
Batch-Adaptive Annotations for Causal Inference with Complex-Embedded Outcomes

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Estimating the causal effects of an intervention on outcomes is crucial to policy and
2 decision-making. But often, information about outcomes can be missing or subject
3 to non-standard measurement error. It may be possible to reveal ground-truth
4 outcome information at a cost, for example via data annotation or follow-up; but
5 budget constraints entail that only a fraction of the dataset can be labeled. In this
6 setting, we optimize *which data points should be sampled for outcome information*
7 and therefore efficient average treatment effect estimation with missing data. We do
8 so by allocating data annotation in batches. We extend to settings where outcomes
9 may be recorded in unstructured data that can be annotated at a cost, such as text or
10 images, for example, in healthcare or social services. Our motivating application is
11 a collaboration with a street outreach provider with millions of case notes, where
12 it is possible to expertly label some, but not all, ground-truth outcomes. We
13 demonstrate how expert labels and noisy imputed labels can be combined into a
14 doubly robust causal estimator. We run experiments on simulated data and two
15 real-world datasets, including one on street outreach interventions in homelessness
16 services, to show the versatility of our proposed method.

17 1 Introduction

18 Evaluating causal effects of a treatment or policy intervention is a challenging problem in its own
19 right, but an added layer of complexity comes when there is missing data. In this paper, we consider
20 a setting of observational causal inference with missing outcomes, where it is possible to obtain
21 information about ground-truth outcomes at a cost, via expert annotation or follow-up. Recent tools
22 in machine learning can label outcomes, but for inferential goals, this can lead to error-prone and
23 biased outputs. With a small budget, one can obtain valid causal effects on a small subsample without
24 using additional contextual information or imputation, but this can be high-variance. We build on
25 doubly-robust causal inference with missing outcomes to determine where to sample additional
26 outcome annotations to minimize the asymptotic variance of treatment effect estimation.

27 Our methodology is motivated by a collaboration with a nonprofit to evaluate the impact of street
28 outreach on housing outcomes, where rich information about outcomes of outreach are embedded
29 in case notes written by outreach workers. Street outreach is an intensive intervention; caseworkers
30 canvass for and build relationships with homeless clients and write case notes after each interaction.
31 These notes are a noisy view on the ground-truth of what happens during the open-ended process of
32 outreach. Was a client progressing on their housing application or their goals, or were they facing
33 other barriers? In our experience, outreach workers can extract structured information, from the
34 unstructured text of case notes. They can provide context and recognize important milestones. But
35 it is simply impossible for under-resourced expert outreach workers to label millions of case notes.
36 While modern natural language processing tools can facilitate annotation at scale, they are often

inaccurate. *Given an annotation budget constraint, how can we strategically collect ground truth data, such as by assigning expert annotation, while leveraging additional data sources or weaker annotation to optimize causal effect estimation?* In this paper, we develop general methodology for optimizing data annotation and we validate our methodological innovations using outcomes with plausible ground-truth information on housing placement.

This problem is not unique to the social work domain and can generally apply to cases of measurement error with misaligned modalities (such as text or images), where it is possible to query the ground truth directly for some portion of the data at a cost. In some settings, we can query other data sources for ground-truth labels directly, while in other settings, outcomes may be recorded in complex information such as text or images. However, due to dimensionality issues, these cannot be directly substituted for ground-truth outcomes Y . Weaker imputation of auxiliary information is feasible at scale, but second-best due to inaccuracies. For example, when an outcome variable, wages, is only observed from self-reported working individuals, surveyors could conduct follow-up interviews with participants to obtain wage data, but this can be expensive. Noisy measures from the same dataset (such as last year’s wages) or transporting prediction models from national wage databases can be predictive. Such trade-offs between expert annotation and scalable, weaker imputation are pervasive in data-intensive machine learning, for example as in the recent “LLM-as-a-judge” framework [56].

