Boosting Randomized Control Trials with Language Models as Synthetic Units

Shiv Shankar
sshankar@cs.umass.edu
University of Massachusetts
USA

Madalina Fiterau
mfiterau@cs.umass.edu
University of Massachusetts
USA

Abstract

A/B Testing or Randomized Control Trials (RCTs) are a keystone of data-driven decision making. However RCT's can be expensive to implement and often need to be run for a long time to get credible estimates. This is especially problematic in online A/B testing, where businesses often run thousands of such experiments in parallel. In this context, large language models with their ability to generalize across tasks might serve as useful proxies for real user behaviour. In this work, we present a framework to leverage blackbox LLMs to augment data from real A/B tests to estimate treatment effects. Our method is guaranteed to reduce estimators asymptotic variance irrespective of the LLM quality. We support our theoretical analysis with experiments on multiple real datasets.

CCS Concepts

• **Applied computing** → **Electronic commerce**; **Marketing**; • **Computing methodologies** → *Machine learning algorithms*; • **General and reference** → *Evaluation*; *Experimentation*.

Reference Format:
Shiv Shankar and Madalina Fiterau. 2024. Boosting Randomized Control Trials with Language Models as Synthetic Units. In *Proceedings of (Neurips '24).* , 12 pages.

1 Introduction

The increasing amount of digital content consumption has intensified competition among businesses to capture and retain user attention. In this landscape, optimizing content and user experience is critical for driving engagement and achieving business objectives. For instance, e-commerce platforms and streaming services continuously experiment with personalized recommendations and interface designs to enhance user satisfaction and retention [Li et al., 2010]. Online bandit based approaches are also used [Larsen et al., 2024] Schwartz et al. .

While randomized A/B tests remain the gold standard for evaluating such interventions, this is inefficient for many media applications. This is because as news and trends have short lifetimes and might become irrelevant by the time a standard A/B test finishes. This problem is further aggravated due to significant democratization of content creation, which has led to shorter feedback cycles and increasing amount of content which needs to be experimented. As a simple numerical example, consider the clickthrough rates (CTRs) for display ads. Consider algorithm A which leads to CTRs of 0.5% and algorithm B with CTRs of 0.525%. A test with 90% p-value and 80% power would require around 300k displays to assess that B has a higher CTR than A.

In this context, large language models (LLMs) and other foundation models can provide powerful tools for improving the efficiency of these methods. LLMs have been demonstrated to have significant potential for processing natural language text, following human instructions and generating high-quality responses [OpenAI, 2024]. This has spurred their use in many applications such as tool learning [Qu et al., 2024] and information retrieval [Zhu et al., 2024b]. LLMs can also provide useful responses for evaluating content. For example, Li et al. [2024] have investigated similarity between LLM-generated perceptual maps and human survey responses produce in car branding analysis. Their study shows a high agreement rate of over 75%. Such methods can provide rapid, low-cost simulations of user behavior and intervention outcomes. However, the reliability of these experiments hinges on the accuracy of the model's predictions, which must faithfully match real-world user responses. For example, a foundation model trained on user reviews may not be very good at predicting news attractiveness. Improving alignment[Goli and Singh, 2024] and calibration [Xiong et al., 2023, Wei et al., 2024] between AI-generated and real-world user behaviour is an active area of research. This trade-off between cost-efficiency and predictive validity underscores the need for robust methodologies that provide valid results while benefiting with integration with such models.

Contribution. In this paper, we propose a framework that can incorporate ratings or other related responses provided by foundation models on A/B tests along with real experimental data to provide efficient estimates of treatment effects. Our proposed method can handle domain shift between the foundation model and experimental data. Our method provides consistent estimators by leveraging the ideas behind 'prediction-powered inference' [Angelopoulos et al., 2023a]. Empirical results on real A/B test datasets shows that estimator is significantly more efficient than the standard estimator using only experimental data.

2 Preliminaries

2.1 Notation

We are given a population of n units. At each unit i we have a treatment assignment $Z_i \in= \{0, 1\}$ which represents whether the unit is treated or not. We use the Neyman potential outcome framework [Neyman, 1923, Rubin, 1974], and denote by $Y_i(z)$ the potential outcome for each $z \in \{0, 1\}$.

Neurips '24, December 2-December 9, 2024, Vancouver, Canada.

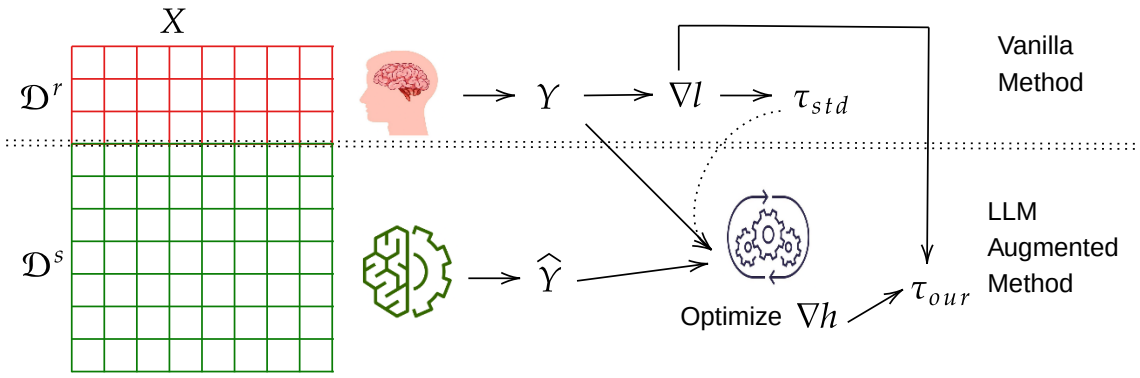


Figure 1: We illustrate two datasets, one real and one obtain using in-silico simulation with LLM. Normally one estimates τ with only real data. We develop a method to use the additional synthetic data to improve estimation. For this we learn an auxiliary function h which is optimized to minimize the variance of the estimator.

We will consider that the data has been obtained via randomized design, each unit i gets allotted the treatment $z_i = 1$ independently with probability $p \in (0, 1)$. For a website performing A/B tests on visitors, this easy to implement, and satisfies standard randomization and positivity assumption in causal inference.

The desired causal effect is the mean difference between the outcomes when $z_i = 1\,\forall i$ and when $z_i = 0\,\forall i$. Under the aforementioned notations, this causal effect is given by:

$$\tau = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - \frac{1}{n}\sum_{i=1}^{n} Y_i(0) \tag{1}$$

A straightforward estimate of τ is $\sum_i \frac{1}{n} Y_i(\frac{Z_i}{p} - \frac{1-Z_i}{1-p})$. This is effectively just the mean outcome for the treated units minus the mean outcome for untreated units, hence it is also called the DM estimate. While the above estimate is unbiased and consistent, usually estimator is done via a regression model using additional covariates available for the units. Let the covariates be denoted by X_i , a common model is the following ANCOVA model

$$Y_i \sim \beta_0 + \beta_X X_i + \hat{\tau} Z_i + \epsilon.$$

If randomization is perfect, then the estimate $\hat{\tau}$ is a consistent estimate of the effect. Usually this method is preferred over DM because of significantly reduced variance.

Another common method when dealing with outcomes like click-through rates (CTRs) is to use logistic regression of the form $Y_i \sim \sigma(\beta_0 + \beta_X X_i + \hat{\tau} Z_i)$ where sigma is the sigmoid function. This also generalizes to a multi-category testing via a multinomial logit model. In such cases the effect corresponds to the change in log-odds of clicking.

