VALIDATED BEHAVIORAL HYPOTHESES AS A LENS FOR EVALUATING PARTICIPANT SIMULATION

Anonymous authors
Paper under double-blind review

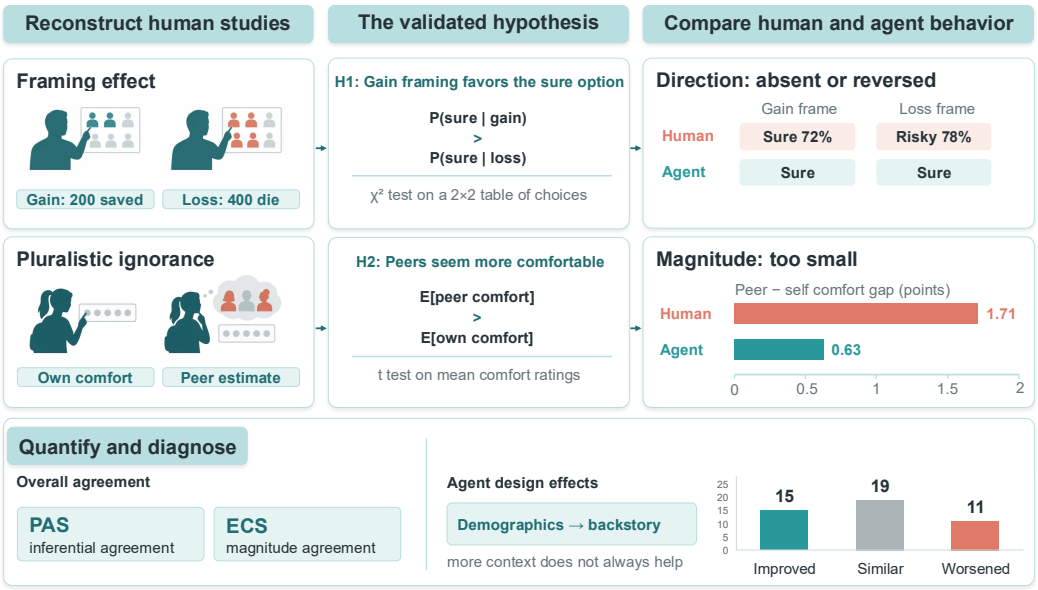


Figure 1: **Behavioral hypotheses reveal where AI agents match and diverge from humans.** HUMANSTUDY-BENCH reconstructs human studies, administers them to agents role-playing as silicon participants, and tests whether agents reproduce the selected behavioral hypotheses. PAS measures inferential agreement; ECS measures effect-magnitude agreement. Top: *framing effect* and *pluralistic ignorance* show discrepancies in effect direction and magnitude. Bottom: agent design improves agreement with some hypotheses while worsening agreement with others.

ABSTRACT

We propose using validated behavioral hypotheses as a lens for evaluating LLM agents as simulated human participants. This approach makes behavioral agreement in participant simulation measurable and decomposable, revealing which human effects agents reproduce, where they diverge, and how agent design changes that agreement. To operationalize this idea, we build HUMANSTUDY-BENCH, an open benchmarking platform that reconstructs experimental protocols from published human studies and administers these protocols to agents serving as *silicon participants*. The benchmark compares population-level effects derived from agent responses with the corresponding published human findings using two metrics: the *Probability Alignment Score* (PAS) for inferential agreement and the *Effect Consistency Score* (ECS) for effect-magnitude agreement. We apply HUMANSTUDY-BENCH across 12 human studies, evaluating 10 models under four agent designs, with 6,588 simulated participants per agent configuration. By making published human studies and their validated behavioral hypotheses reusable for evaluating simulated participants, we hope to make the behavioral assumptions underlying LLM-based social simulations and their sensitivity to agent design explicit and empirically testable.

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

1 INTRODUCTION

Large language models (LLMs) are being used as surrogate participants in social science experiments (Anthis et al., 2025; Aher et al., 2023; Dillion et al., 2023; Gao et al., 2024), as testbeds for simulating user populations (Moon et al., 2024; Zhang et al., 2025; Mou et al., 2024), and as building blocks of multi-agent simulations of organizations and markets (Argyle et al., 2023; Hofmann et al., 2024; Li et al., 2024; Manning et al., 2024). In these settings, models draw on their knowledge of human behavior to play human roles (Shanahan et al., 2023). These applications implicitly assume that agents reproduce relevant aspects of the behavior of the people they represent. Yet evidence is accumulating that current LLM-based agents fall short of this assumption in ways that compromise the validity of the simulations they ground (Gao et al., 2025; Wang et al., 2025). Evaluating agents for these applications therefore requires a principled and diagnostic account of behavioral correspondence: we need to identify which human behavioral patterns agents reproduce, where they diverge, and how their configuration changes those patterns (Puelma Touzel et al., 2026).

The conventional operationalization of the human-likeness test is Turing's *imitation game* (Turing, 1950), in which a human judge interacts with a respondent and decides whether the respondent is human. However, this approach has three key limitations. First, the verdict is subjective and sensitive to the judge's expectations of surface-level fluency (Tjuatja et al., 2024; Xu et al., 2025). Second, the outcome is a coarse binary that says nothing about *which* aspects of human-likeness are missing. Third, the protocol is expensive, requiring fresh human judges for every new agent. Indeed, as LLMs grow more sophisticated, passing such tests has increasingly become a question of stylistic mimicry rather than evidence of any deeper behavioral correspondence with humans (Gui & Toubia, 2023; Kitadai et al., 2025; Grossmann et al., 2023).

Decades of psychology, behavioral economics, and social psychology have produced testable claims about how human populations behave in experimental settings. We use subsequent independent human evidence to screen these behavioral hypotheses at two levels: *direct* support tests the same or a closely matched contrast, whereas *related* support establishes the broader behavioral effect. These hypotheses make quantitative predictions about effects such as framing (Tversky & Kahneman, 1981; Kühberger, 1998), false consensus (Ross et al., 1977; Mullen et al., 1985), trust-game cooperation (Berg et al., 1995; Johnson & Mislin, 2011), and Asch impression formation (Asch, 1946; Nauts et al., 2014). The human studies behind these hypotheses specify the tasks, conditions, and reported effects. Reconstructing those tasks and conditions for agents lets us compare population-level effects derived from their choices and judgments with the human findings, hypothesis by hypothesis. Screening hypotheses against follow-up evidence thus turns an abstract question about human-likeness into a falsifiable one: under corresponding experimental conditions, do agent populations reproduce the selected human effects?

This hypothesis-by-hypothesis comparison reframes human-likeness evaluation from a single holistic judgment into a set of statistical comparisons against behavioral findings screened by follow-up evidence, offering three advantages over the imitation game. *Objective:* the assessment follows a fixed statistical pipeline applied to agents' responses, not a human judge's impression. *Decomposable:* agreement and disagreement can be localized to specific hypotheses and dimensions of each effect. *Scalable:* once an experimental protocol is reconstructed, any new agent can be administered the test end-to-end at low marginal cost.

To operationalize this idea, we build HUMANSTUDY-BENCH, an open platform that turns published human studies into reusable simulation environments. HUMANSTUDY-BENCH introduces a human-in-the-loop pipeline that ingests published articles, extracts each study's hypotheses, experimental protocol, and statistical analysis methods, and compiles them into machine-executable experiment protocols for AI agents. Working with domain experts, we curated an initial suite of 12 studies using independent evidence to guide effect selection (Table 1 shows examples; Appendix E lists the suite). The platform evaluates how well configured agents reproduce reported human behavioral effects under reconstructed experimental protocols.

Given a set of AI agents, we run them through the reconstructed protocols as *silicon participants* and test whether the resulting population-level effects agree with the selected human findings. HUMANSTUDY-BENCH uses two metrics to quantify this agreement. The *Probability Alignment Score* (PAS) measures *inferential* agreement while retaining uncertainty in the human and agent evidence: how closely that evidence agrees about an effect and its direction. The *Effect Consistency*

Table 1: **Selected behavioral hypotheses and supporting evidence.** A subset of hypotheses from the 12 reconstructed studies. Effect sizes for the original and supporting studies follow the respective cited sources (Cohen's d or h). See Appendix E for the full list.

Main hypothesis	**Original**	**Support 1**	**Support 2**	**Support 3**
Gain framing favors the sure option; loss framing favors the risky option.	Tversky & Kahneman (1981) $h = 1.050$, $p < .001$	Kühberger (1998) $d = .31$	Steiger & Kühberger (2018) $d = .52$	Druckman (2001) $h = .97$
Ultimatum-game offers exceed dictator-game offers.	Forsythe et al. (1994) $d = 1.576$, $p \approx 0$	Oosterbeek et al. (2004) $d \approx .40$	Henrich et al. (2001) $d \approx .39$	Engel (2011) $h = 1.12$
Students estimate peers to be more comfortable with campus drinking practices than themselves.	Prentice & Miller (1993) $d = 1.307$, $p < .0001$	Schroeder & Prentice (1998) $d = .44$	Suls & Green (2003) $d = 1.04$	Borsari & Carey (2003) $d = .70$

Score (ECS) measures *effect-magnitude* agreement: how closely the simulated effect sizes match the published human estimates.

We test HUMANSTUDY-BENCH against 10 contemporary LLMs (open-weight and proprietary) under 4 standard agent designs (blank, role-play, demographic, contextualized backstory). The results reveal polarized test-level inferential agreement, limited effect-magnitude agreement, and configuration-dependent changes in agreement. Our evaluation asks three research questions: How closely do agents reproduce human behavioral effects? Where and how do simulated effects differ from human findings? How does agent design change agreement with human findings? Together, these analyses show not only how much agents align with human findings, but which effects align, how others differ, and how those patterns change across agent designs.

We make four contributions.

1. **A behavioral evaluation lens** that uses hypotheses screened against independent human evidence to make specific dimensions of human-likeness objectively testable, decomposable, and scalable.
2. **Two alignment metrics, PAS and ECS**, that quantify *inferential* and *effect-magnitude* agreement between simulated responses and published human findings, with PAS retaining uncertainty in the human and agent evidence.
3. **HUMANSTUDY-BENCH, an open platform** that reconstructs published human-subject experiments end-to-end and administers the test to user-configured agents, with an initial suite of 12 studies and support for community contributions.
4. **An empirical study** of 10 LLMs and 4 agent designs, generating 6,588 simulated participants for each agent configuration, identifying effect-specific discrepancies and the influence of agent design choices.

2 BACKGROUND & RELATED WORK

Behavioral evaluation with human experiments. Prior work evaluates LLMs as simulated human populations through both observational and experimental comparisons. Survey- and dataset-based evaluations compare aggregate model responses with human responses across questions, tasks, and demographic groups, revealing misalignment, reduced response diversity, and distorted relationships between participant characteristics and responses (Santurkar et al., 2023; Tjuatja et al., 2024; Wang et al., 2025; Bisbee et al., 2024; Hu et al., 2026). A complementary line of work administers human-subject experiments to simulated participants and examines whether population-level findings reappear. Turing Experiments establish this paradigm across classic experiments in behavioral economics, psycholinguistics, and social psychology (Aher et al., 2023); subsequent studies evaluate cognitive effects, trust and economic behavior, prosocial biases, and larger collections of scenario-

based and field experiments (Binz & Schulz, 2023; Xie et al., 2024; Kitadai et al., 2025; Liu et al., 2025; Cui et al., 2025; Chen et al., 2024; Ashokkumar et al., 2026).

Our unit of evaluation is the behavioral hypothesis, screened using subsequent independent human evidence. We distinguish direct support for the same or a closely matched contrast from related support for the broader behavioral effect, and score hypotheses with either form of support. For each scored hypothesis, we rerun the original statistical tests on agent responses and standardize agreement with the published human finding in inferential conclusion and, where comparable estimates are available, effect magnitude. This allows us to identify which specific hypotheses an agent reproduces, misses, reverses, or distorts in magnitude.

