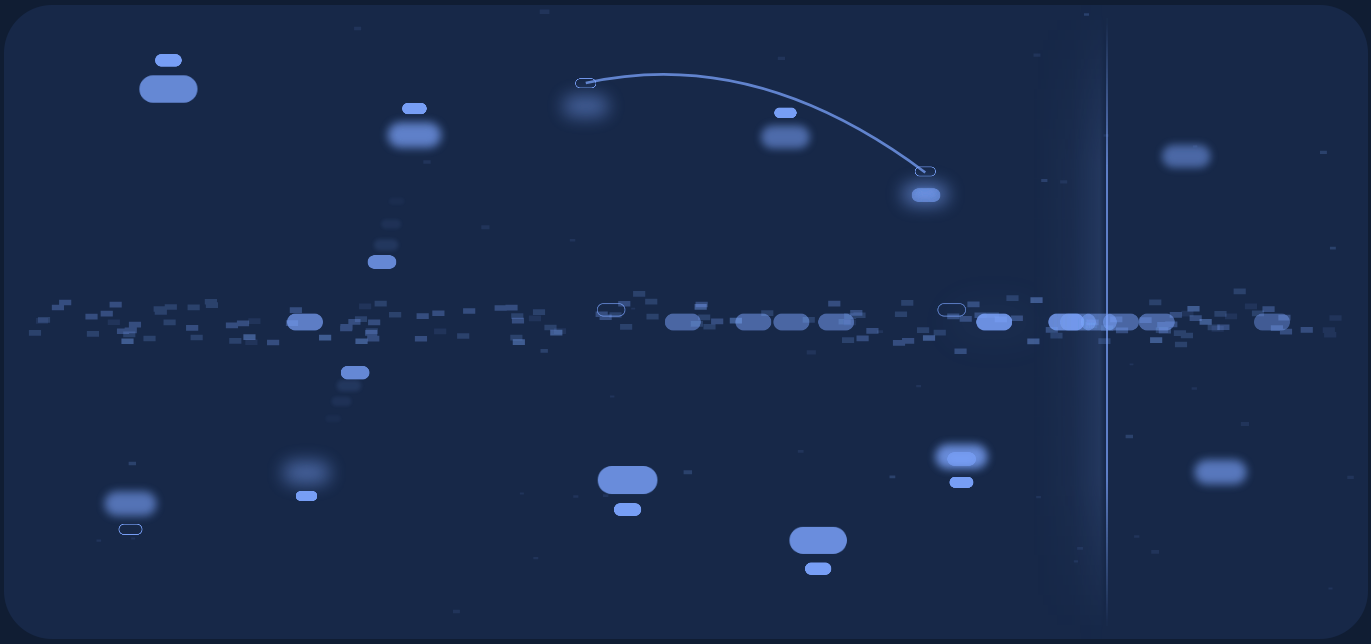
Rillor

RESEARCH REPORT

# How well LLM synthetic respondents match real people

Tests published from 2022 to 2026 check synthetic respondents' accuracy against real people. Many match averages and rankings. Spreads, subgroups and willingness-to-pay figures often differ.



Rillor

# Contents


01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

# Summary

This report sets side by side 26 studies, first released from 2022 to 2026, that compare answers generated by large language models (LLMs) with answers from real people.

- Several tests found averages and rankings close to human data. Vote choice generated from survey participants' backstories correlated with their self-reports at 0.90 to 0.94 in three U.S. elections [1], and predicted effects in 70 survey experiments correlated with actual effects at r = 0.85 [15].
- Synthetic answers varied less than human answers. In 19 studies of LLM twins built from each person's earlier answers, twin standard deviations were lower in 154 of 164 outcomes [24]. Four models portraying 16 identity groups gave flatter answers than 3,200 human participants [25].
- Matching averages can hide errors in subgroups and in relations between variables. LLM-generated 2020 U.S. vote choice correlated with self-reports at 0.94 overall and 0.02 for pure independents [1], and 48% of synthetic regression coefficients differed significantly from survey estimates, with the sign flipped in 32% of those [7].
- In conjoint studies, willingness-to-pay estimates from LLM responses ranged from within 50% of human estimates to about three times too high or wrong-signed [19]. Against incentive-compatible real valuations in four product studies, synthetic and real distributions differed significantly in all four [20].
- Simulated experiments often matched the direction of effects and overstated their size. Predicted effects were roughly twice the actual ones across 70 survey experiments [15], and replications of 156 psychology experiments matched 73% to 81% of main effects while turning 68% to 83% of original null findings significant [13].
- Agents grounded in two-hour interviews predicted 1,052 people's answers at 83% of those people's own consistency two weeks later, against 74% for agents given demographics only [22]. Raw accuracy was 65.67%, against 79.53% self-consistency [22].
- Results shift with model, prompt and date. The same prompt gave significantly different results three months apart [7], and survey guidance notes that models are retrained without notice [4]. Without a small human validation sample, errors in model outputs cannot be assessed [29], and published corrections that use one rest on stated assumptions [27, 28].
- In auctions, simulated bidders matched auction theory while the same model's stated predictions of clearing prices were inaccurate [31]. Three large models kept the human ordering of auction formats by difficulty, while LLM deviations from theory differed from people's in size and often in direction [30].

