This motivates an iterative data-development loop:

generation → expert review → confusion analysis → challenge/training data → model evaluation.

Relationships with high fidelity may require little additional generation, whereas persistent contamination or neighboring-label confusion can guide further review and targeted construction. Any candidates used in training or benchmark data would require expert validation before inclusion.

The present study establishes only the generation and human-evaluation stages of this loop. It does not show that training on the resulting synthetic data improves downstream detection, calibration, or robustness. Such effects require independently held-out, human-authored benchmarks and replication across additional biases, domains, generation runs, model families, and independently blinded reviewers.

8 Conclusion and Future Work

We evaluated whether expert-grounded conditioning changes the construct fidelity of LLM-generated cognitive-bias items. Across 31 matched bias–domain relationships and 857 blindly reviewed items, point estimates favored *Atlas-Full* over *Public* for contamination-free generation ($+8.4$ percentage points), the strict fidelity composite ($+8.9$ points), and the soft composite ($+7.4$ points). These contrasts remained uncertain: confidence intervals included zero and none of the 12 primary comparisons survived Holm correction. Effects were heterogeneous rather than uniformly positive. Confirmation Bias identification increased from 10.7% under *Public* to 75.9% under *Atlas-DJ* and 65.4% under *Atlas-Full*, while several other constructs showed limited need for additional conditioning.

Limitations and future work. The study evaluates a nonrandom subset of 31 relationships and is not representative of the broader Atlas; the primary inferential units are relationships rather than the 857 reviewed items. Evaluation relied on one reviewer who contributed to Atlas curation, so Bias-ID measures fidelity relative to the study's expert construct framework and does not establish inter-rater reliability or open-world identification. The conditions compare complete packages, each relationship–condition cell was generated once with one model version, and Atlas-Full used an additional pre-review overlap safeguard. Finally, the private Atlas permits only partial external reproduction, and the study does not test downstream model utility. Future work should replicate the experiment with independently blinded reviewers, additional relationships, repeated generation runs, broader model coverage, and independently held-out human-authored benchmarks. These studies can also test whether confusion-guided allocation of richer conditioning improves data efficiency and whether gains in construct fidelity translate into better detection, calibration, and robustness.

The findings support expert-grounded generation as a component of iterative dataset development: controlled generation produces candidates, blind review identifies neighboring-construct failures, and confusion analysis can guide validated challenge-set or training-data construction. Establishing downstream utility remains an empirical question for future evaluation.

AI Use Disclosure

We used OpenAI GPT-4o (`gpt-4o-2024-08-06`) to generate the synthetic questionnaire items evaluated in this study. LLM-based tools were also used to assist with research-methodology discussions, manuscript drafting and editing, literature organization, and presentation of the analysis. Human expert review, rather than the generation model, provided the evaluation labels used in the study. The authors reviewed and verified all AI-assisted text, experimental decisions, analyses, citations, and released artifacts and take responsibility for the final contents of the submission.

Ethics Statement

This study involved expert evaluation of model-generated questionnaire items and did not collect behavioral responses, personal information, or clinical data from survey respondents. Released artifacts are screened to remove identifying information, internal keys, and Atlas material outside the approved study-support subset. The generated items are research artifacts and should not be used directly for high-stakes decisions without appropriate domain-specific review and human oversight.

Reproducibility Statement

The supplementary package accompanying this paper provides the analysis code, resolved generation configuration, the release-approved study-support subset for the 31 evaluated bias–domain relationships, de-identified review packets, candidate and review-label subsets, and the supported relationship manifest used in the experiment. The study-support subset contains the condition-relevant Atlas records needed to audit the evaluated Public, Atlas-DJ, and Atlas-Full prompts. The supplement also documents prompt construction, validation procedures, the review rubric, provenance fields, and statistical analyses. The complete Atlas, including its 229 biases, 30 application domains, and 2,385 curated relationships, remains private; unrelated Atlas records, internal keys, and unreleased proprietary materials are not included.

References

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Position: Medical large language model benchmarks should prioritize construct validity. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80991–81004. PMLR, 2025.

Hal R. Arkes and Catherine Blumer. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1):124–140, 1985. doi: 10.1016/0749-5978(85)90049-4.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*, 2024. doi: 10.48550/arXiv.2401.02524.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0590. Datasets and Benchmarks Track.

Shreya Bhandari, Yunting Liu, Yerin Kwak, and Zachary A. Pardos. Evaluating the psychometric properties of ChatGPT-generated questions. *Computers and Education: Artificial Intelligence*, 7: 100284, 2024. doi: 10.1016/j.caeai.2024.100284.

Jocelyn Brickman, Mehak Gupta, and Joshua R. Oltmanns. Large language models for psychological assessment: A comprehensive overview. *Advances in Methods and Practices in Psychological Science*, 8(3), 2025. doi: 10.1177/25152459251343582.

Robert B. Cialdini. *Influence: Science and Practice*. Pearson Education, Upper Saddle River, NJ, 5 edition, 2009. ISBN 9780205609994.

Robert B. Cialdini and Noah J. Goldstein. Social influence: Compliance and conformity. *Annual Review of Psychology*, 55:591–621, 2004. doi: 10.1146/annurev.psych.55.090902.142015.

Lee J. Cronbach and Paul E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

Morton Deutsch and Harold B. Gerard. A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3):629–636, 1955. doi: 10.1037/h0046408.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.

Friedrich M. Götz, Rakoen Maertens, Sahil Loomba, and Sander van der Linden. Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*, 29(3):494–518, 2024. doi: 10.1037/met0000540.

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*, 3:833–838, 2023. doi: 10.1038/s43588-023-00527-x.

Björn E. Hommel, Franz-Josef M. Wollang, Veronika Kotova, Hannes Zacher, and Stefan C. Schmukle. Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2):749–772, 2022. doi: 10.1007/s11336-021-09823-9.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.806.

Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, and Xiangliang Zhang. DataGen: Unified synthetic dataset generation via large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Eric J. Johnson and Daniel Goldstein. Do defaults save lives? *Science*, 302(5649):1338–1339, 2003. doi: 10.1126/science.1091721.