This study makes the following contributions: we propose a two-stage batch-adaptive algorithm for efficient ATE estimation from complex embedded outcomes. We derive the expert labeling probability that minimizes the asymptotic variance of an orthogonal estimator [4]. We design a two-stage adaptive annotation procedure. The first stage estimates nuisance functions for the asymptotic variance on the fully observed data. We use the estimates and functions from the first stage to estimate the optimal labeling probabilities in the second stage. The final proposed estimator combines the model-annotated labels and the expert labels in a doubly robust estimator for the ATE. We show that this two-stage design achieves the optimal asymptotic variance with weaker double-machine learning requirements on nuisance function estimates. We leverage our closed-form characterizations to provide insights on how to improve downstream treatment-effect estimation. We validate and show improvements upon random sampling on semi-synthetic and real-world datasets from retail and street outreach.

2 Related work

Our model is closest to optimizing a validation set for causal inference with missing outcomes, which can be broadly useful for causal inference with non-standard measurement error. Typical distributional conditions for non-standard measurement error [38] are generally inapplicable to text or images, our motivating application. The most related work is that of [20, 58], which leverages the fact that sampling probabilities for data annotation are known to obtain doubly-robust estimation via causal inference. These works generally address non-causal estimands such as mean estimation and M-estimation (therefore without discussion of treatment effect estimation). Our work follows a key approach in adaptive experimental design of optimizing the semiparametric efficiency lower bound, whether via batch or full adaptivity. Hahn et al. [24] studied a two-stage procedure for estimating the ATE with a proportional asymptotic, and show asymptotic equivalence of their batched adaptive estimator to the optimal asymptotic variance. [36] and [46] also considers a double machine learning version of [24], though our estimator is different and we further characterize the closed-form optimal sampling probabilities, yielding additional analysis. Other variants in the same framework include Uehara et al. [46] which optimizes treatment choice or Cook et al. [15] which provides anytime-valid inference for full adaptivity. Armstrong [1] proves the semiparametric efficiency lower bound cannot be beat in general by adaptive designs; so this algorithmic paradigm is the right fit for our goals of efficient statistical inference. Crucially, all these other papers focus on allocating treatments, while we allocate the probability of revealing the outcome for a datapoint (i.e, via expert annotation); this gives us a different optimization problem and different estimation challenges.

Regarding the use of auxiliary information in causal inference, many recent works have studied the use of surrogate or proxy information. Although our context $\tilde{Y}$ aligns with colloquial notions of surrogates or proxies, recent advances in surrogate and proxy methods refer to specific models that differ from our direct measurement/costly observation setting [3, 32, 20]. Surrogates often estimate an outcome that is impossible to measure at the time of analysis [3], such as long-term effects; while we can obtain ground truth outcomes from expert data annotators feasibly but under a binding budget constraint. We do leverage that we can design the sampling probabilities of

outcome observations (ground-truth annotation) or missingness for doubly-robust estimation, like some methods in the surrogate literature or data combination [53, 32]. But we treat the underlying setting as a single unconfounded dataset with missingness. The different setting of proximal causal inference [44, 16] seeks proxy outcomes/treatments that are informative of unobserved confounders; we assume unconfoundedness holds. Recently, [9] study the “design-based supervised learning” perspective of [20] specifically for proxies for unobserved confounding.

Many exciting recent works study adaptive experimentation under different desiderata, such as full adaptivity, in-sample decision regret or finite-sample, non-asymptotic guarantees [21, 54, 15]. Such designs are closely related to covariate-adaptive-randomization; the recent work of [41] studies delayed outcomes. These desiderata are less relevant to our specific setting of data annotation: it’s easier to leverage human annotators with batch annotation rather than full adaptivity, and in-sample regret is less meaningful for data annotation than for treatment allocation. Technical tools from these other works could be applied to our setting as well.

3 Problem setup

Our problem setting is causal inference with missing outcomes. We discuss extensions to a setting where outcomes are measured in a high-dimensional contextual variable $\tilde{Y}$, such as images or text.