# Background

## Synthetic respondents

In this report a synthetic respondent is an LLM prompted to answer a survey question or an experimental task as a person would, usually with a demographic profile, a backstory or a persona. The 2025 edition of the ICC/ESOMAR Code defines a synthetic persona as "a digital representation of a person generated to mimic the behaviours, preferences, and characteristics of real people or groups" [5]. A task force of the American Association for Public Opinion Research (AAPOR), reporting in May 2026, prefers "synthetic responses" to "synthetic samples", because "the method is not a sampling design" [4].

AAPOR names three uses: "(1) pre-field diagnostic testing, (2) post-field augmentation/imputation, and (3) synthetic data collection as a substitute for human respondents" [4]. In pretesting, a model completes the questionnaire repeatedly to find broken skip patterns, ambiguous wording, unexpected branching and excessive completion time [4]. The task force calls synthetic responses "the most risky of the core tasks we consider", and notes that "some vendors now market 'samples' generated entirely from synthetic responses" [4].

## From silicon samples to homo silicus

Argyle and colleagues introduced "silicon samples" in a September 2022 preprint, published in Political Analysis in 2023 [1]. They conditioned a large model on "thousands of socio-demographic backstories from real human participants" from the 2012, 2016 and 2020 American National Election Studies (ANES) and Rothschild et al.'s "Pigeonholing Partisans" survey, and proposed "algorithmic fidelity" with four criteria to assess it [1].

Horton, Filippas and Manning described LLMs as "implicit computational models of humans", a Homo silicus that "can be given endowments, information, preferences, and so on" [2]. Their experiments, derived from published studies such as Charness and Rabin (2002) and Kahneman et al. (1986), "show qualitatively similar results to the original" [2]. The authors disclose financial interests in the field [2]. LLM agents as market participants (https://rillor.com/insights/llm-agents-as-market-participants) reviews that paper with other studies of LLMs in markets.

Aher, Arriaga and Kalai defined a Turing Experiment as one that "requires simulating a representative sample of participants in human subject research" [3]. A 2025 position paper argued that "results to date have been limited, and few social scientists have adopted this method", and that such simulations "can already be used for pilot and exploratory studies" [32].

## Why accuracy is hard to judge

In their willingness-to-pay tests, Brand, Israeli and Ngwe write that a researcher without a human benchmark would have had "no basis for identifying which estimates were unreliable" [19]. AAPOR adds that for fully synthetic data, "individual-level accuracy is inherently unknowable because true values are unobserved" [4].

In AAPOR's account, evaluation moves from whether a model "can match the average response to a single item" to whether synthetic responses "preserve correlations, joint distributions, subgroup interactions, and multivariate structure" [4].

# Method

**What was examined.** Published tests that compare LLM-generated answers, ratings, valuations or choices with those of real people on the same instrument, experiment or population, and what each test found to match and to differ.

**Sources.** Journal articles [7, 14, 20], ICML proceedings papers [3, 6, 32], working papers [2, 19], arXiv and PsyArXiv records [1, 8 to 13, 15 to 18, 21 to 31] and survey research guidance [4, 5]. More than half of the 26 studies are preprints, and the text marks them.

**Selection.** A study is included when it reports a direct comparison with human data: the same survey items, a replicated experiment, an incentive-compatible valuation or published laboratory results. Two methods papers on inference that combines model outputs with a human sample are included for validation [28, 29].

**What counts as a match.** Each study's own metric is kept. The metrics differ: shares and means, correlations (Pearson, tetrachoric and Kendall's τb), distribution similarity measured with the Kolmogorov-Smirnov (KS) statistic, agreement of regression coefficients and their signs, classification accuracy and AUC, coverage of individual respondents, and accuracy normalized by human test-retest consistency. In Park et al.'s definition, "a normalized accuracy of 1 indicates that the generative agent predicts the individual's responses as accurately as the individual replicates their own responses two weeks later" [22]. A match on one metric does not establish a match on another [20], so figures are not converted or averaged across studies.

**How.** Abstracts were read for every source, and full text where a page is cited. Figures come from the passage or table cited. Model and vendor names are replaced by "model family" or a count of models. For Oetzel and Maiberger only the abstract was readable, from the article's Crossref record [20]. For Tigre and Souto the second version, of 25 October 2026, is cited [27].