General Estimation. The above mentioned models (and many others) can be written in terms of an optimization objective The parameter of interest θ^* is usually defined as

$$\theta^\star = \underset{\theta\in\Theta}{\operatorname{argmin}}\, \mathbb{E}[l_\theta(X, Y)], \tag{2}$$

where for practical estimation, the expectation is replaced by empirical mean. In the case of the ANCOVA model we have $l = ||Y - (\beta_0 + \beta_X X + \hat{\tau} Z)||^2$; whereas for logistic regression it is the likelihood loss. We will assume that l is a convex loss function so that θ^* is unique and well defined. We will refer to gradients of a function with respect to estimating parameters θ as $\nabla_\theta f$. When clear in context we will skip the θ on the gradient operator for notational convenience. However our framework also extends to the general estimating equation approach, as described later.

As mentioned earlier, we want to use simulations or treatment ratings by LLMs instead of conducting an actual test on users. Thus along with the n real units for which we observe X_i, Y_i, we have an additional N synthetic units which are obtained by using an LLM to act as users. In other applications it might be using a foundation model to annotate or rate certain inputs. We shall label these synthetic outcomes as $\hat{Y}$. Thus we have N samples of $X_i, \hat{Y}_i$ for $i \in \{n, \cdots, N\}$. We will also assume that we can ask the LLM to also provide $\hat{Y}$ on the real data as well.

2.2 Residual Learning and Control Variates

Since data collection can be costly or time-consuming, leveraging additional variables $\hat{Y}$ that are correlated with the true outcome Y provides a practical strategy to increase the effective sample size. This has become an increasingly important research direction as powerful general purpose ML models have been developed [Angelopoulos et al., 2023c, Wasserman and Lafferty, 2007, Chakrabortty et al., 2022].

A straightforward method is to replace Y by $\hat{Y}$ in Equation (2) ; however, unless $\hat{Y}$ and Y are perfectly correlated, this need not provide a statistically valid answer. A more technically sound approach will account for the relationship between $\hat{Y}$ and Y. For instance, a standard method is to model the joint distribution of $\hat{Y}$ and Y conditional on XX and use this model to impute missing Y values [Tang and Qin, 2012]. Another common technique is residual learning [Galpin and Hawkins, 1984, Pierce and Schafer, 1986]. These methods leverage $\hat{Y}$ in the regression model with the idea that $Y - \hat{Y}$ (or more generally the residual unexplained by $\hat{Y}$) is likely easier to model than Y itself. This can be extended to doubly robust

estimation methods [Lin et al., 2012, Han and Wang, 2013]. One weakness of these residual methods, was that these need both the true outcome Y, and predicted values $\hat{Y}$ on all points, i.e. they in general cannot utilize synthetic units.

2.3 Prediction Powered Inference

With the advent of powerful ML models, researchers are working on methods to leverage purely synthetic data obtained with ML models along with real data [Wang et al., 2020, Gronsbell et al., 2024, Motwani and Witten, 2023]. These approaches called Prediction Based Inference (PBI) [Wang et al., 2020] or Prediction Powered Inference (PPI) [Angelopoulos et al., 2023a] date back to [Robins et al., 1994, Chen and Chen, 2000]. These can be considered to be related to residual learning and doubly robust methods, in that they rely on using a mean zero variable to reduce variance of the estimator. A key difference however arises between these methods and residual or doubly robust methods, in that PPI usually include additional data.

Similar to our setting, PPI methods assume access to two datasets of i.i.d points, one labeled another unlabeled. An ML model f then provides additional predictions $\hat{Y}$. In general , f is used to capture information about Y which is available to us, but is hard to model with statistical guarantees.

While [Angelopoulos et al., 2023a] defined PPI in terms of residuals, many methods in current PPI literature [Angelopoulos et al., 2023c,b, Gronsbell et al., 2024] can be written in the following form:

$$\hat{\theta}^{\text{PPI}} = \underset{\theta}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i, Y_i) - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i, \hat{Y}_i) - \frac{1}{N}\sum_{i=n+1}^{n+N} l_\theta(X_i, \hat{Y}_i)\right)}_{f\text{-dependent}}. \tag{3}$$

Intuitively, since $X_i, \hat{Y}_i$ are iid between $i \in \{1, \cdots, n\}$ and $i \in \{n+1, \cdots, N\}$, the terms in is a mean zero term. Thus if n/N remains constant and $N \to \infty$, the additional term has zero influence and the objective is the same as $\mathbb{E}[l_\theta(X, Y)]$. Thus $\hat{\theta}^{\text{PPI}}$ is asymptotically convergent to θ^*. For more details on the behaviour of $\hat{\theta}^{\text{PPI}}$, we refer the readers to Angelopoulos et al. [2023b].

There are other extensions to this basic framework of PPI which we discuss later in related work.

3 Our Method

In this section, we provide an extension to the basic PPI method. We will then show how to adjust the method to provide asymptotically efficient estimators. Finally we will use this insight to provide a more practical estimation method. We will use the ANCOVA equation as our running example, but the result is more general.

3.1 Extending PPI

The fundamental insight of PPI is the addition of a mean zero term (to the loss or the estimating equation). Thus we suggest the following more general PPI framework

$$\hat{\theta}_g = \underset{\theta}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} l_\theta(X_i, Y_i) - \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} h_\theta(X_i, \hat{Y}_i) - \frac{1}{N}\sum_{i=n+1}^{n+N} h_\theta(X_i, \hat{Y}_i)\right)}_{f\text{-dependent}}. \tag{4}$$

The key difference from PPI approach is not to have the mean zero term be dependent on l but instead to have an arbitrary function h. Placing $h = l$ recovers the PPI objective, but for a wide family of functions h, the above estimator is consistent [1]. If the overall objective is convex, the optimal solution to the above equation is identified and can be obtained by solving the following first order optimality condition:

$$\frac{1}{n}\sum_{i=1}^{n} \nabla l_\theta(X_i, Y_i) = \frac{1}{n}\sum_{i=1}^{n} \nabla h(X_i, \hat{Y}_i) - \frac{1}{N}\sum_{i=n+1}^{N+n} \nabla h(X_i, \hat{Y}_i) \tag{5}$$

Based on classical results from M-estimation methods, if θ^* is the unique solution to $\mathbb{E}[\nabla l_\theta(X, Y)] = 0$, and some mild conditions on l and h, the solution $\hat{\theta}_h$ to empirical FOC condition converges to θ^* and is asymptotically normal. Its distributional properties are summarized in the following result

Proposition 1. *Assume that $n/N = r$ and that both the combined objective function in* (3) *as well as l are strictly convex. Then, $\sqrt{n}(\hat{\theta}_h - \theta^\star)$ converges in distribution to a mean zero Gaussian random variable with covariance Σ_h given by*

$$H^{-1}\left(r\text{Var}(\nabla h(X, \hat{Y})) + \text{Var}(\nabla l(X, Y) - \nabla h(X, \hat{Y}))\right) H^{-1}.$$

where $H = \mathbb{E}[\nabla^2 l_{\theta^\star}(X, Y)]$.

Proof. The estimating equation is:

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(\nabla l_\theta(X_i, Y_i) - \nabla h_\theta(X_i, \hat{Y}_i)\right) + \frac{1}{N}\sum_{i=n+1}^{n+N} \nabla h_\theta(X_i, \hat{Y}_i) = 0.$$

Taking expectations, we get:

$$\mathbb{E}[G_n(\theta)] = \mathbb{E}[\nabla l_\theta(X, Y)] - \cancel{\mathbb{E}[h_\theta(X, \hat{Y})]} + \cancel{\mathbb{E}[h_\theta(X, \hat{Y})]} = \mathbb{E}[\nabla l_\theta(X, Y)].$$

At $\theta = \theta^*$, $\mathbb{E}[\nabla l_{\theta^*}(X, Y)] = 0$. Thus under weak regularity conditions (see [Ross, 2011]) $\hat{\theta} \xrightarrow{p} \theta^*$.