In both cases, we assume the ground-truth data-generating process follows that of standard causal inference. A data instance $(X, Z, Y(Z))$, includes covariates $X \in \mathcal{X}$, a binary treatment $Z \in \{0, 1\}$, and potential outcomes $Y(Z)$ in the Neyman-Rubin potential outcome framework. We only observe $Y(Z)$ for the realized treatment assignment Z and assume the usual stable unit value treatment assumption (SUTVA). If the ground-truth data were observed, we would have a standard causal inference task at hand, so the key challenge is its *missingness*. We let $R \in \{0, 1\}$ denote the presence ($R = 1$) or absence ($R = 0$) of the outcome Y . Therefore, our observational dataset for estimation is (X, Z, R, RY) , i.e. with missing outcomes. For causal identification, we generally proceed under the following assumptions:

Assumption 1 (Treatment ignorability [25, 27, 35]). $Y(Z) \perp\!\!\!\perp Z \mid X$.

Assumption 2 (R -ignorability [37, 4]). $R \perp\!\!\!\perp Y(Z) \mid Z, X$

Assumption 1, or unconfoundedness, posits that the observed covariates are fully informative of treatment. It is generally untestable but robust estimation is possible in its absence, e.g. via sensitivity analysis and partial identification [55, 34]. On the other hand, Assumption 2 is true by design as long as the full corpus of datapoints needing annotation is available from the outset, since we choose what datapoints are annotated for ground-truth labels based on (Z, X) alone.

Although one approach is completely random sampling, we are particularly concerned with *how can we select datapoints for expert annotation for optimal estimation?* We assume the budget is limited for data annotation, but we have control over the missingness mechanism, i.e. assigning data for expert annotation. Define the propensity score and annotation (outcome observation) probability:

$$e_z(X) := P(Z = z \mid X) \text{ (propensity score), and } \pi(Z, X) := P(R = 1 \mid Z, X) \text{ (annotation probability).}$$

We assume positivity/overlap; that we observe treatment and outcome with nonzero probability.

Assumption 3 (Treatment and annotation positivity [25, 27, 35]). $\epsilon < \pi(z, X) \leq 1, z \in \{0, 1\}$ and $1/\nu < e_1(X) < 1 - 1/\nu, \nu > 0$

We define the outcome model, which is identified on the $R = 1$ data by Assumption 2, and the conditional variance:

$$\begin{aligned} \mu_z(X) &:= \mathbb{E}[Y \mid Z = z, X] \stackrel{asn.2}{=} \mathbb{E}[Y \mid Z = z, R = 1, X] \\ \sigma_z^2(X) &:= \mathbb{E}[(Y - \mu_z(X))^2 \mid Z = z, X = x]. \end{aligned}$$

Batch allocation setup. We consider a two-batch adaptive protocol, where n iid observations are randomly split into two batches. We consider a proportional asymptotic regime where the budget and size of first batch n_1 are fixed proportions $\kappa \in (0, 1)$ of the dataset size.

Assumption 4 (Proportional asymptotic [24, 36]). $\lim_{n \rightarrow \infty} \frac{n_1}{n} = \kappa$.

137 In the first batch, we randomly assign annotations according to a small but asymptotically nontrivial
 138 fraction of the budgets. In the first batch, outcomes are realized and observed, and the nuisance
 139 models $(\hat{\mu}_z(x), \hat{e}_z(x), \hat{\sigma}_z^2(x))$ are trained on the observed data. We solve for optimal annotation
 140 probabilities π^* and sample data in the second batch so that the mixture distribution over outcome
 141 observations achieves π^* . We combine the results from both batches and use the data for ATE
 142 estimation, which we describe in the next section.