# Findings

## 1. Opinion surveys and silicon samples

Argyle et al. compared the model's vote with that of ANES respondents whose backstories it was given. The preprint reports [1]:

| ANES year | Model probability of a Republican vote | ANES share | Tetrachoric correlation, whole sample | Pure independents |
|---|---|---|---|---|
| 2012 | 0.391 | 0.404 | 0.90 | 0.31 |
| 2016 | 0.432 | 0.477 | 0.92 | 0.41 |
| 2020 | 0.472 | 0.412 | 0.94 | 0.02 |

The authors describe "a mild amount of overall bias" and write that "the estimates of vote choice do not match well for pure independents, especially in 2020", a group they expected to be hard to predict [1].

Santurkar et al. built OpinionsQA to compare model opinions with those of 60 U.S. demographic groups [6]. They found misalignment "on par with the Democrat-Republican divide on climate change", which persisted after steering models toward particular groups. People aged 65 and over and widowed people were among the groups poorly reflected [6].

Bisbee et al. had a model adopt personas and give feeling thermometer scores for 11 sociopolitical groups, against the 2016 to 2020 ANES [7]. The average scores "correspond closely to the averages in our baseline survey" [7].

Dominguez-Olmedo, Hardt and Mendler-Dünner, at NeurIPS 2024 according to the paper's arXiv record, tested 43 language models on American Community Survey questions [8]. Responses followed ordering and labeling biases, for example toward answers labeled "A". After answer order was randomized, "models across the board trend towards uniformly random survey responses", and models "appear to better represent subgroups whose aggregate statistics are closest to uniform" [8]. Tjuatja et al. found that nine models generally failed to reflect human response biases in survey design, particularly models tuned with reinforcement learning from human feedback, and that they reacted to perturbations that do not move people [9].

Two studies predicted held-out answers. Kim and Lee, in a preprint revised in May 2026, fine-tuned models on the 1972 to 2021 General Social Survey (GSS), with embeddings for questions, respondents and survey periods [10]. Theirs is a fitted model, where most other studies here prompt a persona. Individual-level AUC was 0.868 for imputing missing answers and 0.862 for retrodicting masked opinions, and "performance remains modest for unasked opinion prediction" [10]. Ku et al., a July 2026 preprint, used the Taiwan Election

and Democratization Study 2024 and three open-weight models of 27 billion to 120 billion parameters [11]. Zero-shot accuracy on unseen items was 52%, within 6 percentage points of a supervised random forest trained on the same population, and predictability ranged from 67% for partisan attitudes to 23% for sovereignty [11].

Kumagae and Kaneko, an October 2026 preprint, compared whole respondents [12]. Their respondent coverage gap measures how far each human respondent sits from the nearest generated one. On the U.S. World Values Survey and the 2024 ANES, "all six LLMs we evaluate leave human respondents uncovered" [12]. Under the main settings they fell behind a baseline that samples each answer independently from the human marginals, and each had a larger gap than a Gaussian copula fitted to the human data [12].

## 2. Replications of classic experiments

Aher et al. ran four Turing Experiments: the Ultimatum Game, garden path sentences, the Milgram shock experiment and wisdom of crowds. "In the first three TEs, the existing findings were replicated using recent models", and the fourth revealed a "hyper-accuracy distortion" [3]. The majority of simulated participants gave perfectly correct estimates of obscure quantities such as the melting temperature of aluminum, 660 degrees centigrade, a distortion that "does not seem to be present in earlier" models of the same family [3].

Cui, Li and Zhou, in a preprint last revised in June 2025, replicated 156 scenario-based experiments in psychology and management with three LLMs [13]:

| Measure | Result [13] |
|---|---|
| **Main effects replicated** | 73% to 81% |
| **Interaction effects replicated** | 46% to 63% |
| **Effect sizes (Fisher Z)** | about 2 to 3 times human values |
| **Original null findings returned as significant** | 68% to 83% |

Replication rates were significantly lower for studies on socially sensitive topics such as race, gender and ethics [13].

Ashokkumar, Hewitt, Ghezae and Willer, in Nature in July 2026, used "an archive of 70 preregistered, nationally representative survey experiments in the USA involving 469 experimental effects and 119,330 participants" [14]. Predictions were "strongly correlated with actual treatment effects", with accuracy similar to pooled human forecasts, and correlations stayed high for studies not public by the model's training-data cutoff and for open-weight models [14]. The preprint gives r = 0.85 (95% CI 0.81 to 0.88) and an adjusted r of 0.92 for one large model family [15]. The predictions "systematically overestimated effect sizes" [14], "roughly

twice as large as the actual effect sizes", with a root mean squared error of 11.09 percentage points [15]. In 15 megastudies with 606 effects, correlations were lower, r = 0.39 (0.27 to 0.52) and "poorly calibrated in scale", and comparable to those of pooled expert forecasters [14, 15].

