Analysis

Predicting Who Will Buy Is Not Knowing Who Was Persuaded: Marketing Incrementality in the LLM Era Must Be Anchored by Experiments

HX Research

HX Research · Industry Modelling Practice · November 5, 2026

Every stage of marketing decisions now involves large language models: analytics copilots write queries and reports, synthetic respondents answer surveys, and generated answers are becoming a new exposure channel. The old problem of marketing measurement remains: the effect of advertising and promotion is causal, while company data mostly record correlations. Large field experiments repeatedly show that observational methods overstate advertising effects, and response models that target those most likely to buy can spend the budget on people who would have bought anyway. We argue that in marketing decisions LLMs can accelerate hypotheses, data and priors but cannot replace identification: credible incrementality must be anchored by randomized experiments and then extended by uplift models, experiment-calibrated marketing mix models and prediction-powered inference. We formalize uplift, targeting value, experimental power, marketing mix models, endogeneity bias, experiment calibration with a bias budget and prediction-powered inference, and contribute a five-layer evidence framework, a five-route comparison, an experiment-anchored measurement algorithm, three seeded illustrative simulations and five falsifiable forecasts.

Keywords: incrementality measurement; uplift modelling; marketing mix models; geo experiments; experiment calibration; synthetic respondents; prediction-powered inference

Introduction

The central question of a marketing budget has always been the same: how many purchases did this money bring about that would not otherwise have happened? It is a causal question whose answer is the difference between the treated outcome and a counterfactual that is never observed[1,2]. Economics and marketing have long been clear about this: correlations between advertising and sales in observational data mix in seasonality, demand shocks, targeted delivery and selection, and must be identified through experiments or credible quasi-experimental designs[3,4].

Large language models give this old problem several new forms. First, marketing analytics copilots: work has combined semantic search, prompt engineering and fine-tuning to improve LLMs' domain question answering, SQL generation and tabular analysis for marketing mix modelling and attribution[5]; a benchmark for enterprise data agents shows that the best agent achieves only 57% pass@1 on realistic queries spanning several heterogeneous databases (per the paper)[6]. Second, synthetic respondents: language models conditioned on demographics can reproduce the attitude distributions of particular groups[7]; per the paper, on 57 personal-care product surveys with 9,300 human responses, semantic similarity rating reaches 90% of human test–retest reliability[8]. Third, new objects of measurement: brand exposure in generated answers is becoming a new channel, and a causal framework has been proposed that brings generative engine optimization and generative engine marketing into marketing mix models[9]; neural marketing mix models that use Transformers over search-query and creative embeddings have also appeared[10].

All of this is valuable, but none of it changes the logic of identification. The evidence from large field experiments is consistent. Lewis and Rao analysed 25 large online advertising field experiments and found that individual sales are far more volatile than advertising effects, so that even with very large samples the confidence intervals of return on investment are often too wide to tell profit from loss (per the paper)[11]. Gordon and colleagues used large advertising experiments at Facebook to test common observational methods and found that they often fail to recover accurately the treatment effects obtained from randomized experiments[12]; in a follow-up study covering 663 experiments and more than 5,000 features, median experimental lifts for three conversion outcomes were 29%, 18% and 5%, whereas double machine learning gave 83%, 58% and 24%, and stratified propensity-score matching gave 173%, 176% and 64% (per the paper)[13]. Large-scale search-advertising experiments at eBay showed that brand-keyword ads have no measurable short-term benefit; non-brand keyword ads positively influence new and infrequent users, but most of the advertising spend goes to frequent users whose purchasing is not influenced by ads, making average returns negative (per the paper)[14].

Against this background, current discussion leaves three gaps. First, it conflates "predicting who will buy" with

"knowing who was persuaded". Response models predict the probability of purchase after treatment; uplift models estimate the change the treatment brings about. Targeting high-risk or high-response customers may select precisely those who are unaffected or even put off[15,16]. Second, the relationship between marketing mix models and experiments has not been settled. Leading open-source tools already support calibrating priors with experiment results[17–19], but when the observational model is systematically biased there is still no operational rule for how much weight it and the experiment should each receive. Third, LLM-generated evidence is treated as a substitute for samples rather than as an aid. Studies show that synthetic respondents can reproduce answer distributions without covering real combinations of respondents[20], that small differences in analytic choices can change conclusions[21], and that combining synthetic data with a small amount of human data under statistical rules saves samples while preserving validity[22,23].

Our central thesis is that in marketing decisions LLMs can accelerate hypotheses, data and priors but cannot replace identification; credible incrementality must be anchored by randomized experiments and then extended by uplift models, experiment-calibrated marketing mix models and prediction-powered inference. In short: experiments anchor, models extend, LLMs accelerate. We make three contributions:

1. A unified formalization of uplift and targeting value, experimental power, marketing mix models, endogeneity bias, experiment calibration with a bias budget and prediction-powered inference (equations (1)–(7)), with a notation table, showing why response ranking wastes budget and why sampling-variance weights for an observational model fail in the presence of bias.
2. A five-layer evidence framework organized by source of identification, and a mechanism-level comparison of five measurement routes with evidence for and against.
3. An experiment-anchored marketing measurement algorithm (Algorithm 1), three seeded illustrative simulations and five forecasts, each with a time scale, falsification condition and implications for enterprises.

Section 2 gives the formalization and notation; Section 3 the framework; Section 4 the mechanism analysis and route comparison; Section 5 the algorithm; Section 6 the simulations; Sections 7–10 cover forecasts and research agenda, common misconceptions, limitations and conclusions.

Problem formalization and notation

Uplift and targeting value

Let customer i have features X_i, treatment (a coupon, an ad, a promotion) $T_i \in \{0, 1\}$, outcome Y_i and potential outcomes $Y_i(1)$ and $Y_i(0)$[1]. The conditional average treatment effect, or uplift, is

$$\tau(x) = E[Y(1) - Y(0) \mid X = x] = \mu_1(x) - \mu_0(x) \quad (1)$$

where $\mu_t(x) = E[Y(t) \mid X = x]$. A response model estimates $\mu_1(x)$, "will they buy after treatment"; an uplift model estimates $\tau(x)$, "how much does treatment change"[24]. If each incremental purchase earns gross margin m and each treatment costs c, the value of treating group S is

$$V(S) = \sum_{i \in S} [m \cdot \tau(X_i) - c] \quad (2)$$

Under a budget $|S| = k$, V is maximized by ranking on τ and taking the top k; ranking on $\mu_1 = \mu_0 + \tau$ costs nothing only when μ_0 and τ rank customers the same way. When $\mathrm{Corr}(\mu_0, \tau) < 0$ — when those already likely to buy are least affected — response ranking systematically selects "people who would have bought anyway". The literature's division of customers into persuadables, sure things, lost causes and "sleeping dogs" (who buy less when treated) rests on exactly this distinction[16,24].

