Can AI Agents Simulate A/B Test Outcomes? A Validation Framework for Agentic Experimentation

Stefan Hut*
hutstefa@amazon.com

Lorenzo Masoero*
masoerl@amazon.com

Abstract

A/B testing remains the standard for rolling out new features in the technology industry. Each experiment, however, consumes real traffic, engineering effort, and weeks of wall-clock time. Can AI agents—conditioned on behavioral profiles and contextual descriptions of the intervention—simulate outcomes accurately enough to vet candidate treatments before committing live traffic? We formalize this question as a *Simulated Randomized Controlled Trial* (S-RCT) and derive a two-layer error decomposition that separates agent approximation error from subsampling error, enabling targeted improvements to each. The framework is agent-agnostic: any behavioral model—from a fine-tuned specialist to a general-purpose foundation model—can serve as the simulation engine. Validated on 67 historical marketing A/B tests, a baseline S-RCT using an off-the-shelf foundation model systematically overshoots effect magnitudes. A two-phase pre-period calibration protocol corrects much of this magnitude error, reducing the squared prediction error (after removing irreducible measurement noise) by $\sim$77$\times$; a within-subject design—where each agent is exposed to both arms—further reduces standard errors by $\sim$2.4$\times$. On naive directional metrics the simulator looks promising, agreeing with the historical sign for roughly 70% of tests (sign accuracy 0.70, sign overlap 0.70); but measured against simple uninformed baselines and the sign-replication ceiling of the A/B tests, this agreement is not clear evidence of directional signal on this noisy set. We discuss limitations of the current approach and identify applications where experimenters stand to benefit from agentic signals.

1 Introduction

Online controlled experimentation (A/B testing) is how technology services learn what works for customers, enabling safe, systematic iteration over product experiences at scale Kohavi et al. (2009; 2020). Yet each experiment requires engineering integration, weeks of wall-clock time to accumulate statistical power, and exposure of real customer traffic to the intervention under study. At the scale of a large organization running thousands of experiments per year, the aggregate cost is substantial: each experiment consumes engineering time, traffic allocation, and analysis overhead, while exposing customers to potentially underperforming treatments. Teams that iterate through dozens of candidate treatments before finding a winner bear this cost repeatedly. This affects applied scientists, product managers, and builders alike.

A sufficiently accurate behavioral simulator could fundamentally change this dynamic: letting teams vet dozens of candidates cheaply and reserve live traffic for the most promising few, so that customers benefit from improvements sooner. The advent of large foundation models offers a path toward such a simulator: conditioning a general-purpose model on a behavioral profile and eliciting decisions directly, without the domain-specific engineering bottleneck of hand-crafted agent-based models Epstein & Axtell (1996); Tesfatsion (2006). Importantly, the simulation engine is a pluggable module: while we instantiate it here with an off-the-shelf foundation model for rapid iteration, the framework accommodates any

*Equal contribution.

behavioral model, including fine-tuned specialists and non-language-model architectures. We expect domain-specific models to eventually outperform general-purpose ones for this task.

The idea of using language models as simulated economic agents dates to Horton (2023); recent work on agentic A/B testing Mansour et al. (2025); Castelo et al. (2026); Zhang et al. (2025) takes the next step, proposing to simulate user behavior with AI agents and estimate treatment effects from the simulated outcomes. Early validation found promising aggregate alignment with human baselines but systematic individual-level divergences Maier et al. (2025); Anthis et al. (2025); Liu et al. (2026); subsequent work showed that structured persona pipelines and rich contextual grounding narrow the gap Paglieri et al. (2026); Bougie et al. (2026); Meshi et al. (2026). The most recent wave deploys agents directly to simulate controlled experiments: AgentA/B Lu et al. (2025) sends interactive agents to browse two website variants; PAARS Mansour et al. (2025) conditions agents on structured personas; S-Researcher Wang et al. (2026) scales to 10^5 concurrent agents. None of these systems reports an explicit decomposition of estimation error into independently controllable components, making it difficult to diagnose or reduce error systematically. A complementary line of work uses AI predictions not to replace experiments but to reduce variance within them: Arbour et al. (2026) show that including AI-generated predictions as covariates in standard regression adjustment yields efficiency gains with a "do no harm" guarantee. Our work takes the more ambitious step of simulating entire experiments, but connects to this literature through the calibration protocol (Section 4.2).