143 **Extension to Missing Outcomes with Context.** We extend our missing outcome framework to
 144 cases where we have additional contextual measurements of outcomes. In this setting, our observed
 145 data includes $\tilde{Y}$, a widely available “*complex-embedded outcome*”, i.e. a ground-truth outcome
 146 embedded in more complex information, such as images, text, etc. Though $\tilde{Y}$ is observable for every
 147 datapoint, it is not usable for direct estimation. We assume causal effects operate through a latent
 148 true outcome Y , of which $\tilde{Y}$ is a complex observation (but not deterministic function) thereof. We
 149 assume an exclusion restriction that the direct effect of treatment passes through the latent outcomes
 150 only [25, 49].

Assumption 5 (Complex embedded outcomes: exclusion restriction [42, 28]).

$$\tilde{Y}(Z) = g(Y(Z), X) + \epsilon, \epsilon \neq 0 \text{ a.s.}; \text{ and } Z \perp \tilde{Y} \mid X, Y(Z)$$

151 We don’t make distributional assumptions on the measurement error mechanism, appealing instead
 152 to data annotation/validation measurement which we assume reveals the ground-truth outcome Y ,
 153 although ultimately predicting Y from $\tilde{Y}$ just needs to be consistently estimable, discussed later on.

154 This assumption asserts that treatment assignment does not affect $\tilde{Y}$ beyond affecting latent outcomes
 155 Y . For example, in a medical setting, it holds if treatment affects underlying biological phenomena,
 156 e.g. makes a tumor smaller, these phenomena are recorded via clinical notes or raw pixel images, and
 157 treatment doesn’t change textual or visual expression. It assures that predicting latent outcomes Y
 158 from $\tilde{Y}$ does not introduce collider bias, and is testable, i.e. after the first batch of data.

159 In this setting, we allow the outcome model to depend on the complex embedded $\tilde{Y}$, and denote
 160 $\mu_Z(X, \tilde{Y}) := \mathbb{E}[Y \mid Z, X, \tilde{Y}]$. There are a few variants of estimating this from data. We denote an
 161 ML-prediction based on $\tilde{Y}$ (with X covariates and treatment information) as $f_z(X, \tilde{Y})$; for example
 162 zero-shot prediction using an LLM or pretrained model. Variants include calibrating zero-shot
 163 predictions to ground-truth $\mathbb{E}[Y \mid Z, R = 1, f_z(X, \tilde{Y})]$, predicting Y and including ML predictions
 164 as a covariate alongside X , or various ensembling combinations thereof. This last approach is
 165 suggested in Egami et al. [20]. Later on, we find that in practice, choosing the outcome model that
 166 reduces the mean squared error leads to better numerical results.

167 4 Method

168 This section outlines our proposed methodology. We first recap the AIPW estimator for the missing
 169 outcomes case and provide the lower bound for the asymptotic variance. Then we consider a global
 170 budget optimization problem and solve for the optimal $\pi^*(z, x)$. (In Appendix F we discuss a very
 171 similar case of treatment-specific budget). We describe feasible estimation of the ATE by the AIPW
 172 estimator (with missing outcomes).

173 **Recap: Optimal asymptotic variance for the ATE with missing outcomes.** Our target parameter
 174 of interest is the ATE of a binary treatment vector Z on an outcome Y .

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

175 Bia et al. [4] derives a double-machine learning estimator for ATE estimation with missing outcomes:

$$\mathbb{E}[Y(z)] = \mathbb{E}[\psi_z], \text{ where } \psi_z = \frac{\mathbb{I}[Z = z]R(Y - \mu_z(X))}{e_z(X)\pi(z, X)} + \mu_z(X), \text{ and } \tau_{AIPW} = \mathbb{E}[\psi_1 - \psi_0].$$

176 The outcome model $\mu_z(X)$ is estimated on data with observed outcomes, since SUTVA and assump-
 177 tion 2 give that $\mathbb{E}[Y(z) \mid X] = \mathbb{E}[Y \mid Z = z, X] = \mathbb{E}[Y \mid Z = z, R = 1, X]$.

178 The focus of our work is to optimize the semiparametric efficient asymptotic variance (proven in [4]),
 179 which is closely related to the ATE of [23].