Experimental power

Uplift is causal, and its most reliable estimates come from randomized experiments. But marketing effects are usually small relative to the variability of individual outcomes. With n per arm, outcome standard deviation σ and average effect $\bar{\tau}$ to be detected, at two-sided 5% significance and 80% power

$$n \approx 2(z_{0.975} + z_{0.8})^2 \cdot \sigma^2/\bar{\tau}^2 \approx 15.7 \cdot \sigma^2/\bar{\tau}^2 \quad (3)$$

When $\sigma/\bar{\tau}$ is large, the required sample grows with its square. Lewis and Rao's finding is the empirical version of equation (3): in the experiments they analysed, the coefficient of variation of individual sales was about 10 while the advertising effect was a small fraction of the mean, so even samples of more than ten million person-weeks struggled to estimate return on investment precisely (per the paper)[11]. Variance reduction, ghost ads and geo experiments are all designs for raising efficiency under the constraint of equation (3)[25–27].

Marketing mix models and endogeneity

Marketing mix models estimate channel contributions from weekly or daily aggregate data. A common form is

$$y_t = b_0 + \sum_k \beta_k \cdot h_k(A_{k,t}) + \gamma^T z_t + \varepsilon_t,\ A_{k,t} = x_{k,t} + \delta_k \cdot A_{k,t-1},\ h(A) = A^s/(A^s + \kappa^s) \quad (4)$$

where $x_{k,t}$ is spend in channel k, A is the carryover (adstock) with decay δ, h is a saturation function and z_t are controls such as seasonality and price[28]. The problem is that spend is often set on the basis of expected demand: more in peak season, more before promotions, more for channels that have been performing. If an unobserved demand shock u_t affects both spend and sales, then

$$\mathrm{plim}\ \hat{\beta_k} = \beta_k + \mathrm{Cov}(A_k, u)/\mathrm{Var}(A_k) \quad (5)$$

This bias does not disappear as the sample grows. A recent synthetic benchmark deliberately generates spend from four mechanisms — quarterly budget feedback, anticipatory

spending ahead of promotions, scheduled TV bursts and algorithmic performance chasing — to test whether models can recover true effects under endogenous spend[29]. In Bayesian marketing mix models, priors have a large impact on posteriors when samples are small and may lead to biased estimates (per the paper)[28].

Experiment calibration with a bias budget

Let the observational model give a channel-effect estimate $\hat{\theta}_{obs}$ with sampling variance v_{obs} and systematic bias of magnitude B, and let an experiment give an unbiased estimate $\hat{\theta}_{exp}$ with variance v_{exp}. Weighting by sampling variance alone means that when v_{obs} is small the combination follows the observational model almost entirely and keeps its bias intact. We propose treating the squared bias as additional variance:

$$\hat{\theta} = w\cdot\hat{\theta}_{obs} + (1 - w)\cdot\hat{\theta}_{exp},\ w = (v_{obs} + \hat{B}^2)^{-1} / [(v_{obs} + \hat{B}^2)^{-1} + v_{exp}^{-1}],\ \hat{B}^2 = max(\bar{g}^2 - v_{exp},\ 0) \quad (6)$$

where $\bar{g}^2$ is the mean squared historical gap "observational estimate minus experimental estimate" for comparable channels. The larger $\hat{B}^2$, the more weight shifts to the experiment; if the observational model has historically agreed with experiments, $\hat{B}^2$ is near zero and equation (6) reduces to ordinary variance weighting. The Meridian documentation defines calibration as "using experiment results and other domain knowledge to set channel-specific priors"[17]; structural estimation recovers decay, saturation and effectiveness parameters directly from geo-experimental time series[30]. Equation (6) answers a different question: how much weight experimental information should carry when bias is present.

Synthetic respondents and prediction-powered inference

Let the population quantity of interest be $\theta = E[Y]$ (for example, the mean purchase intent for a new product concept). We have n human samples (X_j, Y_j) and N synthetic samples with only LLM predictions $f(\tilde{X}_i)$. The prediction-powered inference (PPI) estimator is

$$\hat{\theta}_{PP} = (1/N)\sum_i f(\tilde{X}_i) - (1/n)\sum_j [f(X_j) - Y_j],\ Var(\hat{\theta}_{PP}) \approx \sigma_f^2/N + \sigma_r^2/n \quad (7)$$

The second term uses the human sample to estimate and subtract the model's bias, and σ_r^2 is the variance of "prediction minus truth". As long as predictions correlate with the truth, σ_r^2 is smaller than Var(Y), so the interval is narrower than with humans alone, and it remains valid however biased the model is[22]. A stratified version tightens intervals further[31]; in conjoint analysis, combining LLM-generated with human data under statistical rules saves 24.9% to 79.8% of data and cost per the paper, whereas direct substitution saves nothing because of bias[23].

Notation

Table 1 summarizes the notation.

Table 1 | Main notation

Symbol	Meaning	Where used
X, T, Y, Y(t)	features, treatment, outcome, potential outcome	Eq. (1)
$\tau(x)$, $\mu_t(x)$	uplift (conditional average treatment effect), expected outcome under treatment t	Eq. (1)
m, c, k, V(S)	margin per incremental purchase, cost per treatment, budgeted headcount, targeting value	Eq. (2)
n, σ, $\bar{\tau}$	sample size per arm, outcome standard deviation, average effect	Eq. (3)
$x_{k,t}$, $A_{k,t}$, δ_k, h, κ, s	spend, carryover, decay rate, saturation function and its parameters	Eq. (4)
u_t, β_k	unobserved demand shock, channel effect	Eq. (5)
$\hat{\theta}_{obs}$, $\hat{\theta}_{exp}$, v, B, w	observational and experimental estimates, variances, bias budget, combination weight	Eq. (6)
f, N, n, σ_r^2	LLM prediction, synthetic and human sample sizes, variance of the rectifier	Eq. (7)

Framework: five layers of evidence by source of identification

In marketing measurement, the same question — "how much incrementality did this spend bring?" — can be answered in five ways whose credibility comes from different sources of identification. We arrange them in five layers (Table 2). The lower the layer, the closer to causal identification; the higher the layer, the better the coverage and speed, but the more it needs anchoring from below. LLMs can contribute at every layer, but the nature of the contribution differs.