We propose a framework for agentic A/B prediction, formalized as a Simulated Randomized Controlled Trial (S-RCT), and validate it against real historical outcomes. We develop a two-layer error decomposition that separates agent approximation error from subsampling error, and instantiate the framework with pre-period calibration and within-subject estimation. We validate on a benchmark of 67 historical marketing A/B tests. Our results surface both positive findings (calibration and within-subject estimation compress magnitude error substantially) and fundamental bottlenecks, including systematic agent over-responsiveness and the limits of persona completeness. We emphasize lessons learned and failure modes, which may be as informative for the field as the accuracy gains.

Contributions. (1) We formalize agentic A/B prediction as an S-RCT and derive a two-layer error decomposition separating approximation error from subsampling error (Section 2). (2) We validate the framework on 67 historical A/B tests, documenting accuracy gains and fundamental bottlenecks (Section 3). (3) We propose calibration and within-subject techniques that address each error layer independently (Section 4).

2 Method

2.1 Framework and formalism

Consider a universe of M possible users eligible for an experiment. An experiment-specific triggering mechanism determines which users are exposed to the intervention, defining a triggered set $\mathcal{P} \subset \{1, \ldots, M\}$ with $|\mathcal{P}| = N_{\text{exp}} \leq M$. Each triggered customer $j \in \mathcal{P}$ is randomly assigned to a treatment arm T or C and is characterized by two unit-level quantities:

- a *persona* $\Pi_j \in \mathcal{X}$—a vector of observable profile features whose value is independent of treatment assignment;
- an *outcome* $Y_j^t \in \mathbb{R}$, $t \in \{C, T\}$, for a metric of interest.

The quantity of interest is the average treatment effect (ATE), a population-level causal parameter that is a latent state of nature: it can be estimated from data but never directly observed (Neyman, 1990; Rubin, 1978; Imbens & Rubin, 2015). Formally,

$$\text{ATE} := \mathbb{E}_{\Pi \sim Q}\big[\mu^T(\Pi) - \mu^C(\Pi)\big], \tag{1}$$

Table 1: Example S-RCT inputs for a marketing creative A/B test measuring click-through rate on a product page.

Component	Concrete example
Persona Π_j	35-year-old frequent shopper, electronics enthusiast, high brand affinity
Context (control)	Current banner: "Free shipping on orders $25+"
Context (treatment)	New banner: "Members save 20% today"
Task	Does this user click the banner? (yes/no)
Outcome $\widetilde{Y}_j$	1 (click) or 0 (no click)

where $\mu^t(\Pi) := \mathbb{E}[Y^t \mid \Pi]$ is the conditional mean outcome under arm $t \in \{T, C\}$ (treatment or control) and Q is a distribution over personas. The choice of Q determines which population the ATE summarizes: setting Q to the empirical distribution of the triggered set $\mathcal{P}$ yields the finite-population ATE (the effect for the users who actually entered the experiment); treating $\mathcal{P}$ as a sample from a broader superpopulation yields the superpopulation ATE (the effect one would expect for future users) (Imbens & Rubin, 2015).

A *Simulated Randomized Controlled Trial* (S-RCT) replaces live traffic with computational surrogates: given a description of *who* the user is (persona Π_j), *what* the user experiences (context), and *what decision* the user must make (task), a simulator produces a simulated outcome $\widetilde{Y}_j$. Table 1 illustrates these inputs for a marketing creative test. The simulator is characterized by a parameter θ encoding an approximating response distribution $\widetilde{P}_\theta(\cdot \mid \Pi_j, t)$.

The S-RCT estimator in its simplest form is the difference in simulated means:

$$\widehat{\widetilde{\mathrm{ATE}}} \;=\; \bar{\widetilde{Y}}_T - \bar{\widetilde{Y}}_C \;=\; \frac{1}{|\mathcal{S}_T|}\sum_{j\in\mathcal{S}_T}\widetilde{Y}_j \;-\; \frac{1}{|\mathcal{S}_C|}\sum_{j\in\mathcal{S}_C}\widetilde{Y}_j, \tag{2}$$

computed over $N_{\mathrm{agt}} := |\mathcal{S}_T| + |\mathcal{S}_C|$ simulated *agents*—computational units that map (persona, context, task) triples to decisions. Our framework is engine-agnostic: the simulator could be a rule-based model, a learned behavioral surrogate, or a foundation model.

Persona completeness. S-RCTs introduce a representation bottleneck: the simulator receives only the persona Π_j as its window into who user j is. Recall that $\mathcal{X}$ is the space of all observable profile features available to the simulator. If the true response depends on factors not captured in $\mathcal{X}$ (e.g., the user's current intent, mood, or external context), even a perfect simulator incurs an irreducible gap.