Now expanding $G_n(\hat{\theta})$ around θ^* and applying Taylor's theorem:

$$0 = G_n(\theta^*) + \nabla G_n(\theta^*)(\hat{\theta} - \theta^*) + o(1),$$

which gives

$$\sqrt{n}(\hat{\theta} - \theta^*) \approx -\left(\nabla G_n(\theta^*)\right)^{-1} \sqrt{n}\, G_n(\theta^*).$$

Further as $n \to \infty$:

$$\nabla G_n(\theta^*) \xrightarrow{p} \mathbb{E}[\nabla^2 l_{\theta^*}(X, Y)] = H,$$

since the h_θ terms cancel in expectation.

[1] Difference choice of h leads to different estimators proposed in PPI literature

The variance of $G_n(\theta^*)$ is:

$$\begin{aligned}\mathrm{Var}(G_n(\theta^*)) &= \frac{1}{n}\mathrm{Var}(\nabla l_{\theta^*} - \nabla h_{\theta^*}) + \frac{1}{N}\mathrm{Var}(\nabla h_{\theta^*}) \\ &= \frac{1}{n}\left(\mathrm{Var}(\nabla l_{\theta^*} - \nabla h_{\theta^*}) + r\,\mathrm{Var}(\nabla h_{\theta^*})\right).\end{aligned}$$

By CLT we have, $\sqrt{n}\,G_n(\theta^*) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, where:

$$\Sigma = \mathrm{Var}(\nabla l_{\theta^*} - \nabla h_{\theta^*}) + r\,\mathrm{Var}(\nabla h_{\theta^*}).$$

Thus:

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} H^{-1}\sqrt{n}\,G_n(\theta^*) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}\Sigma(H^{-1})\right).$$

where in the last line we use the fact that H being a hessian is symmetric. □

One advantage of moving from the common PPI-variants to Equation 4 is important when considering using LLMs for synthetic simulations. LLMs primarily provide textual outputs, and in most cases Y is not text. While LLMs can be prompted to produce numerical scores it can be very biased [Xiong et al., 2023]. On the other hand LLM as a judge framework [Zheng et al., 2023] has shown promise in directly choosing answers or rating options [Huang et al., 2024a]. Thus as long as we can choose a suitable h, we can input a non-numeric $\hat{Y}$ such as text based or discrete values.

3.2 Optimal Estimation

Another key advantage of replacing l_i in the $f-$dependent term with h, is that it provides a handle to influence different properties of the estimator $\hat{\theta}$. We propose to choose h so as to minimize the variance of the estimator. For this purpose we can use the asymptotic variance structure of the estimator from 2 and minimize this asymptotic variance.

One point to note here is that the asymptotic variance depends on h via its gradient ∇h and not directly on h itself. This implies that h need not be uniquely determined, and it is not immediately clear whether such an optimal function h will even exist. However, as we will demonstrate, a specific form of h can be constructed to satisfy our requirement of minimizing variance. Before introducing this form, we first extend our approach to the general estimating equation (GEE) framework to provide motivation and context for our choice of h.

Extension to GEE framework. Our approach remains applicable even in the generalized estimating equation (GEE) or generalized method of moments (GMM) framework. In such a framework case the optimal θ (labeled θ^*_{EE})is defined as the parameter satisfying the following equation:

$$\mathbb{E}[U_\theta(X, Y)] = 0 \Rightarrow \frac{1}{n}\sum_{i}^{n} U_\theta(X_i, Y_i) = 0 \tag{6}$$

In practice, estimation is done by replacing true expectations with empirical means, and solving the (6) for θ. For the standard loss optimization method U is given by ∇l_θ.

Under our proposed framework, estimating equation is modified as follows

$$\frac{1}{n}\sum_{i=1}^{n}\left(U_\theta(X_i, Y_i) - \eta_\theta(X_i, \hat{Y}_i)\right) + \frac{1}{N}\sum_{i=n+1}^{n+N}\eta_\theta(X_i, \hat{Y}_i) = 0, \tag{7}$$

where η_θ is a suitably chosen vector valued function. Let the optimal solution to be given by θ^*_{EE} and the empirical estimate be $\hat{\theta}_{EE}$. Then similar to Proposition 1, the asymptotic variance of the $\hat{\theta}_{EE}$ is given by

$$\Sigma_\eta = H^{-1}\left(r\mathrm{Var}(\eta(X, \hat{Y})) + \mathrm{Var}(U(X, Y) - \eta(X, \hat{Y}))\right)(H^{-\top})$$

where $H = \mathbb{E}[\nabla U]$ is the Jacobian of U and is assumed to be non-singular definite matrix.

Returning to the optimization-based case, we note that the GEE framework allows for arbitrary vector functions $\eta(X_i, \hat{Y}_i)$ as free variables. Motivating from this we propose take h to be $h(X_i, \hat{Y}_i) = \theta^T\eta_\phi(X_i, \hat{Y}_i)$, where η is another function. Since in such a case h is a linear function, as long l is convex the overall objective in 4 remains convex.

While our primary goal is to minimize the variance of the desired estimator, the unique structure of Σ_h allows us to simultaneously minimize the variance of all the parameters/components of θ. Additionally for this choice, there is indeed a well defined function h which will attain this optimum value. We make this formal in the following proposition

Proposition 2. *Assume the conditions for Proposition 1. Next set* $h(X, \hat{Y}) = \frac{1}{r+1}\theta^T\mathbb{E}[\nabla l(X, Y)|X, \hat{Y}] = h^*$. *Then the asymptotic variance of* θ^* (Σ_{h^*}) *is given by* $H^{-1}_{\theta^\star}\left(\mathrm{Var}(\nabla l_{\theta^\star}(X, Y)) - \frac{1}{1+r}\mathrm{Var}(\mathbb{E}[\nabla l_{\theta^\star}(X, Y)|X, \hat{Y}])\right)H^{-}_{\theta}$ *Additionally,* $\Sigma_{h^*} \preceq \Sigma_h$ *for any other function* h, *where* $\preceq$ *is the Loewner order [Bhatia, 2013] among positive semi definite matrices.*

For the ANCOVA model, $\nabla l(X, Y) = (Y - \mathbb{E}[Y|X])$ which is just the residual of $\mathcal{R}$ of Y on X. The optimal function h is then the same as $\mathbb{E}[\mathcal{R}|X, \hat{Y}]$ which is also the form which sibling regression models take [Shankar et al., 2020]. This is also true in a wider family of models including GLM and Exponential family models. For these family of models, our proposal takes the same form as sibling regression [Schölkopf et al., 2016] and SUR frameworks [Zellner, 1962], connecting these to a wider family of methods used for variance reduction.

This form of optimal h also provides a partial idea for why the original PPI method [Angelopoulos et al., 2023a] can be very effective. Angelopoulos et al. [2023a] use $\nabla h = \nabla l(X, \hat{Y})$ whereas the optimal $\nabla h = \mathbb{E}[\nabla l(X, Y)|X, \hat{Y}]$. Thus both are of the form of the gradient of l. The difference arises from how $\hat{Y}$ effects the gradient.

We note that despite the easy to optimize objective, the actual optimization is tricky because θ^* is unknown. To solve this issue, we propose using a bootstrap like procedure, using an estimate of $\hat{\theta}_0$ obtained from the labeled data. This $\hat{\theta}_0$, is then used to compute any quantity which depends on θ^* such as the Hessian H and the $\nabla_\theta l(X_i, Y_i)$.