Table 2 | Five layers of evidence on marketing incrementality: source of identification, questions answered, typical failures and the role of LLMs

Layer	Representative methods	Source of identification	Question answered	Typical failure	Appropriate role for LLMs
E1 Randomized experiments	user-level lift tests, ghost ads, geo experiments, control-	random assignment	what is the average incrementality of this spend	low power, signal loss, short-term only	design checks, sample-size calculation, report

	group promotions				generation
E2 Experimental extrapolation	uplift models, predicted incrementality (PIE), policy learning	random samples from E1	whom to treat; incrementality of campaigns without experiments	noisy evaluation metrics, extrapolation out of range	feature construction, explanation, code generation
E3 Calibrated marketing mix models	Bayesian MMM with experiment priors, structural estimation from geo experiments	E1 as prior or constraint	how to allocate channel budgets	endogenous spend, collinearity, prior dominance	data preparation, diagnostics, scenario Q&A
E4 Observational and quasi-experimental	difference-in-differences, synthetic control, Bayesian structural time series, double machine learning	assumptions such as parallel trends or unconfoundedness	approximate incrementality when randomization is impossible	assumptions fail and are hard to test	assumption checklists, robustness checks
E5 LLM-generated evidence	synthetic respondents, digital twins, agentic analysis	no independent source of identification	hypothesis screening, pre-testing surveys and concepts	bias, variance collapse, poor coverage of combinations	generate candidates, rectified with human samples

Table 2 is read in two ways. First, moving up requires anchoring below: E2 uplift models must be trained and evaluated on E1 random samples; E3 mix models must take priors or constraints from E1 results; E5 synthetic evidence must be rectified with human samples by equation (7). Second, the role of LLMs is to accelerate, not to identify: they can make experimental design, data preparation, code and reporting faster, but they cannot generate counterfactuals out of nothing.

Mechanism analysis and route comparison

Those who will buy and those who are persuaded

The meaning of equation (2) can be illustrated by a counter-intuitive fact: the people most likely to buy are often those on whom spending is least valuable. Drawing on retention field experiments, Ascarza found that targeting the customers at highest risk of churn may be ineffective; the customers to target are those most responsive to the intervention, who are not necessarily the riskiest (per the paper)[15]. In a field experiment with more than 53,000 college students, Athey and colleagues compared uplift targeting based on causal forests with targeting based on baseline probabilities, noting that "low baseline outcomes first", though common, is not the ideal criterion, and proposed hybrid approaches combining the two[32]. A customer-retention case study in the financial industry likewise found that uplift models deliver higher profits than predictive models[16].

Industrial data give a sense of scale. On the Criteo uplift dataset of 13.98 million records, the top 20% of customers ranked by predicted uplift captured 77.7% of incremental conversions, 3.9 times random targeting; uncertainty analysis with causal forests showed that only 1.9% of customers were confident persuadables at 95% and 0.1% confident "sleeping dogs" (per the paper)[33]. Uplift estimation has a mature family of methods: meta-learners (T-, S- and X-learners)[34], the quasi-oracle R-learner[35], causal forests and generalized random forests[36,37], recursive partitioning[4] and double machine learning[38].

Evaluating uplift models is harder than estimating them

Uplift is unobservable and can only be compared by group on random samples. Qini curves and cumulative-gain curves are the usual metrics, but their variance is high and their signals are to a degree arbitrary; statistical adjustment of outcomes can reduce the variance[39]. Rank-weighted average treatment effects (RATE) bring the Qini coefficient into a unified family of metrics with a central limit theorem, allowing rigorous inference when comparing prioritization rules[40]. UpliftBench (2026) evaluates 12 estimators across seven dataset families and finds that disagreement among benchmarks about "which estimator is best" is largely about metrics rather than models: across 100 IHDP realizations, the mean rank correlation between Qini and effect accuracy is only +0.07, while the area under the cumulative-gain curve aligns better; in settings that require a sign-threshold decision, calibrating the decision threshold removes 81% of Qini-selection regret (per the paper)[41]. When the experimental budget is limited, observational data can serve as priors for actively choosing the most informative units for randomization[42].

Why observational measurement systematically overstates

The evidence in Section 1 shows that observational failures are not accidental. There are three mechanisms. First, selection: ad systems deliver ads to the people most likely to convert, so exposed and unexposed people differ to begin with[12,13]. Second, attribution is not incrementality: in large paid-acquisition systems, new users attributed to paid channels may overlap with organic demand and brand traffic, overstating true incrementality; after an attribution-correction framework anchored by incrementality experiments was deployed across several TikTok markets, the measured cannibalization rate subsequently fell by about 15 percentage points (per the paper)[43]. Third, cross-platform diversion: an upstream content platform causes a purchase but the transaction is booked on a downstream marketplace; attribution-based return on ad spend cannot see these

diversions, and marketing mix models do not know which channel to credit[44].

Experiments are also adapting to new constraints. Ghost ads record, in the control group, users who "would have seen the ad", making experiments cheaper and more precise; per the paper, in one retail retargeting campaign the ads lifted website visits by 17.2% and purchases by 10.5%[26]. Platform lift studies need dedicated power and sample-size calculations[45], and covariate adjustment using pre-experiment data can reduce variance substantially[25,46]. Lower match rates, aggregation thresholds and random noise introduced for privacy weaken experimental signals: one study proves that there is a decision frontier inside which reports contain too little information for any method to distinguish uniformly between "incremental" and "not incremental" (per the paper)[47].

Between experiments and observation, predicted incrementality by experimentation (PIE) offers a middle route: learn a mapping from campaign features to causal effects on a set of randomized experiments, then apply it to campaigns run without experiments. Per the paper, on 2,226 Meta ad experiments PIE achieves an out-of-sample R^2 of 0.88 for incremental conversions per dollar, against 0.19 for industry-standard 7-day last-click attribution; PIE disagrees with experiment-based decisions in 8–12% of campaigns, against 12–20% for last click[48]. This is a model case of "experiments anchor, models extend": the model's credibility comes from having been trained and tested on experiments.

Marketing mix models need experimental anchors

Marketing mix models are the main tool for allocating channel budgets; their value lies in covering every channel, using only aggregate data and not depending on individual-level tracking[18,19]. Their weakness is equation (5): spend is endogenous. Google researchers noted as early as 2017 that Bayesian mix models with carryover and shape effects are strongly influenced by priors in small samples (per the paper)[28]; the synthetic benchmark for endogenous spend uses 156 weeks of weekly data on three channels and records the true causal decomposition of every week's sales to test whether models recover the truth[29].

Leading tools have therefore turned to experiment calibration. Meridian offers a geo-experiment module and uses experiment results to set channel priors[17]; Robyn uses multi-objective optimization to minimize the error between model predictions and "ground truth" such as geo experiments and platform lift tests[18]. Geo experiments themselves have mature methods: Vaver and Koehler described randomized experiments with regions as units[27], Trimmed Match gives robust estimates when regions are few and heterogeneous[49], and its design counterpart integrates optimal pairing and sample splitting[50]. Structural estimation removes confounding by differencing treatment and control regions and recovers decay, saturation and effectiveness parameters together from geo-experimental time series[30]. Mix models are also extending to finer granularity: deriving campaign-level effects with mix-model priors[51], improving stability with hierarchical structure and sign constraints[52], handling search-query and creative embeddings with neural networks[10], and treating brand exposure in generated answers as a new channel[9]. These extensions add expressiveness but no new source of identification; the more expressive the model, the more it needs experimental constraints.

When randomization is impossible, quasi-experimental methods are the next best option: synthetic control builds the counterfactual from weighted control units[53], and Bayesian structural time series infers how the impact of an intervention evolves over time[54]. The assumptions they rely on — parallel trends, control units unaffected by the intervention — must be checked one by one.