Assumption 1 (Persona completeness) $Y_j^t \;\perp\!\!\!\perp\; Z_j \;\big|\; (\Pi_j, t)$ *for all* $Z_j \notin \mathcal{X}$.

Assumption 1 is strong, and likely violated in practice. Violations contribute to the approximation error formalized next.

Two-layer error decomposition. Define the simulator-population ATE as the treatment effect the simulator would produce if every triggered customer were evaluated:

$$\widetilde{\mathrm{ATE}}_\theta \;:=\; \mathbb{E}_{\widetilde{P}_\theta}\Big[\frac{1}{N_{\mathrm{exp}}}\sum_{j\in\mathcal{P}}\big(\widetilde{Y}_j^T - \widetilde{Y}_j^C\big)\Big]. \tag{3}$$

The S-RCT estimator's total error admits:

$$\widehat{\widetilde{\mathrm{ATE}}} - \mathrm{ATE} \;=\; \underbrace{\big(\widetilde{\mathrm{ATE}}_\theta - \mathrm{ATE}\big)}_{\text{approximation error}} \;+\; \underbrace{\big(\widehat{\widetilde{\mathrm{ATE}}} - \widetilde{\mathrm{ATE}}_\theta\big)}_{\text{subsampling error}}. \tag{4}$$

The approximation error is the gap between the simulator's population-level prediction and the true ATE—a function of the behavioral model, the persona rendering, and conditioning.

Table 2: Evaluation metrics. $\hat{\sigma}_i$: standard error of $\widehat{\text{ATE}}_i$; p_i, q_i: posterior probabilities of a positive effect; $\text{dec}(x) = -1$ if $x < 0.33$, 1 if $x > 0.66$, 0 otherwise.

Corrected MSE	Sign overlap	Launch alignment
$I^{-1}\sum_i[(\widehat{\widehat{\text{ATE}}}_i - \widehat{\text{ATE}}_i)^2 - \hat{\sigma}_i^2]$	$I^{-1}\sum_i(1 - \lvert p_i - q_i\rvert)$	$I^{-1}\sum_i \mathbf{1}\{\text{dec}(p_i) = \text{dec}(q_i)\}$

The subsampling error is the finite-sample gap between the realized N_{agt}-agent estimate and the simulator-population ATE. These two layers are independently addressable: calibration targets the first; principled agent selection targets the second (Section 4).

2.2 Instantiation

We use an internal simulation platform as the engine, with three design choices. First, the simulator is a general-purpose foundation model prompted with the customer's behavioral profile; we use this as a practical bootstrapping choice, though the engine-agnostic framework readily accommodates domain-specific models fine-tuned on behavioral data. Second, agents are constructed via one-to-one *agentic-twin* pairing: every agent is bound to a specific real customer who participated in the historical experiment (trigger-based sampling). This yields individual-level outcome pairs $\{(Y_j, \widetilde{Y}_j)\}$ supporting per-customer error analysis. Third, because the simulator is a stateless function, each agent can be queried under *both* arms—and repeatedly—yielding paired individual treatment effects $\widetilde{\Delta}_j = \widetilde{Y}_j^T - \widetilde{Y}_j^C$ with no analogue in real experiments. This full counterfactual access is exploited in Section 4.3.

3 Baseline Results

We apply the S-RCT framework to a benchmark of $I = 67$ historical marketing A/B test treatment pairs run on a large e-commerce service. Each experiment is a marketing creative test measuring click-through rate (CTR) on a high-traffic product surface. The benchmark includes a variety of creative treatments spanning text, imagery, and layout variations. Historical percent impacts span a range typical of marketing creative tests, with most effects small in magnitude. The benchmark contains a roughly even split of launch, harmful, and inconclusive historical decisions.

Each simulation uses $N_{\text{agt}} = 1{,}000$ agents in a within-subject design where every agent is exposed to both treatment and control, with each agent mapping one-to-one to a real customer who participated in the historical experiment. We use a general-purpose foundation model with no hyperparameter tuning; agents are chosen uniformly at random from the triggered population.

Results. Table 3 reports accuracy metrics across all 67 experiment-treatment pairs. Agentic impact estimates are systematically larger in magnitude than historical ATEs (MAE $= 0.0893$, several times the median historical percent impact), and launch alignment is 0.41—near the random floor of 0.33—driven by the magnitude gap pushing agentic posterior probabilities to extremes. On directional metrics the simulator reaches sign accuracy 0.70 and sign overlap 0.70.