Additionally the parameterization of η (or h) should be powerful enough to capture the optimal $\eta^*(orh*)$. We use neural network functions i.e. we represent η by η_ϕ which is a neural network parameterized by parameters ϕ. While there is an optimal h, as given Proposition 2 and a sufficiently powerful neural network can learn the optimal h; when the network cannot represent the optimal h, it is not clear that learning $\mathbb{E}[\nabla l(X, Y)|X, \hat{Y}]$ is the best choice for an efficiency improvement. Directly estimating $\mathbb{E}[\nabla l(X, Y)|X, \hat{Y}]$ places equal importance to all the parameters. Moreover it does not take into account any scaling between them that the Hessian H might

Table 1: Accuracy for different LLM on the task of choosing the better treatment arm on Upworthy. While not much better than random, we can see that LLM responses still encode some signal about user preferences without additional fine-tuning

Model	Accuracy
GPT-4	64.2
Claude	60.1
Llama-3-8b	60.7

induce. Hence instead of choosing ϕ to estimate $\mathbb{E}[\nabla l(X, Y)|X, \hat{Y}]$ we optimize the variance of the desired estimand directly.

The entire process is done in three phases viz estimate $\hat{\theta}_0$, estimate ϕ, estimate $\hat{\theta}$; each stage depends on the output of the previous stage. If these estimations are done on the exact same data, it can lead to optimization bias and over-inflated statistics. To mitigate these, we employ a cross-fitting procedure [Chernozhukov et al., 2018, Newey and Robins, 2018], in which each stage is estimated from a different subsets of data. The final answer is obtained by averaging across different folds.

Remark 1. *While we have presented the approach with having only one $\hat{Y}$ per observation, this is not essential. One can trivially extend our framework to use k different synthetic outcomes by using $h = \theta^T \eta_\phi(X_i, \hat{Y}_i^1, \cdots, \hat{Y}_i^k)$.*

4 Experiments

In this section, we provide experimental results on two different tasks with real datasets to show the efficacy of our model.

Estimating Treatment Effect. We first experiment with the problem of estimating treatment effect in randomized tests.This dataset of A/B tests conducted by Upworthy [Matias et al., 2021]. The data consists of several versions of headlines created by an editorial teams for various articles. Each user was exposed to only one of these headlines article pair, and the clicks were recorded for each pair.

Another dataset we considered was from a social science experiment on Cancel Culture [Fahey et al., 2023]. This was a US based representative survey experiment on assessing whether individuals oppose certain rights of entities bases on their affinity to the entities ideology.

We first consider whether an LLM can correctly evaluate treatments/headlines. We used GPT-4 and Claude as closed source models, and compare them against Llama as an open source LLM. The task here is simply given two treatments, select the one with higher clicks.

These results are presented in Table 1. We see no model better than ($\sim$ 65%) accuracy unlike on general NLP tasks where LLMs have strong performance[Kang et al., 2023, Dai et al., 2022]. This is partly due to the fact that often standard ICL and COT based tasks are concerned about facts or statements [Freestone and Santu, 2024, Kossen et al., 2024]. In this task however, all headlines are generated by human experts and so consistent with a high-level summary of the article. The differences arise from more nuanced aspects of online user behaviour, which can be difficult to model without additional information. However we do see that in general they are more often than not right, highlighting that **LLMs have biased preferences but do contain signal**.

Methodology Next we consider a setup as follows [2] : for each A/B test, provide GPT-4 with all relevant information about the text and treatments, and prompt it to generate two distinct user profiles, each of which is likely to prefer a different arm. Additionally, we request GPT-4 to specify the relative preference (e.g., CTR) of each profile for their respective treatment arms. We then compute the relative proportions $p1$ and $p2$ of the two profiles required to match the ground truth CTRs observed in the real data. Then we obtain the 'real' dataset $\mathcal{D}^r$, by sampling n users from a binomial distribution with proportions $p1$ and $p2$. Then for each user, generate a binary outcome variable Y (e.g., click or no click) based on the corresponding CTR. On the cancel culture dataset we have individual attributes, and hence do not need to do such calibration step. Next we construct the synthetic dataset $\mathcal{D}^s$ of N users. For each such user we obtain a $\hat{Y}$ outcome based on LLM-generated preferences[3].

Baselines As baselines we consider other recent methods designed to use additional data with ml predictions, viz. PPI [Angelopoulos et al., 2023a] and PPI++ [Angelopoulos et al., 2023b]. For all the method we experiment with GPT, Claude and Llama as the in-silico simulators, though we did not explore combining different models. Furthermore since data generation for Upworthy used GPT results were used for calibrating the profiles, we did not experiment with GPT on it.

Evaluation We fix N, and then use different methods to compute the treatment effect for different sizes of n. As metric we use the mean average precision error (or MAPE), with the treatment effect estimated from the real data as ground truth.

Results Our results are presented in Figure 2. These results report MAPE difference between the no-additional data baseline compared with other methods with the synthetic data. The x-axis represents $r = n/N$, the ratio of real to synthetic data. In Figure 2a, we show results with Claude as the simulator on the Upworthy dataset. Results with Llama are in the Appendix. Our results show that in general PPI based methods can use the LLM generated responses to improve the treatment effect. This effect is consistent across different LLMs. We also see that our approach is the overall best, surpassing PPI and PPI++. This is intuitive as we choose the h function to minimize the estimator variance; and unlike PPI++ which uses a constrained form for h, we choose the best h possible. We can see that for low amounts of real data we get large reductions in MAPE. Some of these correspond to upto 40% reduction in sample size, highlighting the potential of using LLMs along with our method to improve treatment effect estimation.

Conjoint Analysis. The previous task was concerned with estimating a single parameters, the difference in click rates. Now we broaden the design to consider estimation of multiple parameters. For this we use a real choice-based conjoint dataset for sports cars [Spencer, 2019]. The setup roughly follows a normal A/B test, except instead of the options being shown to different cohorts, the

[2]This setup was adopted as Upworthy provides aggregated A/B tests i.e. we know average outcome for each treatment but not the features of the cohort over which the experiment was conducted.

[3]We do not need to query the LLM N times, but simple create such data directly by first querying the LLM for its preference and then sampling using the obtained values

Algorithm 1 Algorithm for LLM assisted τ estimation

Require: Dataset $\mathcal{D}$, Neural network model ϕ

1: Split both the real and synthetic dataset into K folds $[(\mathcal{D}_1^r, \mathcal{D}_1^s), \cdots, (\mathcal{D}_K^r, \mathcal{D}_K^s)]$

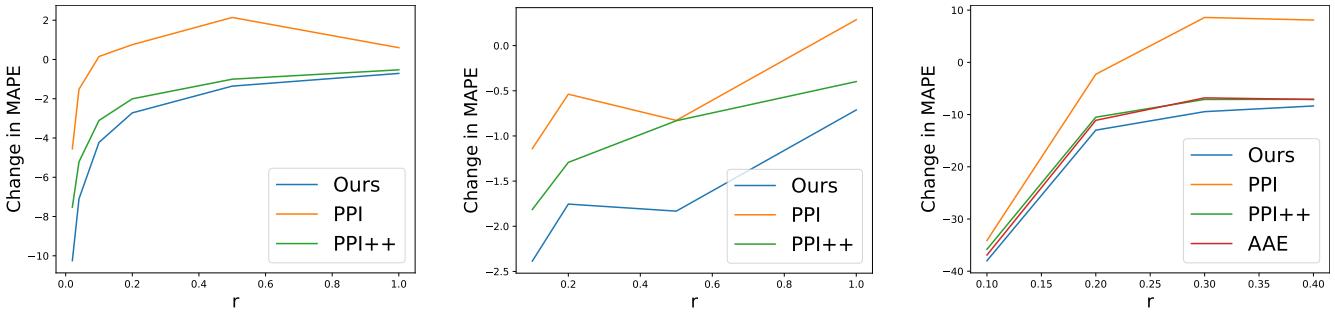
2: $\hat{\theta}_0^i \leftarrow \text{argmin}_\theta \mathbb{E}_{\mathcal{D}_i^r} l(X, Y)$ # preliminary estimate on $\mathcal{D}_i^r$