Erik Jones and Jacob Steinhardt. Capturing failures of large language models via human cognitive biases. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11785–11799, 2022. doi: 10.52202/068431-0856.

Daniel Kahneman, Jack L. Knetsch, and Richard H. Thaler. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98(6):1325–1348, 1990. doi: 10.1086/261737.

Joshua Klayman and Young-Won Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, 94(2):211–228, 1987. doi: 10.1037/0033-295X.94.2.211.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.29.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael Mahoney, Kurt Keutzer, and Amir Gholami. LLM2LLM: Boosting LLMs with novel iterative data enhancement. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6498–6526. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.388.

Philseok Lee, Mina Son, and Zihao Jia. AI-powered automatic item generation for psychological tests: A conceptual framework for an LLM-based multi-agent AIG system. *Journal of Business and Psychology*, 41(1):71–99, 2026. doi: 10.1007/s10869-025-10067-y.

Harvey Leibenstein. Bandwagon, snob, and veblen effects in the theory of consumers' demand. *The Quarterly Journal of Economics*, 64(2):183–207, 1950. doi: 10.2307/1882692.

Chang-Jin Li, Jiyuan Zhang, Yun Tang, and Jian Li. Automatic item generation for personality situational judgment tests with large language models. *Computers in Human Behavior Reports*, 21: 100964, 2026. doi: 10.1016/j.chbr.2026.100964.

Zhicheng Lin. From prompts to constructs: A dual-validity framework for large language model research in psychology. *Annual Review of Psychology*, 78:In press, 2026. doi: 10.1146/annurev-psych-100925-034807. Review in Advance; first posted online August 10, 2026.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11065–11082. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.658.

Simon Malberg, Roman Poletukhin, Carolin M. Schuster, and Georg Groh. A comprehensive evaluation of cognitive biases in LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pp. 578–613, Albuquerque, USA, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.nlp4dh-1.50.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9):741–749, 1995. doi: 10.1037/0003-066X.50.9.741.

Stanley Milgram. Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, 67(4): 371–378, 1963. doi: 10.1037/h0040525.

Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633, 2025. doi: 10.1109/ACCESS.2025.3589503. arXiv:2503.14023.

Nihal Nayak, Yiyang Nan, Avi Trost, and Stephen Bach. Learning to generate instruction tuning datasets for zero-shot task adaptation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12585–12611. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.748.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, 2021. doi: 10.1613/jair.1.12125.

Tanmay Parekh, Yuxuan Dong, Lucas Bandarkar, Artin Kim, I-Hung Hsu, Kai-Wei Chang, and Nanyun Peng. SNaRe: Domain-aware data generation for low-resource event detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20583–20604. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1039.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. DataDreamer: A tool for synthetic data generation and reproducible LLM workflows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3781–3799. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.208.

Stephen Pilli and Vivek Nallur. Predicting biased human decision-making with large language models in conversational settings. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pp. 2084–2119. Association for Computing Machinery, 2026. doi: 10.1145/3742413.3789127.

William Samuelson and Richard Zeckhauser. Status quo bias in decision making. *Journal of Risk and Uncertainty*, 1:7–59, 1988. doi: 10.1007/BF00055564.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive biases in large language models: A survey and mitigation experiments. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pp. 1009–1011. Association for Computing Machinery, 2025. doi: 10.1145/3672608.3707812.

Amos Tversky and Daniel Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973. doi: 10.1016/0010-0285(73)90033-9.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124.

Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683.

Matthias von Davier. Automated item generation with recurrent neural networks. *Psychometrika*, 83 (4):847–857, 2018. doi: 10.1007/s11336-018-9608-y.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.754.

Dirk U. Wulff and Rui Mata. Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, 9:944–954, 2025. doi: 10.1038/s41562-024-02089-y.

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, May Dongmei Wang, Wei Jin, Joyce Ho, and Carl Yang. Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15496–15523. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.916.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J. Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55734–55784, 2023. doi: 10.52202/075280-2433. Datasets and Benchmarks Track.

Ingo Ziegler, Abdullatif Köksal, Desmond Elliott, and Hinrich Schütze. CRAFT your dataset: Task-specific synthetic dataset generation through corpus retrieval and augmentation. *Transactions of the Association for Computational Linguistics*, 13:1693–1721, 2025. doi: 10.1162/tacl.a.56.

A Supplementary Implementation and Reproducibility Details

This appendix documents the implementation, study configuration, conditioning packages, validation workflow, blind-review materials, provenance records, release boundary, and supplementary analyses. The primary statistical framing and reported values are unchanged from the main paper. Private Atlas source text, internal unblinding keys, and other sensitive materials are not reproduced.

A.1 End-to-End Study Workflow

Figure 1 summarizes the implemented study workflow. A versioned study configuration selects supported bias–domain relationships and constructs a matched generation plan. The plan crosses 31 relationships with three conditioning packages, producing 93 relationship–condition cells. Each cell is executed once and requests ten questionnaire items from the language model, yielding 930 planned candidates.

Generated responses are parsed from JSON-mode outputs and passed through deterministic structural validation. Each candidate receives an accepted, warning, or rejected status. Accepted and warning-status candidates remain eligible for expert review, whereas rejected candidates are excluded before packet construction. Validation is non-destructive and non-regenerative: candidate text is not repaired, and a validation result does not trigger a replacement generation request.

Review-eligible candidates are exported into condition-blind packets. Reviewer-facing records contain the respondent-facing item, its application domain, an opaque identifier, and the fields required for evaluation. Hidden target-bias labels, generation conditions, Atlas source text, reference examples, and internal provenance metadata remain outside the reviewer-facing packets. After review is complete, opaque identifiers are joined to internal target and condition metadata for outcome computation and statistical analysis.