Where LLMs fit in measurement

What synthetic respondents can do. Early work showed that language models conditioned on demographics can reproduce the attitude distributions of particular groups[7]. Leading marketing journals then examined the potential of LLMs to substitute for or augment human participants in perceptual analysis and preference measurement[55,56]. Working with a Fortune 500 food company, Arora and colleagues used GPT-4 to replicate a 2019 qualitative and quantitative study and found that a human–LLM hybrid outperformed humans or the model alone, with the model correctly picking the direction and valence of answers in quantitative questions (per the paper)[57]. Using 19,447 personas to replicate 133 experimental findings from 14 papers in the *Journal of Marketing* reproduced 76% of main effects (84 of 111) and 68% of findings overall including interactions (per the paper)[58].

What synthetic respondents cannot do. Asking models directly for numerical ratings produces distributions with unrealistically low variance — "mode collapse"; eliciting text first and mapping it to the scale with embeddings partly repairs this[8,59]. An October psychometric audit tested 14 decision models and 47 generative models: the best generative model reached a psychometric similarity score of 0.813, level with a Gaussian copula fitted only on other respondents (0.824) and below one fitted in sample (0.927) (per the paper)[60]. Another October study proposed the respondent coverage gap: samples generated by all six models left human respondents uncovered and fell behind a baseline that samples each answer independently from human marginals (per the paper)[20]. Across 252 synthetic-sample configurations, recovery of rankings, distributions and between-scale correlations varied substantially, and across 66 alternative configurations of a published study, human–silicon correlations ranged from 0.23 to 0.84 (per the paper)[21]. Gains in individual-level accuracy do not necessarily transfer to the population level[61]; conditioning on

respondents' earlier behavioural answers predicts their later individual choices better than demographic descriptions alone[62]; and a public dataset of 2,058 respondents, 500 questions and four waves with retests provides a benchmark[63].

Why synthetic experiments are confounded. Using demand estimation as the setting and a real experiment with 40 products as the benchmark, Gui and Toubia show that when simulated subjects are blind to the experimental design, variation in treatment systematically shifts variables that should be held constant, violating unconfoundedness; revealing the design to the model mitigates the problem[64]. Synthetic experiments, too, need identification.

LLMs' ability to do causal analysis. On a pure causal-inference task of more than 200,000 samples, 17 models performed close to random[65]; a benchmark of statistical pitfalls such as Simpson's paradox and selection bias shows significant limitations of current models in statistical causal inference, with code-assisted explicit analysis more reliable[66]. The performance of data agents depends heavily on how the execution harness represents tasks, manages state and provides evaluation feedback[67].

The right use of synthetic evidence. Prediction-powered inference gives a unified answer: treat LLM output as a biased prediction, and use a small human sample to estimate and subtract the bias[22,31]; in conjoint analysis this combination saves data and cost compared with direct substitution[23]; and when historical tasks exchangeable with the current one can be found, synthetic data can even carry provable validity guarantees[68].

Counter-evidence and boundaries

"Anchoring by experiments" has limits too. First, experiments are costly and their power limited; equation (3) means that a single experiment by many small and medium brands cannot give a precise return on investment[11]. Second, experiments measure short-term effects in specific settings; long-term brand effects and intertemporal substitution are hard to randomize. Third, privacy-driven signal loss can leave some experimental results in an undecidable zone[47]. Fourth, when variation in spend is itself close to exogenous (for example rule-based adjustments by region or time), quasi-experimental methods can also provide useful evidence[53,54], and not every decision is worth a dedicated experiment. Our claim is therefore "anchor with experiments and let a bias budget set the weights", not "experiment on everything".

Table 3 | Mechanism-level comparison of five routes to measuring incrementality (figures self-reported by the papers)

Route	Main mechanism	Representative evidence	Main risk	Conditions of use
R1 Response and propensity models	predict post-treatment purchase probability μ_1	targeting high-risk customers for retention may be ineffective	selects sure things, wastes budget	prediction only, not targeting decisions
R2 Uplift models on randomized experiments	estimate $\tau(x)$, target by uplift	Criteo top 20% capture 77.7% of incrementality	noisy evaluation metrics, extrapolation out of range	random samples available, repeatable experiments
R3 Uncalibrated observational MMM and attribution	regression on aggregates, rule-based attribution	observational median lifts far above experimental ones	endogenous spend, cannibalization, cross-platform diversion	description only, not budgets on its own
R4 Experiment-calibrated MMM and PIE	experiments set priors; learn features-to-effect mapping	PIE out-of-sample R^2 0.88 vs 0.19 for last click	extrapolation risk when experimental coverage is thin	ongoing experimentation programme
R5 LLM synthetic evidence plus PPI	synthetic samples rectified by humans	conjoint savings of 24.9–79.8%	bias, variance collapse, sensitivity to configuration	concept screening, survey pre-testing

The figures in Table 3 come from different data and settings[13,15,23,33,48] and cannot be added across rows. What the table conveys is that, of the five routes, only R2 and R4 are directly anchored by experiments, R5 is valid when rectified with human samples, and R1 and R3 used on their own are the most likely to misdirect budgets.

Method: experiment-anchored marketing measurement

We now condense the analysis into an executable procedure. The goal of Algorithm 1 is, for a set of marketing decisions (targeting, channel budgets, promotions, new-product concepts), to organize measurement by evidence layer so that every incrementality figure reported can be traced to an experimental anchor, and to keep LLMs working only where they accelerate.

Algorithm 1 | Experiment-Anchored Marketing Measurement (EAMM)

Input: decision list D (targeting, budgets, promotions, concept tests); historical spend and sales data; available experimental budget; paired historical records of observational and experimental estimates; optional LLM synthetic samples

Output: incrementality estimate, interval and evidence layer for each decision; channel priors and bias budget $\hat{B}^2$; targeting lists; a measurement ledger

1: Stratify: label each decision in D with the evidence layer it requires (Table 2); decisions on budgets and targeting need at least E1 or E2/E4 evidence, and E5 is used only for screening candidates

2: Design experiments: compute sample sizes by equation (3), preferring control groups, geo experiments or ghost ads; reduce variance with pre-experiment covariates; when samples are insufficient, pool several experiments or actively choose experimental units

3: Target: estimate $\tau(x)$ on random samples (meta-learners or causal forests), evaluate with RATE or cumulative-gain curves with intervals; rank by $m \cdot \hat{\tau} - c$ and treat only groups with positive expected value

4: Calibrate: estimate $\hat{B}^2$ from paired historical records and combine mix-model and experimental estimates by equation (6); raise experimental frequency for channels with large $\hat{B}^2$

5: Extrapolate: for campaigns without experiments, extrapolate with predicted-incrementality models trained on experiments, and flag anything outside experimental coverage

6: Synthetic evidence: LLM-generated respondents or digital twins serve only as predictions f, rectified by equation (7) with a human sample of at least several hundred; report intervals rather than point estimates

7: Agentic execution: LLMs generate queries, code and draft reports; calculations are done by deterministic code; identification assumptions and conclusions are reviewed by people; every step is written to the measurement ledger so that it can be reviewed and rolled back

Complexity and scope: Experimental cost is set by equation (3); uplift estimation is near-linear in sample size; equations (6) and (7) are closed-form. Applies to companies with some scale of spend that can run control groups or regional randomization; with very small experimental budgets, $\hat{B}^2$ in step 4 can rely only on external benchmarks, which must be noted in the ledger

Algorithm 1 embodies three design intentions. First, steps 3 and 5 base both targeting and extrapolation on random samples, avoiding the "sure things" trap of Section 4.1 and echoing predicted incrementality[32,48]. Second, step 4 estimates the bias budget from historical gaps so that the weight on experiments adjusts automatically to the credibility of the observational model, complementing current practice of "setting priors with experiments"[17,30]. Third, steps 6 and 7 confine LLMs to acceleration: synthetic evidence must be rectified with humans, and agents' conclusions must pass identification checks[22,66].