Manning, Zhu and Horton, in a 2024 preprint, found that an LLM given its proposed causal model "is good at predicting the signs of estimated effects, but it cannot reliably predict the magnitudes" [31].

## 3. Purchase intent and Likert ratings

Maier et al., a preprint revised in October 2025, tested 57 personal care product surveys with 9,300 human responses, each survey with 150 to 400 U.S. participants [17]. They report correlation attainment, the correlation that synthetic results reach as a share of the human test-retest correlation, alongside KS similarity of the response distributions [17].

| Elicitation | Correlation attainment [17] | KS similarity [17] |
|---|---|---|
| **Direct numeric rating, two model families** | about 80% | 0.26 and 0.39 |
| **Semantic similarity rating (SSR), one model family** | 90% | 0.88 |

The second model family's SSR distributions reached a KS similarity of 0.8 [17]. SSR derives the rating from the semantic similarity of the model's answer to reference statements, instead of asking for a number [17, 18]. With direct ratings, "models typically replied with response '3'", a regression to the center of the scale, and almost never with 1 or 5, while real responses were mainly 4 and 5 [17]. Without demographic markers, distributions looked close (KS similarity 0.91, mean 4.0 matched), and correlation attainment fell to 50%, against 92% with demographics [17]. A supervised model trained on in-sample data reached 65%, against 88% for SSR [17]. The authors report that patterns by gender, region and ethnicity "were not consistently replicated", and that the reference sets "were manually optimized for the 57 surveys" [17].

Heath and Alexander, a July 2026 preprint, applied SSR to political attitudes [18]. They attribute the "unrealistically low variance" of silicon samples to LLMs' failure to generate numeric data, and report that SSR improved the fidelity of response distributions with few parameters to calibrate [18].

## 4. Willingness to pay against incentive-compatible benchmarks

Brand, Israeli and Ngwe derived willingness to pay (WTP) from model responses to conjoint studies and compared it with five new human studies of about 300 participants each, run in April 2024 [19]. Estimates "are often inaccurate and in some cases wrong-signed" [19].

| Product attribute | LLM WTP against the human estimate [19] |
|---|---|
| Two ingredients, pilot | Within 50% of a recent survey of real consumers |
| One toothpaste ingredient | Roughly 3 times higher |
| New toothpaste flavors | Wrong sign |
| Larger laptop screens | Negative, opposite to human preferences |

Estimates varied substantially across LLMs [19]. Fine-tuning on human survey data from the same product category "substantially improves alignment, including for new features", and did not help, and in some cases worsened performance, for a different category [19]. Models "better reflect average population WTP and do not provide meaningful across-group estimates" [19]. The authors place LLM responses as a supplement to human studies [19].

Oetzel and Maiberger, in the Journal of Marketing Analytics in September 2026, benchmarked LLM-generated WTP "against incentive-compatible real WTP elicited through the Becker-DeGroot-Marschak mechanism across four consumer product studies" [20]. They compared "means, distributions, price-response functions, profit-maximizing prices, and profit forecasts" [20]. For two more hedonic products, synthetic respondents overstated WTP and implied "overly optimistic demand and profit outcomes". For two more utilitarian products, results ranged "from near mean-level correspondence to substantial WTP underestimation" [20]. "Across all studies, synthetic and real WTP distributions differ significantly, indicating that mean-level similarity is insufficient for validation" [20]. In their reading, synthetic respondents "may support early-stage pricing exploration but should not be treated as direct substitutes" [20].

Reusens et al., a 2026 study, fitted multinomial logit models to models' choices in a travel-assistant setting and compared the implied WTP with human benchmark values from the economics literature [21]. Larger LLMs gave meaningful WTP values with systematic deviations by attribute. They tended to overestimate human WTP overall, particularly when expensive options or business-oriented personas were introduced, and came closer to human benchmarks when conditioned on prior preferences for cheaper options [21].

Gui and Toubia, in a preprint revised in November 2025, estimated demand for 40 products against an actual experiment, varying price from 0 to 200% of the regular price in 20% increments [16]. A price change also shifted variables that should stay constant, such as demographic characteristics, purchase history and the prices of competing products, which produced "implausible demand curves" [16]. Revealing the experimental design to the model, which they call "unblinding", improved performance across all tested models [16].

