

TITLE PAGE

Moral Inflation in Synthetic Investors: When LLM Personas Replace Heterogeneity with Normative Archetypes

Full-Length Research Article

Seung Ho Jeon*

seungho.jeon@smail.astate.edu

Dr. H. Steve Leslie

sleslie@astate.edu

Dr. Hrishikesh Desai

hdesai@astate.edu

Arkansas State University

Corresponding author

Seung Ho Jeon, Arkansas State University, P.O. Box 600, State University, AR 72467, USA; Email: seungho.jeon@smail.astate.edu; Institutional telephone: +1 870-972-2100

Acknowledgments

The authors thank the participating human-sample respondents and the institutional review office for protocol oversight under FY24-25-430.

Funding

This research was supported by the Arkansas State University Middle East Studies Grant, which supported data collection in Qatar.

Declaration of interest

The authors report no competing interests to declare.

* Corresponding author

Moral Inflation in Synthetic Investors: When LLM Personas Replace Heterogeneity with Normative Archetypes

Abstract

Large language models (LLMs) are increasingly used to stand in for investors in surveys, experiments, and product research. The central validity risk may be more systematic than random simulation error: when financial choices carry clear social or moral prescriptions, LLMs may push synthetic investors toward norm-consistent archetypes. We test this possibility in Islamic finance, where formal prescriptions provide an unusually clear benchmark against which simulated behavior can be compared. Seven persona-conditioned LLMs generated 176,229 cleaned responses, benchmarked against 31 human participants on the same eleven-item instrument. Mean alignment is high ($r=0.854$), but it masks substantial distributional and structural distortion. Average sign-standardized Moral Inflation is 0.53 Likert points, AI response variance is roughly half the human variance, and religion effects are sharply amplified, reaching $\eta^2(2)=0.795$ for musharakah endorsement. The LLMs therefore do not simply reproduce investor heterogeneity. They move simulated people toward an idealized normative archetype. Islamic finance supplies the test environment, but the validity problem extends to any behavioral-finance setting in which models can infer how a responsible investor is supposed to act.

Keywords: Synthetic investors ; Large language models ; Moral inflation ; Behavioral finance ; Survey validity ; Islamic finance

Word count: 8,271

Introduction

Persona-conditioned large language models (LLMs) are increasingly asked to stand in for investors in surveys, experiments, product design, and policy research. The appeal is easy to understand. LLM-generated populations scale to millions of respondents at near-zero marginal cost, can be matched to population-demographic ratios, and often reproduce aggregate rankings on general attitudinal items with correlations in the 0.75 to 0.90 range (Argyle et al., 2023; Bisbee et al., 2024; Shrestha et al., 2025). Yet the most consequential error may be systematic rather than random: a model can infer the socially or morally appropriate financial choice and simulate an archetype instead of a heterogeneous person.

A fast-maturing validation literature, however, has documented four boundary conditions that complicate this promise. Synthetic respondents excel on aggregate mean reproduction. Even so, they consistently fail on heterogeneous questions, rare subgroups, individual-level prediction, and multivariate joint distributions (Shrestha et al., 2025). Demographic persona prompting can redistribute or increase errors across subgroups rather than reducing them (Gupta et al., 2024). Direct comparisons of LLM and human financial decisions show that LLMs form a distinctive cluster separate from human national samples. They are generally more rational and risk-neutral than humans, not merely noisy copies (Erdem and Ashok, 2026). Finally, LLM survey responses display measurable social-desirability bias. Responses shift toward morally laudable positions when the model can identify the item as an evaluative probe (Salecha et al., 2024).

Islamic finance offers an unusually clear test environment for this general problem. Financial activity is normatively anchored in principles of fairness (*adl*), risk sharing (*musharakah*), and the prohibition of exploitative interest (*riba*) and excessive uncertainty (*gharar*). Those prescriptions are explicit, while real investor behavior remains shaped by religious literacy, sectarian heterogeneity, communal norms, and the interaction between cultural identity and financial confidence (Din et al., 2021; Bialowolski et al., 2023). Existing LLM research in Islamic finance has focused mainly on knowledge accuracy: reproducing AAOIFI doctrine, matching human mufti rulings, and classifying products as Shariah compliant (Ali et al., 2025). That work does not tell us whether LLMs

reproduce actual human variation. Islamic finance is therefore the test environment, not the boundary of the contribution. The broader question is whether synthetic investors preserve behavioral heterogeneity when a financial domain supplies strong cues about how a responsible person ought to act.

This paper tests that broader synthetic-investor validity problem in the high-clarity environment of Islamic finance. We use an 11-item 7-point Likert instrument covering financial ethics and risk. Survey responses from 30,000 LLM-generated personas yield 176,229 post-cleaning records across seven usable LLMs. We compare them with responses from 31 human participants recruited under Institutional Review Board approval (protocol identifier withheld for anonymous review). Personas were stratified by age, gender, religion, and country using World Population Review (2024a), World Population Review (2024b), and World Population Review (2024c) national distributions, with deliberate over-sampling of the United States and Qatar.

The paper moves beyond the familiar observation that LLM personas differ from humans. It asks a more consequential question:

Do persona-conditioned LLMs preserve investor heterogeneity, or do they move simulated people toward norm-consistent archetypes when financial choices carry explicit moral prescriptions?

To answer that question, we advance three analytical innovations:

1. Validity decomposition into three nested criteria. Mean alignment asks whether average AI and human scores are similar. Distributional alignment asks whether AI responses reproduce the full human distribution, including disagreement, extremal responses, and variance. Structural demographic alignment tests whether religion, confidence, age, education, and financial experience predict AI responses in the same way they predict human responses. This decomposition addresses a recurring critique: mean-level correlation can mask severe subgroup and distributional pathologies (Shrestha et al., 2025).
2. Item-level Moral Inflation (MI_(j)) metric. The metric is sign-standardized so positive values always mean the LLM endorses a position as if it were more ethically conservative than the human benchmark on that dimension. We test whether moral inflation is domain-sensitive, item-sensitive, persona-sensitive, and distributional. Items are further classified on a three-component normative salience scale spanning Shariah relevance, social desirability, and speculation relevance. This lets us test whether moral inflation grows more severe when items explicitly cue religious or social-norm compliance.
3. Mechanism adjudication via six competing testable propositions. We treat the claim that LLMs are stricter than humans not as a one-line finding but as a phenomenon requiring mechanism adjudication. Six competing explanations are framed as testable propositions and evaluated against cross-model and cross-item evidence. The candidates are social-desirability bias, normative substitution of formal Shariah principle for behavioral realism, stereotype amplification of the Muslim-investor persona, safety-alignment effects that reward cautious ethical answers, variance collapse toward the corpus-central answer, and knowledge-behavior confusion in which the model answers what someone should do rather than what they would actually do.

The analysis makes four empirically grounded contributions:

4. It tests a general synthetic-investor validity problem in a financial environment where normative prescriptions are unusually clear. The study provides the first Islamic-finance-specific human-to-AI benchmark and reports demographic-stratum effect sizes for religion, age, gender, and country across all 11 attitudinal items.
5. It documents the coexistence of superficially strong mean alignment with severe distributional pathology. Sign-standardized moral inflation averages +0.53 Likert points, and variance compression reaches 48.3 percent, so AI standard deviations sit at roughly half of human standard deviations across the instrument.
6. It shows that demographic structural alignment is directionally correct but stereotype-amplified. Religion is the dominant variance driver on musharakah endorsement, with $\eta^2(2) = 0.795$. Muslim-persona within-group standard deviations, however, compress to less than one-sixth of Christian counterparts on the same item. This ceiling-effect homogenization appears in no real Muslim sample.

7. It supports a multi-mechanism account of moral inflation. Social-desirability, alignment effects, normative substitution, stereotype amplification, and confusion between knowledge and behavior each operate on different item and persona subsets. No single explanation is sufficient on its own.

Synthetic investors therefore cannot substitute for validated human behavioral data when financial choices are normatively legible. The immediate applications include Shariah-sensitive market research, financial-inclusion modeling, robo-advisory suitability calibration, and policy design, but the concern is broader: the same failure can arise whenever a model knows what a prudent, sustainable, fair, or socially responsible investor is supposed to do. Synthetic data can still serve as a normative audit benchmark for algorithmic drift and as a hypothesis-generation tool before costly fieldwork. In those narrower roles, a consistent idealized baseline is itself informative.

The paper proceeds as follows. Section 2 reviews prior work on LLM synthetic surveys, LLM-to-human financial-decision comparisons, Islamic-finance LLM knowledge audits, social-desirability bias in LLM outputs, cross-cultural behavioral finance, and the stated-preference-to-action gap. It then frames nine hypotheses: three validity-layer hypotheses and six mechanism-adjudication hypotheses. Section 3 describes the survey instrument, model inventory, persona-generation protocol, human-data ethics, the MI_(j) operationalization, and the statistical-analysis plan. Section 4 presents results organized by the three validity layers, reporting Moral Inflation, normative-salience, variance-compression, demographic-structure, cross-model, and OLS regression findings. Section 5 evaluates the six mechanism hypotheses against the evidence, considers the broader research-practice tradeoffs, and outlines explicit prompt-condition experimental designs for future work. Section 6 concludes with actionable principles for the responsible use of LLM personas in ethically and culturally sensitive domains.

Literature Review

Synthetic Respondents and the Validity-Decomposition Problem

Validating LLM survey outputs against human benchmarks is a rapidly growing subfield. The baseline finding is optimistic on its face. Bisbee et al. (2024) show that synthetic responses approximate human group-level means on general attitudinal items, with per-question mean correlations in the 0.80 to 0.90 range. Sun et al. (2024) extend the finding to nationally representative open-weight LLM samples.