Role-playing and agent design. LLMs simulate people by playing roles (Shanahan et al., 2023), and methods for specifying these roles fall into two groups. Prompt-based specifications describe a participant through demographics, backstories, personality profiles, or cognitive biases (Argyle et al., 2023; Hu & Collier, 2024; Lutz et al., 2025; Moon et al., 2024; Jiang et al., 2023; 2024; Serapio-García et al., 2025; Miotto et al., 2022; Liu et al., 2026). Data-based methods train or calibrate agents on character experiences, interviews, surveys, or behavioral records (Shao et al., 2023; Park et al., 2024; Cao et al., 2025; Suh et al., 2025; Kolluri et al., 2025; Zhang et al., 2025; Binz et al., 2025). HUMANSTUDY-BENCH provides open, reusable, and extensible evaluation infrastructure for this line of research. Its reconstructed experimental protocols and evidence-screened behavioral hypotheses can be reused to evaluate new agent-specification methods, while additional studies and agent designs can be incorporated as the field develops.

3 HUMANSTUDY-BENCH

An *agent configuration* pairs a base model with an *agent design* (blank, role-play, demographic, or contextualized backstory); all configurations are evaluated under the same reconstructed protocols. This enables comparisons between agent designs within each reconstructed experimental setting.

Working with domain experts, we curated 12 studies from approximately 100 candidates. Table 1 shows illustrative behavioral hypotheses and independent evidence; Appendix E lists all reconstructions. Each agent configuration includes the same 6,588 simulated participants across the suite, with sample sizes set by the reconstructed studies; the original human samples range from tens to thousands of participants. These studies provide an initial set of behavioral dimensions for diagnostic evaluation. The platform supports adding studies and agent designs (Appendix G.1).

3.1 RECONSTRUCTING AND REPLAYING EXPERIMENTAL PROTOCOLS

The human-in-the-loop **Filter–Extract–Execute** pipeline reconstructs each study, with outputs verified before deployment.

Filter. We use follow-up evidence from independent human studies and meta-analyses to screen the behavioral hypotheses. We select studies with documented stimuli, instructions, conditions, quantifiable outcomes, and statistical analyses. Studies must be feasible to reconstruct without specialized physiological equipment. An LLM-assisted checklist supports screening; human reviewers assess the supporting evidence and make inclusion decisions. Every scored behavioral hypothesis has direct or related support from later independent human studies or meta-analyses; others are reconstructed but not scored (Appendix E.1).

Extract. We formalize the experimental design, hypotheses, statistical tests, and reported human outcomes, including statistics, p-values, and descriptive summaries. Participant sample sizes, demographics, and assignments are extracted where available. These define population-specific references and optional priors for agent specification.

Execute. Study modules encode conditions, stimulus order, and task instructions. The engine presents trial instances to agents and stores their individual responses. Study-specific analyses then aggregate those responses into statistics comparable with the published human summaries; no original human individual records are required. Appendix E describes the reconstructions and adaptations, including text-based versions of laboratory tasks.

3.2 Measuring Agreement on Behavioral Hypotheses

Each study contains behavioral hypotheses, each operationalized by one or more statistical tests. We compare agent outputs with the corresponding published human findings using these analyses. PAS summarizes inferential agreement; ECS summarizes effect-magnitude agreement. The metrics provide complementary summaries of the scored effects: evidence strength depends on effect size, variability, and sample size.

Probability Alignment Score (PAS). We seek a continuous measure of whether agents reach the same inferential conclusions as the published human findings, while retaining uncertainty in each result. PAS compares the human and agent support for a hypothesis's positive, negative, and null states, then aggregates agreement across its tests. Formulating this as a Bayesian estimation problem: in the binary case, a test-specific Bayes factor $\Lambda = BF_{10}$ is mapped to $\pi = P(H_1 \mid D) = \Lambda/(1+\Lambda)$ under equal prior odds. Appendix C.1 specifies the priors and approximations. Assuming the human and agent posterior states factorize, their agreement is

$$\mathrm{PAS}_j = \pi_{h,j}\pi_{a,j} + (1-\pi_{h,j})(1-\pi_{a,j}). \tag{1}$$

PAS_j is the posterior probability that the human and agent hypothesis states agree, conditional on the observed data and the factorized posterior model above. For signed hypotheses we use the directional, three-state form, with posterior vectors over positive, negative, and null effects, (H_+, H_-, H_0) (Appendix C.2):

$$\mathrm{PAS}_j = \pi_{h,+}\pi_{a,+} + \pi_{h,-}\pi_{a,-} + \pi_{h,0}\pi_{a,0}. \tag{2}$$

Thus, strong evidence for opposite directions receives low agreement even if both effects are significant.

To combine test-level PAS within a behavioral hypothesis, we take their arithmetic mean:

$$\mathrm{PAS}_{\mathrm{hypothesis}} = \frac{1}{M}\sum_{j=1}^{M}\mathrm{PAS}_j, \tag{3}$$

Hypothesis scores are averaged within each study in the same way. Because each test-level PAS is the probability that the agent reaches the same conclusion as the human study on that test, a hypothesis or study score is the expected proportion of its tests on which agent and human conclusions agree. We then take the arithmetic mean of study scores for $\mathrm{PAS}_{\mathrm{global}}$, so studies with more hypotheses or larger samples do not dominate the benchmark (Appendix B.3). Scores are interpreted under the specified evidence model and study design.

Effect Consistency Score (ECS). We also ask whether agents reproduce the magnitudes of human effects across tests. ECS compares standardized effect estimates and separates disagreement in their test-level pattern from overall differences in level and scale. We convert reported statistics to standardized effect estimates as described in Appendix C.3, then apply Lin's concordance correlation coefficient (Lawrence & Lin, 1989). Its unweighted form has a simple connection to mean squared error:

$$\mathrm{ECS} = \rho C_b = 1 - \frac{\mathrm{MSE}}{V}. \tag{4}$$

MSE is the mean squared difference between standardized agent and human effects across tests. The normalizer V is their expected squared difference if the two effect vectors were uncorrelated: the independence baseline. ECS therefore measures how much the observed squared difference falls below this baseline. The correlation ρ tracks whether larger human effects also tend to be larger agent effects; C_b penalizes shifts in the mean and differences in spread, which correlation alone misses. Appendix B.3 gives the test-level weighted aggregate used in the benchmark. ECS measures agreement with the published numerical references. Appendix F.3 examines sensitivity to test-family pooling and sampling noise under a specified model.

4 Experiment Setup

4.1 Shared Benchmark Evaluation

Human Studies. We evaluate against 12 human-subject studies (Appendix E) spanning individual cognition, strategic interaction, and social psychology, scoring the 45 behavioral hypotheses with

Table 2: **Main leaderboard.** Probability Alignment Score (PAS) measures agreement about whether effects occur and in which direction; Effect Consistency Score (ECS) measures agreement in effect magnitude. Domain columns report PAS for individual cognition (Cog.), strategic interaction (Stra.), and social psychology (Soc.). Teal highlights the highest values and salmon the lowest. †Study 12 A4 uses Grok 4.3 (Grok 4.1 Fast was retired). Appendix F reports costs.

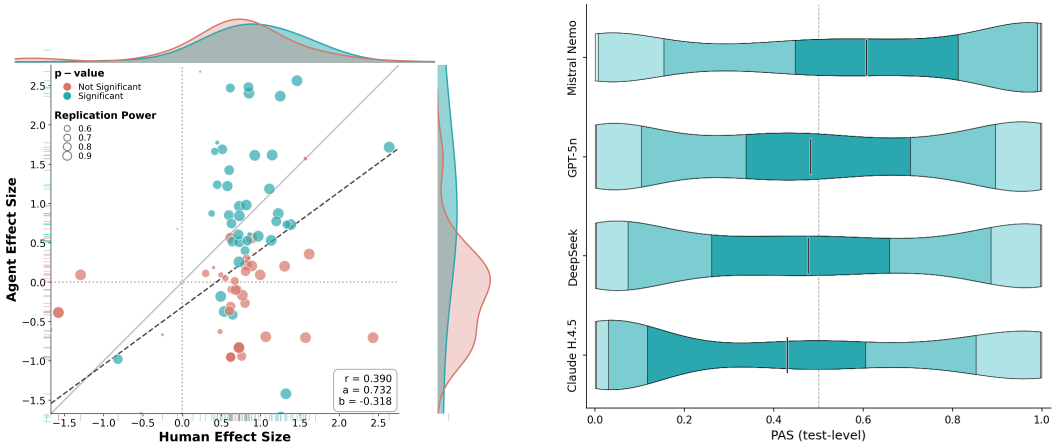
Model		Cog.	Stra.	Soc.	PAS	ECS	Model		Cog.	Stra.	Soc.	PAS	ECS
Claude Haiku 4.5	A1	0.45	0.25	0.62	0.440	0.074	GPT 5 Nano	A1	0.43	0.62	0.50	0.515	0.259
	A2	0.41	0.26	0.62	0.432	0.228		A2	0.55	0.39	0.56	0.501	0.292
	A3	0.47	0.32	0.49	0.427	0.206		A3	0.53	0.65	0.43	0.539	0.266
	A4	0.43	0.21	0.55	0.394	0.127		A4	0.57	0.63	0.57	0.589	0.206
DeepSeek V3.2	A1	0.46	0.21	0.63	0.433	0.104	GPT OSS 120b	A1	0.40	0.22	0.46	0.359	0.037
	A2	0.48	0.25	0.55	0.428	0.148		A2	0.44	0.24	0.51	0.397	0.039
	A3	0.51	0.21	0.63	0.450	0.115		A3	0.50	0.27	0.49	0.422	0.123
	A4	0.47	0.22	0.65	0.448	0.134		A4	0.49	0.25	0.68	0.472	0.080
Gemini 3 Flash	A1	0.50	0.51	0.62	0.542	0.096	GPT OSS 20b	A1	0.50	0.34	0.47	0.437	0.152
	A2	0.51	0.47	0.44	0.475	0.058		A2	0.47	0.30	0.45	0.406	0.137
	A3	0.62	0.45	0.55	0.539	0.146		A3	0.54	0.42	0.43	0.464	0.193
	A4	0.67	0.20	0.56	0.476	0.141		A4	0.52	0.33	0.51	0.453	0.175
Mistral Nemo	A1	0.56	0.30	0.52	0.462	0.114	Qwen 3 Next80b	A1	0.52	0.49	0.53	0.514	0.000
	A2	0.55	0.32	0.41	0.428	0.095		A2	0.50	0.51	0.53	0.516	0.159
	A3	0.58	0.27	0.52	0.458	0.208		A3	0.47	0.54	0.52	0.509	0.112
	A4	0.61	0.32	0.53	0.484	0.205		A4	0.55	0.51	0.50	0.518	0.096
Gemma 4 26b	A1	0.40	0.49	0.58	0.489	0.100	Grok 4.1 Fast	A1	0.44	0.26	0.51	0.404	0.080
	A2	0.37	0.49	0.56	0.473	0.092		A2	0.45	0.24	0.43	0.372	0.050
	A3	0.46	0.50	0.57	0.511	0.146		A3	0.55	0.23	0.57	0.449	−0.008
	A4	0.47	0.24	0.61	0.442	0.130		A4	0.59	0.19†	0.48	0.421†	0.182†

direct or related follow-up support (Appendix E.1). The individual cognition studies focus on cognitive biases and heuristic judgment (Ross et al., 1977; Jacowitz & Kahneman, 1995; Tversky & Kahneman, 1981; Kahneman & Tversky, 1972). Strategic interaction studies are drawn from paradigms in game theory (Nagel, 1995; Shafir & Tversky, 1992; Forsythe et al., 1994; Berg et al., 1995). Social psychology studies examine social cognition, social norms, and group behavior (Knobe, 2003; Asch, 1946; Gagnon & Bourhis, 1996; Prentice & Miller, 1993).

Models and Inference Settings. We evaluate 10 contemporary models, including open-weight (e.g., Mistral, DeepSeek, Qwen) and proprietary APIs (e.g., Claude, GPT, Gemini, Grok). Model details are in Appendix F.1. All models use a temperature $T = 1.0$ to induce the behavioral variance needed for population simulation. Response-bootstrap standard errors are reported in Appendix F.2; these quantify resampling uncertainty within the generated agent responses.