3: $\Sigma_h \leftarrow H_{\theta_0^i}^{-1} \left(r\text{Var}_{\mathcal{D}_{i+1}^s}(\eta_\phi(X, \hat{Y})) + \text{Var}_{\mathcal{D}_{i+1}^r}(\nabla l(X, Y) - \eta_\phi(X, \hat{Y})) \right) H_{\theta_0^i}^{-1}$ on $\mathcal{D}_{i+1}^r, \mathcal{D}_{i+1}^s$

4: $\hat{\phi}^i \leftarrow \text{argmin}_\phi \text{Tr}(\Sigma_h)$ or if only few parameters $\Sigma_h[d, d]$where d is index of desired parameters # optimize Σ_h

5: Estimate $\hat{\tau}^i$ by solving Equation (5) with $h_\theta = \theta \eta_{\phi^i}$ on $\mathcal{D}_{-(i,i+1)}^r, \mathcal{D}_{-(i,i+1)}^s$ # estimate on remaining folds combined

6: Return $\frac{1}{K} \sum_i (\hat{\tau}^i)$ # average over folds



(a) Upworthy **(b) Cancel Culture** **(c) Cars**

Figure 2: Performance comparison of our method against baseline estimators on a) Upworthy, b) Cancel Culture and c) Cars dataset. The x-axis we show $r = n/N$ (the ratio of real to synthetic data) data while the total amount of synthetic data remains fixed. The y-axis plots change in MAPE compared to only real data reference, so negative numbers imply better results.

same user was asked to rate between multiple hypothetical sports cars. Each car is described by a set of 5 attributes. We then follow the procedure of Wang et al. [2024b] in instructing GPT to simulate a random person's preferences.

Baseline We experiment with the PPI and PPI++ baselines mentioned earlier, along with AAE [Wang et al., 2024b], a recently proposed transfer learning based method. Following the protocol of Wang et al. [2024b], we too use 500 synthetic samples from GPT, along with varying number of real data from the survey, and estimate the coefficient of each feature. As metric we consider the average MAPE of all the coefficients together.

Results Similar to the previous experiment, we consider parameters estimated on the full set of real data as the true value. These results are presented in Figure 2(c). Qualitatively we see the same behaviour as in the treatment effect estimation scenario. We also see that in this case PPI can benefit when real data is small, but with sufficient real data it can be even worse than not using any synthetic data at all. This is not surprising by itself and has been noted before by Angelopoulos et al. [2023b].

5 Related Work

PPI Variants. have been proposed to further improve the statistical efficiency of the vanilla PPI model [Angelopoulos et al., 2023b, Gronsbell et al., 2024, Fisch et al., 2024]. As discussed earlier optimal form of h in Proposition 2 provides a partial justification of using the ∇l for the additional terms. However, using $\nabla h = \nabla l(X, \hat{Y})$ (as PPI does) is not the most efficient choice. As such, variations of the PPI method have been proposed that use modified forms of $\nabla l(X, \hat{Y})$. Angelopoulos et al. [2023b] propose using $\lambda \nabla l(X, \hat{Y})$ while Miao et al. [2023] propose using $\Lambda \nabla l(X, \hat{Y})$ where λ, Λ correspond to a learnable scalar and diagonal matrix respectively. Similar to our approach, both these methods tune the learnable parameter to minimize the estimator variance. Both these methods are also guaranteed to improve upon the asymptotic variance of the PPI method.

Integrating additional unlabeled data. is the primary goal of semi-supervised learning. In recent times, there is growing interest in combining principles of semi-supervised learning and causal inference [Kügelgen et al., 2020, Alvari et al., 2019]. PPI based ideas have been leveraged along with semi-supervised learning[Schmutz et al., 2022, Song et al., 2024]. Karlsson et al. [2024] have investigated using an outcome regression estimated from additional observational data as a control variate. Similarly, using patient outcome models as an additional variable has been shown to improve efficiency in randomized trials[Schuler et al., 2022, Liao et al., 2023]. Proximal causal inference [Tchetgen et al., 2020, Cui et al., 2024], based ideas have been applied for estimating average treatment effect (ATE) by combining multiple data sources [Yang and Ding, 2019]. In a related direction, works have explored the use of observational data collected from different environments to improve ATE estimation despite no-knowledge of confounders [Günther et al., 2024, Perry et al., 2022]. More recently, ideas from PPI have also been applied directly to causal inference [Demirel et al., 2024]. Wang et al. [2024b] have proposed a transfer learning based approach

to incorporate LLM based predictions into conjoint analysis for marketing research.

Seemingly Unrelated Regression or SUR [Zellner, 1962] is a powerful method for simultaneously estimating multiple regression models that share correlated errors. Unlike traditional approaches that estimate each model independently, SUR leverages the correlation between the errors of different outcome variables to produce more efficient estimates. SUR and related methods [Swamy and Mehta, 1975, Fiebig, 2001] have been pretty commonly deployed in econometrics [Schmidt, 1977, Foschi, 2004] and ecological applications [Goldberg, 1987, Fu et al., 2016]. A classic version of SUR from Conniffe [1985] is analogous to PPI estimate and under specific conditions, aligns with the method proposed by Chen and Chen [2000]. Gronsbell et al. [2024] have also highlighted connections between Chen and Chen [2000] and PPI methods. More recently, SUR has found applications in PBI methods for genome-wide association studies (GWAS), where researchers analyze large-scale biobank data to identify genetic variants associated with traits or phenotypes. Given the high cost and time required to collect complete phenotypic data, predicted outcomes from pre-trained models can be used in a SUR like fashion [McCaw et al., 2024, Miao et al., 2024]. Both these estimators are variants of the PBI method of Chen and Chen [2000].

LLM simulations. are an increasingly important venue of research, particularly their ability to replicate human-like interactions [Sekulić et al., 2022, Yang et al., 2024a]. These models are being used to simulate human behavior, opening new venues in social and marketing research[Brand et al., 2023, Argyle et al., 2023, Ziems et al., Kim and Lee, 2023, Park et al., 2023]. Many LLM simulations frameworks often involve self-play mechanisms, where LLMs simulate dialogues to study conversational dynamics without human involvement [Wu et al., 2023, Ulmer et al., 2024, Abbasiantaeb et al., 2024]. Guo et al. [2023] compare the performance of ChatGPT against human experts on related task and found LLMS to be quite coherent. LLMs have shown promise as tools for simulating economic agents [Chen et al., 2023, Horton, 2023]. However, their ability to faithfully simulate human preferences is actively debated[Goli and Singh, 2024, Zhu et al., 2024a]. Furthermore, LLM-generated predictions are often miscalibrated and overconfident [Xiong et al., 2023, Wei et al., 2024, Hofer et al., 2024]. Our work shows a robust statistical procedure by which LLMs can offer valuable insights without fully replacing human input.