These directional numbers should be read with care: if the goal is to quantify whether the forecast carries information about the true effect, sign accuracy and sign overlap can both be poor proxies, for at least two reasons.

First, the ground truth is noisy, so we bound the range each metric can occupy. For sign accuracy, a natural upper bound is the sign-replication rate—the probability that two independent repetitions of the same experiment agree in sign, which is 0.69 here and is an upper bound for any forecast no more informative than an independent rerun; from below, uninformed rules already reach 0.50 (a coin flip) and 0.57 (always predict positive). For sign overlap, uninformed forecasts bound it from below: a forecast drawing q_i at random scores

Table 3: Baseline accuracy (within-subject design, $N_{\text{agt}} = 1{,}000$, no calibration). Uninformed floors by metric: sign accuracy 0.50 (coin flip) to 0.57 (always-positive); sign overlap has no single floor (0.66–0.75, depending on the p_i distribution); launch alignment 0.33.

Metric	Value	SE
Corrected MSE	0.0222	0.0108
Sign overlap	0.70	0.03
Sign overlap (BC)	0.80	0.03
Launch alignment	0.41	0.06
MAE	0.0893	0.0178
Sign accuracy	0.70	0.06

0.66, and one that always answers 0.5 scores 0.75. Against these ranges, sign accuracy 0.70 sits near the top of a narrow band (0.57–0.69) and sign overlap 0.70 falls *below* the 0.75 a no-conviction forecast attains, so a high score carries limited information about forecast quality.

Second, sign metrics do not separate genuine predictive information from a systematic directional lean: our simulator tends to over-predict positive effects, which inflates agreement with the mostly-positive historical signs without implying test-level information.

In summary, the baseline simulator is not yet useful for launch decisions without calibration, and its directional agreement should be read against these baselines rather than as evidence of signal, motivating the improvements in Section 4.

4 Improvements

We address the two error layers in Equation (4) with three improvements: principled subsampling (Section 4.1), pre-period calibration (Section 4.2), and within-subject estimation (Section 4.3).

4.1 Principled subsampling

Simulations are cheaper than real experiments, but they are not free. A large service where many teams run A/B tests in parallel, each involving large user populations, faces a clear scaling constraint: simulating every triggered user is infeasible. Smart sampling of *which* users to simulate is therefore critical for this approach to scale.

The subsampling error in Equation (4) arises precisely because we simulate only $N_{\text{agt}} \ll N_{\text{exp}}$ agents. We can always decompose the population response as a mixture of heterogeneous subgroups: $P(Y^t) = \sum_k \pi_k P(Y^t \mid \text{stratum } k)$, where π_k is the population share of stratum k. This decomposition is useful beyond variance reduction. Understanding which subpopulation drives the KPI of interest is valuable in its own right. Teams lacking deep experimental expertise often struggle to identify subgroups where a treatment helps or harms. If the simulation is sufficiently accurate, it offers a cheap way to pre-screen an intervention across subpopulations, identifying potentially harmed groups before live deployment (Wang et al., 2025) and advancing heterogeneous treatment effect estimation at scale (Figure 1c).

For variance reduction specifically, classical survey sampling results (Imbens & Rubin, 2015) show that allocating agents proportionally to $\pi_k \sigma_k$ (where σ_k is the within-stratum standard deviation) minimizes the variance of the stratified estimator (Neyman allocation; Neyman, 1990):

$$\frac{\text{Var}_{\text{uniform}}}{\text{Var}_{\text{optimal}}} = 1 + \frac{\text{Var}_{\pi}(\sigma_k)}{(\mathbb{E}_{\pi}[\sigma_k])^2}. \tag{5}$$

The practical gain depends on how heterogeneous stratum variances are (Figure 1a). For binary outcomes with low base rates (as in our CTR benchmark), within-stratum standard deviations are compressed near $\sqrt{p}$, limiting the gain to a few percent. For continuous