Agent Design Variants. These variants are exploratory probes of how common forms of participant specification affect agreement with human findings. We use them to characterize sensitivity to agent design rather than to identify an optimal simulation method. (1) *Blank (A1)* uses the base model with no additional specification. (2) *Role-Play (A2)* instructs the model to act as a human participant in a psychological study, without specific attributes. (3) *Demographic (A3)* adds attributes (e.g., age, gender, occupation) to the role-playing instruction, based on reported participant demographics where available. (4) *Contextualized Backstory (A4)* augments demographics with a rich natural language narrative about the agent's life history, personality, and daily context. See Appendix D for details.

Evaluation Protocol. The reconstructed protocols assign simulated participants to the original experimental conditions and collect their responses. By default, each agent configuration uses the same participant counts and condition allocation as the original human study. Study-specific evaluators parse these responses, recompute the relevant sample statistics, and compare them with published human summaries. We report the Probability Alignment Score (PAS; $[0, 1]$) for inferential agreement and the Effect Consistency Score (ECS; $[-1, 1]$) for effect-magnitude agreement (1 indicates exact agreement on the scored summaries).



(a) Effect-size correspondence. (b) Test-level PAS.

Figure 2: **Complementary views of agreement.** (a) Effect-size correspondence for GPT-5 Nano A2. The diagonal marks exact agreement; the dashed line is a fitted regression. (b) Test-level PAS distributions. Many agent scores lie near the extremes of inferential agreement on the scored hypotheses.

4.2 Analyses for the Three Research Questions

RQ1 uses PAS and ECS to measure how closely agents reproduce the direction and magnitude of the selected human behavioral effects. For RQ2, the unit of comparison is a statistical test evaluated under one agent configuration. We first ask whether the agent responses support the same effect direction as the published human result and then, for direction-matched comparisons with comparable effect estimates, whether the simulated effect is substantially weaker or stronger. For RQ3, our primary analysis pairs the same model and behavioral hypothesis across adjacent agent designs: A1 to A2, A2 to A3, and A3 to A4. We balance these paired changes across studies and resample studies for exploratory uncertainty intervals.

5 Results

This section organizes the results around the three research questions.

5.1 RQ1: How Closely Do Agents Reproduce Human Behavioral Effects?

Overall, agents sometimes reproduce whether a human effect occurs and in which direction, but reproduce its magnitude much less reliably.

Inferential agreement. Across individual tests, agreement is often polarized: 25% of test-level scores exceed 0.9 and 31% fall below 0.1 (Figure 2(b)). Agreement also varies across domains and agent configurations. For example, Claude Haiku 4.5 with role-play instructions achieves a PAS of 0.62 on the social-psychology tasks but only 0.26 on the strategic-interaction tasks (Table 2). The leading model–design configuration differs by domain: Gemini 3 Flash A4 scores highest in cognition, GPT-5 Nano A3 in strategic interaction, and GPT OSS 120b A4 in social psychology (Table 2). Reproducing one set of human findings therefore does not imply reliable agreement on others.

Effect-magnitude agreement. Agents reproduce effect magnitudes less reliably than effect directions. Even the strongest configuration reaches an ECS of only 0.292 (GPT-5 Nano with role-play instructions), indicating limited correspondence between simulated and human effect magnitudes (Table 2). Figure 2(a) illustrates why: some simulated effects lie close to the line of exact agreement, but many fall well above or below it. For GPT-5 Nano with role-play instructions, the simulation understates some published effects and overstates others rather than shifting all effects in the same

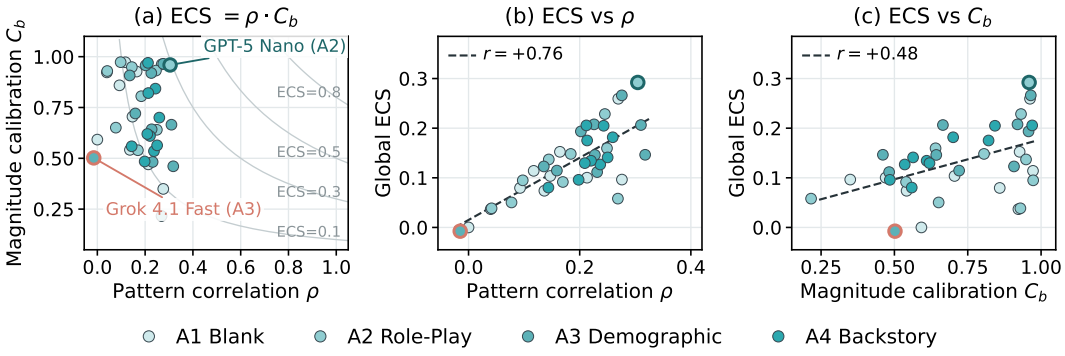


Figure 3: **ECS decomposition.** (a) Pattern correlation ρ and calibration factor C_b for evaluated agents. (b,c) Their associations with ECS. Low ρ indicates limited correspondence in the relative magnitudes of the scored effects.

direction. Thus, an agent may correctly reproduce whether an effect is positive or negative while still giving a misleading account of how large it is.

The same weak correspondence remains when effect magnitudes are compared separately within different statistical test families. Figure 3 summarizes whether ECS is limited by disagreement in the relative pattern of effects or by differences in their overall scale. A mixture of the ten models' responses agrees with the human findings at least as well as most single models (Appendix F.4); sensitivity analyses are in Appendix F.3.

5.2 RQ2: Where and How Do Simulated Effects Differ from Human Findings?

We observe two common forms of disagreement between simulated and human results. First, agents often fail to reproduce the direction of the human effect. Among comparisons where the human study reports a positive or negative effect, 50% of the simulated results show no effect or an effect in the opposite direction. Second, even when the simulated effect points in the same direction as the human effect, its magnitude is often substantially different. After excluding human effects too close to zero for a meaningful comparison, 54% of simulated effects are less than half or more than twice the published estimate, with overestimation occurring more often than underestimation. Table 3 summarizes these two forms of disagreement.

Table 3: **Patterns of behavioral discrepancy.** A simulated effect can differ from a published human finding in its direction or in its magnitude. Appendix F.6 reports the response-level behaviors.

Pattern	**Observable signature in simulated responses**
Direction mismatch	The agent responses provide no evidence for the published effect, or support the opposite direction.
Magnitude mismatch	The direction is retained, but the simulated effect is substantially larger or smaller than the published estimate.

For example, the framing task illustrates a direction mismatch: human choices reversed between gain and loss wording, while agent choices often favored the certain option under both. The peer-comfort task illustrates attenuation: the median simulated self–peer gap was 0.63 points, compared with 1.71 points in the human study. Appendix F.6 provides the classification rules and response-level examples.

5.3 RQ3: How Does Agent Design Change Agreement with Human Findings?

We observe two patterns in how agent configurations change agreement with human findings. First, simply instructing a model to act as a human participant does not reliably improve agreement with

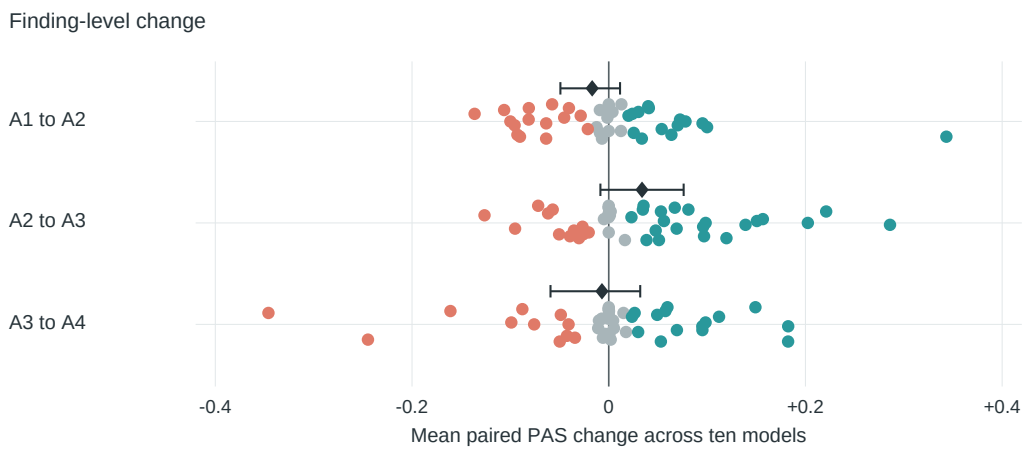


Figure 4: **Hypothesis-level changes across agent designs.** Each dot shows, for one behavioral hypothesis, the average change in PAS across the ten models when one layer of participant description is added. Each comparison covers 45 hypotheses from twelve studies. Black diamonds show averages that give each study equal weight, and the accompanying 95% intervals are obtained by resampling studies. Teal marks hypotheses whose PAS increases by more than 0.02, salmon marks decreases of more than 0.02, and gray marks smaller changes.

the human findings. Moving from no participant specification (A1) to role-playing instructions (A2) slightly lowers PAS on average. Across the 45 behavioral hypotheses, PAS increases by more than 0.02 for 15 hypotheses but decreases by more than 0.02 for 15, with no clear overall improvement.

Second, adding more participant information also produces mixed results. Adding demographics to the role-playing instructions (A2 to A3) raises PAS by 0.034 on average, with A3 exceeding A2 for eight of ten models (Wilcoxon $p = 0.010$; Appendix F.4). Adding contextualized backstories (A3 to A4) produces approximately no average change. This near-zero average masks changes in opposite directions: PAS increases by more than 0.02 for 15 hypotheses and decreases by more than 0.02 for 11 (Figure 4). Thus, richer participant descriptions improve agreement for some human findings but reduce it for others.

Implications for participant simulation. Our results suggest that agreement between simulated and human populations depends on both the agent design and the behavioral hypothesis being tested: a design that reproduces one effect may miss another, and changing the design can improve agreement on some hypotheses while worsening it on others. HUMANSTUDY-BENCH supports these comparisons through reusable experimental protocols and makes it easier to add new hypotheses by reconstructing studies from papers and experimental materials. This open infrastructure enables the community to build on accumulated behavioral knowledge and repeatedly assess simulated populations as models and agent designs evolve.

6 CONCLUSION

Behavioral hypotheses provide a concrete unit for testing whether simulated participants reproduce documented human effects. HUMANSTUDY-BENCH operationalizes this idea by reconstructing experimental protocols from human studies, administering them to LLM agents serving as silicon participants, and comparing the resulting population-level effects with the corresponding published human findings, hypothesis by hypothesis. This hypothesis-level view makes the behavioral assumptions behind AI-based social simulations explicit, testable, and traceable to particular effects. Each scored test remains linked to the agent responses that produced it, allowing disagreement to be inspected at the level of choices and judgments. Through HUMANSTUDY-BENCH, we hope to turn social science experiments and their behavioral hypotheses into open, reusable resources for assessing silicon participants in social simulations.

REFERENCES

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

Norman H. Anderson and Alfred A. Barrios. Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology*, 63(2):346–350, 1961. ISSN 0096-851X. doi: 10.1037/h0046719. URL http://dx.doi.org/10.1037/h0046719.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. Position: LLM social simulations are a promising research method. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=cRBg1dtj7o.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

S. E. Asch. Forming impressions of personality. *Journal of Abnormal and Social Psychology*, 41(3): 258–290, 1946. doi: 10.1037/h0055756.

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 656:115–122, 2026. doi: 10.1038/s41586-026-10742-x.

Daniel Bennett, Stefan Bode, Maja Brydevall, Hayley Warren, and Carsten Murawski. Intrinsic valuation of information in decision making under uncertainty. *PLOS Computational Biology*, 12(7):e1005020, 2016. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005020. URL http://dx.doi.org/10.1371/journal.pcbi.1005020.

Joyce E. Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10:122–142, 1995. URL https://api.semanticscholar.org/CorpusID:144827131.

