EFFICIENT RANDOMIZED EXPERIMENTS USING FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized experiments are the preferred approach for evaluating the effects of interventions, but they are costly and often yield estimates with substantial uncertainty. On the other hand, in silico experiments leveraging foundation models offer a cost-effective alternative that can potentially attain higher statistical precision. However, the benefits of in silico experiments come with a significant risk: statistical inferences are not valid if the model predictions fail to accurately reflect experimental responses to interventions. In this paper, we propose a novel approach that integrates the predictions from multiple foundation models with experimental data while preserving valid statistical inference. Our estimator is consistent and asymptotically normal, with asymptotic variance no larger than the *standard* estimator based on experimental data alone. Importantly, these statistical properties hold even when model predictions are arbitrarily biased.

1 INTRODUCTION

Randomized experiments are widely considered the preferred approach for evaluating the effects of interventions in scientific research. However, obtaining sufficiently large sample sizes can be costly and time-consuming, especially when studying rare outcomes. As a result, there is growing interest in exploring in silico experiments as a potential alternative to randomized experiments. These digital experiments leverage the predictions from foundation models (Bommasani et al., 2021)—machine learning models trained on massive datasets and applicable to many downstream tasks—to simulate the outcome of hypothetical randomized experiments. However, the benefits of in silico experiments come with a significant risk: statistical inferences from such experiments are not valid if model predictions fail to reflect experimental responses to interventions.

In safety-critical fields like medicine, valid statistical inference is an absolute requirement. For instance, the Food and Drug Administration guidelines strongly recommend that any method aimed at improving the efficiency of randomized experiments "should provide valid inference under approximately the same minimal statistical assumptions that would be needed for unadjusted estimation in a randomized trial" (FDA, 2021). This raises a critical question: can we achieve precision gains by using in-silico experiments while preserving valid statistical inference in randomized experiments? In this paper, we introduce the concept of a *hybrid experiment*, a statistical framework that combines predictions from multiple foundation models to improve the efficiency of randomized experiments while preserving valid statistical inference under minimal assumptions (see Figure 1 for an illustration).

2 BACKGROUND ON RANDOMIZED EXPERIMENTS

We observe a dataset $\mathcal{D}$ of size n from a randomized experiment, containing tuples (X, Y, A) of covariates $X \in \mathbb{R}^d$, bounded outcome $Y \in \mathbb{R}$, and treatment variable $A \in \{0, 1\}$. We assume that the data is drawn i.i.d. from $\mathbb{P}$ over $(X, Y(0), Y(1), Y, A)$, where $(Y(0), Y(1)) \in \mathbb{R}^2$ are the potential outcomes.

Our goal is to estimate the average treatment effect (ATE) in the randomized experiment population,

$$\theta := \mathbb{E}[Y(1) - Y(0)],$$

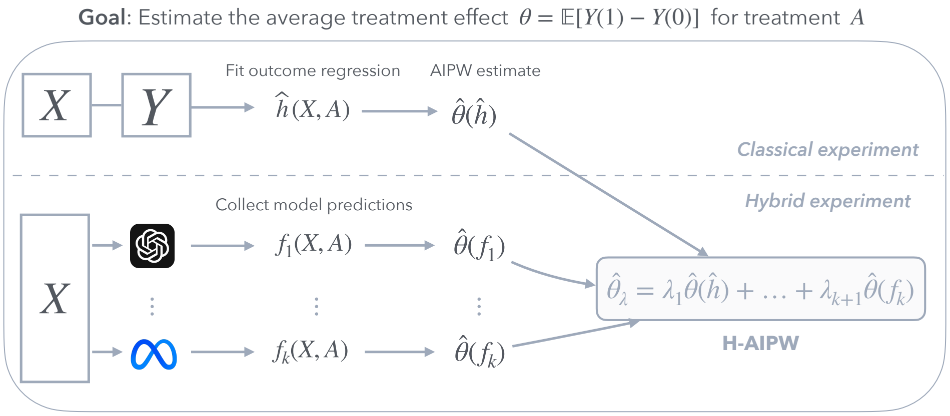


Figure 1: H-AIPW combines the standard AIPW estimator, which relies on experimental data alone, with multiple competing estimators that replace the outcome regression with predictions from foundation models. By leveraging foundation models trained on a much larger sample, rather than estimating the outcome regression with the limited experimental data, H-AIPW significantly reduces the finite sample variance of the average treatment effect estimate.

where the expectation is taken over $\mathbb{P}$. In particular, we want to improve upon the statistical precision of classical ATE estimators by constructing an asymptotically valid confidence interval that is as tight as possible. In randomized experiments, it is well known that the following are sufficient conditions to identify the ATE.

Assumption 2.1 (Randomized experiment assumptions)**.** The data-generating process satisfies

$$
\begin{aligned}
(i)&\;\; Y = Y(A),\ \mathbb{P} - \text{almost surely}.\\
(ii)&\;\; Y(a) \perp\!\!\!\perp A,\ \text{for } a = 0, 1.\\
(iii)&\;\; \pi_a = \mathbb{P}(A = a) > 0,\ \text{for } a = 0, 1.
\end{aligned}
$$

Condition *(i)* holds when the intervention is well-defined, as is typical for protocol-driven treatments in clinical trials. Conditions *(ii)* and *(iii)* are directly supported by randomization in the study design. We further assume that the treatment assignment probability π_a is known by design, as is the case in the overwhelming majority of experiments[1].

Under Assumption 2.1, we can identify the ATE as follows

$$\theta = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y \mid A = 1] - \mathbb{E}[Y \mid A = 0].$$

Therefore, the standard approach is to estimate θ using the difference in means estimator,

$$\widehat{\theta}_{\mathrm{DM}} := \frac{1}{n} \sum_{i \in \mathcal{D}} \left(\frac{Y_i A_i}{\pi_1} - \frac{Y_i (1 - A_i)}{\pi_0} \right).$$

This estimator is consistent and asymptotically normal—see e.g. Wager (2024, Section 1.1):

$$\sqrt{n}(\widehat{\theta}_{\mathrm{DM}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\mathrm{DM}}),$$

where $\rightsquigarrow$ denotes convergence in distribution and V_{DM} denotes the asymptotic variance. Therefore, provided that we can obtain a consistent estimator of the asymptotic variance, $\widehat{V}_{\mathrm{DM}} = V_{\mathrm{DM}} + o_{\mathbb{P}}(1)$, we can construct an asymptotically valid confidence interval

$$\mathcal{C}^{\alpha}_{\mathrm{DM}} = \left(\widehat{\theta}_{\mathrm{DM}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{V}_{\mathrm{DM}}}{n}} \right), \tag{1}$$

such that $\lim_{n \to \infty} \mathbb{P}(\theta \in \mathcal{C}^{\alpha}_{\mathrm{DM}}) \geq 1 - \alpha$, where z_α is the α-quantile of the standard normal distribution. Arguably, $\widehat{\theta}_{\mathrm{DM}}$ is all that is needed to estimate average treatment effects in randomized experiments. However, the confidence interval $\mathcal{C}^{\alpha}_{\mathrm{DM}}$ is often very wide, and it is possible to obtain narrower confidence intervals if we leverage the covariate information, as we will see in the next section.

[1]We note that our framework can easily be extended to allow for covariate-adaptive randomization or settings in which the probability of treatment needs to be estimated.

2.1 A CLASS OF VALID ESTIMATORS: AIPW

Robins et al. (1994) show that all estimators of θ that are consistent and asymptotically normal are asymptotically equivalent (when the propensity score is known) to the AIPW estimator, defined as

$$\widehat{\theta}_{\text{AIPW}}(h) := \frac{1}{n} \sum_{i \in \mathcal{D}} \psi_i^1(h) - \psi_i^0(h),$$

where $h : \mathbb{R}^d \times \{0, 1\} \to \mathbb{R}$ is a square-integrable function, and we define for $a \in \{0, 1\}$:

$$\psi_i^a(h) := \frac{\mathbb{I}\{A_i = a\}(Y_i - h(X_i, a))}{\pi_a} + h(X_i, a).$$

The most efficient estimator within this class can be identified by minimizing the asymptotic variance with respect to the function h. Specifically, the semiparametric efficiency lower bound is attained by choosing $h^\star(x, a) = \mathbb{E}[Y|X = x, A = a]$, which corresponds to the conditional mean of the outcome, also referred to as the *outcome regression*. In other words, the estimator $\widehat{\theta}_{\text{AIPW}}(h^\star)$ attains the smallest asymptotic variance among all consistent and asymptotically normal estimators of θ, and, thus, the smallest possible confidence interval in large samples. In practice, however, we have an estimator of the outcome regression $\widehat{h}$, which achieves the efficiency lower bound only if $||\widehat{h} - h^\star||_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$.

Below, we adapt the standard result that establishes consistency and asymptotic normality of the AIPW estimator to our setting, where the treatment probability is known. The key distinction from the standard setting is that asymptotic normality is achieved irrespective of the convergence rate of the outcome regression estimators. This means that the confidence intervals are valid even when the outcome regression is estimated using complex machine learning models, including those with unknown convergence rates.

Proposition 2.2 (Asymptotic behavior of AIPW). *Let $\widehat{h}$ be the outcome regression estimator, and $h^\dagger$ be its asymptotic limit, i.e. a square-integrable function such that*

$$||\widehat{h}(\cdot, a) - h^\dagger(\cdot, a)||_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1), \textit{ for } a = 0, 1.$$

Assume that $\widehat{h}$ is estimated from an independent sample, e.g. using cross-fitting. Then, it follows that $\widehat{\theta}_{\text{AIPW}}(\widehat{h})$ is root-n consistent and asymptotically normal:

$$\sqrt{n}(\widehat{\theta}_{\text{AIPW}}(\widehat{h}) - \theta) \rightsquigarrow \mathcal{N}(0, V_{h^\dagger}),$$

where $V_{h^\dagger} = \mathbb{E}\left[\left(\psi_i(h^\dagger) - \theta\right)^2\right]$ is the asymptotic variance.

We provide a proof of this result in Appendix A.1. Proposition 2.2 shows that the choice of estimator for the outcome regression does not affect the validity of the inference, provided that it is independent from the experimental data—for example, by using cross-fitting. Under these conditions, we can then construct an asymptotically valid confidence interval $\mathcal{C}^\alpha_{\text{AIPW}}$ as outlined in Equation (1).

However, because the asymptotic variance depends on the limiting function $h^\dagger$, with the smallest variance being achieved by the outcome regression $h^\star$, the choice of the outcome regression estimator is key to obtain precise estimates. The standard machine learning paradigm applied to our setting would first choose an appropriate model class $\mathcal{H}$ (e.g. all linear functions) and loss function $\mathcal{L}$ (e.g. mean squared loss), and minimize the empirical risk separately for each treatment arm a:

$$\widehat{h}(X, a) \in \arg\min_{h \in \mathcal{H}} \frac{1}{n_a} \sum_{i: A_i = a} \mathcal{L}(Y_i, h(X_i)). \tag{2}$$

3 METHODOLOGY

We introduce **H**ybrid **A**ugmented **I**nverse **P**robability **W**eighting (H-AIPW), an estimator that, in contrast to the standard AIPW, leverages the predictions from multiple foundation models to improve statistical precision. Algorithm 1 provides a formal definition of the H-AIPW estimator; here, we first introduce the estimator and then give theoretical results for its asymptotic distribution and variance.