180 **Proposition 1.** *The asymptotic variance (AVar) is:*

$$\text{AVar} = \text{Var}[\mu_1(X) - \mu_0(X)] + \sum_{z \in \{0,1\}} \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z, X)}\right]$$

181 The first term is independent of π ; we focus on optimizing the second term with respect to π .

182 *Remark 1.* We state the results for the base model, though they extend directly for the case with con-
 183 texts. With contexts, by marginalizing over $\tilde{Y}$, the analogous expressions use the estimators $\hat{\mu}_z(X, \tilde{Y})$
 184 instead of $\hat{\mu}_z(X)$ whereas $\hat{\sigma}_z^2(X)$ stays the same (sampling probabilities depend only on (Z, X)
 185 and just correspondingly marginalizes over $\tilde{Y}$, $\hat{\sigma}_z^2(X) = \mathbb{E}[(Y - \hat{\mu}_z(X, \tilde{Y}))^2 | Z = z, X = x]$. In
 186 the setting with noisy measurements $\tilde{Y}$, under the exclusion restriction Assumption 5, the mean
 187 potential outcome is identified by regression adjustment: $\mathbb{E}[Y(z)] = \mathbb{E}[\mathbb{E}[Y|Z = z, R = 1, X, \tilde{Y}]] =$
 188 $\mathbb{E}[\mathbb{E}[Y|Z = z, R = 1, X]]$.

189 **Characterizing the optimal $\pi^*(z, x)$.** We first characterize the population optimal sampling
 190 probabilities $\pi^*(z, x)$, assuming the nuisance functions are known. We optimize the asymptotic
 191 variance over π under a sampling budget. We consider a global budget constraint $B \in [0, 1]$ over
 192 all annotations. The setting is meaningful when the budget binds, $B \ll 1$, which is still practically
 193 relevant.

$$\min_{0 < \pi(z, x) \leq 1, \forall z, x} \sum_{z \in \{0,1\}} \mathbb{E}\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z, X)}\right] \text{ s.t. } \mathbb{E}[\pi(Z, X)] \leq B \quad (\text{OPT (global budget)})$$

194 Note that in the global budget constraint, $\mathbb{E}[\pi(Z, X)] = \mathbb{E}[\pi(1, X)\mathbb{I}[Z = 1] + \pi(0, X)\mathbb{I}[Z = 0]]$.
 195 We can characterize the solution as follows.

196 **Theorem 1.** *The solution to the global budget problem is:*

$$\pi^*(z, X) = \frac{\sqrt{\sigma_z^2(X)}}{e_z(X)} B \left(\mathbb{E}\left[\sqrt{\sigma_1^2(X)} + \sqrt{\sigma_0^2(X)}\right] \right)^{-1}$$

197 Note that sampling probabilities increase in the conditional variance/uncertainty of the model, $\sigma^2(X)$,
 198 and the inverse propensity score. Characterizing the closed-form solution is useful for our analysis
 199 later on, in establishing convergence of estimation to the limiting optimal data annotation probabilities.
 200 For the proof, see Appendix G

201 **Feasible two-batch adaptive design and estimator.** Our characterizations above assume knowl-
 202 edge of the true $\sigma_z^2(x)$ and the propensity scores $e_z(x)$. Since these need to be estimated, we leverage
 203 the double machine learning (DML) framework and conduct a feasible two-batch adaptive design
 204 [10, 4]. Cross-fitting with iid data [10] splits the data, estimates nuisance functions on one fold, and
 205 evaluates the estimator on a datapoint leveraging nuisance functions from another fold of data.