In addition to simulating user behavior, LLMs have been applied to enhance recommendation systems and optimize A/B testing methodologies. For instance, self-play between LLMs has been used to refine recommendation algorithms [Chen et al., 2024], while LLM-assisted approaches have been proposed to warm-start bandit-based methods for online learning and experimentation [Ye et al., 2024]. These applications demonstrate the versatility of LLMs in modeling user interactions, though they often focus on specific aspects such as event transitions or search patterns [Wang et al., 2024a, Kasuga and Yonetani, 2024]. However, concerns persist about their ability to incorporate contextual nuances [Huang et al., 2024b, Yang et al., 2024b]. Additionally, challenges arise in using LLMs for causal inference, as they often fail to account for the underlying causality in decision-making processes [Gui and Toubia, 2023].

6 Conclusion

Given the scale of digital content, speeding up experimentation on digital platforms is an important challenge of better decision making. We propose leveraging LLMs for content experimentation on digital platforms. However, directly using LLMs as synthetic users for this purpose is challenging. To address this challenge, we propose a novel method to integrate LLM based predictions in a PPI-based framework. We show consistency of our estimator, and show that it is the 'least variance' estimator in a wider family of PPI-esque estimators. We then evaluate our approach using a small synthetic dataset and then on two real-life A/B tests. Our results show that our method can significantly improve efficiency by integrating LLM based simulations.

References.

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of WSDM*, pages 8–17, 2024.

Hamidreza Alvari, Elham Shaabani, Soumajyoti Sarkar, Ghazaleh Beigi, and Paulo Shakarian. Less is more: Semi-supervised causal inference for detecting pathogenic users in social media. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 154–161, 2019.

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

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

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. A note on statistical efficiency in prediction-powered inference, 2023c. URL https://web.stanford.edu/~jduchi/projects/AngelopoulosDuZr23w.pdf.

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

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

James Brand, Ayelet Israeli, and Donald Ngwe. Using gpt for market research. *Available at SSRN 4395751*, 2023.

Abhishek Chakrabortty, Guorong Dai, and Raymond J Carroll. Semi-supervised quantile estimation: Robust and efficient inference in high dimensional settings. *arXiv preprint arXiv:2201.10208*, 2022.

Yi-Hau Chen and Hung Chen. A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 62(3):449–460, 2000.

Yiting Chen, Tracy Xiao Liu, You Shan, and Songfa Zhong. The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51):e2316205120, 2023.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024. URL https://arxiv.org/abs/2401.01335.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Denis Conniffe. Estimating regression equations with common explanatory variables but unequal numbers of observations. *Journal of Econometrics*, 27(2):179–196, 1985.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*, 2022.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. *arXiv preprint arXiv:2406.02873*, 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*, 2023.

Denzil G Fiebig. Seemingly unrelated regression. *A companion to theoretical econometrics*, pages 101–121, 2001.

Adam Fisch, Joshua Maynez, R Alex Hofer, Bhuwan Dhingra, Amir Globerson, and William W Cohen. Stratified prediction-powered inference for effective hybrid evaluation of language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Paolo Foschi. *Numerical methods for estimating linear econometric models*. Université de Neuchâtel, 2004.

Matthew Freestone and Shubhra Kanti Karmaker Santu. Word embeddings revisited: Do llms offer something new? *arXiv preprint arXiv:2402.11094*, 2024.

Liyong Fu, Yuancai Lei, Guangxing Wang, Huiquan Bi, Shouzheng Tang, and Xinyu Song. Comparison of seemingly unrelated regressions with error-in-variable models for developing a system of nonlinear additive biomass equations. *Trees*, 30:839–857, 2016.

Jacqueline S Galpin and Douglas M Hawkins. The use of recursive residuals in checking model fit in linear regression. *The American Statistician*, 38(2):94–105, 1984.

Deborah E Goldberg. Neighborhood competition in an old-field plant community. *Ecology*, 68(5):1211–1223, 1987.

Ali Goli and Amandeep Singh. Frontiers: Can large language models capture human preferences? *Marketing Science*, 2024.

Jessica Gronsbell, Jianhui Gao, Yaqi Shi, Zachary R McCaw, and David Cheng. Another look at inference after prediction. *arXiv preprint arXiv:2411.19908*, 2024.

George Gui and Olivier Toubia. The challenge of using LLMs to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023.

Wiebke Günther, Oana-Iuliana Popescu, Martin Rabel, Urmi Ninad, Andreas Gerhardus, and Jakob Runge. Causal discovery with endogenous context variables. *arXiv preprint arXiv:2412.04981*, 2024.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. 2023.

Peisong Han and Lu Wang. Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430, 2013.

R Alex Hofer, Joshua Maynez, Bhuwan Dhingra, Adam Fisch, Amir Globerson, and William W Cohen. Bayesian prediction-powered inference. *arXiv preprint arXiv:2405.06034*, 2024.

John J Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*, 2024a.

Yue Huang, Zhengqing Yuan, Yujun Zhou, Kehan Guo, Xiangqi Wang, Haomin Zhuang, Weixiang Sun, Lichao Sun, Jindong Wang, Yanfang Ye, et al. Social science meets LLMs: How reliable are large language models in social simulations? *arXiv preprint arXiv:2410.23426*, 2024b.

Sungmin Kang, Juyeon Yoon, and Shin Yoo. Large language models are few-shot testers: Exploring llm-based general bug reproduction. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2312–2323. IEEE, 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.

Akira Kasuga and Ryo Yonetani. Cxsimulator: A user behavior simulation using llm embeddings for web-marketing campaign assessment. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3817–3821, 2024.

Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction. *arXiv preprint arXiv:2305.09620*, 2023.

Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Julius Kügelgen, Alexander Mey, Marco Loog, and Bernhard Schölkopf. Semi-supervised learning, causality, and the conditional cluster assumption. In *Conference on uncertainty in artificial intelligence*, pages 1–10. PMLR, 2020.

Nicholas Larsen, Jonathan Stallrich, Srijan Sengupta, Alex Deng, and Ron Kohavi. Statistical challenges in online controlled experiments: A review of a/b testing methodology. *The American Statistician*, 78(2):135–149, 2024. doi: 10.1080/00031305.2023.2257237.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, 2010.

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

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.

NX Lin, JQ Shi, and R Henderson. Doubly misspecified models. *Biometrika*, 99(2):285–298, 2012.

J Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. The upworthy research archive, a time series of 32,487 experiments in US media. *Scientific Data*, 8(1):195, 2021.

Zachary R McCaw, Jianhui Gao, Xihong Lin, and Jessica Gronsbell. Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nature Genetics*, pages 1–10, 2024.

Jiacheng Miao, Xinran Miao, Yixuan Wu, Jiwei Zhao, and Qiongshi Lu. Assumption-lean and data-adaptive post-prediction inference. *arXiv preprint arXiv:2311.14220*, 2023.

Jiacheng Miao, Yixuan Wu, Zhongxuan Sun, Xinran Miao, Tianyuan Lu, Jiwei Zhao, and Qiongshi Lu. Valid inference for machine learning-assisted gwas. *medRxiv*, pages 2024–01, 2024.

Keshav Motwani and Daniela Witten. Revisiting inference after prediction. *Journal of Machine Learning Research*, 24(394):1–18, 2023.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Jerzy Neyman. On the Application of Probability Theory to Agricultural Experiments: Essay on Principles. *Statistical Science*, 5: 465–80, 1923. Section 9 (translated in 1990).

OpenAI. New embedding models and api updates, 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of ACM USIT*, 2023.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Donald A Pierce and Daniel W Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81 (396):977–986, 1986.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. Tool learning with large language models: A survey, 2024. URL https://arxiv.org/abs/2405.17935.

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

Nathan Ross. Fundamentals of stein's method. *Probability Surveys*, 8:210–293, 2011.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Peter Schmidt. Estimation of seemingly unrelated regressions with unequal numbers of observations. *Journal of Econometrics*, 5(3): 365–377, 1977.