## 5. Individual-level agents grounded in interviews

Park et al., a preprint revised in June 2026, built agents for "a diverse national sample of 1,052 Americans" from two-hour semi-structured interviews, from structured surveys including GSS items and the Big Five personality inventory, or from both [22].

| Agent built from | Accuracy as a share of two-week test-retest consistency [22] |
|---|---|
| Interview only | 83% |
| Survey only | 82% |
| Interview and survey | 86% |
| Demographics only | 74% |

The interview agents' "raw accuracy of 65.67%" sits against the participants' "self-consistency of 79.53%" [22]. Their normalized correlation on the Big Five was 0.80, and on economic games 0.66, against 0.38 for survey agents, though no agent type differed significantly on economic games [22]. In five experiments, agent effect sizes "were larger than and highly correlated with those of the participants (ranging from r = 0.91 to r = 0.99)", and the authors note the five-study design "is underpowered for comparing the predicted and replicated effect sizes" [22]. For political ideology on the GSS, the accuracy gap between groups "fell from 13.75% with demographic-based agents to 8.60% with interview-based agents" [22].

Toubia et al.'s Twin-2K-500 dataset, a 2025 preprint, holds answers from "N = 2,058 participants (average 2.42 hours per person) in the US across four waves with 500 questions in total", with a final wave that repeats earlier tasks [23]. Test-retest accuracy averaged 81.72% across 17 tasks, twins reached 88% of that benchmark on held-out tasks, and raw twin accuracy ran from about 68% to 72% by method [23].

Peng et al., a mega-study preprint revised in April 2026, built twins from each person's answers to over 500 questions and tested them in "19 pre-registered studies with 164 diverse outcomes", with 13,506 human participants in total, 1,784 of them unique [24]:

| Measure | Result [24] |
|---|---|
| Average correlation of twin and human responses | 0.197 |
| Outcomes with a positive correlation | 157 of 164 (95.7%), 97 significantly |
| Mean gap, twin against human | 0.352 standard deviations, significant in 105 of 164 (64.0%) |

| Measure | Result [24] |
|---|---|
| **Twin standard deviation below human** | 154 of 164 (93.9%), 140 significantly |

The twins' answers were "only modestly more accurate than those from the (homogeneous) base LLM" [24]. The authors name five distortions: "(i) insufficient individuation, (ii) stereotyping, (iii) representation bias, (iv) ideological biases, (v) hyper-rationality" [24].

## 6. Spread, subgroups and identity flattening

Wang, Morgenstern and Dickerson, in Nature Machine Intelligence in 2025, tested four LLMs against 3,200 human participants across 16 demographic identities on five axes [25]. A model prompted with an identity was more likely to represent what out-group members think of that group than what in-group members think of themselves [25]. The authors found that "all four models on all questions, and across nearly all four measures of diversity we use, generate responses that are flatter than that of humans", and two models tended to cover only 3 of the 5 multiple-choice options in their 100 responses per scenario [25]. Identity-coded names and temperature changes reduced these harms without removing them. The experiments ran in July and August 2023 [25].

Other studies show the same pattern:

| Study | Spread or subgroup result |
|---|---|
| **Bisbee et al.** [7] | Less variation than the ANES. 48% of regression coefficients significantly different, sign flipped in 32% of those |
| **Argyle et al.** [1] | Tetrachoric correlation 0.02 for pure independents in 2020, against 0.94 for the whole sample |
| **Peng et al.** [24] | Twin standard deviation lower in 154 of 164 outcomes |
| **Maier et al.** [17] | Direct ratings at the center of the scale, KS similarity 0.26 and 0.39 |
| **Kumagae and Kaneko** [12] | All six models leave human respondents uncovered |
| **Ku et al.** [11] | "Variance collapse", which also affects supervised models |
| **Heath and Alexander** [18] | "Unrealistically low variance" in silicon samples |

Steering toward a group did not remove misalignment in OpinionsQA [6]. Durmus et al., in a preprint revised in April 2024, found default responses closer to the opinions of populations such as the USA and some European and South American countries, and prompts for a country's perspective shifted answers "but can reflect harmful cultural stereotypes" [26]. Maier et al. write: "Researchers must therefore use caution when interpreting subgroup analyses from synthetic panels" [17].

## 7. Diagnostics and corrections with small human samples

Tigre and Souto, in a preprint revised on 25 October 2026, start from the point that "agreement with aggregate survey outcomes can conceal errors in the subgroup estimates that guide decisions", and write that "a panel can reproduce a topline while reversing subgroup patterns" [27]. They separate two failures: the panel "may represent the wrong customers (wrong WHO), or it may produce the wrong conditional responses for otherwise comparable customers (wrong WHAT)". Reweighting addresses WHO only when overlap is adequate, and "it does not repair WHAT" [27].