But a fast-growing validation literature consistently identifies the same three pathologies. Synthetic respondents consistently fail on heterogeneous questions, rare subgroups, individual-level prediction, and multivariate joint distributions, even when group-level means appear correct (Shrestha et al., 2025). Demographic persona conditioning can increase rather than reduce cross-subgroup errors. Gupta et al. (2024) document gender and religious stereotyping with explicitly neutral prompts. The literature also consistently reports that group-effect sizes in synthetic data often exceed their human counterparts rather than mirroring them. Finally, cross-model studies document that LLM synthetic distributions are substantially narrower than human distributions, inflating statistical certainty and making subgroup comparisons unreliable.

These findings motivate this paper's central methodological move: decompose validity into mean, distributional, and structural layers, reporting all three rather than only the most favorable mean-level statistic.

LLMs and Financial Decision-Making

Recent work has begun directly comparing LLM and human financial decision-making. In the largest study to date, Erdem and Ashok (2026) find that LLMs form a distinctive preference cluster separable from all 50+ national human samples in their elicitation battery. LLMs are generally more rational in the classical-economics sense. They show less loss aversion, less ambiguity aversion, and less risk aversion than human respondents when risk is explicitly parameterized. That finding poses a key tension with the present study's directional expectations. The synthetic-investor literature finds LLMs less risk-averse, while the ethical-survey literature finds LLMs more risk-averse on normatively framed ethics items. Islamic finance is precisely the domain where this ambiguity

surfaces. Classical-lottery risk aversion may be low, while risk aversion on items that cue riba-adjacent speculation should be very high.

Islamic Finance as a High-Clarity Test Environment

Islamic finance is especially useful for studying normative archotyping because the formal prescriptions are visible and can be separated from lived investor behavior. The emerging LLM literature in this area has focused almost exclusively on knowledge-accuracy tasks. Ali et al. (2025) evaluate whether LLMs reproduce AAOIFI-aligned Shariah doctrine and whether their rulings match human mufti outputs, reporting moderate concordance but systematic bias toward leniency on borderline cases. A second line evaluates whether LLMs encode maqasid al-Shariah principles in ethical classification, document recommendation, and product screening, building on the same doctrinal-fidelity framework.

None of these studies test whether LLMs reproduce observed human behavioral variation. A model can know the formal Islamic rule perfectly and still fail to represent how real investors balance ethics, risk, returns, social influence, practical ambiguity, and religious interpretation. The distinction between knowledge accuracy and behavioral fidelity is the substantive gap addressed here, and it travels beyond Islamic finance to other settings in which a model can readily infer the approved answer.

Social-Desirability Bias in LLM Survey Outputs

A dedicated line of work confirms that LLM survey responses are measurably vulnerable to the same social-desirability biases known to affect human self-reports, potentially with larger effect sizes. Salecha et al. (2024) show that GPT-class models shift survey responses toward socially desirable positions when the evaluative framing of the probe is made explicit. Effect sizes on extreme-response shares are comparable to well-documented human social-desirability manipulations. The literature traces the bias to corpus-driven moral norms and safety-training reward shaping. Studies further show that the bias is scale-dependent. Frontier models exhibit stronger moral-saint behavior than smaller open-weight models, which is consistent with the safety-training channel being a major contributor.

The Islamic-finance setting allows a sharper test. If LLM social-desirability bias is a generic moral-saint phenomenon, it should be visible across all normatively cued items and should be strongest in the largest, most-aligned models. If, by contrast, it is an Islamic-finance-specific doctrine substitution effect, it should concentrate on Shariah-specific items and on Muslim-identified personas specifically. Section 5 brings the cross-model and cross-persona evidence to bear on this question.

Culture, Confidence, and the Stated-Preference-to-Action Gap

Cross-cultural behavioral finance identifies two structural regularities. Financial confidence is culturally mediated. Overconfident investors take greater risk, while underconfident investors self-report conservative or ethical portfolios even when realized actions deviate (Aristei and Gallo, 2021a; Aristei and Gallo, 2021b). These patterns interact with gender. Women tend to self-report lower financial confidence yet stronger prosocial and sustainability preferences (Aristei and Gallo, 2021b). Societal-level uncertainty avoidance and individualism to collectivism shape both risk-taking and ethical orientation (Bialowolski et al., 2023). In the Islamic-finance context specifically, religious literacy and social embeddedness moderate not merely knowledge but the moral confidence to act on stated convictions (Din et al., 2021; Adil et al., 2022).

A defining finding of behavioral ethics and sustainable-finance research is that self-reported ethical intentions routinely exceed realized behavior, creating a stated-preference-to-action gap. The literature documents this gap for sustainable-fund holdings. Studies show that overconfidence inflates self-reported ethical orientation relative to realized portfolio choices (Aristei and Gallo, 2021a; Aristei and Gallo, 2021b). Classic behavioral mechanisms such as loss aversion, herding, status-quo preference, and moral licensing drive individuals toward privately optimal choices despite publicly professed ideals (Adil et al., 2022; Bialowolski et al., 2023).

Bender and Koller (2020) offer the methodological bridge. LLMs replicate linguistic and normative patterns of human self-presentation, not necessarily the cognitive dissonance between ideals and action. On this account, synthetic survey data are expected to reproduce the aspirational self-representation layer of survey response more faithfully than the behavioral-disposition layer. The Islamic-finance setting provides a uniquely well-specified test, because the normative content of Shariah-compliant ideals is precisely articulated and empirically distinguishable from observed participation rates (Din et al., 2021).

Hypotheses

Nine testable propositions integrate the review. The first three decompose validity into nested layers, following the synthetic-survey pathology literature. The next six frame the competing mechanistic explanations of moral inflation introduced above. Each generates differential cross-item, cross-model, or cross-persona predictions that can be evaluated against the evidence.

Validity-layer hypotheses

- V1 (Mean alignment).** Per-question mean-level correlations will be high ($r > 0.75$) between the LLM-pooled sample and the human benchmark, consistent with prior synthetic-survey results on general attitudinal items.
- V2 (Distributional non-alignment).** Mean-level alignment will coexist with severe distributional distortion. AI sample standard deviations will be 25 to 50 percent narrower than human standard deviations on most items. Extremal-response shares of 1 and 7 will be systematically biased toward the more ethically stringent pole.
- V3 (Structural direction-but-not-magnitude alignment).** Demographic predictors such as religion, age, gender, risk tolerance, and religious practice will exhibit the same directional effects documented in the behavioral Islamic-finance literature. Their effect sizes, however, will be exaggerated and within-group variance compressed, consistent with stereotype amplification rather than authentic behavioral heterogeneity.

Mechanism-adjudication hypotheses

- M1 (Social-desirability bias).** Moral inflation will be largest on items rated highest on a generic social-desirability scale. It will also be largest in the most heavily safety-aligned frontier-closed models such as GPT-5 and Grok-4 Reasoning relative to smaller or open-weight models such as Llama Maverick and DeepSeek V3.1.
- M2 (Normative Shariah substitution).** Moral inflation will be concentrated on Shariah-specific items: Q9 musharakah, Q10 community welfare, and Q11 fairness-over-return. Among those items, it will be concentrated in Muslim-identified personas specifically rather than across all identities uniformly.
- M3 (Stereotype amplification).** Persona-level effect sizes for religion, gender, and age on religion-specific items will exceed the corresponding human effect sizes reported in the behavioral Islamic-finance meta-literature, consistent with corpus-driven stereotype encoding.
- M4 (Safety-alignment and caution effect).** Moral inflation will be positive across virtually all ethical and prudence items, with no items in the opposite direction. This is consistent with a generic alignment-training reward that favors cautious, ethically laudable, non-speculative answers.
- M5 (Variance collapse).** AI response distributions will be unimodal and centroid-anchored, even for items where the human distribution is clearly bimodal or long-tailed. This pattern is consistent with a corpus-centrality or next-token-regression-to-the-mean mechanism rather than a moral-specific bias.
- M6 (Knowledge-to-behavior confusion).** Moral inflation will be strongest specifically on items where what a devout person should do most diverges from what this person would actually do with real money. The clearest cases are prohibited-behavior items such as Q3, the riba-adjacent Q1, and social-media-susceptibility Q8. It will be weaker on ordinary-prudence items such as Q2, Q5, Q6, and Q7, where the two questions yield similar answers.

All nine hypotheses are evaluated in Section 4 for validity and Section 5 for mechanisms. The following section details the survey instrument, model inventory, demographic-stratification protocol, human-participant ethics framework, MI operationalization, and statistical analysis plan used to test them.

Methodology

Test Environment and Survey Instrument

The eleven-item seven-point Likert instrument, anchored at 1 for strongly disagree and 7 for strongly agree, was adapted from and domain-modified using prior behavioral-finance and Islamic-ethics instruments cited in Din et al. (2021), Aristei and Gallo (2021b), and Adil et al. (2022). Item wording was adjusted for the Islamic-finance domain and refined through iterative review. Domain-expert consultation was obtained during instrument development from Islamic-finance researchers at two Gulf-region postsecondary institutions, including a U.S.-affiliated branch campus and a public national university, plus graduate-level students in finance and Islamic studies at both institutions. Items load on five conceptually distinct dimensions. Table 1 lists the survey items in abbreviated form and their conceptual dimension assignments, providing the key for interpreting Q1–Q11 throughout the results. All items were administered identically to human participants and AI personas without adaptation.