Algorithm 1 Hybrid Augmented Inverse Probability Weighting (H-AIPW)

Require: (i) Dataset $\mathcal{D} = \{(X_i, A_i, Y_i)\}_{i=1}^n$. (ii) Collection of foundation models $f_1, \ldots, f_k$. (iii) Loss function $\mathcal{L}$ and function class $\mathcal{H}$. (iv) π_a for $a = 0, 1$. (v) Significance level α.

1: Use cross-fitting to compute the estimate $\widehat{\theta}_{\text{AIPW}}(\widehat{h})$ from the dataset $\mathcal{D}$, where for each arm a:

$$\widehat{h}(X, a) \in \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{n_a} \sum_{i:A_i=a} \mathcal{L}(h(X_i), Y_i) \right\}.$$

2: Compute $\widehat{\lambda} = \widehat{\Sigma}^{-1}\mathbf{1}/(\mathbf{1}^\top \widehat{\Sigma}^{-1}\mathbf{1})$, where

$$\widehat{\Sigma} := \frac{1}{n} \sum_{i=1}^n \left((\psi_i(\widehat{h}), ..., \psi_i(f_k)) - \bar{\psi} \right)^\top \left((\psi_i(\widehat{h}), ..., \psi_i(f_k)) - \bar{\psi} \right), \text{ and } \bar{\psi} := \frac{1}{n} \sum_{i=1}^n (\psi_i(\widehat{h}), \ldots, \psi_i(f_k)).$$

3: Compute the estimate and its variance

$$\widehat{\theta}_{\hat{\lambda}} := \widehat{\lambda}_1 \widehat{\theta}_{\text{AIPW}}(\widehat{h}) + \sum_{j=1}^k \widehat{\theta}_{\text{AIPW}}(f_j)\, \widehat{\lambda}_{j+1}, \text{ and } \widehat{V}_{\hat{\lambda}} := \widehat{\lambda}^\top \,\widehat{\Sigma}\, \widehat{\lambda}. \tag{3}$$

4: **Return:** $\mathcal{C}^{\alpha}_{\text{H-AIPW}} = \left(\widehat{\theta}_{\hat{\lambda}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{V}_{\hat{\lambda}}}{n}} \right)$, where z_α is the α-quantile of the standard normal.

3.1 Hybrid Augmented Inverse Probability Weighting

With the recent widespread availability of foundation models, we can potentially improve the accuracy of the outcome regression estimator beyond what is obtained from Equation (2) simply by replacing it with a foundation model. Further, as is often the case with language models, multiple competing models may be available, with no clear way to determine the best choice for a given task in advance. Therefore, we propose combining multiple AIPW estimators, each using a different outcome regression estimator.

More formally, we want to estimate the ATE θ based on a collection of several AIPW estimators:

$$\widehat{\theta}_{\text{AIPW}}(\widehat{h}), \widehat{\theta}_{\text{AIPW}}(f_1), \ldots, \widehat{\theta}_{\text{AIPW}}(f_k).$$

Here, $\widehat{h}$ is estimated exclusively from experimental data as shown in Equation (2), while $f_1, \ldots, f_k$ are foundation models trained on independent external data. The problem of dealing with several competing estimators of the same quantity has been extensively studied in the statistics literature; see e.g. Lavancier & Rochet (2016). A common solution is to consider a weighted average of the available estimators, which in our setting corresponds to

$$\widehat{\theta}_\lambda := \lambda_1 \widehat{\theta}_{\text{AIPW}}(\widehat{h}) + \sum_{j=1}^k \widehat{\theta}_{\text{AIPW}}(f_j)\lambda_{j+1}, \text{ for some } \lambda \in \Lambda = \{\lambda \in \mathbb{R}^{k+1} : \sum_{j=1}^{k+1} \lambda_j = 1\}.$$

We restrict the weights to the constraint set Λ so that the combined estimator $\widehat{\theta}_\lambda$ is still in the class of AIPW estimators. We can then choose the weight that minimizes the variance, that is:

$$\lambda^\star = \arg\min_{\lambda \in \Lambda} \operatorname{Var}[\widehat{\theta}_\lambda] = \arg\min_{\lambda \in \Lambda} \lambda^\top \Sigma \lambda = \Sigma^{-1}\mathbf{1}/(\mathbf{1}^\top \Sigma^{-1}\mathbf{1}),$$

with $\Sigma := \operatorname{Cov}[(\psi(h^\dagger), \ldots, \psi(f_k))^\top]$ being the asymptotic covariance matrix of the estimators. However, in practice, we only have access to an estimate $\widehat{\Sigma}$ of the covariance matrix, and thus we use

$$\widehat{\lambda} := \arg\min_{\lambda \in \Lambda} \lambda^\top \widehat{\Sigma} \lambda.$$

Asymptotic validity and efficiency We establish that the H-AIPW estimator is both consistent and asymptotically normal, with an asymptotic variance that is no greater than that of the standard AIPW.

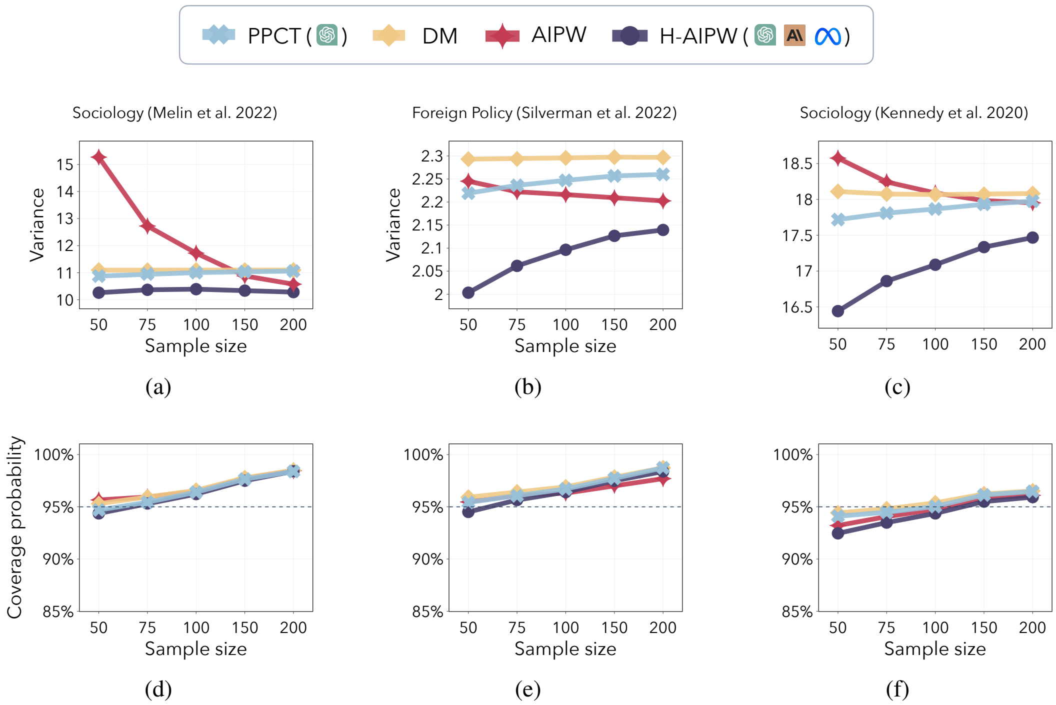


Figure 2: Performance comparison of H-AIPW against baseline estimators (PPCT, DM, AIPW) across three randomized experiments. We randomly subsample each study to obtain the sample sizes shown on the x-axis and report the average over $R = 10k$ repetitions for each metric. The significance level is set to $\alpha = 0.05$. **(First row) Precision:** Figures 2a to 2c show the empirical variance achieved by H-AIPW and the baseline estimators for varying sample sizes. **(Second row) Validity:** Figures 2d to 2f show the empirical coverage probability of each estimator for varying sample sizes; the dashed horizontal line represents the nominal 95% coverage level.

Theorem 3.1 (Asymptotic behavior of H-AIPW)**.** *Let $\widehat{h}$ be an estimator that satisfies the conditions in Proposition 2.2, with asymptotic limit $h^\dagger$. Further, let $\widehat{\theta}_{\widehat{\lambda}}$ be as in Equation (3), and assume that Σ is non-singular and $\widehat{\Sigma}\Sigma^{-1} \xrightarrow{p} I$. Then, it holds that*

$$\sqrt{n}(\widehat{\theta}_{\widehat{\lambda}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\lambda^\star}).$$

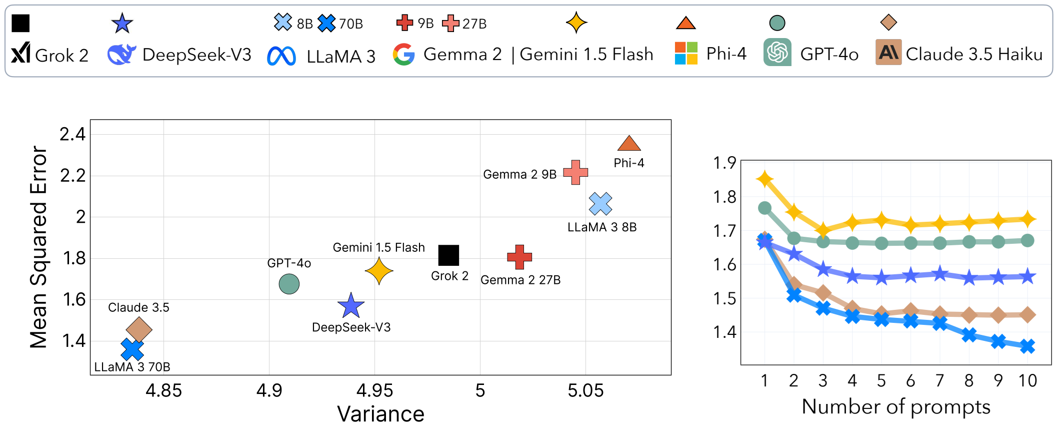
Moreover, the asymptotic variance of the combined estimator is no greater than that of any individual estimator, i.e. it holds that

$$V_{\lambda^\star} \leq \Sigma_{jj}, \text{ for } j = 1, \ldots, k+1.$$

We provide a proof of this result in Appendix A.2. Theorem 3.1 offers a principled approach to combining competing AIPW estimators, ensuring that the resulting estimator is at least as precise as the best estimator in the ensemble. In particular, this approach allows us to leverage the strengths of foundation models without any risks: when these models give accurate outcome predictions, the combined estimator uses their extra information to improve precision. On the other hand, when the foundation models are biased, the final estimator falls back to the standard AIPW estimator.