Illustrative simulations

All numbers below are illustrative and show how equations (1)–(7) interact. All data are synthetic; the script and random seed (20261105) are in src/figs.py. They do not represent measurements of any real brand, campaign or customer.

Response targeting versus uplift targeting. We generate 200,000 customers whose baseline purchase probability is negatively correlated with uplift (correlation −0.642): persuadables cluster at intermediate baselines, while those already likely to buy have uplift near zero or negative. The budget allows treating 20% of customers; each incremental purchase earns a margin of 30 and each treatment costs 2. Random targeting yields 62.2 incremental purchases per 1,000 treated; targeting by an oracle response model (knowing μ_1 exactly) yields only 8.8, with 29% of those selected being "sure things" whose baseline probability exceeds one half; "lowest baseline first" yields 50.9; a T-learner uplift model trained on a random sample of 40,000 yields 97.1, and oracle uplift ranking 122.3 (Fig. 1). In profit per 1,000 treated, response targeting gives −1,736 and the uplift model 914. An accurate response model, in other words, can be the worst targeting rule.

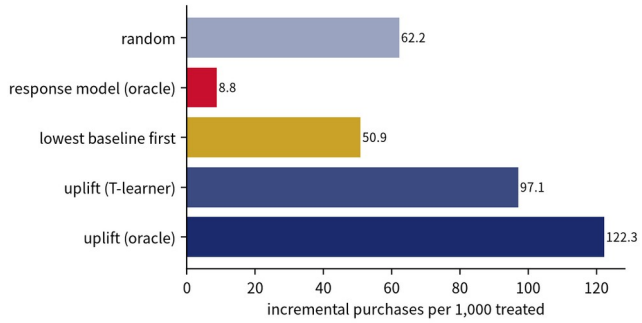


Fig. 1 | Predicting who will buy is not knowing who was persuaded. Incremental purchases per 1,000 treated under five targeting rules with the same budget; the oracle response model delivers the least. Synthetic data, fixed random seed 20261105, computed by the HX Research Team.

Endogenous spend and calibration with a bias budget. We simulate 156 weeks of single-channel data with true return on investment 1.2 and decay 0.5, with spend raised in response to demand shocks the model cannot see. Over 2,000 repetitions, the observational mix model overstates by 1.40 on average with root-mean-square error 1.40; combining the experiment (standard error 0.25) as a prior weighted by sampling variance still leaves a bias of 1.24, because the observational model's sampling variance is small and almost all the weight falls on it; estimating the bias budget from the past 5 "mix model minus experiment" gaps by equation (6) cuts the bias to 0.047 and the RMSE to 0.241, comparable with the experiment alone (0.245) (Fig. 2). As a control, when spend is exogenous the bias-budget method has RMSE 0.238, slightly better than the experiment alone, showing that it automatically uses observational information when the observational model is credible.

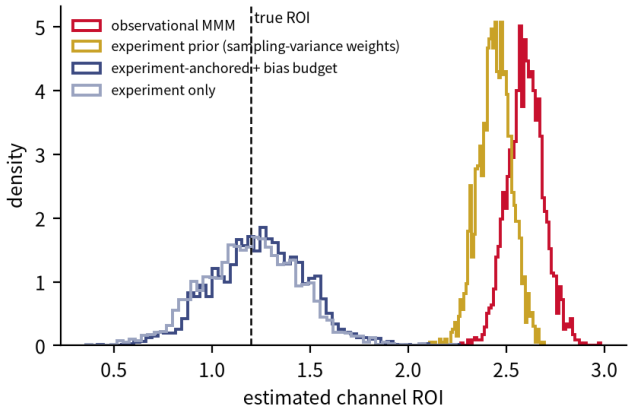


Fig. 2 | With endogenous spend, an experiment prior weighted by sampling variance cannot undo the bias; calibration with a bias budget can. Distributions of four estimators over 2,000 repetitions; the dashed line marks the true return on investment. Synthetic data, fixed random seed 20261105, computed by the HX Research Team.

Synthetic respondents and prediction-powered inference. Let the true mean purchase intent for a new concept on a five-point scale be 3.095; LLM synthetic respondents correlate 0.729 with humans but overstate the mean by 0.414, with a compressed distribution. Over 2,000 repetitions: with 200 humans alone, the actual coverage of the 95% interval is 94.9% with mean width 0.295; with 10,000 synthetic respondents alone, the interval is only 0.019 wide but its coverage is 0% — very precisely wrong; rectified by 200 humans through equation (7), coverage is 95.2% with width 0.218, about a quarter narrower than humans alone (Fig. 3).

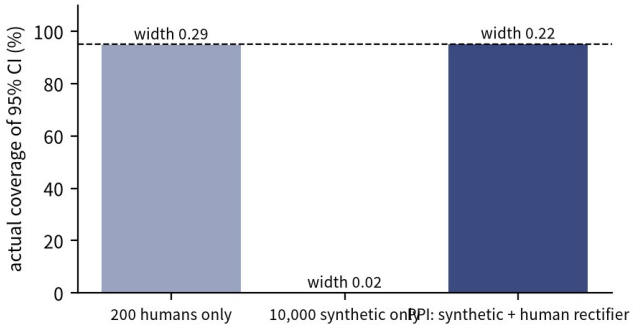


Fig. 3 | Synthetic respondents alone are "precisely wrong"; combined with a few humans they are both valid and narrower. Actual coverage and mean width of 95% intervals under three approaches. Synthetic data, fixed random seed 20261105, computed by the HX Research Team.

Table 4 | Summary of illustrative simulations (fixed random seed 20261105)

Simulation	Key setting	Result	Equations
Targeting rules	200,000 customers, 20% budget, baseline–uplift correlation −0.642	incremental per 1,000: random 62.2, response 8.8, uplift model 97.1, oracle uplift 122.3	Eqs. (1)–(2)
MMM calibration	156 weeks, true ROI 1.2, endogenous spend, 2,000 repetitions	bias: observational 1.40, variance-weighted 1.24, bias budget 0.047	Eqs. (4)–(6)
Synthetic respondents	10,000 synthetic, 200 humans, bias +0.414	coverage: humans 94.9%, synthetic 0%, PPI 95.2%; widths 0.295 / 0.019 / 0.218	Eq. (7)

Table 4 is not a claim about any real system. The mechanism it conveys is that accurate prediction is not correct decision-making and a precise model is not a valid conclusion; numbers become credible only when anchored by experiments or human samples.