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

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, Tankred Saanum, Johannes A. Schubert, Luca M. Schulze Buschoff, Nishad Singhi, Xin Sui, Mirko Thalmann, Fabian J. Theis, Vuong Truong, Vishaal Udandarao, Konstantinos Voudouris, Robert Wilson, Kristin Witte, Shuchen Wu, Dirk U. Wulff, Huadong Xiong, and Eric Schulz. A foundation model to predict and capture human cognition. *Nature*, 644:1002–1009, July 2025. doi: 10.1038/s41586-025-09215-4. URL https://doi.org/10.1038/s41586-025-09215-4.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32 (4):401–416, 2024. doi: 10.1017/pan.2024.5.

Brian Borsari and Kate B Carey. Descriptive and injunctive norms in college drinking: a meta-analytic integration. *Journal of studies on alcohol*, 64(3):331–341, 2003.

Colin F. Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A. Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9):637–644, 2018. doi: 10.1038/s41562-018-0399-z.

Yong Cao, Haijiang Liu, Arnav Arora, Isabelle Augenstein, Paul Röttger, and Daniel Hershcovich. Specializing large language models to simulate survey response distributions for global populations. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3141–3154, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.162. URL https://aclanthology.org/2025.naacl-long.162/.

Yaoyu Chen, Yuheng Hu, and Yingda Lu. Simulating field experiments with large language models. *arXiv preprint arXiv:2408.09682*, 2024. URL https://arxiv.org/abs/2408.09682.

François Cochard, Julie Le Gallo, Nikolaos Georgantzis, and Jean-Christian Tisserand. Social preferences across different populations: Meta-analyses on the ultimatum game and dictator game. *Journal of Behavioral and Experimental Economics*, 90:101613, 2021. ISSN 2214-8043. doi: 10.1016/j.socec.2020.101613. URL http://dx.doi.org/10.1016/j.socec.2020.101613.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988.

Ziyan Cui, Ning Li, and Huaikang Zhou. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5:627–634, 2025. doi: 10.1038/s43588-025-00840-7.

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

James N Druckman. Using credible advice to overcome framing effects. *Journal of Law, Economics, and Organization*, 17(1):62–82, 2001.

Kfir Eliaz and Andrew Schotter. Paying for confidence: An experimental study of the demand for non-instrumental information. *Games and Economic Behavior*, 70(2):304–324, 2010. ISSN 0899-8256. doi: 10.1016/j.geb.2010.01.006. URL http://dx.doi.org/10.1016/j.geb.2010.01.006.

Christoph Engel. Dictator games: A meta study. *Experimental economics*, 14(4):583–610, 2011.

Adrien Fillon, Subramanya P. Chandrashekar, and Gilad Feldman. Asymmetries in attributions of blame and praise, intent, and causality: Free will, responsibility, and the side-effect effect. *Collabra: Psychology*, 11(1), 2025. ISSN 2474-7394. doi: 10.1525/collabra.128423. URL http://dx.doi.org/10.1525/collabra.128423.

Robert Forsythe, Joel L. Horowitz, Nathan Eugene Savin, and Martin Sefton. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6:347–369, 1994. URL https://api.semanticscholar.org/CorpusID:154965582.

André Gagnon and Richard Y. Bourhis. Discrimination in the minimal group paradigm: Social identity or self-interest? *Personality and Social Psychology Bulletin*, 22(12):1289–1301, 1996.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259, 2024. doi: 10.1057/s41599-024-03611-3. URL https://www.nature.com/articles/s41599-024-03611-3.

Yuan Gao, Dokyun Lee, Gordon Burtch, and Sina Fazelpour. Take caution in using llms as human surrogates. *Proceedings of the National Academy of Sciences*, 122(24):e2501660122, 2025. doi: 10.1073/pnas.2501660122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2501660122.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109, 2023.

George Gui and Olivier Toubia. The challenge of using llms to simulate human behavior: A causal inference perspective. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4650172. URL http://dx.doi.org/10.2139/ssrn.4650172.

Erdogan Gunel and James Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557, 1974. doi: 10.1093/biomet/61.3.545.

Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, and Richard McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American economic review*, 91(2):73–78, 2001.

Teck-Hua Ho, Colin Camerer, and Keith Weigelt. Iterated dominance and iterated best response in experimental" p-beauty contests". *The American Economic Review*, 88(4):947–969, 1998.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Ai generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154, Aug 2024. doi: 10.1038/s41586-024-07856-5. URL https://doi.org/10.1038/s41586-024-07856-5.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations, 2024. URL https://arxiv.org/abs/2402.10811.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. Simbench: Benchmarking the ability of large language models to simulate human behaviors. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2fc4d6a6400addda7493b9db258c34e0-Abstract-Conference.html.

Karen E. Jacowitz and Daniel Kahneman. Measures of anchoring in estimation tasks. *Personality and Social Psychology Bulletin*, 21:1161 – 1166, 1995. URL https://api.semanticscholar.org/CorpusID:145430539.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=I9xE1Jsjfx.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. PersonaLLM: Investigating the ability of large language models to express personality traits. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3605–3627, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.229. URL https://aclanthology.org/2024.findings-naacl.229/.

Xu Jiang, Radhika Lunawat, and Brian Shapiro. *The Impact of Financial Histories on Individuals and Societies: A Replication of and Extension of Berg et al. (1995)*, pp. 95–135. Emerald Group Publishing Limited, 2015. ISBN 9781785603501. doi: 10.1108/s0193-230620150000018004. URL http://dx.doi.org/10.1108/s0193-230620150000018004.

Noel D Johnson and Alexandra A Mislin. Trust games: A meta-analysis. *Journal of economic psychology*, 32(5):865–889, 2011.

Daniel Kahneman and Amos Tversky. Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3):430–454, 1972. doi: 10.1016/0010-0285(72)90016-3.

Ayato Kitadai, Sinndy Dayana Rico Lugo, Yudai Tsurusaki, Yusuke Fukasawa, and Nariaki Nishino. Can ai with high reasoning ability replicate human-like decision making in economic experiments? *Group Decision and Negotiation*, 34(6):1303–1326, September 2025. ISSN 1572-9907. doi: 10.1007/s10726-025-09946-9. URL http://dx.doi.org/10.1007/s10726-025-09946-9.

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

Richard A. Klein, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams, Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, Štěpán Bahník, Rishtee Batra, Mihály Berkics, Michael J. Bernstein, Daniel R. Berry, Olga Bialobrzeska, Evans Dami Binan, Konrad Bocian, Mark J. Brandt, Robert Busching, Anna Cabak Rédei, Huajian Cai, Fanny Cambier, Katarzyna Cantarero, Cheryl L. Carmichael, Francisco Ceric, Jesse Chandler, Jen-Ho Chang, Armand Chatard, Eva E. Chen, Winnee Cheong, David C. Cicero, Sharon Coen, Jennifer A. Coleman, Brian Collisson, Morgan A. Conway, Katherine S. Corker, Paul G. Curran, Fiery Cushman, Zubairu K. Dagona, Ilker Dalgar, Anna Dalla Rosa, William E. Davis, Maaike de Bruijn, Leander De Schutter, Thierry Devos, Marieke de Vries, Canay Doğulu, Nerisa Dozo, Kristin Nicole Dukes, Yarrow Dunham, Kevin Durrheim, Charles R. Ebersole, John E. Edlund, Anja Eller, Alexander Scott English, Carolyn Finck, Natalia Frankowska, Miguel-Ángel Freyre, Mike Friedman, Elisa Maria Galliani, Joshua C. Gandi, Tanuka Ghoshal, Steffen R. Giessner, Tripat Gill, Timo Gnambs, Ángel Gómez, Roberto González, Jesse Graham, Jon E. Grahe, Ivan Grahek, Eva G. T. Green, Kakul Hai, Matthew Haigh, Elizabeth L. Haines, Michael P. Hall, Marie E. Heffernan, Joshua A. Hicks, Petr Houdek, Jeffrey R. Huntsinger, Ho Phi Huynh, Hans IJzerman, Yoel Inbar, Åse H. Innes-Ker, William Jiménez-Leal, Melissa-Sue John, Jennifer A. Joy-Gaba, Roza G. Kamiloğlu, Heather Barry Kappes, Serdar Karabati, Haruna Karick, Victor N. Keller, Anna Kende, Nicolas Kervyn, Goran Knežević, Carrie Kovacs, Lacy E. Krueger, German Kurapov, Jamie Kurtz, Daniël Lakens, Ljiljana B. Lazarević, Carmel A. Levitan, Neil A. Lewis, Samuel Lins, Nikolette P. Lipsey, Joy E. Losee, Esther Maassen, Angela T. Maitner, Winfrida Malingumu, Robyn K. Mallett, Satia A. Marotta, Janko Međedović, Fernando Mena-Pacheco, Taciano L. Milfont, Wendy L. Morris, Sean C. Murphy, Andriy Myachykov, Nick Neave, Koen Neijenhuijs, Anthony J. Nelson, Félix Neto, Austin Lee Nichols, Aaron Ocampo, Susan L. O'Donnell, Haruka Oikawa, Masanori Oikawa, Elsie Ong, Gábor Orosz, Malgorzata Osowiecka, Grant Packard, Rolando Pérez-Sánchez, Boban Petrović, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Gabrielle Pogge, Monique M. H. Pollmann, Abraham M. Rutchick, Patricio Saavedra, Alexander K. Saeri, Erika Salomon, Kathleen Schmidt, Felix D. Schönbrodt, Maciej B. Sekerdej, David Sirlopú, Jeanine L. M. Skorinko, Michael A. Smith, Vanessa Smith-Castro, Karin C. H. J. Smolders, Agata Sobkow, Walter Sowden, Philipp Spachtholz, Manini Srivastava, Troy G. Steiner, Jeroen Stouten, Chris N. H. Street, Oskar K. Sundfelt, Stephanie Szeto, Ewa Szumowska, Andrew C. W. Tang, Norbert Tanzer, Morgan J. Tear, Jordan Theriault, Manuela Thomae, David Torres, Jakub Traczyk, Joshua M. Tybur, Adrienn Ujhelyi, Robbie C. M. van Aert, Marcel A. L. M. van Assen, Marije van der Hulst, Paul A. M. van Lange, Anna Elisabeth van 't Veer, Alejandro Vásquez-Echeverría, Leigh Ann Vaughn, Alexandra Vázquez, Luis Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid P. J. Voermans, Marek A. Vranka, Cheryl Welch, Aaron L. Wichman, Lisa A. Williams, Michael Wood, Julie A. Woodzicka, Marta K. Wronska, Liane Young, John M. Zelenski, Zeng Zhijia, and Brian A. Nosek. Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018. ISSN 2515-2467. doi: 10.1177/2515245918810225. URL http://dx.doi.org/10.1177/2515245918810225.

Joshua Knobe. Intentional action and side effects in ordinary language. *Analysis*, 63, 07 2003. doi: 10.1111/1467-8284.00419.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. Finetuning llms for human behavior prediction in social science experiments, 2025. URL https://arxiv.org/abs/2509.05830.

Anton Kühberger. The influence of framing on risky decisions: A meta-analysis. *Organizational behavior and human decision processes*, 75(1):23–55, 1998.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pp. 255–268, 1989.

Lunzheng Li, Zacharias Maniadis, and Constantine Sedikides. Anchoring in economics: A meta-analysis of studies on willingness-to-pay and willingness-to-accept. *Journal of Behavioral and Experimental Economics*, 90:101629, 2021. ISSN 2214-8043. doi: 10.1016/j.socec.2020.101629. URL http://dx.doi.org/10.1016/j.socec.2020.101629.

Mengfei Li and Gilad Feldman. Revisiting mental accounting classic paradigms: Replication registered report of the problems reviewed in thaler (1999). *Royal Society Open Science*, 12(9),

2025. ISSN 2054-5703. doi: 10.1098/rsos.250979. URL http://dx.doi.org/10.1098/rsos.250979.

Peiyao Li, Noah Castelo, Zsolt Katona, and Miklos Sarvary. Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2):254–266, 2024.

Xuan Liu, Jie ZHANG, HaoYang Shang, Song Guo, Yang Chengxu, and Quanyan Zhu. Exploring prosocial irrationality for LLM agents: A social cognition view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=u8VOQVzduP.