4 Experiments

In this section, we evaluate H-AIPW across three randomized experiments in Foreign Policy (Silverman et al., 2022), and Sociology (Kennedy & Horne, 2020; Melin & Merluzzi, 2022), and analyze 5 other studies in Appendix C.1. Moreover, we study the significance of model scale and inference-time compute in the effectiveness of our estimator. For each study, we implement the following subsampling procedure: starting with a full dataset $\mathcal{D}$, we select a target sample size $n \in \{50, \ldots, 200\}$. For each repetition $r \in \{1, \ldots, R\}$, we sample n participants without replacement from $\mathcal{D}$, ensuring the treatment and control groups are balanced, to create a smaller dataset $\mathcal{D}_r$.



(a) (b)

Figure 3: Impact of model scale and inference-time compute on the performance of H-AIPW in the study by Fahey et al. (2023). **(Left) Model scale**: Figure 3a shows the relationship between the empirical estimate of the H-AIPW variance (average on $R = 10k$ repetitions, sample size $n = 50$) and mean squared error (MSE) for LLMs of varying sizes (10 prompts at inference time). **(Right) Inference-time compute**: Figure 3b shows the impact on the MSE of increasing the number of prompts at inference time and averaging the resulting predictions.

Estimators and Metrics We implement H-AIPW by integrating predictions from three popular LLMs: GPT-4o, Claude 3.5 Haiku, and LLaMA 3 70B, unless stated otherwise. We benchmark our estimator against three baselines: $\widehat{\theta}_{\text{DM}}$ (DM), $\widehat{\theta}_{\text{AIPW}}(\widehat{h})$ (AIPW), and the PPCT estimator (Poulet et al., 2025) (see Appendix D.1 for implementation details). To benchmark precision, for each estimator $\widehat{\theta}$, we compute the scaled variance $n\widehat{\text{Var}}[\widehat{\theta}]$, where $\widehat{\text{Var}}$ is the the empirical variance estimate averaged over R subsampling repetitions. To benchmark validity, we compute the fraction of confidence intervals containing the ATE: Coverage $= \frac{1}{R}\sum_{r=1}^{R}\mathbb{I}\{\theta \in \mathcal{C}_r^\alpha\}$, where $\mathcal{C}_r^\alpha$ is the confidence interval obtained from the dataset $\mathcal{D}_r$ and θ is the ground-truth ATE estimate from the full study.

Results Figures 2a to 2c show that H-AIPW consistently achieves lower variance—and hence tighter confidence intervals—than the baselines across all studies and sample sizes. For small sample sizes, H-AIPW yields reductions in variance ranging from 5% to 30%, depending on the study and baseline. This trend aligns with statistical theory, as the outcome regression's estimation error is high in small sample sizes, increasing the finite sample variance of the standard AIPW estimator. For large sample sizes, the gains against the standard AIPW plateau at 2% to 3%. This suggests that beyond finite sample improvements, there are also asymptotic gains due to potential model misspecification in the standard AIPW outcome regression. Finally, while Theorem 3.1 establishes asymptotic validity of the H-AIPW confidence intervals, we confirm that its precision gains do not come at the cost of validity in finite sample settings: Figures 2d to 2f show that H-AIPW maintains coverage comparable to the baselines.

Model scale Figure 3a illustrates the precision gains achieved by H-AIPW when leveraging predictions from LLMs of varying scales. Large models consistently achieve lower MSE and thus lower variance than smaller models, with LLaMA 3 70B excelling despite having fewer parameters than GPT-4o and Claude 3.5 Haiku.

Inference-time compute Figure 3b shows that averaging over many prompts consistently reduces the MSE for the large models—a similar trend is expected for the smaller ones. As smaller MSE is associated with higher precision (see Figure 3a), using multiple prompts is expected to improve the precision of H-AIPW further. We confirm this observation in Appendix C.3, showing that H-AIPW precision improves with more prompts across several randomized studies.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Clara Fannjiang, Michael Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios Angelopoulos, John Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Rishi Bommasani, Drew Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Mark Brandt. Onset and offset deservingness: The case of home foreclosures. *Political Psychology*, 34(2):221–238, 2013.

David Broska, Michael Howes, and Austin van Loon. The mixed subjects design: Treating large language models as (potentially) informative observations. Technical report, 2024.

Peter Caprariello and Harry Reis. To do, to have, or to share? valuing experiences over material possessions depends on the involvement of others. *Journal of personality and social psychology*, 104(2):199, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Issa Dahabreh, Anthony Matthews, Jon Steingrimsson, Daniel Scharfstein, and Elizabeth Stuart. Using trial and observational data to assess effectiveness: trial emulation, transportability, benchmarking, and joint analysis. *Epidemiologic reviews*, 46(1):1–16, 2024.

Piersilvio De Bartolomeis, Javier Abad, Konstantin Donhauser, and Fanny Yang. Detecting critical treatment effect bias in small subgroups. *Uncertainty in Artificial Intelligence*, 2024.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. *International Conference on Machine Learning*, 2024.

Florian Dorner, Vivian Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data. *arXiv preprint arXiv:2410.13341*, 2024.

Naoki Egami, Musashi Hinck, Brandon Stewart, and Hanying Wei. Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems*, 2024.

James Fahey, Damon Roberts, and Stephen Utych. Principled or partisan? the effect of cancel culture framings on support for free speech. *American Politics Research*, 51(1):69–75, 2023.

FDA. Adjusting for covariates in randomized clinical trials for drugs and biological products, 2021.

Adam Fisch, Joshua Maynez, Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William Cohen. Stratified prediction-powered inference for hybrid language model evaluation. *arXiv preprint arXiv:2406.04291*, 2024.

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

Ingar Haaland and Christopher Roth. Beliefs about racial discrimination and support for pro-black policies. *Review of Economics and Statistics*, 105(1):40–53, 2023.

Zeshan Hussain, MingChieh Shih, Michael Oberst, Ilker Demirel, and David Sontag. Falsification of internal and external validity in observational studies via conditional moment restrictions. *International Conference on Artificial Intelligence and Statistics*, 2023.

Rickard Karlsson, Guanbo Wang, Jesse Krijthe, and Issa Dahabreh. Robust integration of external control data in randomized trials. *arXiv preprint arXiv:2406.17971*, 2024.

Emily Kennedy and Christine Horne. Accidental environmentalist or ethical elite? the moral dimensions of environmental impact. *Poetics*, 82:101448, 2020.

Frédéric Lavancier and Paul Rochet. A general procedure to combine estimators. *Computational Statistics & Data Analysis*, 94:175–192, 2016.

Lauren Liao, Emilie Højbjerre-Frandsen, Alan Hubbard, and Alejandro Schuler. Prognostic adjustment with efficient estimators to unbiasedly leverage historical data in randomized trials. *arXiv preprint arXiv:2305.19180*, 2023.

Alex Luedtke, Marco Carone, and Mark van der Laan. An omnibus non-parametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(1):75–99, 2019.

Julia Lee Melin and Jennifer M Merluzzi. When women do “men’s work”: Hybrid femininity and within-gender inequality in job search. In *Academy of Management Proceedings*, pp. 10631, 2022.

Pierre-Emmanuel Poulet, Maylis Tran, Sophie Tezenas du Montcel, Bruno Dubois, Stanley Durrleman, and Bruno Jedynak. Prediction-powered inference for clinical trials. *medRxiv*, 2025.

James Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

James Robins, Andrea Rotnitzky, and Lue Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427): 846–866, 1994.

Alejandro Schuler, David Walsh, Diana Hall, Jon Walsh, and Charles Fisher. Increasing the efficiency of randomized trial estimates via linear adjustment for a prognostic score. *The International Journal of Biostatistics*, 18(2):329–356, 2022.

Eric Shuman, Martijn van Zomeren, Tamar Saguy, Eric Knowles, and Eran Halperin. Defend, deny, distance, and dismantle: A new measure of advantaged identity management. *Personality and Social Psychology Bulletin*, pp. 01461672231216769, 2024.

Daniel Silverman, Daniel Kent, and Christopher Gelpi. Putting terror in its place: An experiment on mitigating fears of terrorism among the american public. *Journal of Conflict Resolution*, 66(2): 191–216, 2022.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer, 2006.

Stefan Wager. Causal inference: A statistical learning approach. Technical report, Stanford University, 2024.

APPENDICES

The following appendices provide deferred proofs, related works, ablation studies, and experimental details.

A PROOFS

A.1 PROOF OF PROPOSITION 2.2

We adapt here a classic result from the semiparametric inference literature to our specific setting where the probability of treatment is known by design. For clarity, we refer to $\widehat{\theta}_{\text{AIPW}}$ as $\widehat{\theta}$.

Let us define the influence function of the AIPW estimator for fixed outcome functions h as:

$$\psi_i(h) = \left(\frac{A_i}{\pi_1}(Y_i - h(X_i, 1)) + h(X_i, 1)\right) - \left(\frac{1 - A_i}{\pi_0}(Y_i - h(X_i, 0)) + h(X_i, 0)\right).$$

We can then decompose the estimation error of the AIPW estimator as follows:

$$\sqrt{n}(\widehat{\theta}(\widehat{h}) - \theta) = \underbrace{\sqrt{n}(\widehat{\theta}(h^\dagger) - \theta)}_{:=T_1} + \underbrace{\sqrt{n}(\widehat{\theta}(\widehat{h}) - \widehat{\theta}(h^\dagger))}_{:=T_2}.$$

The first term, T_1, is an average of i.i.d. random variables with mean zero and finite variance. Therefore, by the Central Limit Theorem, we have:

$$\sqrt{n}(\widehat{\theta}(h^\dagger) - \theta) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \psi_i(h^\dagger) - \theta\right) \rightsquigarrow \mathcal{N}(0, V_{h^\dagger}),$$

where the asymptotic variance is given by $V_{h^\dagger} = \mathbb{E}[\psi_i(h^\dagger)^2]$.

Bounding the Remainder Term We need to show that the second term T_2 is asymptotically negligible, that is $T_2 = o_{\mathbb{P}}(1)$.

We can rewrite this term as:

$$T_2 = \sqrt{n}(\widehat{\theta}(\widehat{h}) - \widehat{\theta}(h^\dagger)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\psi_i(\widehat{h}) - \psi_i(h^\dagger)\right).$$

Further, with some simple algebra we can decompose the difference in the influence functions as:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_i(\widehat{h}) - \psi_i(h^\dagger)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{A_i - \pi_1}{\pi_1}\right)(h^\dagger(X_i, 1)$$
$$- \widehat{h}(X_i, 1)) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{A_i - \pi_1}{1 - \pi_1}\right)(\widehat{h}(X_i, 0) - h^\dagger(X_i, 0))$$

Now, we will show that both terms in the sum above are asymptotically negligible. We focus our proof on the first term; the second follows from symmetric arguments.

Let $\mathbb{P}_n$ denote the empirical measure over $Z_1, \dots, Z_n$, and define the following functions:

$$f(Z_i) := \frac{A_i - \pi_1}{\pi_1}\, h^\dagger(X_i, 1) \;\text{ and }\; \widehat{f}(Z_i) := \frac{A_i - \pi_1}{\pi_1}\, \widehat{h}(X_i, 1).$$

We can rewrite the first term as:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{A_i - \pi_1}{\pi_1}\right)(h^\dagger(X_i, 1) - \widehat{h}(X_i, 1)) = (\mathbb{P}_n - \mathbb{P})(f - \widehat{f}),$$

where we use the fact that $\mathbb{P}(f - \widehat{f}) = 0$, since the treatment probability is known. Since $\widehat{h}$ is estimated from an independent sample, it follows from Chebyshev inequality that

$$(\mathbb{P}_n - \mathbb{P})(\widehat{f} - f) = O_{\mathbb{P}}\left(\frac{||\widehat{f} - f||_{L_2(\mathbb{P})}}{\sqrt{n}}\right) = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right),$$

since it follows from assumptions that $||\widehat{f} - f||_{L_2(\mathbb{P})} = o_{\mathbb{P}}(1)$. Therefore, it also follows that $T_2 = o_{\mathbb{P}}(1)$.