Forecasts and research agenda

The following are views of the HX Research Team, not statements of fact.

Table 5 | Five forecasts (opinion)

No.	Forecast	Time scale	Main basis	Falsification condition	Implication for enterprises
1	Mainstream open-source marketing mix tools will offer diagnostics of "model estimate versus historical experiments", and bias-budget-style weighting of experiments will become common practice	12–24 months	experiment calibration is mainstream[17,18]; structural estimation[30]; endogeneity benchmark[29]	by end-2028, none of Meridian, Robyn and PyMC-Marketing offers systematic diagnostics of model-versus-experiment gaps	start keeping paired records of every experiment and the contemporaneous model estimate now
2	Synthetic respondents in marketing research will move from "replacement samples" to "synthetic plus human rectification", with	12–24 months	PPI[22]; conjoint augmentation[23]; coverage gap[20]; psychometric audit[60]	by end-2028, most empirical studies using synthetic respondents in major marketing journals still report purely synthetic estimates	keep a human anchor sample of several hundred in concept tests

	leading-journal studies generally reporting rectified intervals				
3	Major ad platforms and retail media networks will make "experiment-calibrated incrementality" a standard report alongside attribution	6–18 months	PIE[48]; cannibalization correction[43]; cross-platform diversion[44]	by end-2027, the default reports of major platforms still contain only attributed conversions and attributed returns	ask for incrementality experiments or their calibration when buying media
4	Main evaluations of marketing and data-analysis agents will add causal-pitfall and identification-check dimensions, not just answer or SQL accuracy	12–24 months	data-agent benchmark 57%[6]; causal pitfalls[66]; execution harness[67]	by end-2028, major data-agent benchmarks still contain no causal-identification tasks	let agents write code, have deterministic code compute and people review identification assumptions
5	Brand exposure in generated answers will enter marketing mix models as a new channel, but for two years its credible estimates will still rely mainly on experiments	12–24 months	generative marketing mix models[9]; neural mix models[10]; privacy signal loss[47]	within two years, a method for estimating generative-exposure effects without experiments, with errors within ±20% under independent experimental tests, is widely adopted	run small regional or query-level randomizations for generative channels first

The research agenda has four open questions. First, the bias budget B itself varies with channel, season and spending strategy: how can hierarchical models share information across channels and give a reliable $\hat{B}^2$ when experiments are scarce? Second, aligning uplift evaluation metrics with business objectives: when ranking metrics suffice and when thresholded policy risk is required[40,41]. Third, the conditions for validity of synthetic respondents: which historical tasks are "exchangeable" with the current one, so that synthetic evidence carries guarantees[68]. Fourth, experimental design under privacy constraints: after aggregation and noise injection, how to choose reporting granularity so that experiments remain in the decidable zone[47].

Key points:

1. Use equations (1)–(2) for targeting: rank by uplift, not response; response ranking selects sure things.
2. Use equation (3) for experiments: when effects are small and volatility large, sample size grows with the square of their ratio.
3. Use equations (4)–(5) for mix models: with endogenous spend, bias does not vanish with sample size.
4. Use equation (6) for calibration: treat historical bias as extra variance, and weight shifts automatically to experiments.
5. Use equation (7) for synthetic evidence: synthetic samples rectified by humans are both valid and narrower.

Where common industry views may be wrong

"A highly accurate response model means good targeting." Response models predict whether people buy after treatment; uplift models estimate how much treatment changes. An accurate response model can concentrate the budget on sure things[15,16].

"Conversions in attribution reports are incrementality." Conversions attributed to paid channels may overlap with organic demand or be booked by downstream platforms[43,44]; observational median lifts can be several times experimental ones[13].

"More data make the mix model more accurate." Bias from endogenous spend does not vanish with sample size; priors and experiment calibration decide whether the model can be read causally[28,29].

"LLMs can replace surveys." Synthetic respondents reproduce some distributions but suffer variance collapse, poor coverage of combinations and sensitivity to configuration[20,21,59]; the right use is to combine them with human samples[22].

"Let the LLM answer the causal question directly." On pure causal-inference and statistical-pitfall benchmarks, model performance is limited[65,66]; it is more reliable to have the model write analysis code, let deterministic computation produce results and have people check identification assumptions.

Limitations

This article has four limitations. First, the simulations in Section 6 use simplified synthetic data; they illustrate mechanisms but cannot give numbers for any real brand, and the effect of the bias budget depends on the number and representativeness of historical gap records. Second, much key evidence comes from studies of particular platforms and markets (Facebook, Meta, eBay, TikTok, Criteo), and caution is needed in generalizing to other industries and channels[13,14,33,43,48]. Third, research on synthetic respondents

moves fast, and several 2026 findings come from preprints, as self-reported by the papers[20,60]. Fourth, experiments carry ethical and compliance constraints: a control group means some customers do not receive an offer, and personalized targeting involves automated decision-making. Under Article 24 of China's Personal Information Protection Law, automated decision-making using personal information must ensure transparency and fair, just results, and information push and commercial marketing through automated decision-making must also offer an option not targeting personal characteristics or a convenient way to refuse; Article 55 requires a personal-information protection impact assessment for automated decision-making using personal information[69].

Conclusion

LLMs make marketing analysis faster but do not make causal identification easier. Field experiments repeatedly show that observational methods overstate advertising effects; uplift research shows that those most likely to buy are often least worth treating; and research on synthetic respondents shows that models can reproduce distributions precisely yet cannot replace the combinations and biases of real people. These three lines of evidence point to one conclusion: marketing incrementality must be anchored by experiments and then extended by uplift models, calibrated mix models and prediction-powered inference; the place of LLMs is in acceleration, not identification.

For enterprises the implications are concrete. Target by uplift rather than response; when allocating budgets, keep paired records of experimental and model estimates and let a bias budget set the weight of experiments; in concept tests, combine synthetic respondents with a human anchor sample; let agents write code, run workflows and draft reports, but leave identification assumptions and final conclusions to people. According to the company website, the "response and uplift models" in HX's FMCG solution separate persuadable customers from those who would buy regardless and deliver targeted lists by segment with expected incremental sales, and its "promotion and price-elasticity analysis" combines control-group and difference-in-differences analysis with price-elasticity estimates, all computed by deterministic code[70]. The Harness governance engine of the Enterprise Agent Platform holds that "the model only proposes, tools execute under control, and Harness holds the final say"[71], which can implement the human review, measurement ledger and rollback of Algorithm 1.

Data availability

No real data from any company or customer were used in this article. Figures quoted come from the listed references and the company website; the illustrative simulations were generated by script under the assumptions stated in the text.