Xuan Liu, Haoyang Shang, and Haojian Jin. Cobra: Programming cognitive bias in social agents using classic social science experiments, 2026. URL https://arxiv.org/abs/2509.13588.

Marlene Lutz, Indira Sen, Georg Ahnert, Elisa Rogers, and Markus Strohmaier. The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23212–23237, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1261. URL https://aclanthology.org/2025.findings-emnlp.1261/.

Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Lewend Mayiwar, Kai Hin Wan, Erik Løhre, and Gilad Feldman. Revisiting representativeness heuristic classic paradigms: Replication and extensions of nine experiments in kahneman and tversky (1972). *Quarterly Journal of Experimental Psychology*, 78(4):707–730, 2024. ISSN 1747-0226. doi: 10.1177/17470218241255916. URL http://dx.doi.org/10.1177/17470218241255916.

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is GPT-3? an exploration of personality, values and demographics. In David Bamman, Dirk Hovy, David Jurgens, Katherine Keith, Brendan O'Connor, and Svitlana Volkova (eds.), *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pp. 218–227, Abu Dhabi, UAE, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nlpcss-1.24. URL https://aclanthology.org/2022.nlpcss-1.24/.

Suhong Moon, Marwa Abdulhai, Minwoo Kang, Joseph Suh, Widyadewi Soedarmadji, Eran Kohen Behar, and David M. Chan. Virtual personas for language models via an anthology of backstories, 2024. URL https://arxiv.org/abs/2407.06576.

Richard D. Morey and Eric-Jan Wagenmakers. Simple relation between Bayesian order-restricted and point-null hypothesis tests. *Statistics & Probability Letters*, 92:121–124, 2014. doi: 10.1016/j.spl.2014.05.010.

Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, and Zhongyu Wei. From individual to society: A survey on social simulation driven by large language model-based agents, 2024. URL https://arxiv.org/abs/2412.03563.

Brian Mullen, Jennifer L Atkins, Debbie S Champion, Cecelia Edwards, Dana Hardy, John E Story, and Mary Vanderklok. The false consensus effect: A meta-analysis of 115 hypothesis tests. *Journal of Experimental Social Psychology*, 21(3):262–283, 1985.

Rosemarie Nagel. Unraveling in guessing games: An experimental study. *The American economic review*, 85(5):1313–1326, 1995. URL https://www.jstor.org/stable/2950991.

Sanne Nauts, Oliver Langner, Inge Huijsmans, Roos Vonk, and Daniël HJ Wigboldus. Forming impressions of personality. *Social Psychology*, 2014.

Hessel Oosterbeek, Randolph Sloof, and Gijs Van De Kuilen. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental economics*, 7(2):171–188, 2004.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349 (6251):aac4716, 2015. doi: 10.1126/science.aac4716.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024. URL https://arxiv.org/abs/2411.10109.

Deborah A. Prentice and Dale T. Miller. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of personality and social psychology*, 64 2:243–56, 1993. URL https://api.semanticscholar.org/CorpusID:24004422.

Maximilian Puelma Touzel, Sneheel Sarangi, Aurélien Bück-Kaeffer, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. Position: Time to close the validation gap in LLM social simulations. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/67174.

Lee D. Ross, David Greene, and Pamela House. The "false consensus effect": An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, 13: 279–301, 1977. URL https://api.semanticscholar.org/CorpusID:9032175.

Jeffrey N Rouder, Paul L Speckman, Dongchu Sun, Richard D Morey, and Geoffrey Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, 16(2): 225–237, 2009.

Jeffrey N Rouder, Richard D Morey, Paul L Speckman, and Jordan M Province. Default bayes factors for anova designs. *Journal of mathematical psychology*, 56(5):356–374, 2012.

Kai Ruggeri, Sonia Alí, Mari Louise Berge, Giulia Bertoldo, Ludvig D. Bjørndal, Anna Cortijos-Bernabeu, Clair Davison, Emir Demić, Celia Esteban-Serna, Maja Friedemann, Shannon P. Gibson, Hannes Jarke, Ralitsa Karakasheva, Peggah R. Khorrami, Jakob Kveder, Thomas Lind Andersen, Ingvild S. Lofthus, Lucy McGill, Ana E. Nieto, Jacobo Pérez, Sahana K. Quail, Charlotte Rutherford, Felice L. Tavera, Nastja Tomat, Chiara Van Reyn, Bojana Većkalov, Keying Wang, Aleksandra Yosifova, Francesca Papa, Enrico Rubaltelli, Sander van der Linden, and Tomas Folke. Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 4(6):622–633, 2020. ISSN 2397-3374. doi: 10.1038/s41562-020-0886-x. URL http://dx.doi.org/10.1038/s41562-020-0886-x.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

Dan R. Schley and Evan Weingarten. Fifty years of anchoring effects: A theoretical reintegration and meta-analysis. *Management Science*, 2026. ISSN 1526-5501. doi: 10.1287/mnsc.2023.03238. URL http://dx.doi.org/10.1287/mnsc.2023.03238.

Kathleen Schmidt and R. Grace Drake. Minimal group procedures and outcomes. *Collabra: Psychology*, 9(1), 2023. ISSN 2474-7394. doi: 10.1525/collabra.90187. URL http://dx.doi.org/10.1525/collabra.90187.

Christine M Schroeder and Deborah A Prentice. Exposing pluralistic ignorance to reduce alcohol use among college students 1. *Journal of Applied Social Psychology*, 28(23):2150–2180, 1998.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7:1954–1968, dec 2025. doi: 10.1038/s42256-025-01115-6. URL https://www.nature.com/articles/s42256-025-01115-6.

Eldar Shafir and Amos Tversky. Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology*, 24:449–474, 1992. URL https://api.semanticscholar.org/CorpusID:29570235.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023. doi: 10.1038/s41586-023-06647-8. URL https://www.nature.com/articles/s41586-023-06647-8.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.814/.

Alexander Steiger and Anton Kühberger. A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie*, 2018.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions, 2025. URL https://arxiv.org/abs/2502.16761.

Jerry Suls and Peter Green. Pluralistic ignorance and college student perceptions of gender-specific alcohol norms. *Health Psychology*, 22(5):479, 2003.

Lindia Tjuatja, Valerie Chen, Sherry Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design, 2024. URL https://arxiv.org/abs/2311.04076.

Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950. doi: 10.1093/mind/LIX.236.433.

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

Angelina Wang, Jamie Morgenstern, and John P. Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7:400–411, feb 2025. doi: 10.1038/s42256-025-00986-z. URL https://doi.org/10.1038/s42256-025-00986-z.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip H. S. Torr, Bernard Ghanem, and Guohao Li. Can large language model agents simulate human trust behavior? In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1cb57fcf7ff3f6d37eebae5becc9ea6d-Abstract-Conference.html.

Chenjun Xu, Bingbing Wen, Bin Han, Robert Wolfe, Lucy Lu Wang, and Bill Howe. Do language models mirror human confidence? exploring psychological insights to address overconfidence in llms, 2025. URL https://arxiv.org/abs/2506.00582.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, Baohua Zhou, and Zhongyu Wei. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users, 2025. URL https://arxiv.org/abs/2504.10157.

AI USE STATEMENT

Generative AI enters this work in two ways: as a component of the HUMANSTUDY-BENCH pipeline, and as an aid in conducting the research and writing the paper. The LLMs evaluated as simulated participants are the objects of study.

AI in the HUMANSTUDY-BENCH pipeline. The pipeline has two stages. *Study construction (Filter–Extract–Execute).* An LLM from the Gemini 3 Flash family screens candidate papers against the inclusion checklist, extracts experimental designs, hypotheses, and reported statistics, and drafts the configuration and response-parsing code for each study. Human reviewers make all inclusion decisions, check every extracted parameter against the source paper, and cross-review and test the generated code before deployment. *Simulation and evaluation.* Agents answer the reconstructed trials. For the A4 design, participant backstories are generated by Gemini 3 independently of the experimental task, and reviewers inspect them for task-specific answers. A DeepSeek formatter model rewrites free-text agent answers into the structured answer format; it is instructed to extract only values present in the response and not to supply missing answers. PAS and ECS are then computed by fixed statistical code; no LLM judges agreement between agent and human results.

AI in conducting the research and writing the paper. We used generative AI to write plotting code, edit figures, search related works, assist the targeted literature check that classifies follow-up evidence for each behavioral hypothesis (Appendix E.1), edit the manuscript for readability, provide methodological feedback, and check mathematical claims that the authors formulated and verified. We did not use AI to propose any research ideas or hypotheses, write proofs, or conduct qualitative or thematic analysis. The authors made the evidence classifications, checked citations against source papers, tested AI-assisted code, and reviewed all AI-assisted text and figures. We take responsibility for the final paper and artifacts, including content produced with the aid of AI.

A METRICS THEORETICAL FOUNDATION

A.1 METRIC INTUITION: WHAT ARE WE MEASURING?

PAS summarizes agreement between posterior states of specified hypotheses; ECS summarizes concordance between selected effect estimates. The metrics are related but answer different questions. A high PAS need not imply accurate magnitudes, and matching selected effect estimates does not establish the fidelity of the complete response distribution.

A.2 POSTERIOR FACTORIZATION ASSUMPTION

For human and agent hypothesis states θ_h, θ_a, the probability-of-agreement interpretation assumes

$$P(\theta_h, \theta_a \mid D_h, D_a) = P(\theta_h \mid D_h)P(\theta_a \mid D_a). \tag{5}$$

Separate sampling and independent priors with factorized likelihoods provide one model under which this holds.

A.3 FOUNDATION OF PROBABILITY ALIGNMENT SCORE

Let $\mathrm{PAS}^* = \mathbf{1}(\theta_h = \theta_a)$. Under squared-error loss, its posterior mean minimizes posterior expected loss. With the factorization above and K mutually exclusive hypothesis states,

$$\widehat{\mathrm{PAS}} = \mathbb{E}[\mathrm{PAS}^* \mid D_h, D_a] = \sum_{k=1}^{K} \pi_{h,k}\pi_{a,k}. \tag{6}$$

If the human posterior is uniform over K states, it gives no preferred reference conclusion and the agreement score equals $1/K$ for any agent posterior. Such a test cannot distinguish agents on inferential agreement.

For two states, this becomes $\pi_h\pi_a + (1-\pi_h)(1-\pi_a)$; the implemented directional extension is described in Appendix C.2.

For the binary evidence transform, equal prior odds imply $\pi = \Lambda/(1 + \Lambda)$, where $\Lambda = BF_{10}$. Equal odds are a modeling choice; effect-size priors and approximations still affect the resulting probabilities. Appendix C.1 reports a sensitivity analysis for the tested JZS scales. The continuous transformation avoids a discontinuous significance threshold, and the data-based directional split (Appendix C.2) keeps the three-state PAS continuous when an estimate crosses zero.

B Methodological Rationale and Aggregation

B.1 Methodological Rationale

Our aggregation strategy and metric formulation diverge from standard meta-analytic approaches. We explicitly contrast our choices with alternative methodologies below.

Benchmarking vs. Meta-Analysis. While our framework aggregates results across multiple studies, we intentionally employ unweighted averaging rather than the precision-weighted averaging (inverse-variance weighting) typical of meta-analysis. This decision is grounded in two primary distinctions:

- **Task Independence vs. Parameter Estimation:** Meta-analysis assumes that different studies estimate a shared biological parameter (e.g., a "true" population effect size). In contrast, our goal is benchmarking: evaluating an agent's general capability across a diverse suite of distinct tasks. Weighting by inverse variance would allow a single study with high statistical power to dominate the aggregate score, obscuring the agent's failure on smaller but equally critical tasks.
- **Avoiding Simulation Artifacts:** Although agent sample size (N_{agent}) can be controlled in a simulation, it is fixed by each reconstructed protocol in our evaluation. Precision weighting would allow larger simulated samples to dominate the benchmark score. Equal study weights prevent simulation volume from determining aggregation weights; PAS still uses the fixed sample size of each test when assessing its evidence.