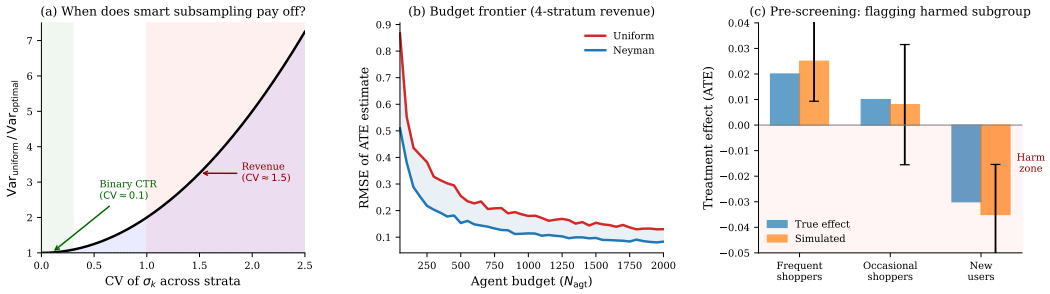

Figure 1: Principled subsampling: (a) efficiency gain from Neyman allocation grows with stratum variance heterogeneity; (b) for a stylized revenue metric, Neyman allocation achieves a given RMSE with fewer agents; (c) simulation pre-screens a treatment across subgroups, flagging the harmed segment before live deployment.

metrics such as revenue, where high-value segments can have 10–100× the variance of low-value ones, the gain reaches 2–5× (Figure 1b).

The key takeaway is that the ratio $\mathrm{Var}_\pi(\sigma_k)/(\mathbb{E}_\pi[\sigma_k])^2$ can be estimated from historical data before any simulation is run, giving practitioners a principled criterion for when smart subsampling is worth the added complexity.

We note a tension here: the estimator we employ (Equation (2), difference in simulated means) is model-free and makes no assumptions about the response surface. One could in principle do better by imposing more structure, for instance modeling the conditional response $\mu^t(\Pi)$ directly and using the simulation budget to refine uncertain regions. However, stronger modeling assumptions carry higher misspecification risk. The stratified difference-in-means estimator strikes a middle ground: it exploits population heterogeneity for efficiency while remaining valid regardless of the within-stratum response model.

4.2 Agent calibration

The magnitude gap documented in Section 3, where agentic ATEs systematically overshoot historical effects, is a manifestation of the approximation-error term in Equation (4). To confidently rely on agent-based simulation, the behavioral model must be calibrated against real activity patterns. A natural solution is to rely on a foundation model trained specifically for this task, whether via post-training (e.g., supervised fine-tuning on historical behavioral data) or by building dedicated behavioral models from scratch. Here we investigate a third, much cheaper (and admittedly less powerful) alternative: a lightweight two-phase calibration protocol, related in spirit to prediction-powered inference (Angelopoulos et al., 2023), that uses pre-exposure user data as a free supervision signal.

The protocol has two phases (Figure 2). In *Phase 1* (calibration), we run the simulator on the pre-period: the pre-period outcome Y_j^{pre} is held out of the feature set, and both arms render the control context (an A/A simulation with zero treatment effect by construction). The simulator produces $\tilde{Y}_j^{\text{pre}}$, and because the real Y_j^{pre} is known, we fit a calibration function:

$$\hat{f} = \arg\min_f \sum_j \ell\big(f(\tilde{Y}_j^{\text{pre}}),\ Y_j^{\text{pre}}\big), \tag{6}$$

where ℓ is a suitable loss (e.g., log-loss for Platt scaling). Figure 3a illustrates this step: the raw simulator systematically overestimates user activity relative to real pre-period outcomes, and Platt scaling learns a monotone correction.

In *Phase 2* (prediction), we run the simulator with full features and real treatment assignment. Raw simulated outcomes are passed through $\hat{f}$ to produce calibrated estimates. The calibration function $\hat{f}$ is fit independently per experiment on Phase-1 data and applied to Phase-2 data; the train/test split is temporal (pre-period vs. treatment period).

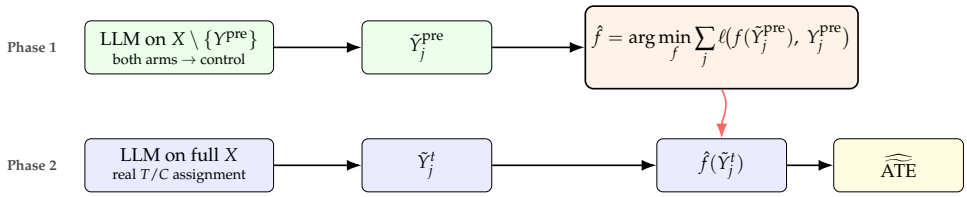

Figure 2: Two-phase calibration pipeline. In Phase 1 we fit the calibration function $\hat{f}$ on pre-period data. In Phase 2 we apply $\hat{f}$ to the raw outputs $\tilde{Y}_j^t$ to produce the calibrated estimate $\widehat{\text{ATE}}$.