Finally, using Slutsky's theorem, we get:

$$\sqrt{n}(\widehat{\theta}(\widehat{h}) - \theta) = \sqrt{n}(\widehat{\theta}(h^{\dagger}) - \theta) + o_{\mathbb{P}}(1) \rightsquigarrow \mathcal{N}(0, V_{h^{\dagger}}),$$

which completes the proof.

A.2 Proof of Theorem 3.1

Recall that $\Sigma := \mathrm{Cov}[(\widehat{\theta}_{\text{AIPW}}(h^{\dagger}), \ldots, \widehat{\theta}_{\text{AIPW}}(f_k))^{\top}]$ and define the oracle weights as $\lambda^{\star} = \arg\min_{\lambda \in \Lambda} \lambda^{\top}\Sigma\lambda$. The corresponding oracle estimator is then

$$\widehat{\theta}_{\lambda^{\star}} = \lambda_1^{\star}\widehat{\theta}_{\text{AIPW}}(\widehat{h}) + \sum_{j=1}^{k} \lambda_{j+1}^{\star}\widehat{\theta}_{\text{AIPW}}(f_j).$$

We now prove the theorem in the following three steps.

First, we observe that $\widehat{\theta}_{\lambda^{\star}}$ can also be written as

$$\widehat{\theta}_{\lambda^{\star}} = \widehat{\theta}_{\text{AIPW}}\left(\lambda_1^{\star}\widehat{h} + \sum_{j=1}^{k} \lambda_{j+1}^{\star} f_j\right),$$

since the constraint set is $\Lambda = \{\lambda \in \mathbb{R}^{k+1} : \sum_{j=1}^{k+1} \lambda_i = 1\}$. Further, it follows from assumptions that $\lambda_1^{\star}\widehat{h} + \sum_{j=1}^{k} \lambda_{j+1}^{\star} f_j$ is also an outcome function estimator that satisfies the conditions in Proposition 2.2, therefore $\widehat{\theta}_{\lambda^{\star}}$ is consistent and asymptotically normal, i.e. it holds that

$$\sqrt{n}(\widehat{\theta}_{\lambda^{\star}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\lambda^{\star}}),$$

where $V_{\lambda^{\star}} = n\lambda^{\star\top}\Sigma\lambda^{\star}$.

Second, we show that the asymptotic variance $V_{\lambda^{\star}}$ satisfies

$$V_{\lambda^{\star}} \leq V_j \;\text{ for }\; j = 1, \ldots, k.$$

By construction, the oracle weights $\lambda^{\star}$ minimize $\lambda^{\top}\Sigma\lambda$, ensuring $\widehat{\theta}_{\lambda^{\star}}$ attains the smallest asymptotic variance among all convex combinations of the initial estimators:

$$\left\{\widehat{\theta}_{\lambda} \coloneqq \lambda_1\widehat{\theta}_{\text{AIPW}}(\widehat{h}) + \sum_{j=1}^{k} \lambda_{j+1}\widehat{\theta}_{\text{AIPW}}(f_j) \,\middle|\, \lambda \in \Lambda\right\}.$$

For any $j \in \{1, \ldots, k\}$, the estimator $\widehat{\theta}_{\text{AIPW}}(f_j)$ corresponds to $\widehat{\theta}_{\lambda'}$, where $\lambda' \in \mathbb{R}^{k+1}$ is the canonical basis vector with $\lambda'_{j+1} = 1$ and $\lambda'_i = 0$ for $i \neq j+1$. Since $\widehat{\theta}_{\lambda'} \in \{\widehat{\theta}_{\lambda} : \lambda \in \Lambda\}$, the optimality of $\lambda^{\star}$ implies:

$$V_{\lambda^{\star}} = n\lambda^{\star\top}\Sigma\lambda^{\star} \leq n\lambda'^{\top}\Sigma\lambda' = V_{j+1} \quad \text{for } \; j = 1, \ldots, k.$$

The same reasoning applies for the estimator $\widehat{\theta}_{\text{AIPW}}(\widehat{h})$.

Finally, we prove that $\widehat{\theta}_{\hat{\lambda}^{\star}}$ and $\widehat{\theta}_{\lambda^{\star}}$ are asymptotically equivalent. This follows directly from Lavancier & Rochet (2016, Proposition 3.3), which implies that if $\widehat{\Sigma}\Sigma^{-1} \xrightarrow{p} I$, $\widehat{\theta}_{\hat{\lambda}^{\star}}$ and $\widehat{\theta}_{\lambda^{\star}}$ have the same asymptotic distribution.

B Related work

Our work draws heavily from the literature on semiparametric inference and double machine learning (Robins et al., 1994; Robins & Rotnitzky, 1995; Tsiatis, 2006; Chernozhukov et al., 2018). In particular, our estimator is an optimal combination of several Augmented Inverse Probability Weighting (AIPW) estimators, whose outcome regressions are replaced with foundation models. Importantly, the standard AIPW estimator, which relies on an outcome regression estimated using experimental data alone, is also included in the combination. This approach allows H-AIPW to significantly reduce finite sample (and potentially asymptotic) variance while attaining the semiparametric *efficiency bound*—the smallest asymptotic variance among all consistent and asymptotically normal estimators of the average treatment effect—even when the foundation models are arbitrarily biased.

Integrating foundation models Prediction-powered inference (PPI) (Angelopoulos et al., 2023a) is a statistical framework that constructs valid confidence intervals using a small labeled dataset and a large unlabeled dataset imputed by a foundation model. PPI has been applied in various domains, including generalization of causal inferences (Demirel et al., 2024), large language model evaluation (Fisch et al., 2024; Dorner et al., 2024), and improving the efficiency of social science experiments (Broska et al., 2024; Egami et al., 2024). However, unlike our approach, PPI requires access to an additional unlabeled dataset from the same distribution as the experimental sample, which may be as costly as labeled data. Recent work by Poulet et al. (2025) introduces Prediction-powered inference for clinical trials (PPCT), an adaptation of PPI to estimate average treatment effects in randomized experiments without any additional external data. PPCT combines the difference in means estimator with an AIPW estimator that integrates the same foundation model as the outcome regression for both treatment and control groups. However, our work differs in two key aspects: (i) PPCT integrates a single foundation model, and (ii) PPCT does not include the standard AIPW estimator with the outcome regression estimated from experimental data. As a result, PPCT cannot achieve the efficiency bound unless the foundation model is almost surely equal to the underlying outcome regression.

Integrating observational data There is growing interest in augmenting randomized experiments with data from observational studies to improve statistical precision. One approach involves first testing whether the observational data is compatible with the experimental data (Dahabreh et al., 2024)—for instance, using a statistical test to assess if the mean of the outcome conditional on the covariates is invariant across studies Luedtke et al. (2019); Hussain et al. (2023); De Bartolomeis et al. (2024)—and then combining the datasets to improve precision, if the test does not reject. These tests, however, have low statistical power, especially when the experimental sample size is small, which is precisely when leveraging observational data would be most beneficial. To overcome this, a recent line of work integrates a prognostic score estimated from observational data as a covariate when estimating the outcome regression (Schuler et al., 2022; Liao et al., 2023). However, increasing the dimensionality of the problem—by adding an additional covariate—can increase estimation error and inflate the finite sample variance. Finally, the work most closely related to ours is Karlsson et al. (2024), that integrates an outcome regression estimated from observational data into the AIPW estimator. In contrast, our approach is not constrained by the availability of well-structured observational data, since it leverages black-box foundation models trained on external data sources.

B.1 Connection with prediction-powered inference

To further study the connection and differences with prediction-powered inference (PPI) Angelopoulos et al. (2023a), it is instructive to consider the simpler problem of estimating the counterfactual mean, $\mathbb{E}[Y(1)]$. For this case, a variant of PPI, referred to as PPI++ (Angelopoulos et al., 2023b), can be shown to be equivalent to an AIPW estimator.

The standard difference in mean estimator in this case is the sample mean of outcomes for the treated group:

$$\widehat{\theta}_{\text{DM}} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i, \text{ where } n_a = \sum_{i=1}^{n} \mathbb{I}\{A_i = a\}.$$

PPI++ improves the difference in mean estimator by incorporating predictions from a black-box model f:

$$\widehat{\theta}_{\text{PPI++}} = \frac{1}{n}\sum_{i=1}^{n} Y_i + \lambda\left(-\frac{1}{n_1}\sum_{i:A_i=1} f(X_i) + \frac{1}{n_0}\sum_{i:A_i=0} f(X_i)\right),$$

where the power-tuning parameter λ is chosen to minimize the variance. Crucially, for $\lambda = \frac{n_0}{n_1+n_0}$ we have equivalence with the AIPW estimator for the counterfactual mean, i.e.

$$\widehat{\theta}_{\text{PPI++}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i(Y_i - f(X_i))}{\pi_1} + f(X_i)\right) = \widehat{\theta}_{\text{AIPW}}(f).$$

A few remarks are in order.

- PPI++ replaces the estimated outcome regression with a black-box model f. However, when $f(x)$ is not equivalent to the outcome regression $\mathbb{E}[Y \mid X = x, A = 1]$, the resulting estimator will not be efficient. In other words, $\widehat{\theta}_{\text{PPI++}}$ will not achieve the smallest asymptotic variance among the regular estimators of the counterfactual mean. In contrast, the standard AIPW will achieve the smallest possible asymptotic variance, assuming that the outcome regression estimator is consistent in L_2-norm. This condition is easy to satisfy in the setting of randomized experiments, since we can use flexible machine-learning models and still have valid confidence intervals as a consequence of Proposition 2.2. In contrast, our estimator is guaranteed to have asymptotic variance no greater than the standard AIPW estimator (see Theorem 3.1). As a result, it is efficient even if the black-box model f is arbitrarily biased.
- Extending PPI and PPI++ to average treatment effect estimation is not straightforward. To do so, Poulet et al. (2025) proposes the following estimator:

  $$\widehat{\theta}_{\text{PPCT}} := \frac{1}{n_1}\sum_{A_i=1}(Y_i - \lambda f(X_i)) - \frac{1}{n_0}\sum_{A_i=0}(Y_i - \lambda f(X_i)).$$

  However, a key limitation of the above estimator is that it forces both outcome regressions, that is $\mathbb{E}[Y \mid X = x, A = 1]$ and $\mathbb{E}[Y \mid X = x, A = 0]$, to be replaced with the same black-box model f. This is particularly problematic when the treatment has a significant effect on the outcome, as a single model f will fail to accurately capture both outcome regressions. In contrast, our approach allows for different black-box models f_1 and f_0 to be plugged-in for the treated and control group, respectively.
- PPI and its variants cannot integrate multiple competing foundation models. This is a key limitation in our setting where model selection is a non-trivial task due to the missingness of potential outcomes. Moreover, it is unclear whether they can be extended to do so. A major hurdle is constructing a consistent estimate of the covariance matrix of the estimators Σ. In contrast, our approach estimates the covariance matrix Σ by exploiting the linear structure of the AIPW estimators in the ensemble.