Code availability

The script for the illustrative simulations and Figs. 1–3 is src/figs.py (Python, fixed random seed 20261105), provided with this article's package for verification.

Competing interests

The authors are employees of Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. The FMCG solution and the Enterprise Agent Platform mentioned in this article are publicly offered products and solutions of the company.

References

1. Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701 (1974); https://doi.org/10.1037/h0037350.
2. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge Univ. Press, 2015); https://doi.org/10.1017/CBO9781139025751.
3. Varian, H. R. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences* **113**, 7310–7315 (2016); https://doi.org/10.1073/pnas.1510479113.
4. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**, 7353–7360 (2016); https://doi.org/10.1073/pnas.1510489113.
5. Gao, Y., Arava, S. K., Li, Y. & Snyder, J. W. Improving the Capabilities of Large Language Model Based Marketing Analytics Copilots With Semantic Search And Fine-Tuning. Preprint at https://arxiv.org/abs/2404.13077 (2024).
6. Ma, R. et al. Can AI Agents Answer Your Data Questions? A Benchmark for Data Agents. Preprint at https://arxiv.org/abs/2603.20576 (2026).
7. Argyle, L. P. et al. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* **31**, 337–351 (2023); https://doi.org/10.1017/pan.2023.2.
8. Maier, B. F. et al. LLMs Reproduce Human Purchase Intent via Semantic Similarity Elicitation of Likert Ratings. Preprint at https://arxiv.org/abs/2510.08338 (2025).
9. Kato, M., Honma, D. & Kato, T. Generative Marketing Mix Modeling: A Causal Inference Framework Linking GEO and GEM to Business Impact. Preprint at https://arxiv.org/abs/2609.11915 (2026).
10. Mulc, T. et al. NNN: Next-Generation Neural Networks for Marketing Measurement. Preprint at https://arxiv.org/abs/2504.06212 (2025).
11. Lewis, R. A. & Rao, J. M. The Unfavorable Economics of Measuring the Returns to Advertising. *The Quarterly Journal of Economics* **130**, 1941–1973 (2015); https://doi.org/10.1093/qje/qjv023.
12. Gordon, B. R., Zettelmeyer, F., Bhargava, N. & Chapsky, D. A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook. *Marketing Science* **38**, 193–225 (2019); https://doi.org/10.1287/mksc.2018.1135.
13. Gordon, B. R., Moakler, R. & Zettelmeyer, F. Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement. *Marketing Science* **42**, 768–793 (2023); https://doi.org/10.1287/mksc.2022.1413.
14. Blake, T., Nosko, C. & Tadelis, S. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica* **83**, 155–174 (2015); https://doi.org/10.3982/ECTA12423.
15. Ascarza, E. Retention Futility: Targeting High-Risk Customers Might be Ineffective. *Journal of Marketing Research* **55**, 80–98 (2018); https://doi.org/10.1509/jmr.16.0163.
16. Devriendt, F., Berrevoets, J. & Verbeke, W. Why you should stop predicting customer churn and start using uplift models. *Information Sciences* **548**, 497–515 (2021); https://doi.org/10.1016/j.ins.2019.12.075.
17. Google. Calibrate treatment priors | Meridian. https://developers.google.com/meridian/docs/advanced-modeling/roi-priors-and-calibration (2026).
18. Meta Marketing Science. Robyn: an AI/ML-powered open-source marketing mix modeling package. https://facebookexperimental.github.io/Robyn/ (2026).