206 We leverage a variant [36] that introduces folds within each batch of data. Figure 3 summarizes the
 207 cross-fitting approach; we leave details to the appendix. First, we split the observations in each batch
 208 $t = 1, 2$ into K folds (e.g. $K = 5$). Let $\mathcal{I}_k$ denote the set of batch and observation indices (t, i)
 209 assigned to fold k and batch t . Then within each fold, we estimate nuisance models on observations in
 210 batch 1. We use cross-fitting to optimize the sampling probabilities, i.e. $\pi^{*, (-k)}$ optimizes asymptotic
 211 variance with out-of-fold nuisances $e^{(-k)}$. Finally we adaptively assign annotation probabilities in
 212 batch 2. This ensures independence, that is the nuisance models only depend on observations in the
 213 previous batch from the same fold. The adaptive procedure with CSBAE cross-fitting procedure to
 214 estimate τ_{AIPW} is summarized in Algorithm 1. See Algorithm 2 for a full description.

215 Therefore the cross-fitted feasible estimator takes the form $\hat{\tau}_{AIPW} =$
 216 $\frac{1}{n} \sum_{t=1}^2 \sum_{k=1}^K \sum_{(t,i) \in \mathcal{I}_k} \hat{\psi}_{1,i} - \hat{\psi}_{0,i}$ where

$$\hat{\psi}_{z,i} = \frac{\mathbb{I}[Z_i = z]R_i(Y_i - \hat{\mu}_z^{(-k)}(X_i))}{\hat{e}_z^{(-k)}(X_i)\hat{\pi}^{(-k)}(z, X_i)} + \hat{\mu}_z^{(-k)}(X_i). \quad (1)$$

217 5 Analysis

218 In this section, we provide a central limit theorem (CLT) for the setting where annotation probabilities
 219 are assigned adaptively and nuisance parameters must be estimated. We provide some insights to
 220 improve estimation as well as an extension to settings with continuous treatments.

Algorithm 1 Batch Adaptive Causal Estimation With Complex Embedded Outcomes

Input: Data $\mathcal{D} = \{(X_i, Z_i)\}_{i=1}^n$, sampling budget $B \in [0, 1]$
Step 1: Partition $\mathcal{D}$ into 2 batches and K folds $\mathcal{D}_1^{(k)}, \mathcal{D}_2^{(k)}$ for $k = 1, \dots, K$
Step 2: On Batch 1, sample $R_1 \sim \text{Bern}(B)$. Estimate nuisances within each k -fold $\hat{\mu}_z^{(k)}(X, \tilde{Y})$, $\hat{\sigma}_z^{2(k)}(X)$, and $\hat{e}_z^{(k)}$.
Step 3: On Batch 2, folds $k = 1, \dots, K$, obtain π^* by optimizing eq. (OPT (global budget)), plugging in nuisance estimates. Solve for $\hat{\pi}_2^{(k)}(X_i) = \frac{1}{1-\kappa}(\pi^*(X_i) - \kappa\pi_1)$
Step 4: On Batch 2, sample $R_2 \sim \text{Bern}(\hat{\pi}_2^{(k)}(X_i))$ and obtain outcomes.
Step 5: Pool data across batches and estimate ATE with AIPW estimator in eq. (1) (or eq. (RZ-plug-in.), or balancing weights) and out of fold nuisances.

221 Denote $\|\cdot\|_2 = (\mathbb{E}[(\cdot)^2])^{1/2}$. The following assumptions can also be found in [36, 10, 49, 46, 4].

222 **Assumption 6** (Consistent estimation and boundedness). Assume bounded second moments of out-
223 comes and errors, $\|Y(z)\|_2 \leq C_1, \|\mu_z(X)\|_2 \leq C_2, \|(Y - \mu_z(X))\|_2^2 \leq 4B_{\sigma^2}, \forall z$; and consistent
224 estimation $\mathbb{E}[(\mu_z(X) - \hat{\mu}_z(X))^2] \leq K_\mu n^{-r_\mu}$ for some constants $C_1, C_2, B_{\sigma^2}, K_\mu, r_\mu \geq 0$.

225 **Assumption 7** (Product error rates [4]). For nuisance functions, assume the products of
226 their mean-square convergence rates vanish faster than $n^{-1/2}$: (i) $\sqrt{n} \|\hat{\mu}_z(X) - \mu_z(X)\|_2 \times$
227 $\|\hat{\pi}(z, X) - \pi(z, X)\|_2 \xrightarrow{p} 0$; (ii) $\sqrt{n} \|\hat{\mu}_z(X) - \mu_z(X)\|_2 \times \|\hat{e}_z(X) - e_z(X)\|_2 \xrightarrow{p} 0$.