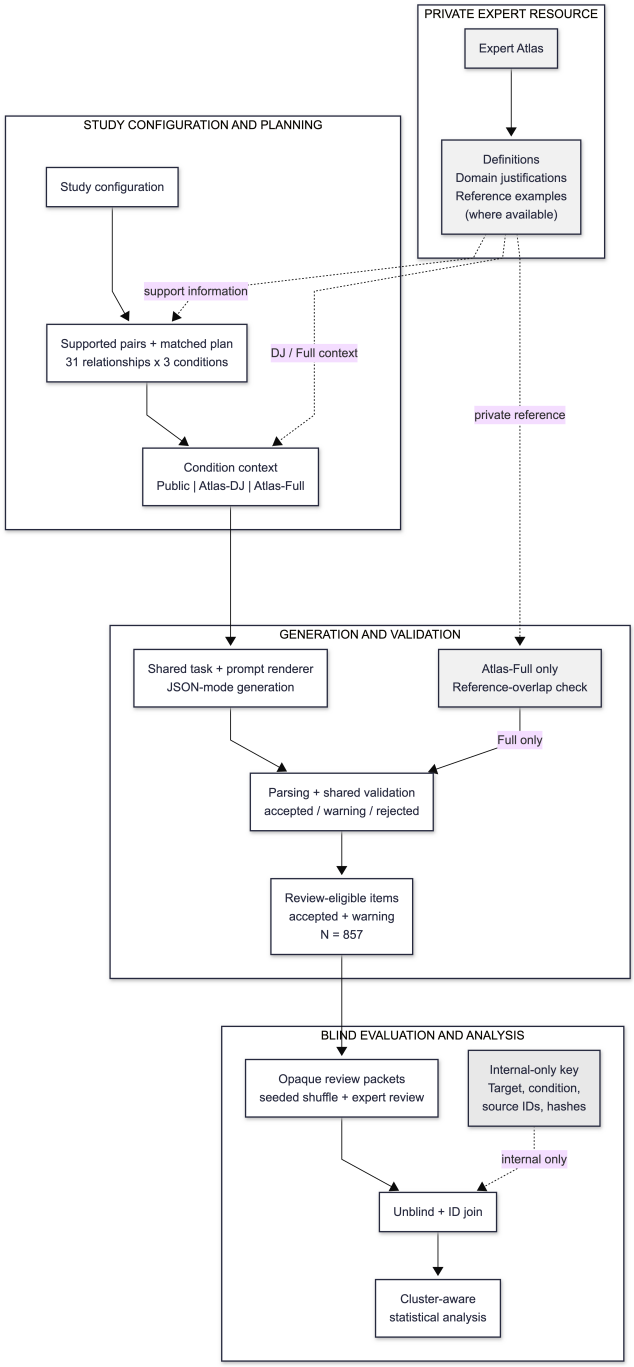


Figure 1: End-to-end study workflow from private expert conditioning and matched generation through automated validation, blind review, unblinding, and relationship-level analysis. Atlas-Full additionally applies reference-overlap validation.

Table 2: Generation configuration used for the study.

Parameter	Value
Provider/interface	OpenAI Chat Completions API
Model	`gpt-4o-2024-08-06`
Temperature	0.7
Top-p	1.0
Maximum completion length	4096 tokens
Generation seed	7
Study version	`v1`
Prompt version	`v1.1`
Candidates per cell	10
Supported relationships	31
Generation cells	93
Response format	JSON mode
Application retries	None

Table 3: Condition-specific information and safeguards.

Component	Public	Atlas-DJ	Atlas-Full
Shared generation task	Yes	Yes	Yes
Bias and domain identifiers	Yes	Yes	Yes
Public definition	Yes	No	No
Expert definition	No	Yes	Yes
Domain-specific justification	No	Yes	Yes
Expert reference example	No	No	Yes
No-copy instruction	No	No	Yes
Reference-overlap validation	No	No	Yes

A.2 Study Configuration and Generation Setup

The study configuration is represented by the research YAML files and the resolved run manifest. Generation uses the OpenAI Chat Completions API with model `gpt-4o-2024-08-06`. The principal configuration is shown in Table 2.

The model receives a single user message containing the rendered generation prompt. JSON mode requests a JSON-formatted response; the implementation does not use JSON-Schema Structured Outputs. For a fixed study configuration, source snapshot, and prompt version, prompt construction and prompt hashing are deterministic. Model generation remains stochastic, so the recorded seed is a repeatability control rather than a guarantee of bit-identical responses.

The generation guard requires the appropriate API configuration and explicit execution authorization. The model can be overridden through the command-line interface, while the remaining study parameters are configuration-driven. Each relationship–condition cell is submitted once. Provider or parsing failures raise an error and may leave a partial run manifest, but validation outcomes do not trigger application-level retries or quality-based regeneration.

A.3 Condition-Specific Prompt Construction

All three conditions use the same respondent-facing generation task, output requirements, model, decoding configuration, requested candidate count, parsing path, and shared structural validators. Their condition-specific components are summarized in Table 3.

The Public condition supplies a literature-grounded public definition. Atlas-DJ supplies an expert-authored definition and a domain-specific justification. Atlas-Full supplies the same expert definition and justification together with an expert reference example, an explicit no-copy instruction, and a reference-specific overlap check.

These are complete conditioning-package contrasts. Public versus Atlas-DJ changes both the definition source and the availability of domain-specific justification. Atlas-DJ versus Atlas-Full introduces

the reference example, no-copy instruction, and reference-overlap validation. Neither contrast is interpreted as an isolated component-level causal ablation.

The shared task requests respondent-facing questionnaire items situated in the selected domain. Items are required to express the target mechanism through observable wording or response options, include an explicit response format, avoid naming the target bias, avoid diagnosing the respondent, and avoid reproducing prompt instructions or placeholder text.

Representative released items. To make the generation target concrete, the following three sanitized items illustrate the respondent-facing output under the Public, Atlas-DJ, and Atlas-Full conditions. Where possible, the examples correspond to the same bias–domain relationship across conditions. They are included for interpretability only and were not selected to support a particular outcome or used to alter the blinded analysis. The complete approved example-item set is included in the supplementary package. The following examples are drawn from the same Anchoring Bias–Finance & Investing relationship under all three conditioning packages. They illustrate the respondent-facing generation target and were selected descriptively as a typical matched triplet; they were not used to select or alter the statistical analysis.