Table 1. Survey items and conceptual behavioral dimensions

Q	Item text (abbreviated)	Dimension
Q1	I prefer investments offering high returns, even with higher risk	Risk & Return
Q2	I prioritize long-term security over short-term gains	Risk & Return
Q3	I will tolerate ethically questionable practices for strong returns	Ethical Flexibility
Q4	Transparency and full disclosure are non-negotiable for me	Transparency
Q5	Environmental impact influences my investment choices	Ethical / Social
Q6	I avoid investments that exploit workers or communities	Ethical / Social
Q7	Friends' and family's opinions influence my investment decisions	Social Influence
Q8	Social media recommendations affect my investment choices	Social Influence
Q9	Profit-loss sharing (musharakah) is preferable to fixed-interest products	Moral / Fairness
Q10	Community welfare should shape corporate investment policy	Moral / Fairness
Q11	Fairness and equitable outcomes matter more than maximum return	Risk & Return

Note. Responses range from 1 (strongly disagree) to 7 (strongly agree). Item wording is abbreviated as indicated in the column heading. Dimension assignments are conceptual; they have not been confirmed through factor analysis.

LLM Model Inventory and Persona Generation Protocol

Eight commercially available LLMs were initially engaged for synthetic-data generation across 30,000 demographic personas. An eighth model, Gemini 2.5 Pro, was excluded a priori after producing structurally errored responses across all 30,000 inquiries. Specifically, every Gemini 2.5 Pro response failed the strict JSON-schema validation step described below, yielding zero usable records. This exclusion explains the difference between the pre-exclusion 176,229-record cleaned total across all eight models and the 147,407 records summed across the seven retained models in Table 2. Per-question post-exclusion usable N falls in the 152,000 to 155,000 range due to additional item-level discard on remaining schema-fail or out-of-range records after master-file consolidation. Table 2 identifies the seven retained models, their architecture classes, and reported response counts, making the composition of the AI ensemble explicit.

Table 2. LLM inventory and retained response counts

Model	Architecture class	Personas generated (approx., post-cleaning)
GPT-5	Closed frontier, multi-modal	2,499
GPT-5 Mini	Closed frontier, lightweight	2,908
GPToss 120B	Open-weight, large	approximately 22,000
Grok-4 NonReasoning	Closed frontier, non-reasoning variant	approximately 30,000
Grok-4 Reasoning	Closed frontier, reasoning variant	approximately 30,000
Llama Maverick	Open-weight frontier	approximately 30,000
DeepSeek V3.1	Open-weight, multilingual	approximately 30,000
Total (seven models, post-cleaning, post-Gemini exclusion)	147,407	

Note. Counts are reported after cleaning, as in the source inventory. Gemini 2.5 Pro was excluded because of structural JSON errors.

Each persona prompt instantiated a named identity including age, gender, religion, and country. It then solicited a set of internally consistent secondary demographic choices covering education, investment priority, investment experience, risk tolerance, and religious-practice frequency. Finally the model was presented with the eleven Likert items. A representative prompt read:

“You are a 49-year-old female from the United States. You practice your religion, Christianity. This is your persona. Answer these questions based on your persona”

Responses were required to conform to a strict JSON schema with four fields: persona (summary), demographics, Likert, and justification. An “END” token was required at the end. Any structural non-compliance, value-range violation outside 1 to 7, or demographic-logic contradiction such as Muslim paired with religious practice of zero days per week triggered discard and regeneration.

Having defined the model inventory and persona-generation protocol, we next describe the demographic-stratification scheme used to assign identities across age, gender, religion, and country strata.

Demographic Stratification

Age, gender, and religion by country were drawn from World Population Review (2024a), World Population Review (2024b), and World Population Review (2024c) country-level population data. Age was grouped into five brackets from 18 to 24, 25 to 34, 35 to 44, 45 to 54, and 55 and older. Gender was binary. Religion was categorized into Muslim, Christian, Unaffiliated, Hindu, Jewish, Buddhist, and Other. Country representation intentionally over-sampled the United States and Qatar for research-design comparability with the human pilot, supplemented by the next ten most populous MENA nations. Education, investment experience, investment priority, risk tolerance, and religious-practice frequency were self-generated by each model conditional on the cued demographic identity. This design preserves empirically anchored population ratios in the stratifying variables while enabling analysis of model-internal demographic reasoning on the derived variables.

The demographic-stratification design for synthetic respondents is complemented by a convenience human pilot whose recruitment, ethical compliance, and representativeness we now describe.

Human Participants and Ethical Compliance

Human data were collected under Institutional Review Board certification (protocol identifier withheld for anonymous review). Participation was anonymous and voluntary. Respondents completed the same 11-item 7-point Likert instrument through an online form with no identifying fields. The sample comprised 31

participants recruited from university students and alumni in affiliated schools. The demographic breakdown of the human sample is reported alongside the AI sample in Table 6 to contextualize representativeness.

Validity Decomposition and Moral Inflation Operationalization

Following the synthetic-survey validity literature (Shrestha et al., 2025; Argyle et al., 2023), the analysis evaluates three nested validity standards rather than reporting only mean-level correlations.

Mean alignment (V1). Pearson r , MAE, and RMSE of per-question AI means against human benchmark means across the 11 items.

Distributional alignment (V2). Per-question variance ratio $SD_{(AI)}/SD_{(H)} = s_{(AI,j)} / s_{(H,j)}$, extremal-response shares at poles 1 and 7, and the sign-standardized Moral Inflation metric defined below to capture directional shifts that may be cancelled out in unsigned variance analysis.

Structural demographic alignment (V3). ANOVA η^2 and ω^2 for religion, age-group, gender, and country strata; OLS regression coefficients for demographic predictors across all eleven items; and within-stratum variance ratios, each compared qualitatively against the directional and magnitude-level benchmarks reported in the behavioral Islamic-finance literature.

The Moral Inflation metric is the key diagnostic for directional ethical distortion. For each question j in $\{1, \dots, 11\}$, define the sign-orientation constant $s_{(j)}$ in $\{-1, 0, +1\}$ as follows. Assign $s_{(j)} = +1$ if higher item agreement corresponds to greater ethical stringency under Islamic-finance ethical framing. Assign $s_{(j)} = -1$ if higher item agreement corresponds to reduced ethical strictness, for example on speculation, exploitative flexibility, or susceptibility. Assign $s_{(j)} = 0$ for items where the ethical direction of agreement is genuinely ambiguous or orthogonal. In this instrument only Q7 receives the neutral classification. The raw Moral Inflation score is:

$$MI_{(j)}^{(raw)} = s_{(j)} \times \text{ft}(Y\text{-bar}_{(AI,j)} - Y\text{-bar}_{(H,j)}) \text{ for } s_{(j)} \neq 0,$$

and the Hedges'-g-standardized version, comparable across items of differing variance, is:

$$MI_{(j)}^{(std)} = s_{(j)} \times g\text{-hat}_{(j)} \text{ for } s_{(j)} \neq 0,$$

where $g\text{-hat}_{(j)}$ is the bias-corrected Hedges' g for question j computed as AI minus human. By construction, positive $MI_{(j)}$ always indicates that LLMs endorse the question as if adopting a more ethically stringent persona than the human benchmark. Questions Q1, Q3, and Q8 are reverse-scored in this construction because higher raw agreement on them signals less ethical strictness: return-chasing, ethical flexibility, and social-media susceptibility respectively. The sign-orientation assignments and corresponding interpretive descriptions are reported in Table 4.

Statistical Analysis

Cleaned data were analyzed in Python using pandas, NumPy, and SciPy. The analysis pipeline proceeded in seven stages, organized by the three validity layers introduced above.

8. Layer 1, Mean alignment (V1): Per-question means and standard deviations for human and AI samples, plus Pearson r , MAE, and RMSE of per-question means across the 11 items.
9. Layer 2, Moral Inflation (V2, directional): Raw $MI_{(j)}^{(raw)}$ and standardized $MI_{(j)}^{(std)}$ per question, item-level classification into strong positive MI, moderate, near-zero, moderate negative, or strong negative, and descriptive averages across the 10 sign-defined items. Q7 is excluded because $s_{(j)} = 0$.
10. Layer 2, Distributional distortion (V2, variance): Per-question variance ratio $SD_{(AI)}/SD_{(H)}$ with severity classification, averaged across the instrument to test the variance-compression prediction shared by V2 and M5.
11. Effect sizes: Hedges' g for every human-to-AI question-level comparison, complemented by Cliff's Delta for ordinal robustness and 10,000-sample bootstrapped 95 percent confidence intervals.
12. Normative-salience association: Each item is classified on a three-component scale covering Shariah relevance from 0 to 2, generic social desirability from 0 to 2, and speculation or prohibited-behavior relevance

- from 0 to 2. The Pearson and Spearman associations between MI magnitude and the 0 to 6 normative-salience sum are reported to test the item-sensitivity claims in M1, M2, and M6.
13. Layer 3, Demographic heterogeneity (V3): ANOVA with η^2 and ω^2 effect sizes for religion collapsed to Christian, Muslim, Other, gender, age group, and country collapsed to United States, Qatar, Other. Religion-stratum means and within-stratum standard deviations are reported explicitly to test the stereotype-amplification predictions of M2 and M3.
 14. Cross-model variance and OLS regression: A model-by-question Hedges' g matrix plus ANOVA for inter-model and demographic-choice variance, testing M1, M3, and M5. Eleven OLS models, one per Likert outcome, are regressed on categorical predictors for country, education, gender, investment experience, investment priority, religion, religious-practice frequency, and risk tolerance, plus linear age. Significance indicators are $^*(p)<0.05$, $^{**}(p)<0.01$, and $^{***}(p)<0.001$.