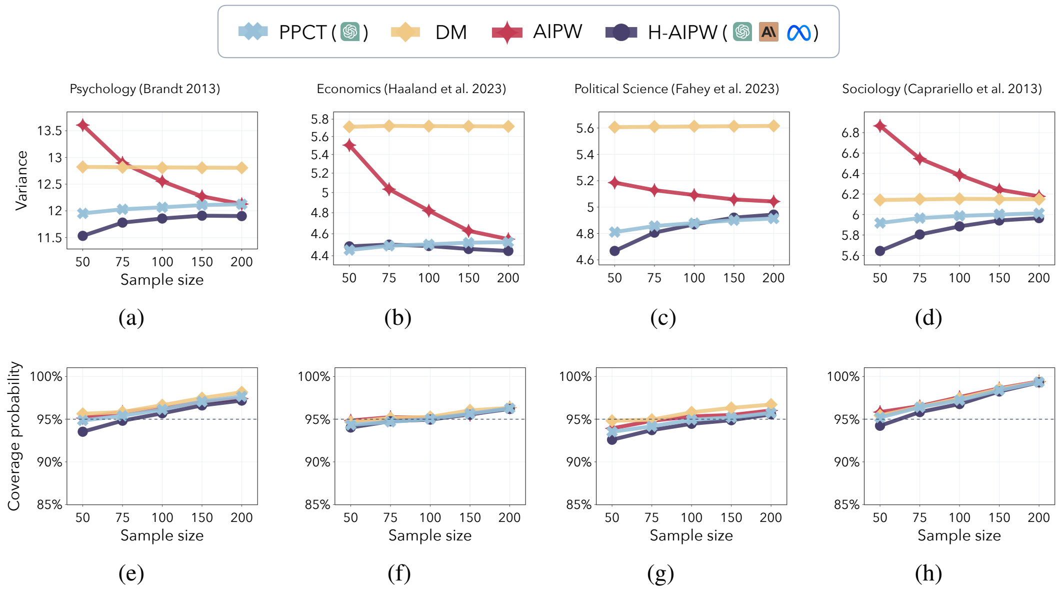


Figure 4: Performance comparison of H-AIPW with baseline estimators (PPCT, DM, AIPW) across four additional randomized studies—Brandt (2013), Haaland & Roth (2023), Fahey et al. (2023), and Caprariello & Reis (2013)—spanning Psychology, Economics, Political Science, and Sociology. We randomly subsample each study to obtain the sample sizes shown on the x-axis and report the average over $R = 10k$ repetitions for each metric. The significance level is set to $\alpha = 0.05$. **(First row) Precision:** Empirical estimate of the variance of H-AIPW and baselines for varying sample sizes. **(Second row) Validity**: Empirical coverage probability of each estimator, with the dashed line marking nominal 95% coverage. Results confirm that H-AIPW improves precision while maintaining valid coverage.

C Additional experiments

We present here additional experiments on randomized studies and ablations of our method. The results reinforce the general trends observed in the main experiments: H-AIPW achieves better precision than the baselines, particularly in the small sample regime, while maintaining comparable coverage. The ablation studies provide insights into the number of models that can be incorporated into our estimator without significantly compromising validity (due to finite sample effects). Additionally, they offer further evidence of the advantages of increasing inference-time compute.

C.1 Evaluation on additional scientific studies

In the main text (Figure 2), we demonstrated the effectiveness of our estimator across three studies. Here, we extend our analysis to five additional studies spanning diverse fields: Economics Haaland & Roth (2023), Psychology Brandt (2013), Sociology Caprariello & Reis (2013), Political Science Fahey et al. (2023), and Social Behavior Shuman et al. (2024). The experimental setup remains consistent with the main part of the paper (see Section 4 for details).

Figure 4 presents results for four of these studies, which align with findings from the main experiments. H-AIPW achieves variance gains often exceeding 20% over the baseline estimator in the small-sample regime. In the large-sample regime, H-AIPW performs similarly to PPCT, while still improving upon standard AIPW by $3\% - 6\%$. As for validity, H-AIPW maintains comparable empirical coverage across studies.

Study with a visual treatment and out of GPT-4o training dataset The study by Shuman et al. (2024) is particularly relevant for two reasons. First, its data was published in December 2024, after the last known training cutoff for GPT-4o, ensuring it was not included in the model's training set. Second, the study's treatment is an image rather than text, allowing us to evaluate our statistical

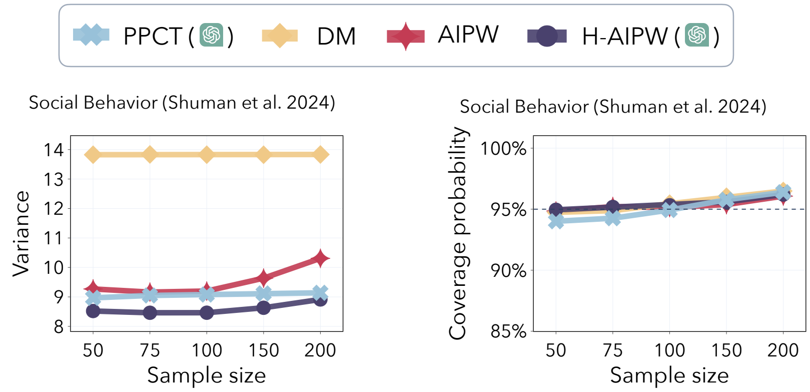


Figure 5: Performance comparison of H-AIPW with baseline estimators (PPCT, DM, AIPW) for the randomized study by Shuman et al. (2024). We randomly subsample each study to obtain the sample sizes shown on the x-axis and report the average over $R = 10k$ repetitions for each metric. The significance level is set to $\alpha = 0.05$. The same experimental configuration as in Figure 2 is maintained, except that predictions are limited to three prompts at inference time.

framework beyond the text modality. As shown in Figure 5, H-AIPW maintains strong performance in both precision and validity, achieving a reduction in variance of up to 37% over the DM estimator and between 5% and 12% compared to others. The empirical coverage is also comparable with the baselines.

C.2 Impact of adding more foundation models to H-AIPW

In this section, we study the impact of increasing the number of models in H-AIPW. Specifically, Algorithm 1 requires integrating predictions from multiple foundation models, which are combined with the standard AIPW to minimize the variance of the resulting estimator. In Figure 6, we show how increasing the number of language models from 1 to 7 affects the precision and validity of H-AIPW in the study by Fahey et al. (2023). Models are incorporated in the estimator sequentially, starting from those with the lowest mean squared error (MSE) (i.e. LLaMA 3 70B) to those with the highest (stopping at Gemma 2 27B), following Figure 3a. We also include the standard AIPW estimator for reference.

Increasing the number of models improves precision compared to the standard AIPW estimator. In the small-sample setting with 50 observations, a single model improves variance by approximately 6%, while using 4 models increases this gain to nearly 12%, and 7 models yield an improvement of around 16%. However, the marginal benefits diminish with larger sample sizes: at 200 observations, the variance difference between using 1 and 7 models shrinks to 4%.

However, adding more models weakens empirical coverage. With 50 samples, combinations of 5 to 7 models exhibit undercoverage of 2%–4% relative to AIPW or H-AIPW with 1–2 models, failing to reach the nominal 95% coverage until the sample size reaches 200. In contrast, combinations of 1 to 3 models maintain coverage levels comparable to AIPW. This is expected as there is a finite sample error term associated with estimating the weights, as discussed in **??**. Therefore, practitioners should carefully determine the number of models to include in the ensemble based on the available sample size.

C.3 Impact of inference-time compute on the precision

In Section 4, we demonstrated that increasing inference-time compute improves the precision of H-AIPW. This was established by studying the relationship between lower mean squared error (MSE) and reduced variance, as well as by showing that a higher number of prompts generally leads to

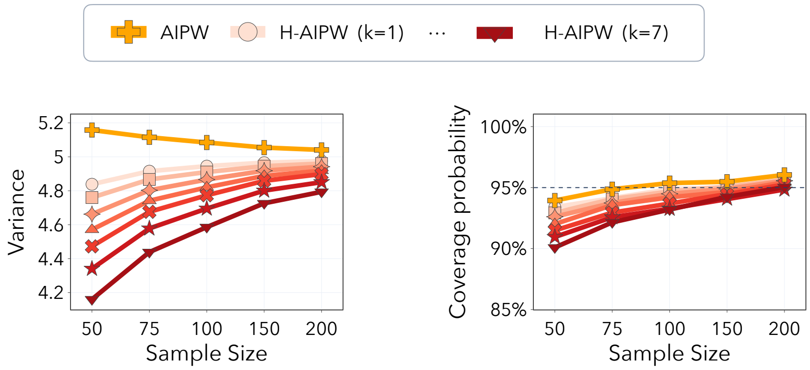
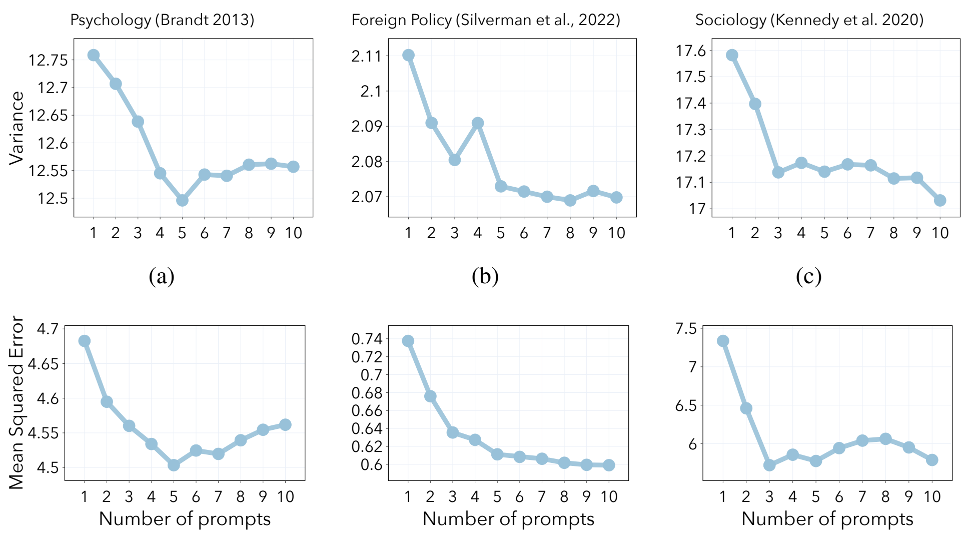


Figure 6: Impact of increasing the number of models in H-AIPW on precision and validity in the study by Fahey et al. (2023). Models are sequentially incorporated based on their mean squared error (MSE), starting with LLaMA 3 70B (lightest red, $k = 1$) and ending with Gemma 2 27B (darkest red, $k = 7$), following Figure 3a. The left panel shows the empirical variance, while the right panel shows empirical coverage. The standard AIPW estimator is included for reference. Each experiment is averaged over $R = 10k$ repetitions, with significance level set to $\alpha = 0.05$.