B.2 Metric Selection Rationale

PAS focuses on evidence for specified hypotheses, while effect-size and distributional metrics answer complementary fidelity questions. Response variance can itself be behaviorally meaningful. Reported statistical summaries permit evaluation where original human microdata are unavailable, but limit checks of individual-level distributions. A common score range supports summarization across tests without removing differences in their statistical models or power.

B.3 Aggregation Implementation

Scores are aggregated across the hierarchy (Test $\rightarrow$ Behavioral Hypothesis $\rightarrow$ Study $\rightarrow$ Benchmark) using the following scoring convention:

1. **Hypothesis & Study Level (Arithmetic Mean):** We average test-level PAS within each behavioral hypothesis and hypothesis-level PAS within each study:

$$\mathrm{PAS}_{\mathrm{hypothesis}} = \frac{1}{M} \sum_{j=1}^{M} \mathrm{PAS}_j. \tag{7}$$

   Because each test-level PAS is the probability that the agent reaches the same conclusion as the human study on that test, each mean is the expected proportion of tests on which agent and human conclusions agree. Unlike averages on the log-odds scale, they are not dominated by scores near zero or one and need no clipping constant.
2. **Benchmark Level (Arithmetic Mean):** We compute the unweighted arithmetic mean of study-level PAS. Given the heterogeneity of the studies—which span diverse cognitive and social domains—we treat each study as an independent unit of capability. This approach ensures equal representation across domains and prevents any single study with distinct statistical properties (e.g., large N) from dominating the global benchmark score.

For ECS, only tests with recoverable agent and human effect pairs enter. Each such test i in hypothesis q of study s receives weight $\tilde{w}_i = 1/(KH_sT_{sq})$, where K is the number of studies with at least one such test, H_s is the number of hypotheses in study s with such tests, and T_{sq} is the number of such tests in hypothesis q. The weights sum to one, so each study contributes $1/K$ and each hypothesis within a study an equal share, matching the aggregation of PAS. With weighted means $\bar{\delta}$ and centered effects $u_{a,i}, u_{h,i}$, the implemented concordance is

$$\mathrm{ECS}_{\mathrm{global}} = \frac{2\sum_i \tilde{w}_i u_{a,i} u_{h,i}}{\sum_i \tilde{w}_i u_{a,i}^2 + \sum_i \tilde{w}_i u_{h,i}^2 + (\bar{\delta}_a - \bar{\delta}_h)^2}. \tag{8}$$

For the unweighted form in the main text, $\mathrm{MSE} = E[(a-h)^2]$. Expanding it gives $\sigma_a^2 + \sigma_h^2 - 2\sigma_{ah} + (\mu_a - \mu_h)^2$. Thus $V = \sigma_a^2 + \sigma_h^2 + (\mu_a - \mu_h)^2$ is the expected squared difference when the covariance is zero, and $\mathrm{ECS} = 1 - \mathrm{MSE}/V$.

C Metrics Implementation Details

C.1 Evidence Transformation (Priors)

To compute posterior probabilities π, we calculate Bayes Factors (BF_{10}) using priors tailored to the test type. For t-tests and ANOVA, we employ the JZS prior (Cauchy distribution on effect size) with default scales $r = \sqrt{2}/2$ and $r = 0.5$, respectively (Rouder et al., 2009; 2012). These account for the majority of the tests. For contingency tables, we utilize a BIC-style approximation ($BF_{10} \approx \exp((\chi^2 - \mathrm{df}\ln n)/2)$), while binomial tests use an exact conjugate Beta-Binomial prior (Beta(1,1)).

Sensitivity Analysis. To ensure that the benchmark rankings are driven by agent capability rather than specific prior choices in the evidence transformation, we conducted a sensitivity analysis on the prior scale r used in the JZS Bayes Factor computation. The default value $r = 0.707$ assumes a medium effect size distribution. We re-evaluated the raw responses of the forty agent configurations on the scored behavioral hypotheses, varying r from 0.5 (small effects) to 1.0 (large effects).

As shown in Table 4, while the absolute magnitude of the posterior probabilities shifts slightly with the prior width, the relative ranking of agents remains highly stable (Spearman's $\rho \geq 0.998$; mean PAS changes by at most 0.006).

We also varied the structural choices of PAS, for human and agent posteriors alike (lower block of Table 4): prior odds of $1/3$ and 3; the sign probability $P(\mathrm{sign}\,\delta \mid D, H_1)$ taken under the JZS prior of the Bayes factor rather than a noninformative prior (t tests and single-df ANOVAs); a sign probability of 1 for the observed sign; and Gunel–Dickey Bayes factors with uniform priors (Gunel & Dickey, 1974) for contingency tables whose counts are available to the evaluator (64% of χ^2 tests). Rankings remain stable (Spearman's $\rho \geq 0.95$). A3 exceeds A2 and social exceeds strategic PAS under every variant.

Table 4: **Sensitivity of agent rankings to PAS choices.** Upper block: Cauchy prior scale r. Lower block: one structural choice changed at a time. ρ is the Spearman correlation of the forty configuration PAS values with the default.

Setting	Spearman's ρ	Mean Δ PAS	Max Δ PAS	Status
$r = 0.500$	0.9991	0.002697	0.005821	Stable
$r = 0.600$	0.9994	0.001347	0.002826	Stable
$r = \mathbf{0.707}$ (Default)	**1.0000**	**0.0000**	**0.0000**	**Baseline**
$r = 0.800$	0.9992	0.000967	0.002217	Stable
$r = 0.900$	0.9991	0.002053	0.004436	Stable
$r = 1.000$	0.9989	0.003069	0.006488	Stable
Prior odds $1/3$	0.9629	0.016529	0.039455	Stable
Prior odds 3	0.9537	0.032426	0.058871	Stable
JZS sign probability	0.9998	0.000401	0.001323	Stable
Sign probability $= 1$	0.9878	0.004973	0.019651	Stable
Gunel–Dickey	0.9976	0.007902	0.012221	Stable

C.2 Generalization to Multiple Hypotheses

Theoretically, PAS generalizes to K hypotheses via the inner product of posterior vectors $\vec{\pi}_h \cdot \vec{\pi}_a$. In our implementation, we specifically operationalize this as a 3-way split (H_+, H_-, H_0). The resulting score is the dot product of the agent and human posterior vectors over these three outcome categories:

$$S = \pi_{h+}\pi_{a+} + \pi_{h-}\pi_{a-} + \pi_{h0}\pi_{a0} \tag{9}$$

Each test compares $H_0 : \delta = 0$ with $H_+ : \delta > 0$ and $H_- : \delta < 0$ (for binomial tests, $\delta = p - p_0$). With equal prior odds for H_0 and $H_1 = H_+ \cup H_-$, and since $\pi_\pm = P(H_1 \mid D)\, P(\operatorname{sign}\delta = \pm \mid D, H_1)$,

$$\pi_0 = \frac{1}{1 + BF_{10}}, \qquad \pi_\pm = \frac{BF_{10}}{1 + BF_{10}}\, P(\operatorname{sign}\delta = \pm \mid D, H_1). \tag{10}$$

For a prior symmetric about the null value this matches the one-sided Bayes factors, $BF_{\pm 0} = 2\, BF_{10}\, P(\operatorname{sign}\delta = \pm \mid D, H_1)$ (Morey & Wagenmakers, 2014). By default, $P(\operatorname{sign}\delta \mid D, H_1)$ is the sign posterior under a noninformative prior: the Student-t posterior for t tests, single-df F tests, and correlations (via their t statistic), a normal approximation with $z = \sqrt{\chi^2}$ for 2×2 tables and z-scored statistics, and the Beta$(1, 1)$ posterior mass on the observed side of p_0 for binomial tests. Taking it under the JZS prior of the Bayes factor instead changes PAS by at most 0.0013 (Table 4). Human and agent posteriors of a test use the same test and prior, so each test-level PAS is the probability that both fall in the same state. When the hypothesis has no sign, as for tests with more than one numerator degree of freedom, PAS compares only the effect and no-effect states, $\pi_{h1}\pi_{a1} + \pi_{h0}\pi_{a0}$ with $\pi_1 = 1 - \pi_0$.

C.3 Standardized Effect Size Recovery

We recover Cohen's d from reported statistics using the following conversions:

- **T-family:** Independent t-tests use $d = t\sqrt{(n_1 + n_2)/(n_1 n_2)}$; paired/one-sample tests use $d = t/\sqrt{n}$. F-tests ($df_1 = 1$) are converted to t-equivalents ($t = \sqrt{F}$) and processed similarly.
- **Correlation-family:** Pearson's r, Fisher's z, and Mann-Whitney U (via rank-biserial r_{rb}) are converted to d using the relationship $d = 2r/\sqrt{1 - r^2}$.
- **Discrete:** 2×2 contingency tables are converted via the Log Odds Ratio ($d \approx \ln(OR)\sqrt{3}/\pi$). Binomial proportions use Cohen's $h = 2\arcsin\sqrt{p} - 2\arcsin\sqrt{p_0}$ against the null proportion p_0.

D Agent Design Variants

All four designs receive the same study trial; only the agent specification changes. The role and demographic fields are assigned before the trial response is generated.

D.1 Blank (A1)

No additional system-level participant description is provided.

D.2 Role-Play (A2)

The system message asks the model to participate in a psychology experiment as a human participant, without demographic attributes.

D.3 Demographic (A3)

The role instruction is supplemented with available age, gender, and education or background attributes sampled from the study's reported participant profile. The model is asked to follow the trial instructions and answer in the required format.

D.4 CONTEXTUALIZED BACKSTORY (A4)

The demographic profile is extended with a generated narrative about everyday life, relationships, routines, and personality. The narrative is placed in the system message before the same trial instructions. Backstory generation is independent of the specific experimental problem, so the specification adds participant context rather than a suggested answer.

E STUDY RECONSTRUCTIONS

The original papers supply protocols and reported human statistics, not participant-level human records. Each study module generates agent trials, stores individual agent responses, and computes study-specific summaries from those responses. The summaries are compared with the published human references. Table 5 describes the task in each reconstruction and the contrast it tests; adaptations to text-based interaction remain part of the interpretation of every comparison.

Two cases clarify the distinction between generated agent data and unavailable human records. In the birth-sequence task within Study 004, agents produce individual numerical answers; the evaluator counts how often those answers fall below the normative value of 72 and compares that proportion with the published human result. In Study 008, agents give individual self, peer, and friend ratings, which the evaluator combines into paired and group-level contrasts. Neither comparison requires the original human participants' response files.

E.1 FOLLOW-UP EVIDENCE FOR SCORED BEHAVIORAL HYPOTHESES

Each scored behavioral hypothesis was classified by the subsequent independent human evidence for its contrast. *Direct* support means that a later independent experiment or meta-analysis tests the same or a very close manipulation–outcome contrast and reports a same-direction result. *Related* support means that later human evidence establishes the broader effect but does not test this specific contrast. Hypotheses with mixed or contrary follow-up results, or without located follow-up evidence, are excluded from every reported score. Table 6 gives, for each study, the number of scored hypotheses with direct and related support (45 in total) and the follow-up evidence.

F COMPLEMENTARY EXPERIMENTAL RESULTS

F.1 MODEL NAMES AND OPENROUTER IDENTIFIERS

All single-model evaluations use the OpenRouter API. Table 7 lists the display name used in this report and the exact OpenRouter model ID (provider/series) used for each run.

F.2 BOOTSTRAP STANDARD ERRORS: METHODOLOGY AND JUSTIFICATION

Uncertainty for PAS (Probability Alignment Score) is quantified by resampling generated agent records. This describes uncertainty conditional on those records; no original human participant records enter the bootstrap.

Resampling. For each study and agent configuration, the archived procedure resamples generated agent response entries with replacement and recomputes the study-specific statistics. The reported results use $B = 200$ unless stated otherwise. In nested multi-round or paired tasks, the implementation flattens responses before resampling and then groups selected entries back by agent ID. This is a response-level bootstrap, not a participant- or pair-cluster bootstrap; its SEs do not preserve within-agent dependence and should be read as descriptive uncertainty for the generated responses.