Condition	Representative generated item
Public	Given that some investors have set a target of $100,000 for their emergency fund, what do you consider a sufficient amount?
Atlas-DJ	The stock you bought last year was at $100 per share. Based on this information, what is your estimate of its current value?
Atlas-Full	A real estate investment trust (REIT) used to yield 7% annually. How do you evaluate its attractiveness if it now yields 4%?

A.4 Atlas Study Subset and Private-Source Boundary

The complete private Atlas contains 229 cognitive biases, 30 application domains, and 2,385 curated bias–domain relationships. The experiment uses a fixed subset of 31 supported relationships spanning 10 biases and five application domains. The subset is selected from relationships for which curated conditioning records are available and is not intended to be statistically representative of the complete Atlas.

The complete Atlas workbook may contain expert-authored definitions, domain-specific justifications, reference examples, internal identifiers, and other annotations. These records remain private outside the evaluated study subset. For reproducibility, the supplementary package releases only the approved study-support records associated with the 31 evaluated bias–domain relationships. The released study subset is limited to the Atlas fields used to construct the evaluated condition prompts, together with permitted identifiers, condition metadata, generated candidates, review outcomes, and non-sensitive provenance fields. An unsupported bias–domain pair is not treated as psychologically invalid or inapplicable.

The Atlas support structure determines where expert-grounded generation is available. It is not used as a new cognitive-bias taxonomy or graph-learning method. In Atlas-Full, the relevant reference example is supplied during prompt construction and used for the no-copy and overlap safeguards. The study-support release includes the reference examples used for the evaluated Atlas-Full relationships for which a reference example was supplied during prompt construction. The complete Atlas and all unrelated reference examples remain private.

A.5 Public Definitions and Domain Specifications

The Public condition uses short, study-authored definitions stored in `configs/research/subset_v1.yaml`. These definitions are concise condensations of public psychological sources and were written independently of the proprietary Atlas definitions. They do not reproduce Atlas definitions, domain-specific justifications, or reference examples. The public source records and bibliographic locators are listed in Table 4; each record was verified against a public bibliographic source.

Table 4: Public sources underlying the study-authored definitions used in the Public condition. The definitions are concise study specifications written independently of the private Atlas source text.

Bias label	Primary public source	DOI or locator
Anchoring Bias	Tversky and Kahneman (1974), *Science*	10.1126/science.185.4157.1124
Framing Effect	Tversky and Kahneman (1981), *Science*	10.1126/science.7455683
Confirmation Bias	Nickerson (1998), *Review of General Psychology*	10.1037/1089-2680.2.2.175
Authority Bias	Milgram (1963), *Journal of Abnormal and Social Psychology*; Cialdini (2009), *Influence: Science and Practice*, as a secondary synthesis	10.1037/h0040525; ISBN 9780205609994
Bandwagon Effect	Leibenstein (1950), *Quarterly Journal of Economics*	10.2307/1882692
Availability Heuristic	Tversky and Kahneman (1973), *Cognitive Psychology*	10.1016/0010-0285(73)90033-9
Default Effect	Johnson and Goldstein (2003), *Science*	10.1126/science.1091721
Endowment Effect	Kahneman, Knetsch, and Thaler (1990), *Journal of Political Economy*	10.1086/261737
Status Quo Bias	Samuelson and Zeckhauser (1988), *Journal of Risk and Uncertainty*	10.1007/BF00055564
Sunk Cost Fallacy	Arkes and Blumer (1985), *Organizational Behavior and Human Decision Processes*	10.1016/0749-5978(85)90049-4

Table 5: Candidate-flow counts before blind review.

Status	Count
Raw candidates	930
Accepted	803
Warning-status	54
Rejected	73
Review-eligible	857

The five domain specifications are study-authored applied-context descriptions aligned to Atlas domain names only. They describe Negotiation & Pricing, Hiring & Recruitment, Finance & Investing, Healthcare & Medicine, and Consumer Behavior & Retail. They do not expose or reproduce private Atlas content.

The supplementary package provides the public-definition records and the approved study-support subset used in the experiment. The public definitions are concise study-authored condensations of public sources and do not reproduce private Atlas text. The study-support subset contains the 31 evaluated bias–domain relationships and the condition-relevant Atlas records needed to audit the evaluated prompts, including the definitions, domain-specific justifications, and reference examples used in the evaluated prompts. The package also provides de-identified example items, candidate and review-label subsets, analysis code, resolved configuration, and documentation required to audit the reported analysis. The complete Atlas, unrelated relationships, internal join keys, and unreleased source records remain private.

A.6 Candidate Flow and Automated Validation

The 31 supported relationships crossed with three conditioning packages produced 93 generation cells. Each cell requested ten candidates, yielding 930 raw candidates. All 93 generation requests completed successfully. Automated validation produced the candidate flow in Table 5.

The 857 review-eligible candidates were distributed as follows: 288 Public, 282 Atlas-DJ, and 287 Atlas-Full. The unequal condition totals result from pre-review validation rather than post hoc sampling or removal after unblinding. Atlas-Full additionally performs reference-overlap validation, so pre-review filtering is not perfectly symmetric across conditions.

Shared validation checks include empty or overlong text, placeholder or prompt fragments, respondent orientation, diagnosis or intent language, missing question structure, target-bias-name leakage, and exact or near-duplicate items. The implemented maximum character threshold is 900. The implementation uses a near-duplicate threshold of 0.92 for the relevant duplicate check.

Validation statuses are assigned without modifying candidate text:

$$v(x) \in \{\text{accepted}, \text{warning}, \text{rejected}\}.$$

Warnings remain review-eligible unless another rule produces rejection. Diagnosis mentions, missing second-person language, or distinctive number reuse may produce warnings, while structural failures, prompt leakage, target-label leakage, and hard duplicate conditions may produce rejection. No warning or rejection triggers a replacement generation request.