Their framework combines diagnostics for covariate overlap, conditional calibration and stability across models with an augmented inverse probability weighting estimator that treats LLM predictions as auxiliary information [27]. Correction requires "a complete invitation frame, response indicators for everyone invited, and human outcomes for respondents", with assumptions of missing at random and positivity [27]. In simulations with known truth, point bias stayed small when either the response model or the outcome model was correct, while nominal 95% confidence intervals covered only 88% to 94% in some regimes [27]. When nonresponse depended on missing human outcomes, "substantial bias persists and coverage approaches zero" [27]. Controlled masking of the paired ANES and Twin-2K-500 responses reproduced this contrast [27]. As reported by Tigre and Souto, Krsteski et al. (2025) found that population-level rectification increased female bias by 58.5% while decreasing Non-Hispanic White bias by 44.7% [27].

Prediction-powered inference, from Angelopoulos et al., gives "provably valid confidence intervals" for means, quantiles and linear and logistic regression coefficients when a dataset is supplemented with machine-learning predictions, without assumptions on the algorithm, and more accurate predictions give smaller intervals [28]. Ludwig, Mullainathan and Rambachan, in a preprint revised in December 2025, argue that for estimation, "valid downstream inference requires combining LLM outputs with a small validation sample" [29]. Without one, researchers cannot assess errors in LLM outputs, and the choice of model or prompt "can produce dramatically different parameter estimates" [29].

AAPOR's task force considers it ideal for a data collection to include "at least some human respondents, or a mixed-subjects design", in "validate-then-simulate" or "simulate-then-validate" workflows, and says a mixed design "enables 'rectification' of synthetic estimates using human responses" [4]. Its example disclosure reads: "A margin of error cannot be produced from synthetic responses" [4]. Models "are continually retrained, updated, or realigned without notice" [4]. Bisbee et al. found results varying with minor changes in prompt wording, and the same prompt yielding significantly different results between April and July 2023 [7].

## 8. Stated answers and revealed bids

Manning, Zhu and Horton ran an auction among simulated agents. "In the auction experiment, the in silico simulation results closely match the predictions of auction theory, but elicited predictions of the clearing prices from the LLM are inaccurate" [31]. In their words, "the LLM knows more than it can (immediately) tell" [31].

Shah et al., a preprint revised in September 2026, evaluated "five LLMs across seven laboratory settings against human benchmarks reconstructed from published experiments" [30]. Deviations from theory differed "in magnitude and often in direction: humans overbid in second-price auctions, whereas most models that deviate underbid" [30]. Three large models without extended reasoning preserved key orderings of auction formats by deviation from theory, and Kendall's τb between one model family's difficulty ranking and the human ranking was 0.60 [30]. The reasoning model "bids almost at equilibrium", and all five models reproduced the stronger first-price winner's curse [30].

The runs covered more than 1,000 auctions with more than 5,000 bidder participants from one model family, at a total cost below $400 [30]. The authors set that against roughly $15,000 for a comparable human study and call the low cost "a practical advantage for exploration, not evidence that they replace human experiments" [30]. They add: "Whether those effects carry over to humans requires separate validation" [30]. Measuring LLM agents in simulated auctions (https://rillor.com/insights/measuring-ai-agents-in-simulated-auctions) describes how bidders are tested with induced values and human reference agents.

# Implications

**For survey and market researchers.** Means matched while coefficients, distributions or subgroups differed in several studies [1, 7, 17, 20], so each check fits the quantity a decision uses: a topline share, a ranking, a coefficient, a subgroup estimate or a price-response curve [4, 20, 27]. Spread needs its own measure, such as KS similarity or respondent coverage, because lower variance appeared in opinion surveys, ratings, twins and identity groups [7, 12, 17, 24, 25]. Simulated experiments often matched direction and overstated size [13, 15], and in one study a model predicted the signs of effects but not their magnitudes [31].

**For pricing teams.** Synthetic WTP missed signs in conjoint tests [19] and differed in distribution from incentive-compatible WTP in all four product studies [20]. Both sets of authors describe a role in early exploration or as a supplement to human studies [19, 20]. Simulated demand curves absorbed shifts in variables that should stay fixed, and unblinding improved performance [16].

**For validation.** Without a human sample, errors cannot be assessed [29], and for fully synthetic data individual accuracy is unknowable [4]. With one, prediction-powered inference gives valid intervals [28]. In simulations, a weighting correction kept point bias small when either of its models was correct, with interval coverage of 88% to 94% in some regimes [27]. One reported correction moved two subgroups' bias in opposite directions [27]. The ICC/ESOMAR Code requires that clients be told when synthetic data and personas are used, with the extent of human oversight stated [5].

**For agent simulation.** In auctions, simulated results matched theory where stated predictions did not, and LLM bidders preserved human orderings of formats while deviating from people differently [30, 31]. Rillor's market simulation work builds human reference agents from published experimental data and gives each agent an error measure for how far its behavior sits from theirs.