19. PyMC Labs. PyMC-Marketing: open source Bayesian marketing mix modeling & CLV in Python. https://www.pymc-marketing.io/en/stable/ (2026).
20. Kumagae, Y. & Kaneko, M. Silicon Samples Match the Answers but Not the Respondents. Preprint at https://arxiv.org/abs/2610.30900 (2026).
21. Cummins, J. The threat of analytic flexibility in using large language models to simulate human data. Preprint at https://arxiv.org/abs/2509.13397 (2025).
22. Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I. & Zrnic, T. Prediction-powered inference. *Science* **382**, 669–674 (2023); https://doi.org/10.1126/science.adi6000.
23. Wang, M., Zhang, D. J. & Zhang, H. Large Language Models for Market Research: A Data-augmentation Approach. Preprint at https://arxiv.org/abs/2412.19363 (2024).
24. Gutierrez, P. & Gérardy, J.-Y. Causal inference and uplift modelling: a review of the literature. In *Proc. 3rd International Conference on Predictive Applications and APIs* 1–13 (PMLR 67, 2017); https://proceedings.mlr.press/v67/gutierrez17a.html.
25. Deng, A., Xu, Y., Kohavi, R. & Walker, T. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. *Proceedings of the sixth ACM international conference on Web search and data mining*, 123–132 (2013); https://doi.org/10.1145/2433396.2433413.
26. Johnson, G. A., Lewis, R. A. & Nubbemeyer, E. I. Ghost Ads: Improving the Economics of Measuring Online Ad Effectiveness. *Journal of Marketing Research* **54**, 867–884 (2017); https://doi.org/10.1509/jmr.15.0297.
27. Vaver, J. & Koehler, J. Measuring ad effectiveness using geo experiments. Google Research, https://research.google/pubs/measuring-ad-effectiveness-using-geo-experiments/ (2011).
28. Jin, Y., Wang, Y., Sun, Y., Chan, D. & Koehler, J. Bayesian methods for media mix modeling with carryover and shape effects. Google Research, https://research.google/pubs/bayesian-methods-for-media-mix-modeling-with-carryover-and-shape-effects/ (2017).
29. Heusch, N. A Synthetic Benchmark Dataset with Endogenous Marketing Spend for Validating Marketing Mix Models. Preprint at https://arxiv.org/abs/2608.21130 (2026).
30. Heusch, N. Structural Estimation of Marketing Mix Model Parameters from Geo-Experiments. Preprint at https://arxiv.org/abs/2608.21128 (2026).
31. Fisch, A. et al. Stratified Prediction-Powered Inference for Hybrid Language Model Evaluation. Preprint at https://arxiv.org/abs/2406.04291 (2024).
32. Athey, S., Keleher, N. & Spiess, J. Machine Learning Who to Nudge: Causal vs Predictive Targeting in a Field Experiment on Student Financial Aid Renewal. Preprint at https://arxiv.org/abs/2310.08672 (2023).
33. Singh, A. A Large-Scale Empirical Comparison of Meta-Learners and Causal Forests for Heterogeneous Treatment Effect Estimation in Marketing Uplift Modeling. Preprint at https://arxiv.org/abs/2604.06123 (2026).
34. Künzel, S. R., Sekhon, J. S., Bickel, P. J. & Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* **116**, 4156–4165 (2019); https://doi.org/10.1073/pnas.1804597116.
35. Nie, X. & Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108**, 299–319 (2021); https://doi.org/10.1093/biomet/asaa076.
36. Wager, S. & Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* **113**, 1228–1242 (2018); https://doi.org/10.1080/01621459.2017.1319839.
37. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *The Annals of Statistics* **47**, 1148–1178 (2019); https://doi.org/10.1214/18-AOS1709.
38. Chernozhukov, V. et al. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**, C1–C68 (2018); https://doi.org/10.1111/ectj.12097.
39. Bokelmann, B. & Lessmann, S. Improving uplift model evaluation on RCT data. Preprint at https://arxiv.org/abs/2210.02152 (2022).
40. Yadlowsky, S., Fleming, S., Shah, N., Brunskill, E. & Wager, S. Evaluating Treatment Prioritization Rules via Rank-Weighted Average Treatment Effects. Preprint at https://arxiv.org/abs/2111.07966 (2021).
41. Li, B. UpliftBench: Revealing Outcome-Regime and Objective Mismatch in Uplift Evaluation. Preprint at https://arxiv.org/abs/2608.00915 (2026).
42. Gao, J. et al. Budgeted Active Experimentation for Treatment Effect Estimation from Observational and Randomized Data. Preprint at https://arxiv.org/abs/2602.22021 (2026).
43. Li, D., Yuan, B., Yang, Z., Chen, Q. & Song, L. Attributed, But Not Incremental: Cannibalization-Corrected Attribution for Large-Scale Advertising. Preprint at https://arxiv.org/abs/2606.26690 (2026).
44. Konitzer, T. Media Measurement and the Assisted Own Goal: Attribution, Marketing-Mix Models, and Individual-Level Incrementality. Preprint at https://arxiv.org/abs/2607.09608 (2026).
45. Liu, C. H. B., Bettaney, E. M. & Chamberlain, B. P. Designing Experiments to Measure Incrementality on Facebook. Preprint at https://arxiv.org/abs/1806.02588 (2018).
46. Kohavi, R., Tang, D. & Xu, Y. *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge Univ. Press, 2020); https://doi.org/10.1017/9781108653985.
47. Shekhar, P. & Howard, C. Privacy-Robust Incrementality Measurement for Advertising Systems under Signal Loss. Preprint at https://arxiv.org/abs/2606.03878 (2026).
48. Gordon, B. R., Moakler, R. & Zettelmeyer, F. Predicted Incrementality by Experimentation (PIE) for Ad Measurement. Preprint at https://arxiv.org/abs/2304.06828 (2023).
49. Chen, A. & Au, T. C. Robust causal inference for incremental return on ad spend with randomized paired geo experiments. *The Annals of Applied Statistics* **16**, 1–20 (2022); https://doi.org/10.1214/21-AOAS1493.
50. Chen, A., Longfils, M. & Remy, N. Trimmed Match Design for Randomized Paired Geo Experiments. Preprint at https://arxiv.org/abs/2105.07060 (2021).
51. Bhat, M. R. et al. Integrated Marketing Attribution: A Bayesian Framework for Privacy-Safe Granular Measurement Anchored in MMM. Preprint at https://arxiv.org/abs/2606.16878 (2026).
52. Chen, H., Zhang, M., Han, L. & Lim, A. Hierarchical Marketing Mix Models with Sign Constraints. Preprint at https://arxiv.org/abs/2008.12802 (2020).
53. Abadie, A., Diamond, A. & Hainmueller, J. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* **105**, 493–505 (2010); https://doi.org/10.1198/jasa.2009.ap08746.
54. Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N. & Scott, S. L. Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics* **9**, 247–274 (2015); https://doi.org/10.1214/14-AOAS788.
55. Li, P., Castelo, N., Katona, Z. & Sarvary, M. Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Marketing Science* **43**, 254–266 (2024); https://doi.org/10.1287/mksc.2023.0454.
56. Goli, A. & Singh, A. Frontiers: Can Large Language Models Capture Human Preferences? *Marketing Science* **43**, 709–722 (2024); https://doi.org/10.1287/mksc.2023.0306.
57. Arora, N., Chakraborty, I. & Nishimura, Y. AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators. *Journal of Marketing* **89**, 43–70 (2025); https://doi.org/10.1177/00222429241276529.
58. Yeykelis, L., Pichai, K., Cummings, J. J. & Reeves, B. Using Large Language Models to Create AI Personas for Replication, Generalization and Prediction of Media Effects: An Empirical Test of 133 Published Experimental Research Findings. Preprint at https://arxiv.org/abs/2408.16073 (2024).
59. Heath, O. & Alexander, R. Correcting Mode Collapse in Silicon Sampling with Semantic Similarity Rating. Preprint at https://arxiv.org/abs/2607.28550 (2026).
60. Lukauskas, M. & Šarkauskaitė, V. Probable but Not Valid: A Psychometric Audit of Decision Models as Synthetic Survey Respondents. Preprint at https://arxiv.org/abs/2610.14912 (2026).
61. Wang, P., Wang, L., Li, Y. & Chen, X. Understanding Reliability in LLM-based Human Behavior Simulation. Preprint at https://arxiv.org/abs/2609.25066 (2026).
62. Pourtaheri, K. & Zareei, A. Behavioral History Outperforms Descriptions of the Person for LLM Synthetic Personas. Preprint at https://arxiv.org/abs/2610.03998 (2026).

63. Toubia, O. et al. Twin-2K-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. Preprint at https://arxiv.org/abs/2505.17479 (2025).
64. Gui, G. & Toubia, O. The Challenge of Using LLMs to Simulate Human Behavior: A Causal Inference Perspective. Preprint at https://arxiv.org/abs/2312.15524 (2023).
65. Jin, Z. et al. Can Large Language Models Infer Causation from Correlation? In *Proc. International Conference on Learning Representations* (2024); https://arxiv.org/abs/2306.05836.
66. Du, J. et al. Ice Cream Doesn't Cause Drowning: Benchmarking LLMs Against Statistical Pitfalls in Causal Inference. Preprint at https://arxiv.org/abs/2505.13770 (2025).
67. Liu, F. & Liu, H. DS-Lighting: Making Agent Harnesses Explicit for Data-Science Automation. Preprint at https://arxiv.org/abs/2608.28590 (2026).
68. Tan, L. & Zrnic, T. Valid Inference with Synthetic Data via Task Exchangeability. Preprint at https://arxiv.org/abs/2606.13629 (2026).
69. 全国人民代表大会常务委员会. 中华人民共和国个人信息保护法（Personal Information Protection Law of the People's Republic of China）. 国家互联网信息办公室网站转载, https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm (2021).
70. 弘犀智能（Great Unicorn Intelligence）. 快消与营销解决方案（FMCG & marketing solutions）. https://www.hxallai.com/solutions/fmcg/ (2026).
71. 弘犀智能（Great Unicorn Intelligence）. 企业智能体平台 · Harness（Enterprise Agent Platform · Harness）. https://www.hxallai.com/platform/agent/ (2026).

Copyright Notice: © 2026 Great Unicorn Intelligence Technology (Shenzhen) Co., Ltd. All rights reserved. Any reposting, quotation in commentary, or adaptation must credit the authors (Great Unicorn Intelligence Research Team) and cite the source (Great Unicorn Intelligence, https://www.hxallai.com).