Atlas-Full additionally performs an in-memory comparison between each candidate and the private reference example. Exact or sufficiently high overlap can produce rejection, while lower overlap can produce a warning. The reference example and overlap comparison text are not stored in reviewer-facing packets. The overlap calculation is a safeguard against copying and is not interpreted as a psychological-quality measure.

A.7 Blind-Review Materials and Packet Construction

All review-eligible items were evaluated by one expert reviewer with graduate psychology training through the doctoral level and university research and teaching experience. The reviewer had contributed to Atlas curation. Review was conducted before access to the internal condition and target-label mapping.

The packet-construction procedure assigns the three conditions to opaque packet identifiers using a seeded randomization procedure. The default packet-randomization seed is 7, and item order is independently shuffled within each packet. Reviewer-facing share records contain the opaque share identifier, item text, application domain, model field, and blank evaluation fields. They do not contain the target bias, generation condition, Atlas source text, reference example, prompt hash, or internal provenance key.

The same rubric and interface were used across conditions. For every item, the reviewer selected the best-represented bias from the study's closed 10-label vocabulary and independently evaluated neighboring-bias contamination, domain relevance, survey usability, and the retain–revise–reject editorial decision. Human judgments were used only as evaluation labels; they did not trigger item repair, regeneration, or removal.

All ratings were finalized before unblinding. Opaque identifiers were then rejoined with hidden targets, conditions, and provenance metadata for analysis. No item was rescored, reclassified, removed, or otherwise modified after unblinding.

As an intra-rater consistency check, a stratified sample of 30 items underwent a subsequent blind re-review. The second ratings agreed with the original ratings for all 30 selected bias labels, 29 of 30 auxiliary target-presence judgments, and 27 of 30 editorial decisions. The target-presence field was not used to compute Bias-ID, which was defined exclusively by exact equality between the selected canonical label and the hidden target label. The three editorial disagreements were all retain-versus-revise rather than rejection. This check supports within-reviewer consistency but does not establish inter-rater reliability.

A.8 Review-Label Schema and Outcome Construction

The review-label schema contains the fields required to assess construct fidelity, neighboring-bias contamination, applied-domain relevance, practical usability, and editorial suitability. The reviewer selected the best-represented bias from the study's closed 10-label vocabulary and independently assigned the remaining evaluation labels. The operational rubric is summarized in Table 6.

For an item with hidden target bias b and reviewer-selected canonical label $\hat{b}$, Bias-ID is defined by exact label equality:

$$\mathrm{BiasMatch} = \mathbf{1}\{\hat{b} = b\}. \tag{4}$$

This implementation compares the canonical reviewer-selected label with the canonical target label. It does not use a separate target-presence field. Bias-ID accuracy is the proportion of reviewed items satisfying $\mathrm{BiasMatch} = 1$. Because the reviewer selects from a closed 10-label vocabulary, uniform random selection has a nominal expected accuracy of 10%.

The primary contamination outcome is:

Table 6: Operational review rubric used for blind evaluation. Categories were assigned before unblinding and were not used to repair, regenerate, or remove candidate items.

Field	Allowed categories	Operational definition and analytic use
Bias identification	One label from the closed 10-label vocabulary	The reviewer selected the single bias best represented by the item. Bias-ID is scored as an exact match between the selected canonical label and the hidden target label.
Auxiliary target presence	`yes`, `no`	A separate target-presence judgment recorded for the intra-rater consistency check; not used in Bias-ID or any primary outcome.
Neighboring-bias contamination	`none_apparent`, `possible`, `likely`	The reviewer assessed whether a competing or additional mechanism was present. The primary contamination outcome equals one only for `none_apparent`; `likely` contamination is excluded from the soft composite.
Domain relevance	`strong`, `acceptable`, `weak`, `incorrect`	The reviewer assessed how clearly the item expressed the target mechanism in the specified application domain. The proportion receiving `strong` was treated as a secondary outcome.
Survey usability	`usable`, `usable_with_revision`	The reviewer assessed whether the item could be used directly or after limited revision. The retained review set contained 808 directly usable items and 49 items usable with revision.
Editorial decision	`retain`, `revise`, `reject`	The reviewer assigned an overall editorial disposition. Retain and revise decisions contribute to the soft composite; only retain contributes to the strict composite.
Optional fields	Naturalness, originality, reviewer notes	These fields were retained for supplementary review context and were not included among the four primary outcomes.

$$\text{ContaminationNone} = \mathbf{1}\{\text{contamination rating} = \texttt{none_apparent}\}. \tag{5}$$

The strict and soft composites are defined as:

$$\text{Strict} = \text{Retain} \wedge \text{BiasMatch} \wedge \text{ContaminationNone} \wedge \text{Usable}. \tag{6}$$

$$\text{Soft} = (\text{Retain} \vee \text{Revise}) \wedge \text{BiasMatch} \wedge \neg\text{LikelyContam} \wedge (\text{Usable} \vee \text{UsableWithRevision}). \tag{7}$$

Across the 857 reviewed items, contamination ratings were `none_apparent` for 573 items, `possible` for 238 items, and `likely` for 46 items. Editorial decisions were retain for 605 items, revise for 223 items, and reject for 29 items. These category counts are descriptive and do not constitute independent inferential observations.

The four primary outcomes are Bias-ID accuracy, Contamination none, the strict composite, and the soft composite. Domain relevance, usability, and editorial decisions are secondary or descriptive outcomes. These measures assess item-level construct fidelity and dataset utility; they do not independently validate respondent behavior or the underlying Atlas taxonomy.

A.9 Provenance, Identifiers, and Release Boundary

Generated records retain identifiers and fingerprints that connect each item to its generation context. Depending on the artifact, these fields include:

- record and request identifiers;
- bias and domain identifiers;
- internal generation-condition metadata;
- prompt version and prompt hash;
- effective configuration and configuration hash;
- source-record identifiers;
- Atlas snapshot checksum and Atlas-context hash;
- public-definition hash;
- whether an Atlas reference example was supplied;

- requested seed and model metadata;
- validation statuses and validation messages; and
- generation and processing timestamps, where retained.