# Limits of this analysis

- Every figure belongs to one setting, set of models, date and metric, and models are retrained without notice [4, 7]. No source here supports a general accuracy figure for synthetic respondents.
- More than half of the studies are preprints. Argyle et al.'s figures come from the 2022 preprint, which predates the published article [1].
- For Oetzel and Maiberger only the abstract was read, so their effect sizes and sample size are left out [20].
- The human benchmarks come mainly from U.S. samples, such as the ANES, the GSS and national panels, with Taiwan [11] and cross-national surveys [26] as exceptions.
- Shah et al.'s human benchmarks are reconstructed from published experiments [30], and Park et al. call their five-experiment comparison underpowered [22].
- Some authors have commercial ties to the field. The homo silicus authors disclose financial interests [2], one working paper lists an author affiliation with a model developer [19], and the purchase-intent surveys came from a company in that market [17].
- Models may have seen published results in training. Ashokkumar et al. tested studies not public by the training cutoff [14], and Ludwig et al. require no training leakage for prediction problems [29].
- No systematic search or meta-analysis was done. Studies were chosen for a direct comparison with human data.

# Sources


1. Argyle, Busby, Fulda, Gubler, Rytting and Wingate. Out of One, Many: Using Language Models to Simulate Human Samples. Preprint arXiv:2209.06899, published in Political Analysis 31(3), 337-351, DOI 10.1017/pan.2023.2. arXiv, 14 September 2022, viewed 8 November 2026. https://arxiv.org/abs/2209.06899

2. Horton, Filippas and Manning. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? NBER Working Paper 31122, revised February 2026. National Bureau of Economic Research, April 2023, viewed 8 November 2026. https://www.nber.org/papers/w31122

3. Aher, Arriaga and Kalai. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. ICML 2023, PMLR 202, 337-371. Proceedings of Machine Learning Research, 2023, viewed 8 November 2026. https://proceedings.mlr.press/v202/aher23a.html

4. Rothschild, Marlar et al. Responsible AI Integration in Survey Research. Report of the AAPOR Task Force on Responsible AI Integration in Survey Research. American Association for Public Opinion Research, May 2026, viewed 8 November 2026. https://aapor.org/wp-content/uploads/2026/05/Responsible-AI-Integration-In-Survey-Research.pdf

5. ICC/Esomar International Code on Market, Opinion and Social Research and Data Analytics, 2025 edition. International Chamber of Commerce and Esomar, July 2025, viewed 8 November 2026. https://standards.esomar.org/assets/documents/icc-esomar-code-2025.pdf

6. Santurkar et al. Whose Opinions Do Language Models Reflect? ICML 2023, PMLR 202, 29971-30004. Proceedings of Machine Learning Research, 2023, viewed 8 November 2026. https://proceedings.mlr.press/v202/santurkar23a.html

7. Bisbee, Clinton, Dorff, Kenkel and Larson. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. Political Analysis 32(4), 401-416, DOI 10.1017/pan.2024.5. Cambridge University Press, 17 May 2024, viewed 8 November 2026. https://www.cambridge.org/core/journals/political-analysis/article/synthetic-replacements-for-human-survey-data-the-perils-of-large-language-models/B92267DC26195C7F36E63EA04A47D2FE

8. Dominguez-Olmedo, Hardt and Mendler-Dünner. Questioning the Survey Responses of Large Language Models. arXiv:2306.07951 version 4. arXiv, 9 December 2024, viewed 8 November 2026. https://arxiv.org/abs/2306.07951

9. Tjuatja et al. Do LLMs exhibit human-like response biases? A case study in survey design. arXiv:2311.04076 version 5. arXiv, 6 February 2024, viewed 8 November 2026. https://arxiv.org/abs/2311.04076

10. Kim and Lee. AI-Augmented Surveys: Leveraging Large Language Models and Surveys for Opinion Prediction. Preprint arXiv:2305.09620 version 4. arXiv, 19 May 2026, viewed 8 November 2026. https://arxiv.org/abs/2305.09620

11. Ku et al. Silicon Sampling via Cross-Survey Transfer. Preprint arXiv:2607.03091. arXiv, 3 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.03091

12. Kumagae and Kaneko. Silicon Samples Match the Answers but Not the Respondents. Preprint arXiv:2610.30900. arXiv, 22 October 2026, viewed 8 November 2026. https://arxiv.org/abs/2610.30900

13. Cui, Li and Zhou. Can Large Language Models Replace Human Subjects? A Large-Scale Replication of Scenario-Based Experiments in Psychology and Management. Preprint arXiv:2409.00128 version 3. arXiv, 20 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2409.00128

14. Ashokkumar, Hewitt, Ghezae and Willer. Large language models can predict the results of social science experiments. Nature 656(8126), 115-122, DOI 10.1038/s41586-026-10742-x. Nature, 8 July 2026, viewed 8 November 2026. https://www.nature.com/articles/s41586-026-10742-x