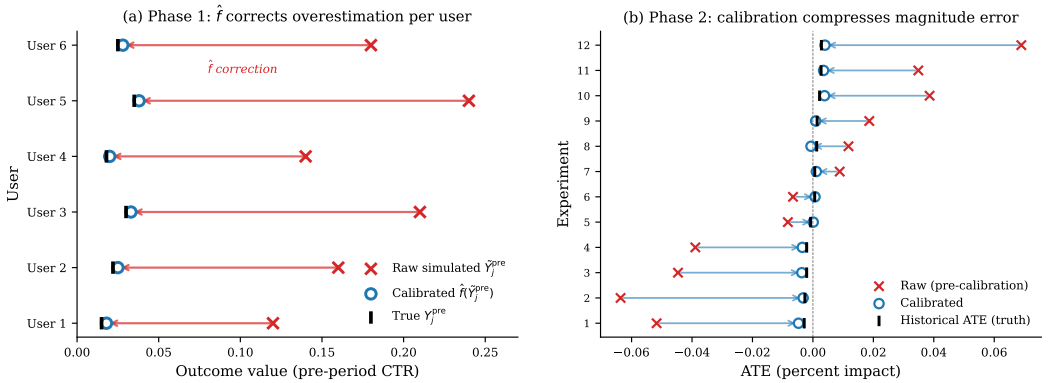

Figure 3: Two-phase calibration. (a) Phase 1: the raw simulator overestimates user activity; Platt scaling learns a correction from pre-period data. (b) Phase 2 result: per-experiment ATE estimates before (red crosses) and after (blue circles) calibration, relative to the historical ATE (black bars). Calibration compresses the magnitude gap by $\sim$77$\times$.

On 16 experiments from the benchmark using Platt scaling, calibration compresses the squared prediction error by $\sim$77$\times$, from systematic magnitude overshoot to estimates near the historical noise floor (Figure 3b).

4.3 Within-subject estimates

With agents split across arms, a between-subject design is susceptible to random composition imbalances. For subtle effects—where the true ATE is small—the estimated difference can be dominated by chance differences in who was assigned to each arm rather than by the treatment itself. We address this with a within-subject design where every agent is exposed to *both* treatment and control. This eliminates between-agent composition differences entirely: each agent serves as its own control, and the estimator becomes the within-subject paired difference. The paired structure removes the persona-level variance component while targeting the same population ATE. To address ordering effects, we randomize the order in which treatment and control are presented to each agent.

The within-subject design's primary advantage is variance reduction: standard errors shrink by $\sim$2.4$\times$ on average. Sign accuracy rises from 65% (between-subject) to 0.70 (within-subject), consistent with the elimination of random arm-composition imbalance. Launch alignment improves from 0.33 to 0.41, though the magnitude gap continues to push posterior probabilities toward extremes. Combining within-subject estimates with calibration (Section 4.2) is the natural next step.

5 Discussion and Conclusion

This paper presents a framework for agentic A/B prediction formalized as Simulated Randomized Controlled Trials (S-RCTs) with a two-layer error decomposition separating behavioral approximation error from subsampling error. Validation on 67 historical marketing A/B tests reveals directional agreement that stays within the range of uninformed baselines on this noisy set, and a systematic magnitude gap where agentic estimates overshoot historical effects. Within-subject estimation improves sign accuracy; calibration addresses magnitude error; principled subsampling targets the remaining variance.

Behavioral implications. The systematic over-responsiveness of agent behavior—predicting larger effects than observed in real customers—suggests that current general-purpose foundation models exhibit a form of *behavioral amplification*: when conditioned on a persona and asked to choose, they respond more decisively to treatment differences than real customers do. Understanding and correcting this behavioral gap is, we believe, a core challenge for the agent behavior community: it reflects not just miscalibration but a fundamental mismatch between how these models represent decision-making and how humans actually behave in low-stakes product contexts. Bridging this gap may require behavioral fine-tuning on revealed preference data, structured constraints on agent response distributions, or hybrid architectures that combine language model reasoning with learned behavioral priors.