These fields are divided between reviewer-facing, analysis, study-support, and internal artifacts. Reviewer-facing packets contain only the fields required for evaluation. Internal packet keys and join files contain the hidden condition-to-packet mapping, true target labels, and provenance keys. The complete Atlas workbook, unrelated Atlas relationships, internal packet keys, private join files, and unreleased source records remain outside the release package.

The supplementary package provides the analysis code, resolved generation configuration, public-definition records, the 31-relationship study-support manifest, de-identified reviewer-facing packets, generated candidate subsets, review labels, and non-sensitive provenance fields. Each released review packet is checked to ensure that target labels, condition identities, packet mappings, reviewer identifiers, private Atlas text, reference examples, and internal join keys are absent.

Because Atlas-conditioned prompts depend on private expert-authored material, the supplementary package supports auditing and analysis without permitting independent reconstruction of the complete Atlas-conditioned generation context. Version identifiers, configuration fingerprints, and prompt-level hashes provide traceability while preserving the privacy of the underlying expert source.

A.10 Statistical Implementation

The primary inferential unit is the matched bias–domain relationship. The 857 reviewed items are nested within 31 relationships and are not treated as independent inferential observations.

For each binary outcome, the analysis first computes a success proportion within every relationship–condition cell. Pairwise condition contrasts are calculated as within-relationship differences in these proportions and averaged across the 31 relationships, giving each relationship equal weight. Item-weighted condition rates are reported separately as descriptive summaries.

Uncertainty is quantified using a nonparametric cluster bootstrap that resamples complete bias–domain relationships with replacement. The analysis uses 10,000 bootstrap replicates and percentile two-sided 95% confidence intervals.

Matched permutation tests exchange condition labels within relationships under the null hypothesis of no condition difference. The analysis uses 100,000 Monte Carlo draws. The reported permutation p-values are based on the same relationship-level matched estimand used for the primary contrasts.

The primary inferential family contains four outcomes and three prespecified pairwise contrasts:

- Atlas-DJ versus Public;
- Atlas-Full versus Public; and
- Atlas-Full versus Atlas-DJ.

This produces 12 primary comparisons. Raw permutation p-values and Holm-adjusted p-values for all 12 comparisons are reported in Table 7. The Holm adjustment is applied jointly across this family.

Logistic generalized estimating equation models with exchangeable working correlation are reported as robustness analyses. These models use item-level observations clustered by matched bias–domain relationship. The GEE analyses are not treated as replacements for the primary relationship-weighted permutation analysis and are not included in the primary Holm family.

Bias-specific analyses, Confirmation Bias analyses, neighboring-bias confusion patterns, leave-one-bias-out analyses, and leave-one-domain-out analyses are exploratory. The leave-one-bias-out and leave-one-domain-out analyses are reported as item-weighted Atlas-Full–Public point-estimate sensitivity analyses. These analyses are descriptive and are not accompanied by cluster-bootstrap confidence intervals, alternative condition contrasts, or relationship-weighted estimates.

Table 7: Complete primary inferential results. Matched differences are relationship-weighted percentage-point contrasts. Confidence intervals are based on 10,000 relationship-cluster bootstrap replicates. Raw p-values are based on 100,000 matched permutation draws; Holm-adjusted values use the joint family of 12 primary comparisons.

Outcome	Contrast	Δ (pp)	95% CI	Raw p	Holm p
Bias-ID accuracy	Atlas-DJ − Public	+3.15	[−5.64, 12.96]	0.54804	1.00000
Bias-ID accuracy	Atlas-Full − Public	+4.17	[−3.20, 13.07]	0.37087	1.00000
Bias-ID accuracy	Atlas-Full − Atlas-DJ	+1.02	[−4.90, 7.23]	0.75388	1.00000
Contamination none	Atlas-DJ − Public	+2.06	[−6.37, 10.75]	0.64314	1.00000
Contamination none	Atlas-Full − Public	+8.43	[−1.35, 18.64]	0.11852	1.00000
Contamination none	Atlas-Full − Atlas-DJ	+6.36	[−1.34, 14.60]	0.13300	1.00000
Strict composite	Atlas-DJ − Public	+4.02	[−6.26, 14.14]	0.45645	1.00000
Strict composite	Atlas-Full − Public	+8.90	[−0.27, 18.65]	0.07590	0.91080
Strict composite	Atlas-Full − Atlas-DJ	+4.88	[−3.42, 14.18]	0.30326	1.00000
Soft composite	Atlas-DJ − Public	+2.25	[−6.76, 12.25]	0.67016	1.00000
Soft composite	Atlas-Full − Public	+7.35	[−1.65, 17.88]	0.17163	1.00000
Soft composite	Atlas-Full − Atlas-DJ	+5.11	[−1.40, 11.76]	0.15167	1.00000

Table 8: Bias-level sample sizes and Bias-ID accuracy by condition. Rates are item-weighted within each bias–condition group.

Bias	n_{rel}	Public n	Public ID	DJ n	DJ ID	Full n	Full ID
Anchoring Bias	5	48	95.8%	45	80.0%	48	95.8%
Authority Bias	2	20	100.0%	18	100.0%	18	100.0%
Availability Heuristic	3	28	89.3%	25	92.0%	27	100.0%
Bandwagon Effect	2	20	90.0%	20	95.0%	19	84.2%
Confirmation Bias	3	28	10.7%	29	75.9%	26	65.4%
Default Effect	3	28	96.4%	30	100.0%	30	93.3%
Endowment Effect	3	23	100.0%	21	90.5%	23	91.3%
Framing Effect	4	39	92.3%	36	100.0%	38	97.4%
Status Quo Bias	3	26	100.0%	29	86.2%	30	86.7%
Sunk Cost Fallacy	3	28	92.9%	29	93.1%	28	96.4%

A.11 Primary Results and Robustness Outputs

The main paper reports item-weighted descriptive condition rates and the principal relationship-weighted Atlas-Full–Public contrasts. The complete matched inferential results for all four primary outcomes and all three prespecified contrasts are reported in Table 7.