15. Ashokkumar, Hewitt, Ghezae and Willer. Large language models can predict the results of social science experiments. PsyArXiv preprint, DOI 10.31234/osf.io/3svep_v1. OSF, 17 June 2026, viewed 8 November 2026. https://osf.io/preprints/psyarxiv/3svep_v1

16. Gui and Toubia. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. Preprint arXiv:2312.15524 version 3. arXiv, 23 November 2025, viewed 8 November 2026. https://arxiv.org/abs/2312.15524

17. Maier et al. LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings. Preprint arXiv:2510.08338 version 3. arXiv, 27 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2510.08338

18. Heath and Alexander. Correcting Mode Collapse in Silicon Sampling with Semantic Similarity Rating. Preprint arXiv:2607.28550. arXiv, 30 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.28550

19. Brand, Israeli and Ngwe. Using LLMs for Market Research. Harvard Business School Working Paper 23-062, version of 30 April 2026. Harvard Business School, viewed 8 November 2026. https://www.hbs.edu/ris/Publication%20Files/23-062_1f58623a-ee21-44b9-a262-276047bc5543.pdf

20. Oetzel and Maiberger. How price sensitive is AI? Benchmarking LLM-generated willingness to pay in pricing analytics. Journal of Marketing Analytics, DOI 10.1057/s41270-026-00542-7, abstract from the Crossref record. Springer Nature, 28 September 2026, viewed 8 November 2026. https://doi.org/10.1057/s41270-026-00542-7

21. Reusens, Goethals, Calders and Martens. Would a Large Language Model Pay Extra for a View? Inferring Willingness to Pay from Subjective Choices. arXiv:2602.09802 version 2. arXiv, 16 June 2026, viewed 8 November 2026. https://arxiv.org/abs/2602.09802

22. Park et al. LLM Agents Grounded in Self-Reports Enable General-Purpose Simulation of Individuals. Preprint arXiv:2411.10109 version 3. arXiv, 28 June 2026, viewed 8 November 2026. https://arxiv.org/abs/2411.10109

23. Toubia et al. Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. Preprint arXiv:2505.17479. arXiv, 23 May 2025, viewed 8 November 2026. https://arxiv.org/abs/2505.17479

24. Peng et al. Digital Twins as Funhouse Mirrors: Five Key Distortions. Preprint arXiv:2509.19088 version 5. arXiv, 19 April 2026, viewed 8 November 2026. https://arxiv.org/abs/2509.19088

25. Wang, Morgenstern and Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. Accepted version arXiv:2402.01908 version 3, published in Nature Machine Intelligence 7(3), 400-411, DOI 10.1038/s42256-025-00986-z. arXiv, 3 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2402.01908

26. Durmus et al. Towards Measuring the Representation of Subjective Global Opinions in Language Models. Preprint arXiv:2306.16388 version 2. arXiv, 12 April 2024, viewed 8 November 2026. https://arxiv.org/abs/2306.16388

27. Tigre and Souto. When Can You Trust Your Synthetic Users? Diagnostics and Corrections for LLM Consumer Panels. Preprint arXiv:2609.13148 version 2. arXiv, 25 October 2026, viewed 8 November 2026. https://arxiv.org/abs/2609.13148

28. Angelopoulos, Bates, Fannjiang, Jordan and Zrnic. Prediction-Powered Inference. arXiv:2301.09633 version 4. arXiv, 9 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2301.09633

29. Ludwig, Mullainathan and Rambachan. Large Language Models: An Applied Econometric Framework. Preprint arXiv:2412.07031 version 4. arXiv, 5 December 2025, viewed 8 November 2026. https://arxiv.org/abs/2412.07031

30. Shah et al. LLM Bidders Preserve the Mechanism-Level Orderings of Human Bidders. Preprint arXiv:2507.09083 version 2. arXiv, 28 September 2026, viewed 8 November 2026. https://arxiv.org/abs/2507.09083

31. Manning, Zhu and Horton. Automated Social Science: Language Models as Scientist and Subjects. Preprint arXiv:2404.11794 version 2. arXiv, 25 April 2024, viewed 8 November 2026. https://arxiv.org/abs/2404.11794

32. Anthis et al. Position: LLM Social Simulations Are a Promising Research Method. ICML 2025, PMLR 267, 81005-81034. Proceedings of Machine Learning Research, 2025, viewed 8 November 2026. https://proceedings.mlr.press/v267/anthis25a.html

## About this report

Published by Rillor on 14 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How well LLM synthetic respondents match real people. 14 November 2026. https://rillor.com/insights/llm-synthetic-respondents

## Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

## Work with Rillor

Send us the agent you want tested, the market or auction format and the behavior you need measured. Contact us at rillor.com/contact?type=simulation or contact@rillor.com. More on market simulation: rillor.com/markets/simulation