Results

Layer 1, Mean Alignment (V1)

Table 3 reports per-question means, human to AI mean differences, and effect sizes for the full AI dataset. Sample sizes after stratum-specific cleaning range from roughly 152,417 to 155,125 per question, aggregated from the 176,229-record master file. V1 is clearly supported. The 11-question Pearson correlation of per-question means reaches $r = 0.854$, with MAE of 0.851 and RMSE of 1.047 on the 7-point Likert scale. Five of the 11 questions fall within 0.7 Likert points of the human benchmark on a point-estimate basis. The superficial headline of the synthetic-survey literature is replicated here for an Islamic-finance-specific instrument. Figure 1 plots these item means against the line of equality. The positive association shows similar item rankings, while departures from the diagonal show why correlation alone does not establish agreement in response levels.

Table 3. Human and pooled AI responses by survey item

Q	Human mean	AI mean	Delta (H - AI)	Hedges' g	Cliff's Delta	95% CI (bootstrap)		Direction
						Lo	Hi	
Q1	4.58	6.40	-1.81	-3.19	-0.07	-2.44	-1.17	AI more return-seeking
Q2	5.55	6.47	-0.92	-1.64	0.07	-1.37	-0.47	AI more long-term
Q3	2.94	1.30	+1.63	+3.32	0.60	0.95	2.35	Human more ethically flexible
Q4	5.10	5.82	-0.71	-0.81	0.19	-1.36	-0.10	AI more transparency-focused
Q5	5.26	5.05	+0.20	+0.30	0.45	-0.32	0.69	Near-equivalent
Q6	6.32	6.61	-0.28	-0.57	0.17	-0.62	0.01	AI more anti-exploitation
Q7	3.61	3.61	approximately 0	+0.01	0.34	-0.63	0.65	Near-identical peer influence
Q8	4.19	2.48	+1.72	+1.80	0.71	1.10	2.33	Human more social-media susceptible
Q9	4.97	5.36	-0.39	-0.25	0.30	-0.81	0.06	AI more musharakah-favoring
Q10	5.74	6.42	-0.68	-0.97	0.09	-1.16	-0.28	AI more community-focused
Q11	4.94	6.45	-1.52	-2.64	-0.04	-2.15	-0.88	AI more fairness-over-return
Overall Pearson r at the question level			0.854					
Avg. absolute Hedges' g			1.45					
Avg. Hedges' g (signed)			-0.422					

Note. Human N = 31; pooled AI sample sizes vary by item (approximately 152,000–155,000). Delta is the human mean minus the AI mean. Positive Hedges' g indicates higher human agreement; negative values indicate higher AI agreement. The 95% confidence intervals for mean differences use 10,000 bootstrap resamples. Q denotes survey item; Lo and Hi denote the lower and upper confidence limits.

Read together, Table 3 and Figure 1 distinguish agreement in item rankings from differences in response levels. Three patterns stand out. The 11-question mean-level correlation is high at $r = 0.854$, confirming that the LLM ensemble captures human directional ranking on the survey instrument. The signed mean Hedges' g of -0.422 indicates medium to large systematic AI to human divergence in the direction of AI over-endorsement of conservative and ethically stringent positions. Finally, the divergence is highly item-specific. The high but imperfect correlation also emphasizes that mean-level alignment alone cannot certify synthetic-survey validity, a point the Layer 2 analysis below now formalizes.

Extreme divergence with $|g| > 2.5$ appears for Q3 on ethically-flexible tolerance at $+3.32$, Q1 on return-chasing at -3.19 , and Q11 on fairness-over-return at -2.64 . Moderate divergence in the range of 0.8 to 2.0 appears for Q2, Q8, Q4, and Q10. The remaining six items, specifically Q2, Q4, Q5, Q6, Q7, and Q9, show moderate to near alignment, with Q7 on peer and family social influence sitting at just $+0.01$ Hedges' g where AI and human means are effectively indistinguishable. Translated into the validity-decomposition language, this pattern already signals the central result. Layer 1 mean alignment passes, but Layer 2 distributional and directional alignment fails substantially. It is a case study in how a high mean-level correlation can mask severe ethical-direction distortion. The next three subsections formalize the Layer 2 diagnostics.

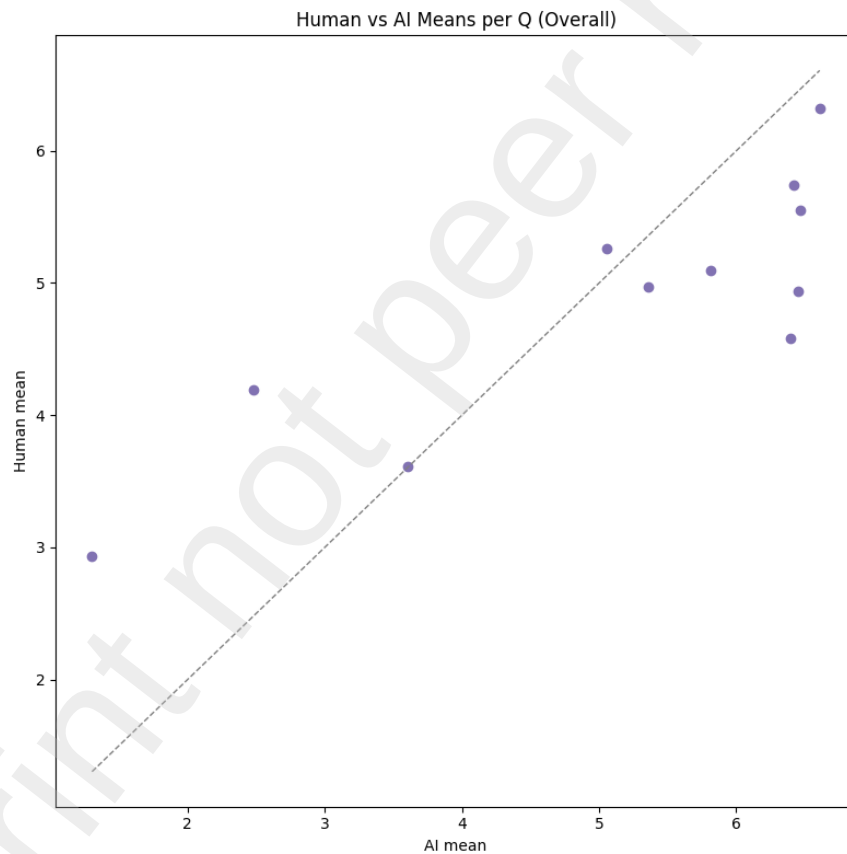


Figure 1. Agreement between human and pooled AI item means

Note. Each point represents one of the 11 survey items. The diagonal marks equal human and AI means. Pearson $r = 0.854$ summarizes association across item means; distance from the diagonal indicates disagreement in their levels.

Layer 2a, Moral Inflation for Directional Alignment (V2)

Table 4 reports the sign-orientation assignments used to operationalize $MI_{(j)}$ in Equations 1 and 2. The scoring directions distinguish greater agreement from greater ethical stringency, while the three salience components provide the item classifications used in the Layer 2c analysis.

Table 4. Item scoring direction and normative salience

Q	Dimension	s _(j)	Ethical direction of plus agree	Sh	SD	Sp	Sum
Q1	Risk & Return	-1	Less strict (return-chasing, riba-adjacent)	1	1	2	4
Q2	Risk & Return	+1	More strict (prudence, long-termism)	1	1	0	2
Q3	Ethical Flex.	-1	Less strict (exploitation for returns)	1	2	2	5
Q4	Transparency	+1	More strict (adl principle)	1	2	0	3
Q5	Ethical/Social	+1	More strict (environmental)	0	2	0	2
Q6	Ethical/Social	+1	More strict (anti-exploitation, adl/maqasid)	1	2	0	3
Q7	Social Infl.	0	Ambiguous (peer influence neutral)	0	1	0	1
Q8	Social Infl.	-1	Less strict (social-media susceptibility)	0	0	1	1
Q9	Moral/Fair.	+1	More strict (musharakah / riba)	2	2	0	4
Q10	Moral/Fair.	+1	More strict (maqasid communal welfare)	2	2	0	4
Q11	Risk & Return	+1	More strict (fairness over return, adl)	1	2	1	4

Note. The scoring direction s_j is +1 when higher agreement indicates greater ethical stringency, -1 when it indicates less stringency, and 0 for the ambiguous item Q7. Sh = Shariah relevance; SD = social-desirability relevance; Sp = speculation or prohibited-behavior relevance. Each component ranges from 0 to 2, and the sum ranges from 0 to 6. Q7 is excluded from Moral Inflation averages.

Table 5 reports the per-question Moral Inflation results computed with these orientations. Its mean columns show where the samples differ, the MI columns express the ethical direction and magnitude of those differences, and the final column compares response dispersion.