None of the 12 primary comparisons survived Holm correction, and every primary confidence interval included zero. The observed contamination and composite-fidelity differences are therefore interpreted as potentially practically relevant but uncertain package-level contrasts rather than established population-level effects.

Complete item-weighted bias-level results are reported in Tables 8 and 9. The number of matched relationships and reviewed items per bias and condition are shown in Table 8. The outcome rates in Table 9 are reported in the order Bias-ID, Contamination none, strict composite, and soft composite.

Confirmation Bias relationship-level breakdown. The aggregate Confirmation Bias result is distributed across three verified bias–domain relationships. The following breakdown reports the number correct, denominator, and percentage within each condition for each relationship. These values are descriptive and make explicit the small relationship-level support underlying the aggregate comparison.

Confirmation Bias–domain relationship	Public	Atlas-DJ	Atlas-Full
Finance & Investing	0/10 (0.0%)	10/10 (100.0%)	10/10 (100.0%)
Healthcare & Medicine	0/10 (0.0%)	6/9 (66.7%)	2/8 (25.0%)
Hiring & Recruitment	3/8 (37.5%)	6/10 (60.0%)	5/8 (62.5%)
Total	**3/28 (10.7%)**	**22/29 (75.9%)**	**17/26 (65.4%)**

Table 9: Complete item-weighted bias-level outcome rates. Each cell reports percentages in the order Bias-ID, Contamination none, strict composite, and soft composite. Denominators are given in Table 8.

Bias	Public ID/C/St/So	Atlas-DJ ID/C/St/So	Atlas-Full ID/C/St/So
Anchoring Bias	95.8/43.8/37.5/77.1	80.0/60.0/48.9/73.3	95.8/64.6/47.9/89.6
Authority Bias	100.0/95.0/95.0/100.0	100.0/94.4/94.4/100.0	100.0/83.3/72.2/88.9
Availability Heuristic	89.3/96.4/60.7/71.4	92.0/84.0/40.0/68.0	100.0/85.2/77.8/96.3
Bandwagon Effect	90.0/75.0/45.0/85.0	95.0/75.0/60.0/90.0	84.2/63.2/47.4/84.2
Confirmation Bias	10.7/46.4/0.0/7.1	75.9/72.4/27.6/58.6	65.4/80.8/34.6/61.5
Default Effect	96.4/42.9/39.3/96.4	100.0/66.7/53.3/100.0	93.3/60.0/53.3/86.7
Endowment Effect	100.0/78.3/30.4/87.0	90.5/57.1/38.1/85.7	91.3/69.6/47.8/87.0
Framing Effect	92.3/69.2/51.3/92.3	100.0/75.0/47.2/97.2	97.4/89.5/52.6/97.4
Status Quo Bias	100.0/65.4/53.8/92.3	86.2/44.8/34.5/65.5	86.7/66.7/40.0/83.3
Sunk Cost Fallacy	92.9/46.4/32.1/89.3	93.1/41.4/41.4/89.7	96.4/57.1/53.6/89.3

Table 10: Weighting-sensitivity results for the Atlas-Full–Public contrast. Relationship-weighted values are matched estimates; item-weighted values are descriptive differences.

Outcome	Contrast	Relationship-weighted	Item-weighted
Bias-ID accuracy	Atlas-Full − Public	$+4.17$	$+4.83$
Contamination none	Atlas-Full − Public	$+8.43$	$+8.58$
Strict composite	Atlas-Full − Public	$+8.90$	$+8.86$
Soft composite	Atlas-Full − Public	$+7.35$	$+7.94$
Contamination none	Atlas-DJ − Public	$+2.06$	$+2.41$
Contamination none	Atlas-Full − Atlas-DJ	$+6.36$	$+6.17$

These relationship-level values are descriptive. Atlas-DJ and Atlas-Full exceed Public in all three Confirmation Bias–domain relationships, although the magnitude of the difference varies across relationships.

For Confirmation Bias, blind identification accuracy was 3/28 (10.7%) under Public, 22/29 (75.9%) under Atlas-DJ, and 17/26 (65.4%) under Atlas-Full. These items came from three Confirmation Bias–domain relationships. Under Public conditioning, 13 items were identified as Bandwagon Effect, 10 as Authority Bias, two as Availability Heuristic, and three as Confirmation Bias.

The item-weighted Full–Public Bias-ID difference was $+4.83$ percentage points across all biases and -0.75 points after excluding Confirmation Bias. After that exclusion, the other item-weighted Full–Public differences remained positive: $+5.88$ points for contamination-free generation, $+5.95$ points for the strict composite, and $+2.73$ points for the soft composite.

Weighting sensitivity results are reported in Table 10. The relationship-weighted matched estimate is the primary estimand; item-weighted estimates are descriptive sensitivity results.

The item-weighted contamination-none difference was $+8.58$ points compared with $+8.43$ points under relationship weighting. The strict-composite difference was $+8.86$ points under item weighting and $+8.90$ points under relationship weighting.

Complete GEE results are reported in Table 11. All reported models converged. The GEE models used item-level observations with matched bias–domain relationship as the clustering unit and exchangeable working correlation.

The Atlas-DJ versus Public GEE association for strong domain relevance was OR $= 2.23$, 95% CI $[1.39, 3.58]$, $p = 0.0008$. This result is secondary and is not included in the primary Holm-adjusted family.

The complete item-weighted leave-one-bias-out results are reported in Table 12 as descriptive point estimates for the Atlas-Full–Public contrast. The analysis assesses sensitivity to omitting each bias; results are intentionally limited to item-weighted Atlas-Full–Public point estimates.