Hugo Schmutz, Olivier Humbert, and Pierre-Alexandre Mattei. Don't fear the unlabelled: safe semi-supervised learning via debiasing. In *The Eleventh International Conference on Learning Representations*, 2022.

Bernhard Schölkopf, David W Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 113(27):7391–7398, 2016.

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 M Schwartz, Eric T Bradlow, and Peter S Fader. Customer acquisition via display advertising using multi-armed bandit experiments.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. Evaluating mixed-initiative conversational search systems via user simulation. In *WSDM*, 2022.

Shiv Shankar, Daniel Sheldon, Tao Sun, John Pickering, and Thomas G Dietterich. Three-quarter sibling regression for denoising observational data. *arXiv preprint arXiv:2101.00074*, 2020.

Shanshan Song, Yuanyuan Lin, and Yong Zhou. A general m-estimation theory in semi-supervised framework. *Journal of the American Statistical Association*, 119(546):1065–1075, 2024.

Vic Spencer. Choice modeling sports cars. https://github.com/spensorflow/Marketing-Analytics---Choice-Modeling-Sports-Car-Sales, 2019.

Paravaster AVB Swamy and Jatinder S Mehta. Bayesian and non-bayesian analysis of switching regressions and of random coefficient regression models. *Journal of the American Statistical Association*, 70(351a):593–602, 1975.

Cheng Yong Tang and Yongsong Qin. An efficient empirical likelihood approach for estimating equations with missing data. *Biometrika*, 99(4):1001–1007, 2012.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Dennis Ulmer, Elman Mansimov, Kaixiang Lin, Justin Sun, Xibin Gao, and Yi Zhang. Bootstrapping llm-based task-oriented dialogue agents via self-talk. *arXiv preprint arXiv:2401.05033*, 2024.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. User behavior simulation with large language model-based agents for recommender systems. *ACM Transactions on Information Systems*, 2024a.

Mengxin Wang, Dennis J. Zhang, and Heng Zhang. Large language models for market research: A data-augmentation approach, 2024b.

Siruo Wang, Tyler H McCormick, and Jeffrey T Leek. Methods for correcting inference based on outcomes predicted by machine learning. *Proceedings of the National Academy of Sciences*, 117 (48):30266–30275, 2020.

Larry Wasserman and John Lafferty. Statistical analysis of semi-supervised regression. *Advances in Neural Information Processing Systems*, 20, 2007.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi

Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Dayu Yang, Fumian Chen, and Hui Fang. Behavior alignment: A new perspective of evaluating llm-based conversational recommendation systems. In *Proceedings of the 47th International ACM SIGIR*, pages 2286–2290, 2024a.

Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, and Hui Liu. Are large language models (LLMs) good social predictors? *arXiv preprint arXiv:2402.12620*, 2024b.

Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2019.

Zikun Ye, Hema Yoganarasimhan, and Yufeng Zheng. Lola: Llm-assisted online learning algorithm for content experiments, 2024.

Arnold Zellner. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368, 1962.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Lixi Zhu, Xiaowen Huang, and Jitao Sang. How reliable is your simulator? analysis on the limitations of current llm-based user simulators for conversational recommendation, 2024a.

Yutao Zhu, Huaying Yuan, Shuting Wang, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval, 2024b. URL https://arxiv.org/abs/2308.07107.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*.

Residual Regression

For exponential family models $p(Y|X) \propto g(Y)\exp(T(Y) - \theta^T X)$ where g is a base measure, T is a sufficient statistic for Y an. The score function for such models is given by

$$s_\theta(y,x) = \nabla_\theta \log p(Y|X) = T(y) - \mathbb{E}[T(y)|x,\theta]$$

Thus when trained using log-likelihood, the optimal h takes the same residual form as in ANCOVA models, with the residual computed on $T(Y)$.

Similarly GLMs parameterize their means as a non-linear link function applied on a linear model i.e. $\mathbb{E}[Y] = g(\theta^T X)$. A similar argument applies in this case.

Proof

Proposition 3. *Assume the conditions for Proposition 1. Next set* $h(X,\hat{Y}) = \frac{1}{r+1}\theta^T\mathbb{E}[\nabla l(X,Y)|X,\hat{Y}] = h^*$. *Then the asymptotic variance of* θ^* (Σ_{h^*}) *is given by* $H^{-1}_{\theta^\star}\left(\mathrm{Var}(\nabla l_{\theta^\star}(X,Y)) - \frac{1}{1+r}\mathrm{Var}(\mathbb{E}[\nabla l_{\theta^\star}(X,Y)|X,\hat{Y}])\right)H_{\theta}$ *Additionally,* $\Sigma_{h^*} \preceq \Sigma_h$ *for any other function* h, *where* $\preceq$ *is the Loewner order among positive semi definite matrices.*

Proof. Consider

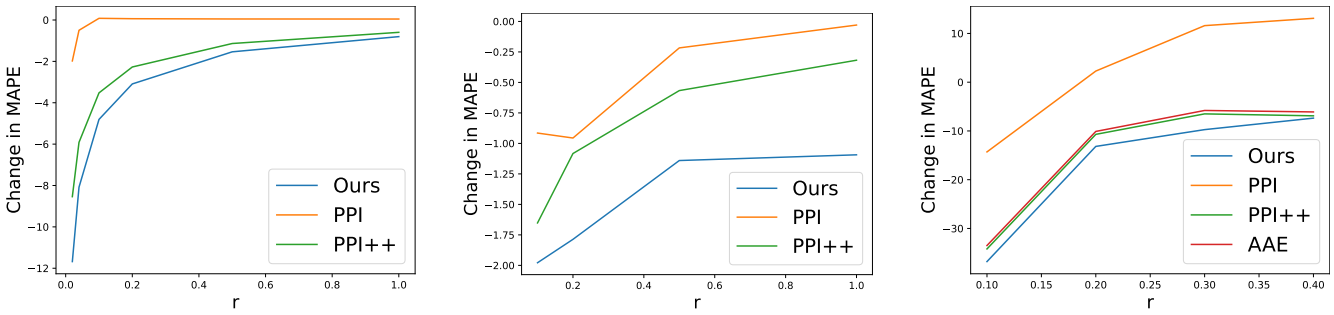
$$\begin{aligned}
A &= r\mathrm{Var}(\nabla h(X,\hat{Y})) + \mathrm{Var}(\nabla l(X,Y) - \nabla h(X,\hat{Y})) \\
&= r\mathrm{Var}(\nabla h(X,\hat{Y})) + \mathrm{Var}(\nabla l(X,Y)) + \mathrm{Var}(\nabla h(X,\hat{Y})) \\
&\quad - 2\mathrm{Cov}(\nabla l(X,Y), \nabla h(X,\hat{Y})) \\
&= (r+1)\mathrm{Var}(\nabla h(X,\hat{Y})) + \mathrm{Var}(\nabla l(X,Y)) \\
&\quad - 2\mathrm{Cov}(\nabla l(X,Y), \nabla h(X,\hat{Y})) \\
&= \mathrm{Var}(\sqrt{r+1}\nabla h(X,\hat{Y}) - \frac{1}{\sqrt{r+1}}\nabla l(X,Y)) \\
&\quad + \frac{r}{r+1}\mathrm{Var}(\nabla l(X,Y))
\end{aligned}$$