Table 5. Moral Inflation and response dispersion by survey item

Q	s _(j)	Y-bar _(H)	Y-bar _(AI)	MI _(j) ^(raw)	MI _(j) ^(std) (s _(j) x g)	SD _(AI) /SD _(H)
Q1	-1	4.58	6.40	-1.82	-3.19	0.319
Q2	+1	5.55	6.47	+0.92	+1.64	0.426
Q3	-1	2.94	1.30	+1.63	+3.32	0.244
Q4	+1	5.10	5.82	+0.72	+0.81	0.487
Q5	+1	5.26	5.05	-0.20	-0.30	0.441
Q6	+1	6.32	6.61	+0.29	+0.57	0.555
Q7	0	3.61	3.61	(excluded)	(excluded)	0.529
Q8	-1	4.19	2.48	+1.72	+1.80	0.549
Q9	+1	4.97	5.36	+0.39	+0.25	1.272
Q10	+1	5.74	6.42	+0.68	+0.97	0.543
Q11	+1	4.94	6.45	+1.51	+2.64	0.321
Average across 10 sign-defined items, Q7 excl.				+0.53	+0.77	0.517

Note. MI = Moral Inflation. Scores use the orientations in Table 4; positive MI indicates greater ethical stringency in AI responses. H = human; AI = artificial intelligence; Y-bar denotes the sample mean. Raw MI is expressed in Likert points; standardized MI uses Hedges' g computed as AI minus human. SD_{AI}/SD_H is the ratio of sample standard deviations; values below 1 indicate narrower AI response dispersion. Moral Inflation averages exclude Q7.

Average MI_(j)^(raw) reaches +0.53 Likert points and MI_(j)^(std) sits at +0.77 in Hedges-g-standardized units across the 10 sign-defined items. This confirms a systematic and directionally uniform LLM over-endorsement of ethical strictness. The four items with |MI_(j)^(raw)| at or above 1.5 are exactly the four prohibited-behavior, speculation, and fairness-over-return items highlighted in hypothesis M6: Q1 reverse-coded return-chasing, Q3 tolerance of ethical compromise for returns, Q8 reverse-coded social-media susceptibility, and Q11 fairness over maximum return.

The remaining six sign-defined items all fall in the -0.2 to +0.9 range. This concentration near zero is more consistent with the knowledge-to-behavior confusion mechanism of M6 than with a generic omnibus rightward shift. The two Shariah-core items Q9 on musharakah and Q10 on communal welfare show moderate positive raw MI of +0.39 and +0.68 respectively. They are not extreme on mean endorsement, but Section 4.4 shows the pathology for these items concentrates in within-group variance compression for Muslim personas rather than overall mean shift.

The sole negative-MI items Q1 at -1.82 raw MI and Q5 at -0.20 reflect the tension flagged in the literature review between generic LLM risk neutrality in classical lotteries and LLM ethical strictness in normatively framed items (Erdem and Ashok, 2026). In the present scoring, Q1's return-chasing item functions simultaneously as a risk-seeking equals less strict normative cue and as a classical high-return high-risk preference cue. It is the only item where the two competing logics pull in opposite directions, and the classical-risk-neutrality channel wins. Finally Q7 with $s_j=0$ on peer influence shows literally zero raw MI. This confirms that the MI signal is not an artifact of generic LLM regression to the scale midpoint or other central-answer tendencies.

Layer 2b, Variance Compression for Distributional Width (V2 and M5)

The last column of Table 5 reports the per-question standard-deviation ratio $SD_{(AI)}/SD_{(H)}$. V2 and M5 together predict that AI variance should be substantially narrower than human variance on most items. Centroid-anchored unimodal AI distributions should appear even on items where humans are clearly bimodal or long-tailed. The evidence is strongly supportive.

Mean $SD_{(AI)}/SD_{(H)}$ across all 11 items reaches 0.517. On the typical question, the AI sample has roughly half the standard deviation of the human sample. Severe compression below 0.5, meaning narrower by more than a factor of two, appears on 6 of 11 items: Q3 at 0.24, Q1 at 0.32, Q11 at 0.32, Q2 at 0.43, Q5 at 0.44, and Q4 at 0.49. Moderate compression in the range of 0.5 to 0.75 appears on 4 of 11 items: Q7 at 0.53, Q10 at 0.54, Q8 at 0.55, and Q6 at 0.56. The single outlier is Q9 on musharakah at $SD_{(AI)}/SD_{(H)} = 1.27$, the only question where AI appears more spread out than humans. As Layer 3 shows, however, this is a Simpson's-paradox artifact. Within-stratum Muslim Q9 standard deviation approaches 0.15 at ceiling with extreme compression. Within-stratum Christian Q9 standard deviation sits near 0.90. The full-sample inflated AI standard deviation arises purely from the between-stratum Muslim versus Christian split. Within every religious stratum individually, Q9 variance is more compressed in the AI sample than in the human benchmark.

The V2 prediction of distributional non-alignment is unambiguously confirmed. The combination of the Layer 1 high mean-correlation result of $r=0.854$ with Layer 2's 48.3 percent average variance compression is precisely the pathology identified in the synthetic-survey pathology literature. Surface-level mean agreement coexists with severe distributional compression that, if left uncorrected, would massively inflate statistical significance on any within-item subgroup comparison. This would be true whether the study is a published market-segmentation piece or an in-production robo-advisory calibration.

Layer 2c, Normative Salience Association for Item Sensitivity (M1 / M2 / M6)

To disentangle M1 on generic social-desirability bias, M2 on Shariah-specific normative substitution, and M6 on knowledge-to-behavior confusion concentrated on speculation and prohibited items, we computed the Pearson and Spearman association between per-item absolute Moral Inflation $|MI_j^{(raw)}|$ and the 0 to 6 normative-salience sum from Table 4.

Pearson r between absolute MI and salience sum reaches -0.008. Spearman rank correlation reaches +0.191 on the 10 sign-defined items, again excluding Q7. Both correlations are near zero. This disconfirms the simple unidimensional prediction that higher salience always leads to higher moral inflation, which would hold if a single generic mechanism such as social desirability alone or Shariah substitution alone were sufficient. Instead, the data support a multi-mechanism reading.

Items high on speculation or prohibited-behavior relevance only, specifically Q1, Q3, and Q8 with Sp scores of 2, 2, and 1 respectively, show the largest raw Moral Inflation magnitudes of 1.7 or more. This is consistent with M6 on knowledge-to-behavior confusion when what someone should do diverges from what they would actually do. Items high on social-desirability relevance but low on both Shariah core and speculation content, such as Q5 and Q7 with SD scores of 2 and 1 and no Shariah or speculation weight, show near-zero MI below 0.21 in magnitude. This counts against a pure M1 generic social-desirability account. If generic SD were sufficient, these items would show large positive MI. Items high on Shariah core but moderate on speculation or prohibition, specifically Q9 and Q10 with Sh equals 2 and Sp equals 0, show moderate positive MI of 0.39 and 0.68, visible but not extreme. When coupled with the within-stratum Muslim Q9 ceiling effect documented in Layer 3, this pattern is consistent with a constrained version of M2. Normative Shariah substitution operates within the religious-identity stratum on the Muslim personas specifically, rather than uniformly inflating all AI respondents regardless of identity.

The near-zero aggregate r and the clear dimension-level divergence together support a multi-mechanism rather than single-mechanism reading.

Layer 3, Structural Demographic Alignment (V3)

Layer 3 evaluates whether demographic predictors such as religion, age, gender, risk tolerance, and religious practice operate in the same directional and magnitude sense for the AI synthetic population as they do for the human benchmark and the published behavioral Islamic-finance literature. V3 predicts directional correctness but magnitude exaggeration, with stereotype amplification and within-group variance compression. We begin with demographic representativeness of the AI strata relative to the human pilot, then report religion-stratum means, ANOVA effect sizes, and within-group variances.

Demographic Representativeness and Stratum-level Bias

Table 6 compares AI-ensemble and human-sample distributions on four core demographic variables. The AI sample is more male-skewed at 59.9 versus 48.4 percent, more Qatari-represented at 44.9 versus 22.6 percent, and substantially more risk-conservative at 47.6 versus 41.9 percent than the human pilot. These imbalances are expected. The AI sample was stratified from World Population Review (2024a), World Population Review (2024b), and World Population Review (2024c) global data rather than matched to the convenience pilot, and they are adjusted for in the regression analysis below.

Table 6. Demographic composition of the human and AI samples

Variable	AI %	Human %
Gender		
Male	59.92	48.39
Female	40.08	51.61
Risk tolerance		
Conservative	47.56	41.94
Moderate	24.89	48.39
High	0.18	9.68
Country share		
United States	45.19	29.03
Qatar	44.86	22.58
Religion share		
Muslim	43.20	58.06

Variable	AI %	Human %
Christian	35.28	16.13
Unaffiliated or Other	21.52	25.81

Note. AI percentages pool the seven retained models; human N = 31. Percentages describe the samples used here and should not be interpreted as population estimates.

Religion as Dominant Variance Driver (V3 direction-but-not-magnitude test)

Table 7 presents AI-sample question means stratified by the three religion classes used in the consolidated human comparison: Christian, Muslim, Other. Religion is empirically the dominant variance driver in the dataset. ANOVA η^2 reaches 0.795 for Q9 on musharakah preference and 0.456 for Q10 on community-welfare orientation. These effect sizes are among the largest reported in any cross-demographic synthetic-survey validation, and they substantially exceed the 0.10 to 0.25 religion- η^2 range typically observed in human behavioral Islamic-finance meta-samples (Din et al., 2021; Bialowolski et al., 2023).

Table 7. AI response means and effect sizes by religion

Q	Christian	Muslim	Other	η^2 (ANOVA)	Notable contrasts
Q1	6.42	6.51	6.15	0.112	Muslim highest return-seeking
Q3	1.36	1.06	1.68	0.201	Muslim lowest ethical flexibility
Q4	5.63	6.38	4.99	0.298	Muslim most transparency-focused
Q6	6.56	6.80	6.31	0.157	Muslim most anti-exploitation
Q9	3.93	6.98	4.49	0.795	Muslim near-ceiling on musharakah
Q10	6.16	6.95	5.82	0.456	Muslim most community-focused
Q11	6.46	6.57	6.19	0.089	Moderate spread

Note. Responses use a 1–7 agreement scale. Approximate AI sample sizes per item are 54,000 Christian, 66,000 Muslim, and 33,000 Other personas. ANOVA = analysis of variance; η -squared measures the share of response variance associated with the three religion strata.