The Bias-ID contrast is most sensitive to Confirmation Bias: the item-weighted Atlas-Full–Public difference changes from $+4.83$ points with all biases retained to -0.75 points after its exclusion. The other three outcomes remain positive after excluding Confirmation Bias and after excluding each of the other individual biases.

Table 11: Cluster-aware logistic GEE robustness analyses. The reference condition is Public for Atlas-DJ and Atlas-Full comparisons and Atlas-DJ for the Atlas-Full–Atlas-DJ comparison.

Outcome	Contrast	OR	95% CI	p
Bias-ID accuracy	Atlas-DJ vs Public	1.42	[0.58, 3.49]	0.447
Bias-ID accuracy	Atlas-Full vs Public	1.59	[0.71, 3.56]	0.263
Bias-ID accuracy	Atlas-Full vs Atlas-DJ	1.12	[0.56, 2.23]	0.750
Contamination none	Atlas-DJ vs Public	1.11	[0.77, 1.62]	0.566
Contamination none	Atlas-Full vs Public	1.51	[0.95, 2.38]	0.079
Contamination none	Atlas-Full vs Atlas-DJ	1.35	[0.93, 1.97]	0.119
Strict composite	Atlas-DJ vs Public	1.17	[0.78, 1.76]	0.449
Strict composite	Atlas-Full vs Public	1.44	[0.97, 2.13]	0.070
Strict composite	Atlas-Full vs Atlas-DJ	1.23	[0.85, 1.77]	0.269
Soft composite	Atlas-DJ vs Public	1.18	[0.64, 2.16]	0.599
Soft composite	Atlas-Full vs Public	1.72	[0.87, 3.41]	0.118
Soft composite	Atlas-Full vs Atlas-DJ	1.46	[0.88, 2.44]	0.144
Strong domain relevance	Atlas-DJ vs Public	2.23	[1.39, 3.58]	0.0008
Strong domain relevance	Atlas-Full vs Public	1.32	[0.78, 2.25]	0.302
Strong domain relevance	Atlas-Full vs Atlas-DJ	0.59	[0.33, 1.05]	0.075

Table 12: Item-weighted leave-one-bias-out descriptive sensitivity analysis for the Atlas-Full–Public contrast. Values are percentage-point differences; the first row retains all biases. Results are point estimates only.

Excluded bias	Bias-ID	Contam. none	Strict	Soft
All biases retained	+4.83	+8.58	+8.86	+7.94
Anchoring Bias	+5.79	+6.14	+8.55	+7.03
Authority Bias	+5.26	+10.18	+11.38	+9.38
Availability Heuristic	+4.23	+10.77	+8.08	+6.15
Bandwagon Effect	+5.60	+10.07	+9.33	+8.58
Confirmation Bias	**−0.75**	**+5.88**	**+5.95**	**+2.73**
Default Effect	+5.67	+7.77	+8.29	+9.85
Endowment Effect	+6.01	+10.08	+8.12	+8.63
Framing Effect	+4.82	+6.83	+10.04	+8.43
Status Quo Bias	+6.72	+9.40	+11.32	+9.69
Sunk Cost Fallacy	+4.97	+8.36	+7.51	+8.80

The complete item-weighted leave-one-domain-out results are reported in Table 13 as descriptive point estimates for the Atlas-Full–Public contrast. The analysis assesses sensitivity to omitting each application domain. It is a leave-one-domain-out sensitivity analysis, not a domain-by-condition interaction model.

Table 13: Item-weighted leave-one-domain-out descriptive sensitivity analysis for the Atlas-Full–Public contrast. Values are percentage-point differences.

Excluded domain	Bias-ID	Contam. none	Strict	Soft
Consumer Behavior & Retail	+3.85	+8.81	+10.63	+4.14
Finance & Investing	+5.09	+10.63	+6.24	+11.29
Healthcare & Medicine	+5.28	+6.42	+9.95	+9.12
Hiring & Recruitment	+4.47	+7.88	+8.69	+7.08
Negotiation & Pricing	+5.40	+9.31	+8.61	+8.48

The leave-one-domain-out estimates remain positive for contamination and both composite outcomes across all five domain exclusions. The Bias-ID estimates also remain positive, although their magnitudes vary across omitted domains.

The omnibus Friedman results are reported in Table 14. These tests operate on relationship-level condition proportions and are not part of the primary Holm-adjusted permutation family.

Table 14: Relationship-level Friedman robustness tests with 31 matched relationships and two degrees of freedom. No multiplicity adjustment was applied.

Outcome	Statistic	p
Bias-ID accuracy	0.37	0.831
Contamination none	0.69	0.708
Strict composite	2.23	0.328
Soft composite	0.91	0.633
Strong domain relevance	6.60	0.037

A.12 Reproduction and Release Instructions

A complete reproduction should use the analysis code, resolved study configuration, the approved study-support subset for the 31 evaluated bias–domain relationships, the de-identified candidate and review-label subsets, and the analysis outputs. The generation workflow requires an appropriately configured model-provider account and the required environment variables. The analysis workflow can be run independently on the released review labels and does not require access to the complete private Atlas. The released study-support subset contains the condition-relevant Atlas records used by the evaluated Atlas-Full prompts, allowing reconstruction of the prompt inputs while leaving the complete Atlas private.

The run manifest records the effective generation configuration, environment dependency versions, source identifiers, prompt and configuration hashes, request counts, validation summaries, and output paths. The generation manifest for the reported run records a modified working-tree state. Exact regeneration should therefore use a clean source snapshot and record the corresponding commit identifier and tree status.

The implementation components relevant to reproduction include:

- generation entry points and study planning;
- prompt rendering and condition-specific context construction;
- provider and JSON-response handling;
- candidate parsing and validation;
- provenance and hashing;
- blind-packet construction and review-label joining; and
- cluster-aware statistical reanalysis.

The supplementary package provides dependency specifications, execution commands, expected intermediate counts, validation summaries, analysis seeds, and final tables. The complete Atlas workbook, unrelated Atlas relationships, internal packet keys, private join files, and unreleased source text remain outside the public package.