Bootstrap distribution and SE. On each bootstrap sample we recompute the study-level PAS (or the same metric used in the main analysis). The bootstrap distribution of that metric is summarized by its standard deviation, which we take as the *bootstrap standard error* for that study. Formally, $\widehat{\mathrm{SE}}_{\mathrm{boot}} = \mathrm{SD}(\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B)$, where $\hat{\theta}^*_b$ is the statistic from the b-th bootstrap sample. No distributional assumption is made beyond the data-generating process implied by the empirical participant set.

Table 5: **What each reconstructed study asks a simulated participant to do.** Studies are grouped by domain. The last column names the manipulation and the measured outcome behind the scored behavioral hypotheses (Table 6). Human references are published summary results; original human microdata are not used.

Study	Task for the simulated participant	Manipulation → outcome
Individual cognition		
False consensus (001) Ross et al. (1977)	Reads short dilemmas (e.g., whether to walk around campus wearing an advertising sign), chooses an action, estimates how many peers would choose each option, and rates the traits of people who would.	Own choice → estimated peer agreement and trait ratings
Anchoring (002) Jacowitz & Kahneman (1995)	Says whether an uncertain quantity is above or below a given number, then estimates it and rates confidence. Variants call the anchor uninformative or ask for willingness to pay.	High vs. low anchor → numeric estimates
Framing (003) Tversky & Kahneman (1981)	Chooses between options in classic decision problems: the Asian disease problem, concurrent gambles, a pseudo-certain prize, and a lost theater ticket vs. lost cash.	Gain vs. loss wording and related reframings → choices
Representativeness (004) Kahneman & Tversky (1972)	Judges how likely samples are: birth-order sequences, school-program majorities, marble distributions, averages from small and large samples, and heights.	How representative a sample looks, or its size → probability judgments
Strategic interaction		
Beauty contest (009) Nagel (1995)	Guesses a number from 0 to 100; the winner is closest to p times the average guess ($p = 1/2$, $2/3$ or $4/3$). Four rounds with feedback.	Value of p and round → guesses and convergence
PD and Newcomb (010) Shafir & Tversky (1992)	Plays one-shot prisoner's dilemmas with the opponent's move known or unknown, faces Newcomb's two-box problem with a computer predictor, and decides whether to pay to learn the opponent's move.	Known vs. unknown opponent move → cooperation; paying for information
Bargaining (011) Forsythe et al. (1994)	Proposes how to split \$5 or \$10 with an anonymous partner, who must accept (dictator game) or may reject (ultimatum game); pay is real or hypothetical.	Game type, pay and stake → offers
Trust game (012) Berg et al. (1995)	As sender, chooses how much of \$10 to send; the amount is tripled and the receiver chooses how much to return. One condition shows results from an earlier session.	Social history → amounts sent and returned
Social psychology		
Intentional action (005) Knobe (2003)	Reads about a company chairman (or an army lieutenant) who knowingly causes a harmful or helpful side effect, then judges whether it was intentional and how much blame or praise is deserved.	Harmful vs. helpful side effect → intentionality, blame and praise
Impression formation (006) Asch (1946)	Reads a list of personality traits (e.g., intelligent, skillful, warm or cold), describes the person, and picks between paired traits such as generous/ungenerous.	Swapped, removed or reordered traits → overall impression
Minimal groups (007) Gagnon & Bourhis (1996)	Is assigned to one of two arbitrary groups by a coin toss, then allocates course points to anonymous in-group and out-group members.	Interdependence and group identification → in-group favoritism
Pluralistic ignorance (008) Prentice & Miller (1993)	Rates own comfort with campus drinking and the comfort of the average student and of friends, and answers questions about a campus keg ban.	Self vs. peer ratings → perceived-norm gap

Table 6: **Follow-up evidence for scored behavioral hypotheses.** Number of scored hypotheses per study with direct or related support, and the follow-up human evidence.

Study	Direct	Related	Follow-up evidence
False consensus (001)	1	4	Mullen et al. (1985); Klein et al. (2018)
Anchoring (002)	2	2	Klein et al. (2014); Schley & Weingarten (2026); Li et al. (2021)
Framing (003)	2	2	Klein et al. (2018); Li & Feldman (2025); Ruggeri et al. (2020)
Representativeness (004)	6	0	Mayiwar et al. (2024)
Intentional action (005)	2	2	Klein et al. (2018); Fillon et al. (2025)
Impression formation (006)	2	3	Nauts et al. (2014); Anderson & Barrios (1961)
Minimal groups (007)	0	1	Schmidt & Drake (2023)
Pluralistic ignorance (008)	1	2	Schroeder & Prentice (1998)
Beauty contest (009)	3	3	Ho et al. (1998)
PD and Newcomb (010)	0	1	Eliaz & Schotter (2010); Bennett et al. (2016)
Bargaining (011)	0	2	Cochard et al. (2021)
Trust game (012)	2	2	Johnson & Mislin (2011); Jiang et al. (2015)
Total	21	24	

Table 7: Model display names and OpenRouter model IDs. All runs use the OpenRouter API; the "OpenRouter model ID" is the exact `model` string sent to the API.

Model	OpenRouter model ID
Claude Haiku 4.5	`anthropic/claude-haiku-4.5`
DeepSeek V3.2	`deepseek/deepseek-v3.2`
Gemini 3 Flash	`google/gemini-3-flash-preview`
Mistral Nemo	`mistralai/mistral-nemo`
Gemma 4 26b	`google/gemma-4-26b-a4b-it`
GPT 5 Nano	`openai/gpt-5-nano`
GPT OSS 120b	`openai/gpt-oss-120b`
GPT OSS 20b	`openai/gpt-oss-20b`
Qwen3 Next 80b	`qwen/qwen3-next-80b-a3b-instruct`
Grok 4.1 Fast	`x-ai/grok-4.1-fast`

Aggregation across studies. For each agent configuration, a single "Total PAS" is computed as the mean of per-study PAS over the K studies. Under the assumption that study-level estimates are approximately independent, the SE of the mean is propagated as

$$\widehat{\mathrm{SE}}(\overline{\mathrm{PAS}}) = \frac{1}{K}\sqrt{\sum_{k=1}^{K} \widehat{\mathrm{SE}}_k^2}\,.$$

F.3 Extended Distributional Analysis

The main text reports polarized test-level inferential agreement. Here we examine the underlying p-values and variability across tasks and agent configurations.

Benchmark Sensitivity and Model Disentanglement To interrogate the source of variance, we decompose performance by task difficulty in Figure 6. The distribution of mean PAS (Left) demonstrates that the benchmark covers a broad difficulty spectrum rather than collapsing into a binary distribution.

The mean–variance plot shows more cross-configuration variation at intermediate mean PAS. This identifies tests on which evaluated configurations disagree, without attributing the variation to a particular architectural mechanism.

Table 8: **Total PAS and Total SE.** Each cell reports Total PAS with its bootstrap Total SE in parentheses; Total PAS is the simple mean of per-study raw PAS, and Total SE is propagated from per-study bootstrap SEs as $\sqrt{\sum \mathrm{SE}_i^2}/N$. Both are computed on the scored behavioral hypotheses; Total PAS matches Table 2.

Model	A1	A2	A3	A4
Claude Haiku 4.5	0.440 (0.0166)	0.432 (0.0159)	0.427 (0.0107)	0.394 (0.0083)
DeepSeek V3.2	0.433 (0.0141)	0.428 (0.0154)	0.450 (0.0130)	0.448 (0.0144)
Gemini 3 Flash	0.542 (0.0103)	0.475 (0.0193)	0.539 (0.0086)	0.476 (0.0306)
Mistral Nemo	0.462 (0.0122)	0.428 (0.0182)	0.458 (0.0136)	0.484 (0.0189)
Gemma 4 26b	0.489 (0.0064)	0.473 (0.0156)	0.511 (0.0074)	0.442 (0.0094)
GPT 5 Nano	0.515 (0.0168)	0.501 (0.0296)	0.539 (0.0198)	0.589 (0.0157)
GPT OSS 120b	0.359 (0.0169)	0.397 (0.0123)	0.422 (0.0132)	0.472 (0.0116)
GPT OSS 20b	0.437 (0.0135)	0.406 (0.0222)	0.464 (0.0331)	0.453 (0.0160)
Qwen3 Next 80b	0.514 (0.0040)	0.516 (0.0063)	0.509 (0.0086)	0.518 (0.0098)
Grok 4.1 Fast	0.404 (0.0071)	0.372 (0.0034)	0.449 (0.0080)	0.421 (0.0085)

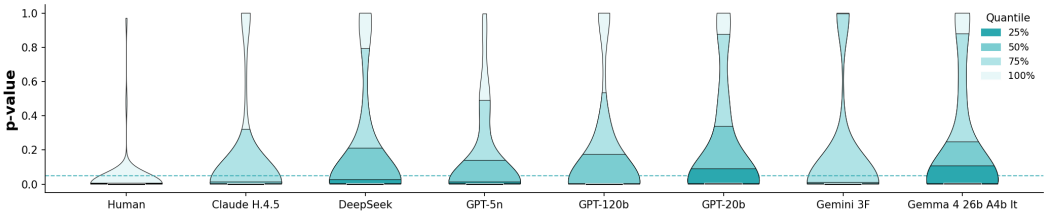


Figure 5: **Distribution of p-values across human and agents (A4 context).** The blue dashed line marks the standard statistical significance threshold ($p = 0.05$). The human data (far left) consistently yields highly significant results, with almost all probability mass tightly clustered near $p \approx 0$. In contrast, agent simulations fail to consistently reproduce these effects, showing a much wider spread of p-values with a substantial portion falling into the non-significant range ($p > 0.05$).

Configuration-Specific Agreement Figure 7 shows the item-level scores across model–prompt combinations. No configuration uniformly dominates, motivating diagnosis at the effect level alongside aggregate scores.

Robustness to Effect-Size Heterogeneity. A natural concern is that the low cross-hypothesis pattern correlation ρ in our RQ1 analysis of how closely agents reproduce human behavioral effects (§5) could stem from converting heterogeneous effect-size metrics (d, h, η^2, r) onto a common scale. As a sensitivity check, we recompute ECS separately within each test family (t, χ^2, F), avoiding any cross-family conversion; the correlation family has only three scored tests and is omitted. Averaged over agent configurations, the resulting ECS values range from 0.06 (F) to 0.22 (t), within the range of global ECS values in Table 2. These results suggest that the reported low concordance is not solely a consequence of pooling test families.

Figure 3 separates pattern correlation (ρ), which asks whether effects that are relatively large in the human studies are also relatively large in the simulations, from the calibration factor (C_b), which asks whether the two sets of effects have similar overall averages and spread. For GPT-5 Nano with role-play instructions, calibration is high ($C_b \approx 0.96$) but pattern correlation is much lower ($\rho \approx 0.30$). Its effects therefore occupy a broadly similar numerical range while frequently assigning the wrong relative magnitude to particular tests.

A Sampling-Noise Reference. We simulated a replication of each scored test at its original design and sample size, treating the published result as the true effect, and scored it against the human reference as for agents ($B = 1000$). The mean PAS was 0.848, below one because not all published tests are strongly significant; the corresponding ECS was 0.972.