The magnitude of the Q9 gap, Muslim at 6.98 versus Christian at 3.93 on a 7-point scale, exceeds the human to AI gap on every item except Q3 and Q1. It comports with the directional pattern of the behavioral Islamic-finance literature. However, the ceiling-level Muslim mean of 6.98 and the tight within-group standard deviation of 0.15 for Muslim Q9 against 0.90 for Christian Q9 also confirm the variance-compression prediction of H2. Synthetic Muslim personas converge nearly uniformly on full endorsement of musharakah. In real human samples this position is distributed with substantial intra-Muslim heterogeneity by sect, religious literacy, and national context (Din et al., 2021; Adil et al., 2022).

Cross-Model Effect-Size Variation

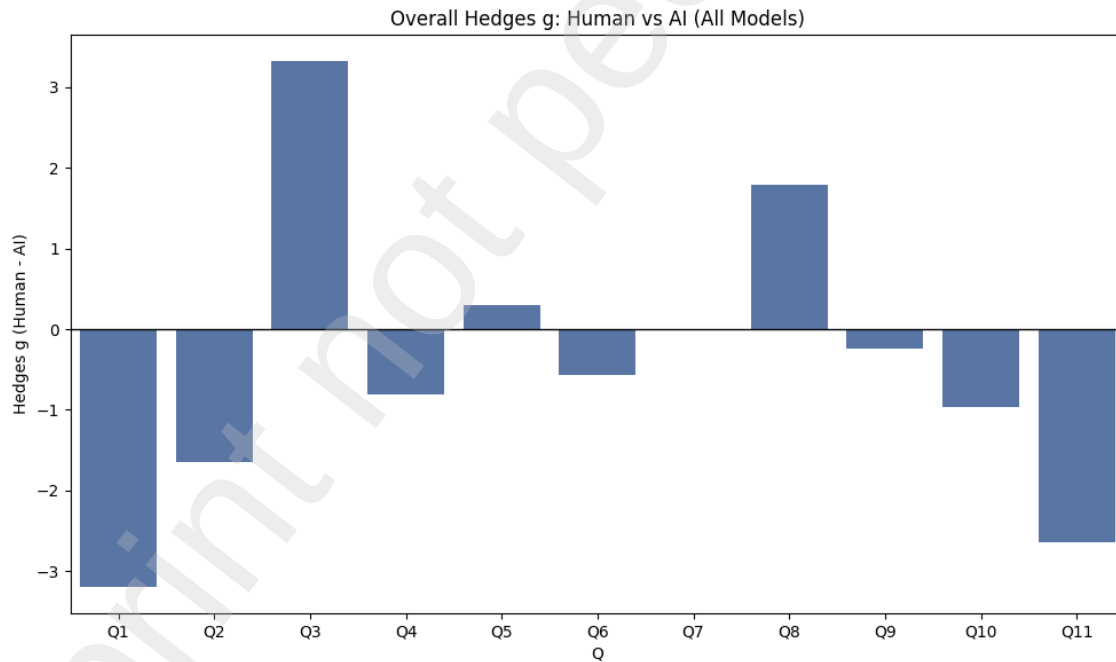
Table 8 reports the Hedges' g matrix comparing humans versus each individual model for the six high-divergence items displayed in the table. Three patterns are notable. Llama Maverick and Grok-4 NonReasoning are the most extreme on Q1 and Q11, with Llama Maverick reaching $g = -6.27$ on Q1. All models converge on the same direction for every item. They show positive g on Q3 and Q8 where humans are more flexible and more influenced by social media, and negative g on Q1, Q2, and Q11 where AI is more conservative, long-term, and fairness-focused. Ethical inflation is a cross-architecture invariant. Finally, GPT-5 is consistently the closest to the human benchmark across all items, consistent with its superior r and RMSE statistics.

Table 8. Human–AI effect sizes across seven LLMs

Q	DeepSeek	GPT-5	GPT-5 Mini	GPToss 120B	Grok-4 NR	Grok-4 R	Llama Mav.
Q1	-4.02	-2.78	-2.66	-3.04	-4.18	-3.16	-6.27
Q3	+4.13	+2.99	+3.75	+3.09	+5.59	+2.79	+2.80
Q8	+3.00	+1.19	+1.06	+1.30	+2.82	+1.21	+2.57
Q11	-3.26	-2.95	-2.43	-2.17	-4.44	-2.91	-4.53
Q2	-3.48	-1.79	-1.91	-1.17	-1.77	-1.60	-1.96
Q7	+0.11	+0.47	-1.01	-2.33	+0.48	-0.45	+2.21
Avg g	3.00	2.03	2.14	2.18	3.21	2.19	3.39

Note. The six displayed items are compared across seven retained models. Hedges' g is oriented as human minus AI: positive values indicate higher human agreement, and negative values indicate higher AI agreement. Grok-4 NR = NonReasoning; Grok-4 R = Reasoning; Llama Mav. = Llama Maverick. Gemini 2.5 Pro was excluded for structural errors.

At the question level, cross-model ANOVA indicates mean $\eta^2(2)$ of roughly 0.20 and $\omega^2(2)$ of roughly 0.20 across all eleven items. This confirms moderate structural divergence between models on absolute positioning despite their shared directional ethics. Demographic-choice variance across models is similarly ordered with mean $\eta^2(2)$ approximately 0.18. Risk tolerance and investment-priority choices exhibit greater model sensitivity than education-level choices. Figure 2 complements the model-specific comparisons in Table 8 by showing pooled effect-size point estimates for all 11 items. The near-zero result for Q7 and the large differences for Q1, Q3, and Q11 show that disagreement is concentrated in particular attitudes, rather than spread evenly across the instrument.

**Figure 2. Pooled human–AI effect sizes across the 11 survey items**

Note. Human $N = 31$. Bars show Hedges' g point estimates. Positive values indicate higher human agreement; negative values indicate higher AI agreement. Q7 is close to zero, while Q1, Q3, and Q11 show large differences. These effect sizes retain the human-minus-AI orientation and are not the sign-adjusted Moral Inflation scores in Table 5.

OLS Regression: Demographic Predictors Across Key Questions

Table 9 reports key predictors from the OLS regressions for three focal questions: Q1 on return-seeking, Q9 on musharakah preference, and Q8 on social-media influence. Complete regression output covering 325 or more predictor rows across all 11 OLS models is provided in the Supplementary Online Appendix. The three outcome columns allow the reader to compare how the same demographic predictors relate to return-seeking, musharakah preference, and social-media influence.

Table 9. Demographic predictors of three investment attitudes

Predictor	Q1 coef (SE)	Q9 coef (SE)	Q8 coef (SE)
Intercept	6.643 ^{***} (0.178)	-	-
Risk tolerance vs Conservative/Low			
Moderate	-0.620 ^{***} (0.003)	+0.338 ^{***} (0.008)	+0.514 ^{***} (0.006)
High	-1.491 ^{***} (0.030)	+0.184 ^{***} (0.072)	+0.836 ^{***} (0.108)
Religion vs Unaffiliated/Buddhist reference composite			
Muslim	+0.217 ^{***} (0.013)	+2.681^{***} (0.030)	-0.248 ^{***} (0.020)
Christian	+0.067 ^{***} (0.012)	-0.297 ^{***} (0.028)	-0.431 ^{***} (0.018)
Religious practice vs 0 days/week			
3 to 7 days per week	-0.218 ^{***} (0.022)	+1.207 ^{***} (0.057)	-0.856 ^{***} (0.036)
Country vs Qatar			
United States	+0.007 (0.010)	-1.622 ^{***} (0.023)	+0.355 ^{***} (0.015)
Gender vs Female			
Male	-0.084 ^{***} (0.003)	-0.041 ^{***} (0.006)	-0.109 ^{***} (0.004)
Investment experience vs Experienced			
Novice	-0.420 ^{***} (0.006)	-0.177 ^{***} (0.014)	+0.071 ^{***} (0.009)
Age (linear, per year)	+0.0002 [*] (0.0001)	-0.0032 ^{***} (0.0003)	-0.0091 ^{***} (0.0002)
Adj. R ² (2)	0.278	0.715	0.229
N (per model)	approximately 155,125	approximately 152,123	approximately 154,925

Note. OLS = ordinary least squares; coef = coefficient; SE = standard error (in parentheses). Outcomes are Q1 (return-seeking), Q9 (musharakah preference), and Q8 (social-media influence). Reference categories are those stated in the table rows: conservative/low risk tolerance, unaffiliated/Buddhist religion, 0 days of religious practice per week, Qatar, female, and experienced investor. * $p < .05$; ** $p < .01$; *** $p < .001$.

The Q9 regression validates the religion-dominance finding inferentially. Muslim identity carries a coefficient of +2.68 on the 7-point musharakah endorsement scale with $p < 0.001$. Frequent religious practice adds +1.21, and U.S. residence relative to Qatar subtracts -1.62. Together these three predictors explain the vast majority of fitted variance at Adj. $R^2 = 0.715$. For Q1 on return-seeking, the strongest predictor is risk-tolerance level of -1.49 for High tolerance relative to Conservative or Low. This is directionally plausible. The Muslim coefficient is modestly positive at +0.22, consistent with the mean-level results.