Applications beyond ATE estimation. Estimating the average treatment effect is arguably the hardest target for agentic simulation: it requires both correct direction and accurate magnitude. However, the framework produces several secondary outputs that are useful even when magnitude accuracy is unreliable. *Directional screening* (flagging likely losers before committing live traffic) requires only correct sign rather than accurate magnitude, a lower bar than ATE estimation. On the present benchmark set, however, sign accuracy and sign overlap are only partially informative: as discussed above, benchmark noise limits how much they reveal about whether the simulator's directional predictions carry real signal. Progress on this application will therefore require both larger, less noisy benchmarks and better evaluation metrics—itself a direction for future work. *Reasoning traces* (qualitative explanations of why a simulated agent chose one option over another) offer experimenters a novel interpretive lens with no analogue in classical A/B testing. *Flipper analysis* (identifying which agent personas flip their decision between treatment and control) surfaces treatment-effect heterogeneity and can guide segment-level analysis in real experiments. *Agentic priors for early decisions*: even a noisy directional signal, when combined with early real-experiment data, can accelerate go/no-go decisions by serving as an informative prior in a Bayesian framework. These applications lower the accuracy bar required for practical value and represent the nearest-term path from the current system to practical deployment.

Limitations. Near-term limitations include magnitude overshoot even with correct sign; a single-domain benchmark (marketing CTR); potential correlation in agent responses from a shared model that could understate variance; and retrospective-only evaluation against historical targets. The benchmark evaluates a prediction against a known historical outcome—in a prospective setting, the triggering population is unknown and no ground truth is available for calibration. Predicting the triggering population is the main missing piece for deployment; it is best understood as a thin upstream layer that can be developed independently of the behavioral simulator.

Broader considerations. If the framework is used to pre-screen which experiments to run, treatments benefiting customer segments poorly captured by the behavioral model—low-frequency users, underrepresented demographics—may be filtered out before reaching real measurement. The two-phase calibration protocol partially mitigates this by surfacing per-segment bias on the pre-period, but practitioners must remain attentive to whose behavior the simulator represents faithfully and whose it does not. The goal is not to replace real experiments but to make the experimentation pipeline faster and more informed, so that customers benefit from product improvements sooner.

Acknowledgments

We thank Ryan Kessler and Hammaad Adam for useful discussions that significantly helped the development of this paper.

References

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Prediction-powered inference. *Science*, 382(6671):669–674, 2023. URL https://doi.org/10.1126/science.adi6000.

Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. LLM social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025. URL https://arxiv.org/abs/2504.02234.

David Arbour, Eli Ben-Michael, Avi Feller, Apoorva Lal, and Lo-Hua Yuan. AI-assisted variance reduction in randomized experiments. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2026. URL https://arxiv.org/abs/2606.08853.

Nicolas Bougie, Gian Maria Marconi, Xiaotong Ye, and Narimasa Watanabe. Beyond offline A/B testing: Context-aware agent simulation for recommender system evaluation. *arXiv preprint arXiv:2604.09549*, 2026. URL https://arxiv.org/abs/2604.09549.

Alberto Castelo, Zahra Zanjani Foumani, Ailin Fan, Keat Yang Koay, Vibhor Malik, Yuanzheng Zhu, Han Li, Meysam Feghhi, Ronie Uliana, Shuang Xie, et al. SimGym: Traffic-grounded browser agents for offline A/B testing in e-commerce. *arXiv:2602.01443*, 2026. URL https://arxiv.org/abs/2602.01443.

Joshua M. Epstein and Robert Axtell. *Growing Artificial Societies: Social Science from the Bottom Up*. MIT Press, 1996.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023. URL https://www.nber.org/papers/w31122.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. URL https://doi.org/10.1017/CBO9781139025751.

Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: Survey and practical guide. *Data Mining and Knowledge Discovery*, 18:140–181, 2009. URL https://doi.org/10.1007/s10618-008-0114-1.

Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing*. Cambridge University Press, 2020. URL https://doi.org/10.1017/9781108653985.

Xuan Liu, Haoyang Shang, Zizhang Liu, Xinyan Liu, Yunze Xiao, Yiwen Tu, and Haojian Jin. HumanStudy-Bench: Towards AI agent design for participant simulation. *arXiv preprint arXiv:2602.00685*, 2026. URL https://arxiv.org/abs/2602.00685.

Yuxuan Lu, Ting-Yao Hsu, Hansu Gu, Limeng Cui, Yaochen Xie, William Headden, Bingsheng Yao, Akash Veeragouni, Jiapeng Liu, Sreyashi Nag, Jessie Wang, and Dakuo Wang. AgentA/B: Automated and scalable web A/B testing with interactive LLM agents. *arXiv preprint arXiv:2504.09723*, 2025. URL https://arxiv.org/abs/2504.09723.

Benjamin F. Maier, Ulf Aslak, Luca Fiaschi, Nina Rismal, Kemble Fletcher, Christian C. Luhmann, Robbie Dow, Kli Pappas, and Thomas V. Wiecki. LLMs reproduce human purchase intent via semantic similarity elicitation of Likert ratings. *arXiv preprint arXiv:2510.08338*, 2025. URL https://arxiv.org/abs/2510.08338.