(a) (b) (c)

(d) (e) (f)

Figure 7: Impact of the number of prompts on the empirical variance and MSE. Results are reported for studies by Brandt (2013); Silverman et al. (2022); Kennedy & Horne (2020). We randomly subsample each study to obtain the sample sizes shown on the x-axis and report the average over $R = 10k$ repetitions for each metric. **(First row)** Reduction in variance as the number of prompts increases. **(Second row)** Reduction in MSE as the number of prompts increase. These results suggest that increasing inference-time compute improves the precision of H-AIPW by reducing the MSE.

lower MSE. For completeness, Figure 7 explicitly visualizes the connection between the number of prompts, MSE, and variance.

We present results for three studies—Brandt (2013); Silverman et al. (2022); Kennedy & Horne (2020)—using H-AIPW with predictions from GPT-4o. Figures 7a to 7c show the empirical estimate of the variance as a function of the number of prompts, while Figures 7d to 7f illustrate the corresponding changes in MSE. The findings reinforce the conclusions from the main text: increasing inference-time compute through multiple prompts generally reduces the variance of H-AIPW.

D EXPERIMENTAL DETAILS

D.1 IMPLEMENTATION DETAILS

For all experiments, we begin with a feature selection step that identifies the five features most correlated with the outcome variable. The AIPW estimator is implemented using cross-fitting with 30 folds and ridge regression with a regularization parameter of $\lambda = 1.0$ for outcome function estimation. For PPCT, we follow the implementation by Poulet et al. (2025), using GPT-4o's predictions for the control scenario as the prognostic score. The correlation coefficients for the optimal combination are computed using standard `Python` libraries. Finally, the DM estimator requires no hyperparameter tuning.

Implementation of H-AIPW Our estimator integrates synthetic outcomes generated by multiple LLMs. Unless stated otherwise, we use predictions from LLaMA 3 70B, GPT-4o, and Claude 3.5 Haiku for all experiments in Section 4. Additional models, such as Gemma 2, Grok 2, and Gemini 1.5 Flash, are used in specific cases. We leverage both proprietary and open-source LLMs. For open-source models, we apply nucleus sampling with a temperature of 1.2, top-p of 0.9, and a maximum of 100 new tokens. For proprietary models, we use default decoding settings, except for Claude 3.5 Haiku, where we set the temperature to 1.

In summary, H-AIPW extends the classic AIPW estimator by incorporating multiple AIPW estimators that integrate LLM predictions; see Algorithm 1 for full details.

D.2 STEP-BY-STEP RECIPE WITH LARGE LANGUAGE MODELS

We now provide a step-by-step guide for practitioners to implement H-AIPW using Large Language Models (LLMs). Our guide focuses on LLMs as they are both widely accessible and have demonstrated strong accuracy in predicting human behavior (Grossmann et al., 2023). As a concrete example, we present a political science survey experiment conducted by Fahey et al. (2023), which evaluates the effect of free speech framings on opposition to cancel culture among Americans. We provide simplified prompts here and refer readers to Appendix D.3 for the full LLM prompts.

1. **Extract participant information.** Extract the tuples $Z_i = (X_i, Y_i, A_i)$ for each participant i in the study. In Fahey et al. (2023), covariates include age, gender, ideology, income, and religion. The treatment represents a scenario where an Antifa protest is banned: for safety reasons only ($A = 0$), or for safety reasons and cancel culture ($A = 1$). The outcome is measured on a scale from 1 to 5, indicating the level of agreement with the statement: *"Cancel culture is a big problem in today's society."*
2. **Construct system prompts.** For each participant i, create a *persona* that matches X_i and guides the LLM in simulating participant responses. In this study, personas summarize the participant's demographics. The constructed persona is then used as the *system* prompt for the LLM; see Figure 8a for an example.
3. **Construct user prompts.** The *user* prompt includes the experimental treatment, the outcome question, and instructions to guide the LLM (see an example in Figure 8b). We prompt the LLM to generate a synthetic outcome for both conditions (treatment and control). The final instruction is sampled from a predefined pool to introduce variability in the LLM's responses; we provide examples in Appendix D.3.9.
4. **Simulate outcome responses.** Query the LLM using the user and system prompts. Validate that the responses are numeric and conform to the specified outcome scale. For experiments where multiple instructions are sampled, compute the average response.
5. **Estimate treatment effects.** Compute the confidence interval $\mathcal{C}^{\alpha}_{\text{H-AIPW}}$ following Algorithm 1. We find that using cross-fitting to the classic AIPW estimator is key for coverage in small-sample settings.

Example System Prompt

You are a 35-year-old female, politically Democrat, holding liberal views. Additionally, your religion is Christianity, and you once or twice a month attend religious services. You reside in a building with two or more apartments, and your household has a yearly income of $85,000 to $99,999.
You are responding to a scenario reflecting a debate involving college campus events and broader social issues.

(a)

Example User Prompt

Treatment: *A student organization denied Antifa's request for a rally, citing safety concerns due to altercations at similar events. Antifa plans to appeal the decision.*
Outcome question: Do you agree or disagree with the statement: *"Cancel culture is a big problem in today's society."* Choose an integer between 1 (definitely agree) and 5 (definitely disagree).
Instruction: Reflect on the scenario and use your reasoning to assign a value.

(b)

Figure 8: Examples of a system prompt and a user prompt used to generate synthetic responses in the study by Fahey et al. (2023).

D.3 Preprocessing of scientific studies and prompt design

In this section, we describe the preprocessing steps, selected outcomes, and control and treatment scenarios for the studies used in our experiments. We also provide an example prompt, including both system and user components, used to query the LLMs. The studies are sourced from the Time-sharing Experiments for the Social Sciences (TESS) repository, with findings published in peer-reviewed journals. These studies span various fields, demonstrating the versatility of our methodology.

D.3.1 Cancel Culture for Friends, Consequence Culture for Enemies: The Effects of Ideological Congruence on Perceptions of Free Speech (Fahey et al., 2023)

Abstract: Political scientists have long been interested in the effects that media framings have on support or tolerance for controversial speech. In recent years, the concept of cancel culture has complicated our understanding of free speech. In particular, the modern Republican Party under Donald Trump has made "fighting cancel culture" a cornerstone of its electoral strategy. We expect that when extremist groups invoke cancel culture as a reason for their alleged censorship, support for their free speech rights among Republicans should increase. We use a nationally representative survey experiment to assess whether individuals' opposition to cancel culture is principled or contingent on the ideological identity of the speaker. We show that framing free speech restrictions as the consequence of cancel culture does not increase support for free speech among Republicans. Further, when left-wing groups utilize the cancel culture framing, Republicans become even less supportive of those groups' free speech rights.

Data availability: The study is publicly available at: `https://www.tessexperiments.org/study/faheyS78`

Data pre-processing: The primary outcome variable is `CC_1`. The treatment condition is defined as `P_GROUP` $= 2$ (safety reasons + cancel culture), and the control condition is defined as `P_GROUP` $= 1$ (safety reasons). The following variables are included as covariates: `PARTYID7, IDEO, RELIG, ATTEND, GENDER, AGE, HOME_TYPE, INCOME`. The final processed dataset contains $n = 998$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 35-year-old male, politically Democrat, holding liberal views. Additionally, your religion is Christianity, and you once or twice a month attend religious services. You reside in a building with two or more apartments, and your household has a yearly income of $85,000 to $99,999. You are responding to a scenario reflecting a debate involving college campus events and broader social issues.

Treatment Condition:

We are now going to ask you to imagine you have read about the following scenario, describing a debate on a recent College Campus.

Local Group Denied Permit to Protest on Campus, Provoking Debate About "Cancel Culture"

A debate on the merits of free speech erupted recently when the student chapter of the controversial far-left group Antifa attempted to obtain a permit to conduct a demonstration on the main quad of Rutgers University in New Jersey. Citing safety concerns, the president of the organization in charge of Registered Student Organizations (RSOs) initially denied the organization the right to conduct their rally, arguing that their presence would endanger college students. They cited a recent incident in Berkeley, CA where three Antifa members and two bystanders were injured by rocks thrown in an altercation between the group and counter protesters. A member of the local Antifa group, Luke Vargas, is appealing the decision, arguing that the permit denial represented "cancel culture run amok," and the University was simply "afraid to hear the truth." When asked to comment, the University Ombudsman's Office promised that a final decision on whether the rally would be permitted would be made by this Thursday, three days before the march is scheduled to take place on Sunday.

Control Condition:

We are now going to ask you to imagine you have read about the following scenario, describing a debate on a recent College Campus.

Local Group Denied Permit to Protest on Campus

A debate on the merits of free speech erupted recently when the student chapter of the controversial far-left group Antifa attempted to obtain a permit to conduct a demonstration on the main quad of Rutgers University in New Jersey. Citing safety concerns, the president of the organization in charge of Registered Student Organizations (RSOs) initially denied the organization the right to conduct their rally, arguing that their presence would endanger college students. They cited a recent incident in Berkeley, CA where three Antifa members and two bystanders were injured by rocks thrown in an altercation between the group and counter protesters. A member of the local Antifa group, Luke Vargas, promised to bring an appeal to the desk of the University President. When asked to comment, the University Ombudsman's Office promised that a final decision on whether the rally would be permitted would be made by this Thursday, three days before the march is scheduled to take place on Sunday.

Question:

Generally speaking, do you agree or disagree with the following statement: "Cancel culture is a big problem in today's society." Reply using numbers between 1 (definitely agree) and 5 (definitely disagree).

D.3.2 Can Factual Misperceptions be Corrected? An Experiment on American Public Fears of Terrorism (Silverman et al., 2022)

Abstract: An American's yearly chance of being killed by a terrorist attack sits at about 1 in 3.5 million. Yet over 40% of the American public consistently believes that they or their family members are likely to be the victim of a terror attack. Can these inflated estimates of the risks of terrorism be brought closer to reality? With trillions of dollars spent on the War on Terror since 9/11, this question is not just theoretically but practically important. In order to investigate, we field a nationally representative survey experiment containing a brief vignette with corrective information about the actual risks of terrorism vs. other dangers facing Americans. Additionally, we vary whether there is a political elite endorsement accompanying the information, with either a Democratic politician, Republican politician, or senior military officer driving home the message.

Data availability: The study is publicly available at: https://tessexperiments.org/study/silverman1035

Data pre-processing: The primary outcome variable is `Q5`. The treatment condition is defined as `P_TESS031` $= 1$ (corrective information), and the control condition is defined as `P_TESS031` $= 0$ (no corrective information). The following variables are included as covariates: `PARTYID7, IDEO, RELIG, ATTEND, GENDER, AGE, RACETHNICITY, EDUC4, INCOME`. The final processed dataset contains $n = 503$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 33-year-old, ethnicity White, gender Male, strong Democrat. You hold very liberal views and college education. Additionally, your religion is Catholic, and you attend religious services nearly every week. Your household has a yearly income of $75,000 to $84,999. Your answer must be a single integer without additional text, in JSON format with a key-value pair.