Discussion

The results point to a clear but easily missed validity problem. At the level of item means, the synthetic investors look convincing: mean alignment reaches $r = 0.854$. Beneath that surface, however, the distributions and demographic structure diverge. Moral Inflation is positive on 9 of 10 sign-defined items, averaging +0.53 raw Likert points, while variance compresses by 48.3 percent ($SD_{(AI)}/SD_{(H)} = 0.517$). Religion effects are also directionally plausible but much too strong. Muslim-persona identity drives Q9 endorsement with $\eta^2 = 0.795$, and the Muslim mean reaches 6.98 compared with 3.93 for Christian personas. Within the Muslim group, the standard deviation falls to 0.15. Taken together, these results describe archotyping: the models preserve the recognizable outline of a group while stripping away much of the variation within it.

The discussion first evaluates the six competing mechanisms against the cross-item and cross-model evidence. It then considers what normative archotyping means for behavioral-finance research and financial decision-making, addresses the central limitation of the $N=31$ human benchmark, and proposes a five-condition prompt experiment to separate descriptive prediction from normative prescription.

Evaluating the Six Mechanism Hypotheses

Table 10 brings together the evidence for the six mechanism hypotheses introduced in Section 2.6. It links each verdict to the relevant item patterns and model comparisons, allowing the proposed explanations to be assessed side by side. These comparisons inform interpretation but do not isolate causal mechanisms.

Table 10. Evidence for six proposed mechanisms of Moral Inflation

Hypothesis	Verdict	Key evidence
M1 Social-desirability bias	partially supported partial	The generic SD-only model is ruled out because pure SD-high items such as Q5 and Q7 show near-zero MI. Closed-frontier larger models do show greater average absolute $ g $ than open-weight models. GPT-5 Avg $ g = 2.03$ against Grok-4 NR 3.21 and Llama Mav 3.39. This is consistent with a scaled SD and alignment contribution in addition to other mechanisms.
M2 Normative Shariah substitution	strongly supported constrained	Confirmed as a within-persona mechanism. The Muslim Q9 ceiling of 6.98 with SD 0.15 is a clear example of doctrine substitution on the Muslim persona specifically. Disconfirmed as an omni-identity mechanism. The all-identity mean MI on Shariah-core items is only moderate at 0.39 to 0.68 raw. Shariah substitution operates in interaction with identity, not uniformly across all respondents.
M3 Stereotype amplification	strongly supported	Confirmed. Religious-persona effect sizes on Shariah items substantially exceed effect-size benchmarks in human Islamic-finance meta-samples, which typically report religion η^2 in the 0.10 to 0.25 range on identical constructs. Within-group variances collapse for stereotype-consistent persona groups.
M4 Safety-alignment and caution effect	strongly supported cross-item generality	Confirmed as a contributing factor. Of 10 sign-defined items, 9 show positive raw MI in the more-ethical direction, versus only 1 item Q1 showing negative MI. Near-zero MI on the only directionally ambiguous item Q7 with $s_{(j)} = 0$ supports the view that the generic caution gradient acts as a contributing floor under all normatively framed items.

Hypothesis	Verdict	Key evidence
M5 Variance collapse and corpus centroid	strongly supported	Confirmed. Mean $SD_{(AI)}/SD_{(H)} = 0.517$, severe compression on 6 of 11 items, and the Simpson's-paradox pattern on Q9 all point to centroid-anchoring rather than a moral-specific effect. The between-groups spread is inflated while within-groups variances collapse. This mechanism explains width failures independently of direction failures.
M6 Knowledge-to-behavior confusion	strongly supported strong	Best single explanation for item-level MI variance. The four items with $ MI^{(raw)} \geq 1.5$ are exactly Q1, Q3, Q8, Q11. They are the four items where what a devout or responsible person should do most diverges from what that person would actually do with real money. Ordinary-prudence items where the two answers roughly agree, specifically Q2, Q5, Q6, and Q7, show MI within the range $[-0.20, +0.92]$.

Note. Verdicts summarize the evidence discussed in the text. Strong support indicates consistency with the observed patterns, rather than causal identification. Partial support indicates conditional or incomplete consistency. A disconfirmed explanation is ruled out as a sufficient explanation on its own.

No single mechanism is sufficient. The best global reading of the data is a multi-mechanism additive model. M5 on variance collapse uniformly narrows all item distributions. M4 on caution and alignment shifts most items weakly positive in the more-ethical direction. M6 on knowledge-to-behavior confusion adds very large positive MI on the four should-versus-would diverging speculation and prohibition items. M1 on SD bias contributes an additional gradient in the largest, most-aligned closed-frontier models. Finally M2 and M3 on Shariah substitution and stereotype amplification operate jointly on religious-identity and Shariah-core item pairs, producing the extreme Muslim-persona ceiling effects and compressed within-group variances on Q9 and Q10. This additive reading is consistent with the near-zero aggregate salience-to-MI correlation reported in Section 4.2.3. The aggregate correlation is small because different salience components map to different mechanisms, not because salience is irrelevant.

Implications for Synthetic-Investor Research and Financial Decision-Making

For behavioral and experimental finance, the main risk is not that synthetic investors make random mistakes. It is that they make coherent, norm-consistent mistakes that survive aggregate validity checks. A high mean correlation can create false confidence while disagreement, subgroup heterogeneity, and trade-off behavior disappear. This matters wherever a model can infer the "responsible" answer, from sustainable investing and risk-taking to suitability assessment and financial inclusion.

Islamic finance makes this failure mode unusually visible because knowledge of a formal rule can be separated from behavioral fidelity. The literature to date asks whether LLMs know the Shariah rule (Ali et al., 2025). The present study asks whether they can reproduce an investor who does not always follow that rule, who trades off devotional ideals against return incentives, and who belongs to a religious community with substantial internal heterogeneity.

In the present pipeline, the answer is no. A bank or fintech using synthetic Muslim personas for Shariah-product suitability testing, customer segmentation, or financial-inclusion modeling would obtain systematically biased estimates. It would overestimate demand for strict musharakah products: the Q9 mean is 6.98, compared with a human-pilot mean nearer 5.0. It would underestimate tolerance for riba-adjacent products among younger, less-observant Muslim subsegments. It would also overstate confidence because synthetic within-group variances are about 50 percent of realistic values or lower. Inflated point estimates and deflated variance would combine to produce overconfident segmentation claims.

On a more constructive reading, the standard maqasid-aligned architectural framework receives direct empirical support. That framework recommends explicit rule-based Shariah constraints layered on top of LLM outputs rather than reliance on emergent moral intuitions, precisely the design recommended in doctrinal-fidelity audits such as Ali et al. (2025). The present results show that emergent LLM ethics in this domain are not merely imprecise. They are systematically directionally biased toward idealized moral responses, and they are especially

biased exactly where cultural and religious subgroups have the highest internal heterogeneity. This architectural recommendation therefore draws support not only from first-principles Islamic legal reasoning but from empirical behavioral validation.

The N=31 Human Benchmark: Limitation and Mitigation Agenda

This study's most consequential limitation is the size of the human validation sample: 31 respondents, recruited from a single U.S. university, for a pooled-AI comparison of roughly 155,000 records per item. The resulting statistical asymmetry is not inherently fatal for the directional claims advanced here. The extremal items Q3 with $g=3.32$ and Q1 with $g=-3.19$ sit so far from zero that the bootstrapped 95 percent CIs exclude any interpretation of chance. The Layer 2 variance-compression results are also independent of the human sample mean entirely. However, three categories of claim are materially weakened by the small human sample.

A second material limitation concerns instrument psychometric validation. No formal reliability or construct-validation tests were conducted on either the human N=31 sample or the synthetic sample. Specifically, Cronbach's internal-consistency estimates, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) were not computed. The instrument's dimensional structure is therefore reported on a conceptual basis derived from the source instruments in Din et al. (2021), Aristei and Gallo (2021b), and Adil et al. (2022), and the dimension labels in Table 1 should be interpreted as interpretive rather than empirically confirmed via factor analysis. A follow-up administration on the planned 150 to 300 respondent sample will enable Cronbach's and EFA/CFA tests, which in turn will allow MI computation to be conducted at the factor-score level in addition to the current item-level operationalization.

Quantitative calibration of $MI_{(j)}$ magnitudes is uncertain. The raw $MI^{(raw)} = +0.53$ average is reported with considerable uncertainty on the low-variance, close-to-zero items such as Q5, Q6, and Q9. A 150 to 300 respondent sample would reduce per-item standard error of the mean by roughly a factor of 2.5 times. This would enable confident separation of genuinely near-zero-MI items from statistically indistinguishable-from-zero items.

Structural subgroup benchmarks against human subgroup coefficients are similarly limited. The V3 claim that LLMs exaggerate religion effect sizes rests on qualitative comparison with published behavioral Islamic-finance samples such as Din et al. (2021) rather than direct within-study comparison. The cleanest next-step design would include a simultaneous 150 to 300 respondent human collection from the same United States, Qatar, and other-MENA geographic mix as the synthetic persona sample.

Cross-country and cross-sect within-Muslim heterogeneity remains largely unmeasured. A sufficiently large human sample would enable direct comparison of Sunni, Shia, Ibadi, and non-denominational Muslim within-group variances against the synthetic Q9 Muslim SD ceiling at 0.15. This is the single most diagnostic test for the M2 and M3 stereotype-amplification hypothesis. If real within-Muslim Q9 SD approaches the 0.9 to 1.1 range suggested by the Christian AI stratum and the overall human pilot, the LLM Muslim-persona homogeneity is not merely a conservative bias. It is a factual misrepresentation of lived religious community diversity.