Saab Mansour, Leonardo Perelli, Lorenzo Mainetti, George Davidson, and Stefano D'Amato. PAARS: Persona aligned agentic retail shoppers. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 143–159, 2025.

Ofer Meshi, Krisztian Balog, Sally Goldman, Avi Caciularu, Guy Tennenholtz, Jihwan Jeong, Amir Globerson, and Craig Boutilier. ConvApparel: A benchmark dataset and validation framework for user simulators in conversational recommenders. *arXiv preprint arXiv:2602.16938*, 2026. URL https://arxiv.org/abs/2602.16938.

Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990. URL https://doi.org/10.1214/ss/1177012031. Translated and edited by D. M. Dabrowska and T. P. Speed from the 1923 Polish original.

Davide Paglieri, Logan Cross, William A. Cunningham, Joel Z. Leibo, and Alexander Sasha Vezhnevets. Persona generators: Generating diverse synthetic personas for arbitrary contexts. *arXiv preprint arXiv:2602.03545*, 2026. URL https://arxiv.org/abs/2602.03545.

Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978. URL https://doi.org/10.1214/aos/1176344064.

Leigh Tesfatsion. Agent-based computational economics: A constructive approach to economic theory. In *Handbook of Computational Economics*, volume 2, pp. 831–880. Elsevier, 2006.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025. URL https://www.nature.com/articles/s42256-025-00986-z.

Lei Wang, Yuanzi Li, Jinchao Wu, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. LLM agents as social scientists: A human-AI collaborative platform for social science automation. *arXiv preprint arXiv:2604.01520*, 2026. URL https://arxiv.org/abs/2604.01520.

Yimeng Zhang, Tian Wang, Jiri Gesi, Ziyi Wang, Yuxuan Lu, Jiacheng Lin, Sinong Zhan, Vianne Gao, Ruochen Jiao, Junze Liu, et al. Shop-r1: Rewarding LLMs to simulate human behavior in online shopping via reinforcement learning. *arXiv:2507.17842*, 2025.

A Evaluation Metrics: Detailed Definitions

Corrected MSE. The noise-corrected MSE removes the irreducible historical sampling variance from the squared error:

$$\text{CMSE} = \frac{1}{I}\sum_{i=1}^{I}\left[(\widehat{\widehat{\text{ATE}}}_i - \widehat{\text{ATE}}_i)^2 - \hat{\sigma}_i^2\right],$$

where $\hat{\sigma}_i$ is the standard error of the historical ATE estimate for experiment i. Without this correction, MSE would be dominated by the historical noise floor rather than the simulator's error.

Sign overlap. Sign overlap measures the agreement between historical and agentic posterior probabilities of a positive effect:

$$\text{SO} = \frac{1}{I}\sum_{i=1}^{I}(1 - |p_i - q_i|),$$

where p_i and q_i are the posterior probabilities of a positive effect under the historical and agentic estimates, respectively. A value of 1.0 indicates perfect calibration of directional confidence. There is no single random floor: for an uninformative agent drawing $q_i \sim$

Uniform$[0, 1]$, the expected per-experiment score is $\frac{1}{2} + p_i(1 - p_i)$, which equals 0.50 only as $p_i \to 0$ or 1 and rises to 0.75 at $p_i = 0.5$. Averaged over our p_i this floor is 0.66, and a constant $q_i = 0.5$ agent scores 0.75; because our posteriors cluster near 0.5, these uninformed baselines sit close to the achievable range, so a high sign overlap is not on its own evidence of directional quality.

Launch alignment. Launch alignment discretizes the posterior into three decisions (harmful: $p < 0.33$; inconclusive: $0.33 \leq p \leq 0.66$; launch: $p > 0.66$) and measures agreement:

$$\text{LA} = \frac{1}{I} \sum_{i=1}^{I} \mathbf{1}\{\text{dec}(p_i) = \text{dec}(q_i)\}.$$

The random floor is 0.33 (three equally likely categories).

B Benchmark Details

The benchmark comprises 67 treatment-control pairs from marketing creative tests on a high-traffic e-commerce product surface. All experiments measure click-through rate (CTR) as the primary metric. Historical percent impacts are small, consistent with the subtle nature of marketing creative changes. The distribution of historical decisions is approximately balanced across harmful, inconclusive, and launch outcomes. Each experiment draws from a large triggered population, representing a broad cross-section of customer segments.