Treatment Condition:

The number of people who say that acts of terrorism against Americans are imminent is up 3% from last year, according to a new poll released this week. In the wake of attacks in San Bernardino, Orlando, Paris, and London, the Pew Research Center found that 63% of Americans think major terrorist attacks are likely to occur soon on American soil. Government officials have echoed these concerns. "We are issuing a new advisory that the terror threat is now elevated across the country," said Undersecretary for Homeland Security Stephen Krause. "We have to remain vigilant and we have to stay alert. Terrorists can strike anytime, anywhere."

But does terrorism really pose a critical threat to us? Below is a figure showing the average American's risk of death from different sources. As can be seen, around 90 Americans are killed each year by terrorism on U.S. soil. This means the risk of being a victim of terrorism in a given year is about 1 in 3.5 million. In comparison, the risk of being killed by cancer is 1 in 540, the risk of being killed in a car accident is 1 in 8,000, and the chance of being killed by your own home appliances is 1 in 1.5 million. These numbers provide some essential context when thinking about the different threats to our public safety.

Control Condition:

The number of people who say that acts of terrorism against Americans are imminent is up 3% from last year, according to a new poll released this week. In the wake of attacks in San Bernardino, Orlando, Paris, and London, the Pew Research Center found that 63% of Americans think major terrorist attacks are likely to occur soon on American soil. Government officials have echoed these concerns. "We are issuing a new advisory that the terror threat is now elevated across the country," said Undersecretary for Homeland Security Stephen Krause. "We have to remain vigilant and we have to stay alert. Terrorists can strike anytime, anywhere."

Question:

How likely do you think it is that another terrorist attack causing large numbers of American lives to be lost will happen in the near future? Choose an integer between 1 (very likely) and 5 (not likely at all).

D.3.3 Accidental Environmentalists: Examining the Effect of Income on Positive Social Evaluations of Environmentally-Friendly Lifestyles (Kennedy & Horne, 2020)

Abstract: Many US households have adopted behaviors aimed at reducing their environmental impact. Existing scholarship examines antecedent variables predicting engagement in these pro-environmental behaviors. But little research examines the effect of making efforts to reduce environmental impact on positive evaluations. Based on our qualitative pilot data, we suspect that income may be an important factor in the extent to which green lifestyles earn social approval. We predict that a household that reduces its environmental impact will be viewed more positively if that household has a high (rather than low) income. We manipulate household income (high vs low) and proenvironmental behavior (green vs typical). We then measure participants' approval of the household, how socially close they feel to the household, as well as their evaluations of the household's competence, morality, and environmental commitment. This research allows us to identify the bases for social approval of green lifestyles and examine how social approval for a household's green lifestyle varies with that household's income.

Data availability: The study is publicly available at: https://tessexperiments.org/study/kennedy1017

Data pre-processing: The primary outcome variable is `Q5`. The treatment condition is defined as `P_TESS23 = 4` (green lifestyle), and the control condition is defined as `P_TESS23 = 2` (typical lifestyle). The following variables are included as covariates: `PartyID7`, `IDEO`, `ATTEND`, `GENDER`, `AGE`. The final processed dataset contains $n = 1276$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 45-year-old, lean Democrat, gender Female, and hold slightly conservative views. Additionally, you attend religious services several times a year. We are going to give you some information about a family. Please read the information very carefully, as we will be asking you questions about it. Your answer must be in JSON format with a single key-value pair.

Treatment condition:

A family with two children lives in a neighborhood nearby to yours. You chat with them sometimes when you see them in the neighborhood. As far as you can tell, they make a huge amount of money and seem to have plenty of extra money to spend. Their house is small and they often take public transit or walk to avoid driving. They also dry their clothes on a clothesline and don't have air conditioning in their home. This family has a much lower environmental impact than other people in their neighborhood.

Control condition:

A family with two children lives in a neighborhood nearby to yours. You chat with them sometimes when you see them in the neighborhood. As far as you can tell, they make very little money and seem to have no extra money to spend. Their house is small and they often take public transit or walk to avoid driving. They also dry their clothes on a clothesline and don't have air conditioning in their home. This family has a much lower environmental impact than other people in their neighborhood.

Question:

How much is the environment a high priority for this family? Choose an integer between 1 (not at all) and 11 (very much).

D.3.4 Beliefs about Racial Discrimination (Haaland & Roth, 2023)

Abstract: This paper provides representative evidence on beliefs about racial discrimination and examines whether information causally affects support for pro-black policies. Eliciting quantitative beliefs about the extent of hiring discrimination against blacks, we uncover large disagreement about the extent of racial discrimination with particularly pronounced partisan differences. An information treatment leads to a convergence in beliefs about racial discrimination but does not lead to a similar convergence in support of pro-black policies. The results demonstrate that while providing information can substantially reduce disagreement about the extent of racial discrimination, it is not sufficient to reduce disagreement about pro-black policies.

Data availability: The study is publicly available at: https://www.tessexperiments.org/study/Haaland874

Data pre-processing: The primary outcome variable is `Q2`. The treatment condition is defined as `GROUP = 1` (statistics of white-sounding and black-sounding names), and the control condition is defined as `GROUP = 2` (statistics of white-sounding names). The following variables are included as covariates: `PartyID7`, `INCOME`, `ATTEND`, `RELIG`, `GENDER`, `AGE`, `REGION9`, `RACETHNICITY`. The final processed dataset contains $n = 1539$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 60-year-old, politically Independent, gender Female, ethnicity Hispanic. Additionally, your religion is just Christian and you never attend religious services. You live in a state of the West South Central region. Your household has a yearly income of $30,000 to $34,999. You are responding to a survey experiment collecting data on people's beliefs about racial discrimination and whether these beliefs affect people's views on affirmative action policies.

Treatment condition:

Researchers from Harvard University conducted an experiment to study racial discrimination in the labor market. They did so by sending out fictitious resumes to help-wanted ads in Boston newspapers. The resumes were exactly the same except for one thing: the name of the job applicant. Half of the resumes had typically white-sounding names like "Carrie" and "Todd". The other half of the resumes had typically black-sounding names like "Tanisha" and "Kareem". The idea was to make sure that the applicants were seen as having identical qualifications, but that the employers would use the applicants' names to infer whether they were white or black. Resumes with white-sounding names had to be sent out on average 10 times to get one callback for an interview.

Further, the researchers found that resumes with black-sounding names on average had to be sent out 15 times to get one callback for an interview. Since resumes with white-sounding names on average only had to be sent out 10 times to get one callback for an interview, this means that employers were 50 percent more likely to give callbacks to applicants with white-sounding names compared to applicants with black-sounding names.

Control condition:

Researchers from Harvard University conducted an experiment to study racial discrimination in the labor market. They did so by sending out fictitious resumes to help-wanted ads in Boston newspapers. The resumes were exactly the same except for one thing: the name of the job applicant. Half of the resumes had typically white-sounding names like "Carrie" and "Todd". The other half of the resumes had typically black-sounding names like "Tanisha" and "Kareem". The idea was to make sure that the applicants were seen as having identical qualifications, but that the employers would use the applicants' names to infer whether they were white or black. Resumes with white-sounding names had to be sent out on average 10 times to get one callback for an interview.

Question:

In the United States today, do you think that racial discrimination against blacks in the labor market is a serious problem? Reply with a JSON numerical answer using one of these numbers: 1 (A very serious problem), 2 (A serious problem), 3 (A problem), 4 (A small problem), or 5 (Not a problem at all).

D.3.5 To Do, to Have, or to Share? Valuing Experiences and Material Possessions by Involving Others (Caprariello & Reis, 2013)

Abstract: Recent evidence indicates that spending discretionary money with the intention of acquiring life experiences-events that one lives through-makes people happier than spending money with the intention of acquiring material possessions-tangible objects that one obtains and possesses. We propose and show that experiences are more likely to be shared with others, whereas material possessions are more prone to solitary use and that this distinction may account for their differential effects on happiness. In 4 studies, we present evidence demonstrating that the inclusion of others is a key dimension of how people derive happiness from discretionary spending. These studies showed that when the social-solitary and experiential-material dimensions were considered simultaneously, social discretionary spending was favored over solitary discretionary spending, whereas experiences showed no happiness-producing advantage relative to possessions. Furthermore, whereas spending money on socially shared experiences was valued more than spending money on either experiences enacted alone or material possessions, solitary experiences were no more valued than material possessions. Together, these results extend and clarify the basic findings of prior research and add to growing evidence that the social context of experiences is critical for their effects on happiness.

Data availability: The study is publicly available at: `https://www.tessexperiments.org/study/caprariello130`

Data pre-processing: The primary outcome variable is `Q7A`. The treatment condition is defined as `XTESS086 = 1` (spend money with people), and the control condition is defined as `XTESS086 = 2` (spend money alone). The following variables are included as covariates: `XPARTY7`, `XREL1`, `XREL2`, `XIDEO`, `PPAGE`, `PPGENDER`. The final processed dataset contains $n = 397$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 53-year-old, not so strong Republican, gender Male, and hold moderate views. Additionally, regarding religion you are Buddhist and you more than once a week attend religious services. You are responding to a survey on how you spend your discretionary money. Your answer must be a single integer without additional text, in JSON format with a key-value pair.

Treatment condition:

We are interested in ways you spend your discretionary money. Discretionary money refers to money that is spent on anything that is NOT essential to basic activity (that is, essentials refer to things like tuition and textbooks, groceries, transportation, rent, gas for a car, health care, etc.). We'd like you to answer the questions that follow for money that you spent on something discretionary. Please think of the last time you spent at least $10 (but no more than $10,000) of your discretionary money in order TO DO SOMETHING WITH AT LEAST ONE OTHER PERSON. The primary focus of this expense should have been on an activity – doing something with at least one other person – and not on buying something that could be kept. Maybe you bought tickets to see a movie with some people, maybe you paid to visit an art museum with friends, maybe you and some other people went to a spa together . . . any of these would be legitimate examples of spending money to do something with others.

Control condition:

We are interested in ways you spend your discretionary money. Discretionary money refers to money that is spent on anything that is NOT essential to basic activity (that is, essentials refer to things like tuition and textbooks, groceries, transportation, rent, gas for a car, health care, etc.). We'd like you to answer the questions that follow for money that you spent on something discretionary. Please think of the last time you spent at least $10 (but no more than $10,000) of your discretionary money in order TO DO SOMETHING BY YOURSELF. The primary focus of this expense should have been on an activity – doing something by yourself – and not on buying something that could be kept. Maybe you bought a ticket to see a movie by yourself, maybe you paid to enter an art museum, maybe you went to a spa by yourself . . . any of these would be legitimate examples of spending money to do something by yourself.

Question:

Think about the last time you used your possession. To what extent did it help you feel loved and cared about? Reply with a JSON numerical answer using one of these numbers: 1 (not at all), 2 (slightly), 3 (moderately), 4 (very), or 5 (extremely).