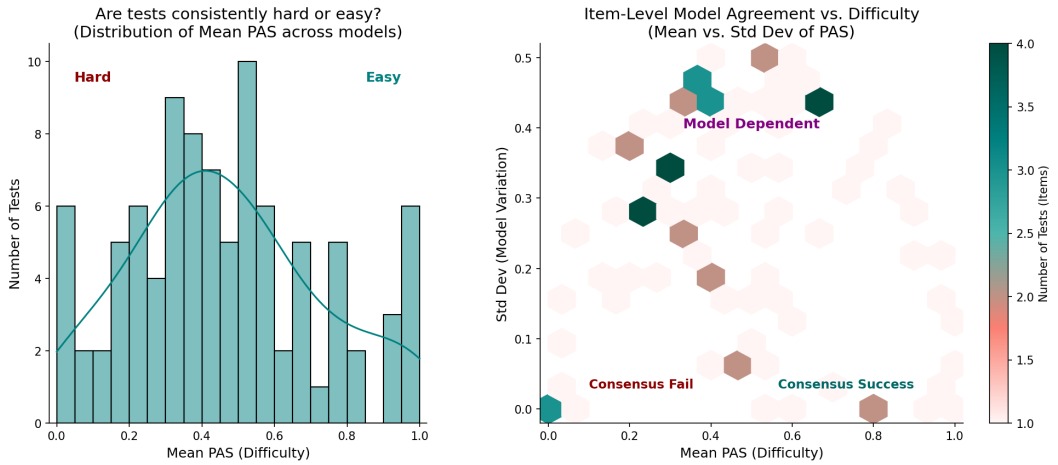


Figure 6: **Decomposing Variance via Task Difficulty. (Left)** The benchmark spans a wide spectrum of difficulty levels, with variation in mean PAS. **(Right)** The Hexbin analysis identifies a "Zone of Disagreement" (red center), where variance peaks. Configurations differ most at intermediate average scores.

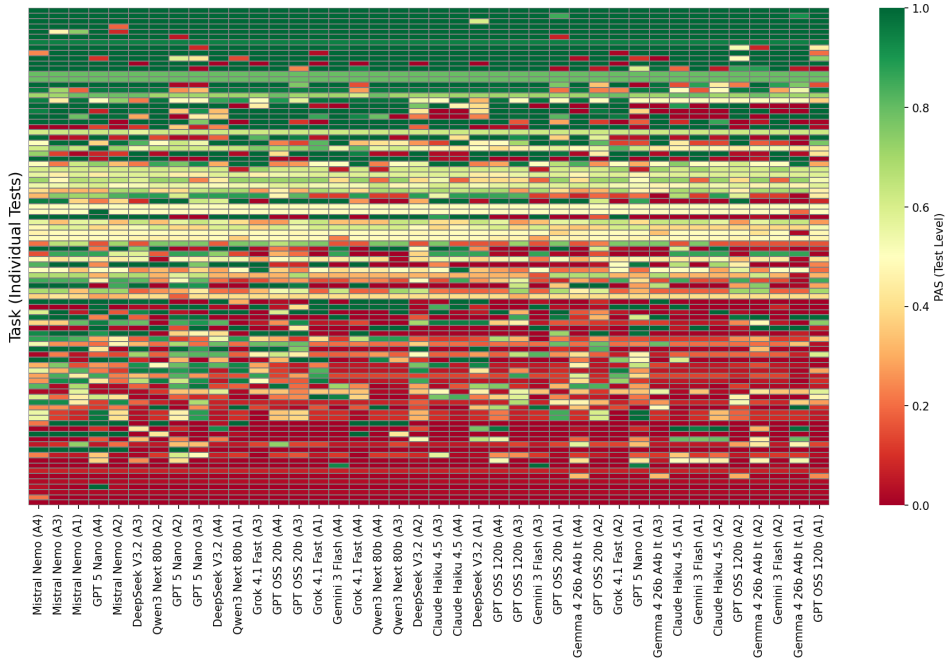


Figure 7: **The Landscape of Idiosyncratic Capabilities.** The scattered distribution of high (green) and low (red) alignment scores illustrates that capabilities are fragmented. No single agent universally dominates; instead, performance is highly specific to the interaction between agent and task type.

F.4 Hypothesis Testing Details

Table 9 reports exploratory comparisons of aggregate PAS. Agent-design contrasts use two-sided Wilcoxon signed-rank tests over the ten models; Δ is the mean paired difference. Domain comparisons are descriptive because the studies and effect types differ.

In the response-mixing check, each simulated participant is drawn at random from one of the ten models and the mixture is scored like a single configuration (100 draws). The mixture reached PAS

Table 9: **Summary of Hypothesis Tests.** Results based on aggregated PAS scores. Exploratory, uncorrected tests; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Study-bootstrap intervals for the same design changes include zero (Table 11).

ID	Hypothesis Description	Method	Δ	Statistic	Result
Group A: Agent Design					
H1	A3 (Demo) > A1 (Empty)	Wilcoxon	$+0.017$	$W = 10$	n.s.
H1b	A3 (Demo) > A2 (Role-Play)	Wilcoxon	$+0.034$	$W = 3^{**}$	Significant
H2	A3 (Demo) > A4 (Narrative)	Wilcoxon	$+0.007$	$W = 21$	n.s.
H3	A1 > A2 (Role-Play)	Wilcoxon	$+0.017$	$W = 10$	n.s.
Group B: Domain Specificity					
H4	Social > Cognition	Pooled t	$+0.034$	$t = 2.28^{*}$	Significant
H5	Social > Strategic	Pooled t	$+0.180$	$t = 7.52^{***}$	Significant
Group C: Model Scaling					
H6	Proprietary $\neq$ Open-weight	Mann-Whitney	$+0.012$	$p = 1.00$	n.s.

0.522 and ECS 0.192; one of ten single models exceeded it on each metric (mean PAS difference -0.060, study-bootstrap 95% interval -0.142 to 0.008).

F.5 Inference Cost Analysis

Table 10: **Inference Cost Analysis (USD).** Total cost to replicate the full experimental suite. For each model, teal marks the cheapest specification and salmon marks the most expensive.

	Claude H.4.5	DeepSeek V3.2	Gemini 3 Flash	Mistral Nemo	Gemma 4 26b	GPT 5 Nano	GPT OSS 120b	GPT OSS 20b	Qwen3 Next80b	Grok 4.1 Fast	Mixed Models
A1	9.55	0.81	2.99	0.39	0.62	6.45	1.70	1.42	0.82	0.59	2.35
A2	10.37	0.81	2.82	0.38	0.55	2.83	1.69	1.38	0.84	0.51	2.15
A3	6.55	1.07	2.84	0.44	0.59	2.57	1.74	1.18	0.87	0.85	1.65
A4	8.37	3.06	5.33	0.20	0.61	0.40	1.01	1.24	0.97	1.29	1.71

F.6 Effect-specific comparisons for RQ2

The direction analysis uses each combination of a statistical test and an agent configuration for which the human posterior favors a positive or negative effect. We classify the agent posterior as favoring the reported direction, the opposite direction, or no effect. The magnitude analysis uses the subset in which the direction matches, both standardized effects are finite, and the published human effect satisfies $|d_h| \geq 0.2$, the conventional lower bound for a small effect (Cohen, 1988). Within this subset, an absolute agent-to-human effect ratio below 0.5 is classified as attenuation and a ratio above 2 as exaggeration. The factor-of-two threshold brackets the average ratio of replication to original effect sizes reported by large replication projects (Open Science Collaboration, 2015; Camerer et al., 2018). Comparisons without a directional human reference are excluded rather than counted as disagreement.

The examples in Table 3 were checked against parsed final answers and published human summaries. In the framing task, most agent configurations favored the certain option under both wordings, rather than showing the human gain/loss reversal. For the intentional-action task, the configuration-level harm/help medians were 95.7%/0% versus 82%/23% in the published human sample. A further attenuation example is peer comfort: the median simulated self–peer gap was 0.63 points versus the published 1.71-point gap.

F.7 Hypothesis-level agent-design comparisons for RQ3

This exploratory analysis pairs the same model and behavioral hypothesis across adjacent agent designs, so each difference measures the change associated with adding one layer of participant specification. We first average paired changes across models within each hypothesis and then balance

the hypothesis-level changes across studies, so studies with more hypotheses do not dominate. The exploratory intervals resample studies.

Table 11: **Hypothesis-level changes in PAS.** Mean paired differences and study-bootstrap 95% intervals. Hypothesis counts describe the ten-model mean difference: increase > 0.02, decrease < -0.02, or within ± 0.02.

Change	Mean ΔPAS	95% interval	Increase	Decrease	Within ± 0.02
A1 $\rightarrow$ A2	-0.017	$[-0.049,\ 0.012]$	15	15	15
A2 $\rightarrow$ A3	$+0.034$	$[-0.008,\ 0.076]$	21	12	12
A3 $\rightarrow$ A4	-0.007	$[-0.059,\ 0.032]$	15	11	19

The effect of added context is model-specific: relative to A1, backstories (A4) raised PAS for GPT OSS 120b and GPT-5 Nano but lowered it for Claude Haiku 4.5, Gemini 3 Flash, and Gemma 4 26b, each by more than twice the response-bootstrap standard error of the difference.

F.8 Prompt Sensitivity Ablation

We assessed sensitivity to prompt wording in the A3 (Demographic) specification. We created two paraphrased variants by rewriting all three components—role-playing framing, demographic formatting, and task instructions—while preserving informational content (Table 12). We measured within-paraphrase PAS variation (max–min range across the three A3 variants) on Study 2 ($N = 256$) and Study 8 ($N = 468$), selected for their moderate sample sizes, and compared it with the A1–A4 range from the same rerun (Table 13).

Table 12: **Prompt paraphrase design.** Three A3 variants with equivalent information but different surface forms.

	Role frame	Demographics	Task instructions
A3	"You are participating…"	Bullet list	"Follow the experimenter's…"
Para. A	"Imagine you are a person…"	Inline	"Respond to each task…"
Para. B	"You have been recruited…"	Indented block	"Answer every question…"

Table 13: **Prompt sensitivity results.** PAS range (percentage points) across the three A3 wordings and across the four agent designs, computed on the scored behavioral hypotheses.

Model	Study	Paraphrase range	A1–A4 range
Mistral Nemo	Study 2 (N=256)	2.4	3.2
Mistral Nemo	Study 8 (N=468)	6.0	13.0
DeepSeek V3.2	Study 2 (N=256)	1.3	8.1
DeepSeek V3.2	Study 8 (N=468)	13.3	24.4

Across the four tested model–study combinations, PAS variation from paraphrasing stays within 13.3 percentage points and is smaller than the A1–A4 range in every combination (3.2 to 24.4 percentage points).

G Broader Impacts

G.1 Potential Benefits

HumanStudy-Bench makes published experimental protocols reusable for evaluating AI agents. Researchers can use it to test whether an agent reproduces a specific behavioral finding, identify where its responses differ from human evidence, and examine how those differences change with model or agent design. The platform may also reduce the repeated effort required to reconstruct experimental materials, conditions, and analyses. By evaluating population-level effects rather than surface-level fluency, it provides evidence that is more diagnostic than a single overall judgment of

whether an agent appears human-like. The benchmark is available for research use and is designed to support contributions of additional studies and human reference evidence.

G.2 Risks and Responsible Use

Limits of interpretation and use. Agreement on the included findings concerns particular effects under particular experimental conditions. Our results show that agents can miss or reverse a human effect, preserve its direction while distorting its magnitude, and change unevenly with agent design. Benchmark scores should therefore not be used on their own for consequential decisions in areas such as public policy, health, hiring, education, or market research; new uses require validation against relevant human data.

Representativeness and stereotypes. The benchmark draws on 12 original studies published between 1946 and 2003 (Table 1 shows examples). Their samples are predominantly WEIRD (Western, educated, industrialized, rich, and democratic)—mainly US, UK, Canadian, and Dutch university students, with one US public-park sample (Knobe, 2003) as the only non-student exception. The follow-up evidence extends to 2026 and includes some cross-national evidence, notably a 19-country replication (Ruggeri et al., 2020), but is still drawn largely from Western, educated populations. Findings may also change over time. In addition, demographic attributes and backstories may lead models to rely on stereotypes rather than represent the people described. Benchmark agreement should therefore not be interpreted as universal human agreement, and representational harms should be evaluated separately from PAS and ECS.

Materials and reproducibility. The benchmark uses materials derived from published, peer-reviewed studies for non-commercial research. We recruited no new human participants and collect no personally identifiable information. Released materials should follow applicable licenses and publisher restrictions, with sensitive or deceptive materials documented in context. Reconstructed protocols and study specifications undergo human review, and the statistical analyses are deterministic for fixed inputs. Model responses, however, can vary with sampling, model updates, and provider-side changes; results should therefore report the model version, agent configuration, prompts, sampling settings, and execution date.