Only the first term here depend on h. Therefore we can focus only on that. Moreover given its a Var term which is always positive semi-definite, its minima is obtained by moment matching i.e. when the learnable parameter takes the mean value. Thus the optimal h is given by $\sqrt{r+1}\nabla h(X,Y) = \frac{1}{\sqrt{r+1}}\mathbb{E}\nabla l(X,Y)$, which gives

$$\nabla h = \frac{1}{1+r}\mathbb{E}[\nabla l(X,Y)]$$

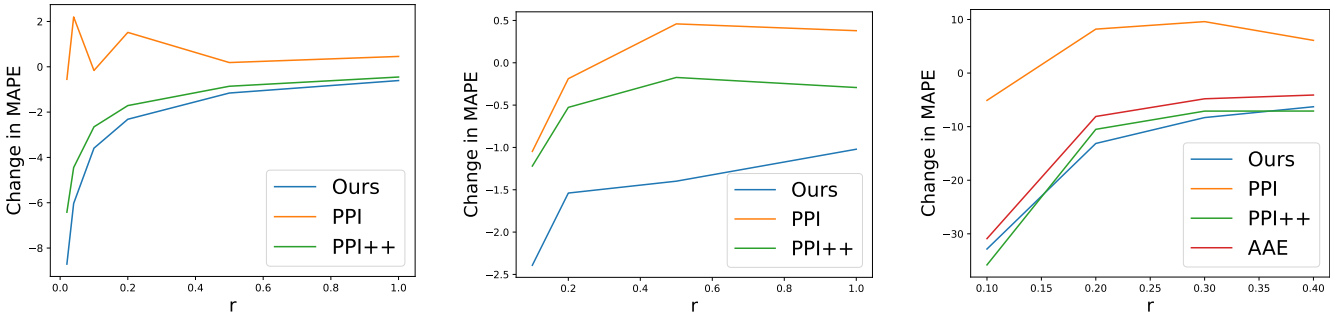
Plugging in this value of h, in the expression for Σ_h, gives the final value of Σ_{h*}. □

A Additional Results



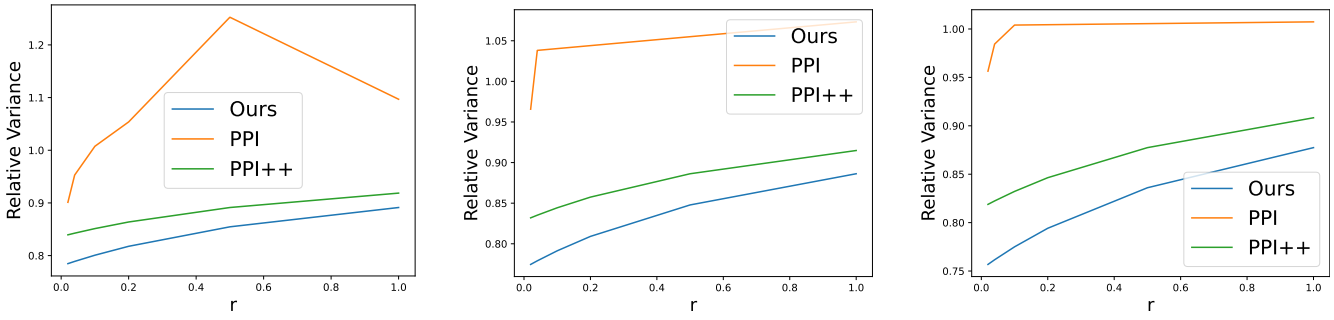
(a) Upworthy (b) Cancel Culture (c) Cars

Figure 3: Performance comparison of our method against baseline estimators on a) Upworthy, b) Cancel Culture and c) Cars with Claude COT prompting. The x-axis we show $r = n/N$ (the ratio of real to synthetic data) data while the total amount of synthetic data remains fixed. The y-axis plots change in MAPE compared to only real data reference, so negative numbers imply better results.



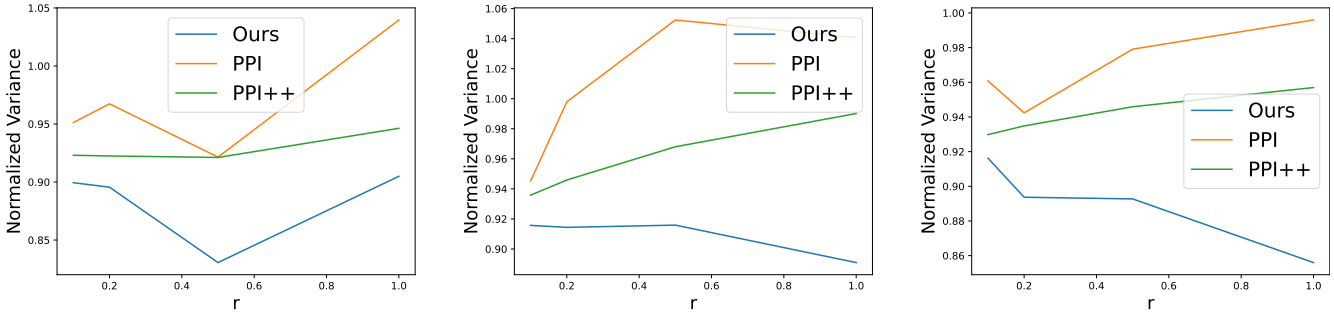
(a) Upworthy (b) Cancel Culture (c) Cars

Figure 4: Performance comparison of our method against baseline estimators on a) Upworthy, b) Cancel Culture and c) Cars dataset with Llama prompting. The x-axis we show $r = n/N$ (the ratio of real to synthetic data) data while the total amount of synthetic data remains fixed. The y-axis plots change in MAPE compared to only real data reference, so negative numbers imply better results.



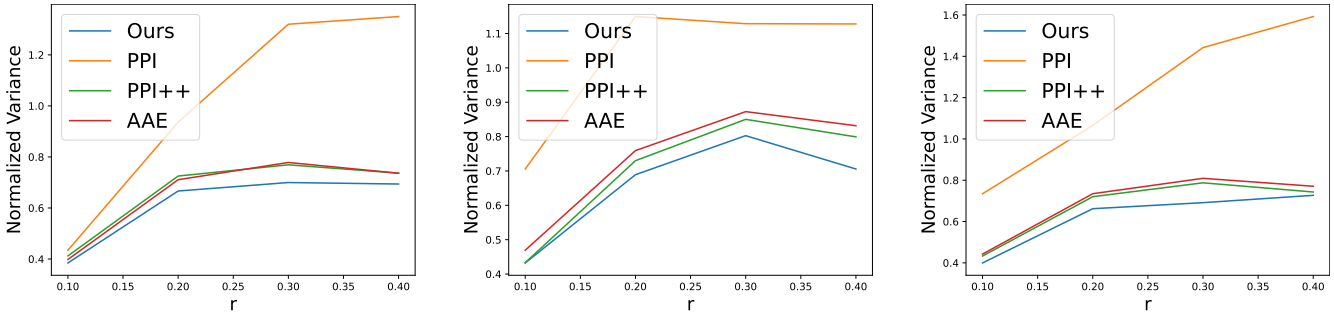
(a) GPT (b) Llama (c) Claude

Figure 5: Variance comparison of different methods against baseline on Upworthy with different LLMs. The x-axis we show $r = n/N$ (the ratio of real to synthetic data) data while the total amount of synthetic data remains fixed. The y-axis plots the ratio of variance between the estimator and the baseline of using only real data.



(a) GPT (b) Llama (c) Claude

Figure 6: Variance comparison of different methods against baseline on Cancel Culture with different LLMs. The x-axis we show $r = n/N$ (the ratio of real to synthetic data) data while the total amount of synthetic data remains fixed. The y-axis plots the ratio of variance between the estimator and the baseline of using only real data.



(a) GPT (b) Llama (c) Claude

Figure 7: Variance comparison of different methods against baseline on Cars with different LLMs. The x-axis we show $r = n/N$ (the ratio of real to synthetic data) data while the total amount of synthetic data remains fixed. The y-axis plots the ratio of variance between the estimator and the baseline of using only real data.