D.3.6 Onset and Offset Controllability in Perceptions and Reactions to Home Mortgage Foreclosures (Brandt, 2013)

Abstract: The circumstances and rhetoric surrounding home foreclosures provide an ideal and timely backdrop for an extension of research on attributional judgments. While people face foreclosure for many reasons, the current debate surrounding the mortgage crisis has highlighted reasons that are either onset or offset controllable; that is, the initial cause, or the subsequent solution may be seen as controllable.In the current study, I examine how people use attributional evidence from multiple time points to determine affective reactions and helping intentions for people undergoing foreclosure, as well as ideological differences in these attributional processes. Participants read about people who were undergoing foreclosure for onset and offset controllable or uncontrollable reasons and then answer questions about their perceptions of these targets. The results suggested that both onset and offset controllable information contributed to the emotional reactions and helping intentions of the participants with the participants experiencing more negative affect and less helping intentions when the target was in a controllable onset or offset situation. Conservatives primarily relied on onset controllability information to decide who should receive government aid, while liberals updated their initial attributions with offset controllability information.

Data availability: The study is publicly available at: `https://www.tessexperiments.org/study/brandt708`

Data pre-processing: The primary outcome variable is `Q7`. The treatment condition is defined as `XTESS003 = 1` (family can afford the mortgage), and the control condition is defined as `XTESS003 = 2` (family might not afford the mortgage). The following variables are included as covariates: `XPARTY7`, `XREL1`, `XREL2`, `PPAGE`, `PPGENDER`. The final processed dataset contains $n = 624$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 75-year-old, not so strong Democrat, gender Female. Additionally, regarding religion you are a Muslim and you once a week attend religious services. You are responding to a survey on perceptions towards people who are facing foreclosure. Your answer must be a single integer without additional text, in JSON format with a key-value pair.

Treatment condition:

Recently the growing number of home foreclosures has put a strain on the financial system, which has weakened the United States economy. Foreclosure occurs when a person is behind on home mortgage payments to their bank and the bank decides to repossess (i.e., take back) the home. People may go into foreclosure for a variety of reasons. We are interested in your perceptions towards people who are facing foreclosure. In the following section you will be presented with a situation that describes some people facing foreclosure. Please carefully read the situation and answer the following questions about your reactions to the situation. Some people have a large monthly mortgage payment because they wanted to purchase a larger house than they needed. Now they are facing foreclosure because they do not want to continue paying the mortgage, even though they are able to afford the payments.

Control condition:

Recently the growing number of home foreclosures has put a strain on the financial system, which has weakened the United States economy. Foreclosure occurs when a person is behind on home mortgage payments to their bank and the bank decides to repossess (i.e., take back) the home. People may go into foreclosure for a variety of reasons. We are interested in your perceptions towards people who are facing foreclosure. In the following section you will be presented with a situation that describes some people facing foreclosure. Please carefully read the situation and answer the following questions about your reactions to the situation. Some people have a large monthly mortgage payment because they wanted to purchase a larger house than they needed. Now they are facing foreclosure because the primary income earner in the household lost their job due to their company closing and they can no longer afford payments.

Question:

Do you strongly oppose or strongly support the following statement: The government should offer help (e.g., time, money, resources, etc.) in an effort to help people in this situation. Reply with an integer from 1 (Strongly Oppose) to 7 (Strongly Support), where 4 is a Neutral stance.

D.3.7 Understanding White Identity Management in a Changing America (Shuman et al., 2024)

Abstract: This paper examines how White Americans manage their identity amidst societal shifts using a new measure of advantaged identity management, representative data (N = 2648), and latent profile analysis. The findings reveal five subgroups of White Americans, each managing their identity differently. Four profiles correspond to the main advantaged identity management strategies (defend, deny, distance, dismantle), with a fifth using strategies flexibly. Of 15 predictions regarding how valuing hierarchy, meritocracy, and egalitarianism predict profile membership, 13 were supported. These profiles show contrasting attitudes toward social change, with defender-deniers opposing, denier-distancers moderately opposing, distancers remaining neutral, and dismantlers supporting change. These findings provide some of the first empirical evidence for a theorized model of white identity management and suggest that how White Americans manage their identity has important implications for social change.

Data availability: The study is publicly available at: https://www.tessexperiments.org/study/melin1066

Data pre-processing: The primary outcome variable is `Q5D`. The treatment condition is defined as `RND_01 = 1` (disadvantage back people), and the control condition is defined as `RND_01 = 0` (advantage white people). The following variables are included as covariates: `AGE`, `GENDER`, `RACETHNICITY`, `EDUC5`, `REGION9`, `IDEO`, `PartyID7`, `RELIG`, `ATTEND`, `INCOME`. The final processed dataset contains $n = 1623$ observations.

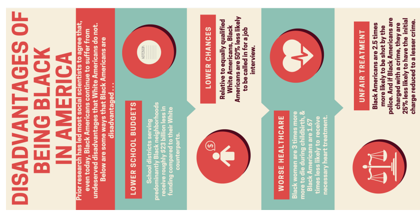
Prompting details: An example prompt is provided below.

Example Prompt

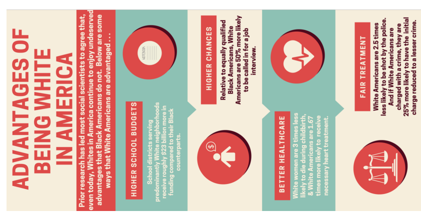
System Prompt:

> You are a 41-year-old individual with gender Male, ethnicity Asian, and with Bachelor's degree education. You live in a state of the New England region. You hold Moderate views and are not so strong Democrat. Additionally, your religion is Atheist and you attend religious services never. Your household has a yearly income of $175,000 to $199,999.

Treatment condition: The general purpose of this study is to examine the attitudes of people regarding social issues in America today. You will now be presented with an infographic:



Control condition: The general purpose of this study is to examine the attitudes of people regarding social issues in America today. You will now be presented with an infographic:



Question:

> Rate the extent to which you agree with the following statement from 1 (STRONGLY DISAGREE) to 7 (STRONGLY AGREE): "There should be large scale criminal justice reform to address racial inequalities in the justice system." Your answer must be in JSON format with a single key-value pair.

D.3.8 TESTING A THEORY OF HYBRID FEMINITY (MELIN & MERLUZZI, 2022)

Abstract: Although men experience advantages working in highly feminized occupations, they are commonly stigmatized as lesser men by outsiders—the people they meet outside of their occupations—for doing "women's work." This experiment is designed to assess whether a woman who has worked in a hypermasculine occupation would similarly be stigmatized as a lesser woman by workers outside of her hypermasculine occupation, or alternatively, whether she would be viewed more favorably by such outsiders for doing "men's work." Specifically, this study aims to develop and empirically test a theory of hybrid femininity, which specifies the conditions under which hypermasculinity as signaled through occupation creates status and reward distinctions among women in external labor markets. The experiment asks respondents to provide recommended compensation and status ratings for a woman candidate while manipulating the gender-typing of her occupational history as well as her intended target job. By disentangling the underlying mechanisms driving these predicted status and reward differences, this study seeks to shed light on how gender inequality persists, even among women, through the privileging of masculinity over femininity, with important implications for the labor market and society at large.

Data availability: The study is publicly available at: https://www.tessexperiments.org/study/melin1066

Data pre-processing: The primary outcome variable is `Q7_1`. The treatment condition is defined as `P_41 = 3` (applicant has experience in the Army), and the control condition is defined as `P_41 = 6` (applicant has experience in the Cosmetics industry). The following variables are included as covariates: `P_IDEO`, `P_ATTEND`, `P_RELIG`, `RELIG`, `GENDER`, `AGE`, `REGION9`, `RACETHNICITY`, `INCOME`, `P_PARTYID`. The final processed dataset contains $n = 545$ observations.

Prompting details: An example prompt is provided below.

Example Prompt

System Prompt:

You are a 30-year-old, politically Independent, gender Male, ethnicity Hispanic. Your ideology is slightly liberal. Additionally, your religion is Protestant and you about once a month attend religious services. You live in a state of the Pacific region. Your household has a yearly income of $85,000 to $99,999. This task is part of a larger study on the design of Human Resources (HR) recruiting practices to pre-screen job applicants.

Treatment condition:

Please imagine you work for a prominent management consulting company. You will be provided with a job description and an applicant's résumé who is applying for a Senior Manager position. After thoroughly reviewing the job description and the applicant's résumé, you will be asked to provide your immediate and uncensored opinion. Job description for your review:

`[Job description, check original paper for details]`

Applicant's résumé for your review:

Name: Amy Decker Motivated Project Manager with 5 years of experience working in military and defense. Education: Rutgers University (New Brunswick, NJ), May 2017 (Graduated) B.A. in Business Administration, GPA: 3.72/4.00 Work Experience: U.S. Army Project Manager (Active-duty Enlisted), 2014 - Present Fort Dix Military Base (Fort Dix, NJ) - Plan and track progress of entire life-cycle of military and defense projects. - Build and maintain project plans, including actual and forecasted activities and timelines. - Ensure project staffing and timely communications throughout project lifecycle. - Identify and manage project risks. Skills and Interests: Computer: Proficient in Microsoft Office (including Word, Excel, Outlook, and PowerPoint). Interests: Running and traveling.

Control condition:

Please imagine you work for a prominent management consulting company. You will be provided with a job description and an applicant's résumé who is applying for a Senior Manager position. After thoroughly reviewing the job description and the applicant's résumé, you will be asked to provide your immediate and uncensored opinion. Job description for your review:

`[Job description, check original paper for details]`

Applicant's résumé for your review:

Name: Amy Decker Motivated Project Manager with 5 years of experience working in military and defense. Education: Rutgers University (New Brunswick, NJ), May 2017 (Graduated) B.A. in Business Administration, GPA: 3.72/4.00 Work Experience Cosmetics Project Manager 2014 - Present Precious Cosmetics (Lodi, NJ) - Plan and track progress of entire life-cycle of cosmetics and beauty product projects. - Build and maintain project plans, including actual and forecasted activities and timelines. - Ensure project staffing and timely communications throughout project lifecycle. - Identify and manage project risks. Skills and Interests: Computer: Proficient in Microsoft Office (including Word, Excel, Outlook, and PowerPoint). Interests: Running and traveling.

Question:

On a scale from 1 "Not at all" to 7 "Extremely", to what extent do you perceive this applicant as MASCULINE.

D.3.9 INTRODUCING VARIABILITY IN MULTI-PROMPT EXPERIMENTS

The user prompts described in the previous section include a final question or instruction sampled from a predefined pool to introduce variability in the multi-prompt settings. Below are some examples of such instructions:

- “Consider all relevant factors and place this on the scale.”
- “Reflect on the scenario and use your reasoning to assign a value.”
- “From your understanding of the situation, quantify this feeling.”
- “Given your insights and the context described, provide your evaluation.”
- “With the provided details in mind, rate your feeling on the scale.”
- “Consider all the information and your perspective to choose a suitable score.”
- “Evaluate the feeling here and align a number with your reasoning.”
- “Use the scale provided and your judgment to determine your feeling.”
- “Judge this scenario thoughtfully, considering the context and the details shared.”
- “Reflect on the key aspects provided and numerically assess your feeling.”