A targeted follow-up round of at least 150 to 300 respondents, drawn from multiple countries, with religiosity, Islamic-finance familiarity, and investor-experience stratification, ideally covering student, ordinary-investor, finance-professional, and mufti subgroups, would materially strengthen claims requiring quantitative calibration. Data-collection planning for the next round is already under way. The present paper should be read as a first-pass diagnostic audit that is highly informative on directional and qualitative claims, and deliberately conservative on quantitative-calibration claims where the N=31 benchmark imposes material uncertainty.

Future Research: Experimental Prompt-Condition Decomposition

A particularly publishable next-step design would experimentally compare five prompt conditions within the same LLM and same demographic-persona battery.

15. Descriptive condition. Ask how this person would most likely answer.
16. Normative condition. Ask how this person should answer under Islamic-finance principles.

17. First-person persona condition. This is the present study's baseline: tell the model it is this person and to answer the survey.
18. Behavioral prediction with stakes. Ask it to predict what this person would actually choose if their answer committed them to a real 1,000-dollar portfolio allocation.
19. Anonymous or private condition. State explicitly that this response will never be audited, evaluated, or shown to any human or AI judge, and to answer the survey truthfully.

The prediction motivated by the present results is that the current first-person baseline in condition 3 sits between the descriptive and normative prompts of conditions 1 and 2. It should be more conservative than descriptive prompts and less conservative than explicit normative prompts. Conditions 4 and 5 provide, respectively, the stakes treatment and the anonymity treatment that have historically moved human responses away from socially desirable positions toward behaviorally realistic positions in behavioral-finance and stated-preference studies (Aristei and Gallo, 2021a; Aristei and Gallo, 2021b). If these conditions also move LLM personas, it would confirm that the Moral Inflation documented here is not an immutable property of LLMs in this domain but a manipulable outcome of prompt framing. That is precisely the finding the responsible-synthetic-AI research agenda needs.

Beyond the prompt experiment, future work should pursue four broader priorities in sequence:

20. Targeted human-recruitment round. Collect 150 to 300 respondents from the United States, Qatar, and other MENA jurisdictions. Include explicit religiosity, Islamic-finance-literacy, and investor-experience stratification and, where feasible, subgroup disaggregation across students, ordinary investors, finance professionals, and muftis.
21. Five-condition prompt decomposition. Execute the five-condition prompt experiment described above to experimentally decompose the descriptive-to-normative conflation documented in this study.
22. Hybrid-validation designs. Develop methods that combine synthetic stratification with small-N human calibration samples and post-stratification MI-debiasing weights.
23. Domain-specific bias audits. Extend the MI framework to audits with explicit *riba*, *gharar*, and *maqasid* rule sets rather than general Likert ethics scales, enabling Shariah-granular diagnostic testing.

Conclusion

This study identifies Moral Inflation as a general validity risk in synthetic-investor research. When a financial setting makes the normatively approved response easy to infer, an LLM may simulate how a responsible investor is supposed to act rather than preserve how real investors differ. Islamic finance provides a high-clarity test of that problem because its formal prescriptions are explicit while lived financial behavior remains heterogeneous.

Across seven LLMs and 176,229 post-cleaning persona records, compared with 31 human participants, the three layers of validity give sharply different answers. Mean alignment is high ($r = 0.854$), but average sign-standardized Moral Inflation reaches $+0.53$ raw Likert points and variance compresses by 48.3 percent ($SD_{(AI)}/SD_{(H)} = 0.517$). Demographic structure is directionally plausible but exaggerated: religion reaches $\eta^2(2) = 0.795$ on Q9 *musharakah*, while Muslim personas cluster at a mean of 6.98 with a within-group standard deviation of only 0.15.

The central failure is not simply inaccurate simulation. It is archotyping. The LLMs retain recognizable group-level norms while moving simulated people toward an idealized moral centroid and removing much of the disagreement that defines actual investor populations. Variance collapse, safety-alignment caution, knowledge-to-behavior confusion, social desirability, and identity-linked Shariah substitution each contribute to this pattern.

Synthetic investors should therefore be treated as complements to human evidence, not replacements for it. They remain useful for hypothesis generation, normative stress tests, and algorithmic audits, but behavioral claims require human calibration at the mean, distributional, and structural levels. Although Islamic finance is the empirical setting, the lesson applies wherever financial choices are tied to strong ideas about what a prudent, ethical, sustainable, or socially responsible investor ought to do.

Declaration of Generative AI Use

Generative AI tools were used in four capacity areas during the conduct of this study and the preparation of this manuscript. First, AI-assisted prose copyediting was used during drafting and revision to check sentence-level grammar, passive-voice rebalancing, and readability. Second, AI-assisted bibliography auditing was used to cross-reference the 17-entry reference corpus against DOI, venue, and author metadata during the reference-verification step described in the project audit trail. Third, Python-based automated pre-processing scripts using the pandas, NumPy, and SciPy libraries, developed with AI code-generation assistance, were used to compute synthetic-data descriptive statistics, per-item Hedges' g effect sizes, bootstrapped confidence intervals, and ANOVA η^2 and ω^2 across the seven-model and per-question datasets. Fourth, LaTeX macro generation and figure-caption templating were assisted by generative tools during table and figure assembly. All AI-generated content, including prose suggestions, bibliography cross-checks, statistical-computation code, and LaTeX macro output, was manually reviewed, validated, and substantively re-edited by the human co-authors before submission. No AI tool was used to generate substantive empirical claims, interpret demographic or cross-model findings, or write the mechanism-adjudication discussion. The human co-authors accept full responsibility for all content, arguments, and errors remaining in the final manuscript.

Data Availability Statement

The data supporting the findings of this study are available from the corresponding author upon reasonable request, subject to ethical and privacy restrictions applicable to the human-participant data.

Disclosure of Interest

The authors report no competing interests to declare.

References

- Adil, M., Singh, Y., Ansari, M., 2022. How financial literacy moderate the association between behaviour biases and investment decision? *Asian Journal of Accounting Research* 7, 17-30. doi: 10.1108/AJAR-09-2020-0086.
- Ali, H., Zafar, M., Aysan, A., 2025. Can AI think like a mufti? evaluating the semantic and jurisprudential fidelity of large language models in Islamic finance. *SSRN Electronic Journal* doi: 10.2139/ssrn.5232339.
- Argyle, L., Busby, E., Fulda, N., Gubler, J., Rytting, C., Wingate, D., 2023. Out of one, many: using language models to simulate human samples. *Political Analysis* 31, 337-351. doi: 10.1017/pan.2023.2.
- Aristei, D., Gallo, M., 2021a. Financial knowledge, confidence and sustainable financial behaviour: a cross-country analysis in Europe. *Sustainability* 13, 10926. doi: 10.3390/su131910926.
- Aristei, D., Gallo, M., 2021b. The relationship between financial knowledge, overconfidence and sustainable investment: gender differences across European countries. *Finance Research Letters* 46, 102200. doi: 10.1016/j.frl.2021.102200.
- Bender, E., Koller, A., 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185-5198. doi: 10.18653/v1/2020.acl-main.463.
- Bialowolski, P., Xiao, J., Weziak-Bialowolska, D., 2023. How do cultural dimensions affect financial capability? a cross-country analysis of 137 countries. *Social Indicators Research* 170, 877-891. doi: 10.1007/s11205-023-03221-7.
- Bisbee, J., Clinton, J., Dorff, C., Kenkel, B., Larson, J., 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis* 32, 401-416. doi: 10.1017/pan.2024.5.
- Din, S.M.U., Mehmood, S., Shahzad, A., Ahmad, I., Davidyants, A., Abu-Rumman, A., 2021. The herding and behavioural biases in investment intention: does financial literacy moderates the role of religious personalities? *Frontiers in Psychology* 11, 600570. doi: 10.3389/fpsyg.2020.600570.
- Erdem, O., Ashok, R., 2026. LLMs are not WEIRD: comparing AI and human financial decision-making. *Journal of Behavioral and Experimental Economics* 120, 102505. doi: 10.1016/j.socec.2025.102505.

- Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., Khot, T., 2024. Bias runs deep: probing persona-based LLM surrogates for human-subject surveys, in: Proceedings of ICLR 2024. doi: 10.48550/arXiv.2311.04892. arXiv: 2311.04892.
- Salecha, A., Ireland, M., Subrahmanya, S., Sedoc, J., Ungar, L., Eichstaedt, J., 2024. Large language models show human-like social desirability biases in survey responses. Working paper doi: 10.48550/arXiv.2405.06058. arXiv: 2405.06058.
- Shrestha, P., Krpan, D., Koaik, F., Schnider, R., Sayess, D., Binbaz, M., 2025. Beyond WEIRD: can synthetic survey participants substitute for humans in global policy research? Behavioral Science & Policy 10, 26-45. doi: 10.1177/23794607241311793.
- Sun, S., Lee, E., Nan, D., Zhao, X., Lee, W., Jansen, B., Kim, J., 2024. Random silicon sampling: simulating human sub-population opinion using LLM based on group-level demographic information. Working paper doi: 10.48550/arXiv.2402.18144. arXiv: 2402.18144.
- World Population Review, 2024a. Age structure by country. Available at: <https://worldpopulationreview.com/country-rankings/age-structure-by-country>. Accessed 15 October 2024.
- World Population Review, 2024b. Countries by sex ratio. Available at: <https://worldpopulationreview.com/country-rankings/countries-by-sex-ratio>. Accessed 15 October 2024.
- World Population Review, 2024c. Religion by country. Available at: <https://worldpopulationreview.com/country-rankings/religion-by-country>. Accessed 15 October 2024.