228 **Assumption 8** (VC dimension for nuisance estimation[2]). The nuisance estimation of e_z , and σ_z^2
229 occurs over function classes with finite VC-dimension.

Assumption 9 (Sufficiently weak dependence across batches).

$$\sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \left\| \mathbb{E} \left[\hat{\psi}_i(R; \hat{e}^{(-k)}, \hat{\pi}^{(-k)}, \hat{\mu}^{(-k)}) - \psi_i(R; e^{(-k)}, \pi^{(-k)}, \mu^{(-k)}) \mid \mathcal{I}^{(-k)}, X_i \right] \right\|^2} = o_p(n^{-\frac{1}{4}})$$

230 **Theorem 2.** Given Assumptions 1 to 3, suppose that we construct the feasible estimator $\hat{\tau}_{AIPW}$
231 (Equation (1)) using the CSBAE crossfitting procedure in Figure 3 with estimators satisfying Assump-
232 tions 6 and 7 (consistency and product error rates). Then

$$\sqrt{n}(\hat{\tau}_{AIPW} - \tau) \Rightarrow \mathcal{N}(0, V_{AIPW}),$$

233 where $V_{AIPW} = \sum_{z \in \{0,1\}} \mathbb{E} \left[\frac{\sigma_z^2(X)}{e_z(X)\pi^*(z,X)} \right] + \text{Var} [\mu_1(X) - \mu_0(X)]$. Here τ is the ATE.

234 For the proof, see Appendix G. The main result from Theorem 2 shows that the batch adaptive design
235 and feasible estimator has an asymptotic variance equal to the variance of the true ATE under missing
236 outcomes and the optimal π^* . This implies that our procedure successfully minimizes the asymptotic
237 variance bound. With this, we can also quantify the uncertainty of our treatment effect estimates by
238 producing level- α confidence intervals for τ that achieve coverage with $1 - \alpha$ probability.

239 Insights and improvements

240 **When is our method much better than uniform sampling?** Prior works of [20, 59], though they
241 do not study treatment effect estimation, obtain valid inference with uniform sampling (i.e. with
242 the budget probability). When do optimized data annotation probabilities improve upon uniform
243 sampling? To answer this, we analyze the relative efficiency.

Corollary 1 (Relative efficiency). *The relative efficiency of estimation with optimized sampling probabilities π vs. uniform sampling, for the same budget, is*

$$\text{RelEff} = \frac{\text{AVar of estimation with } \pi^*}{\text{AVar of estimation with uniform prob. } B} = \frac{\frac{1}{B} \left(\mathbb{E} \left[\sqrt{\sigma_1^2(X)} + \sqrt{\sigma_0^2(X)} \right] \right)^2 + \text{Var}[\tau(X)]}{\frac{1}{B} \mathbb{E} \left[\frac{\sigma_1^2(X)}{e_1(X)} + \frac{\sigma_0^2(X)}{e_0(X)} \right] + \text{Var}[\tau(X)]}$$

244 By construction, $\text{RelEff} \leq 1$; the smaller it is, the larger the improvement from our method. The
245 above expression reveals our method improves further as the budget grows smaller ($B \downarrow$) or if there

are imbalanced propensities where $e_1(X)$ close to 0 or 1. On the other hand, improvements from our method are limited for large budgets, $B \rightarrow 1$, or when variances in treated/control group are similar.

Direct estimation of $(e\pi^*)^{-1}$ mitigates estimation stability. It is well known that estimating propensities and then inverting estimates can be unstable in practice. This problem is doubly-so for causal inference with missing outcomes. We find many papers on adaptive treatment allocation note this challenge and mix their optimized allocation probabilities with uniform in the experimental sections [18, 59, 15]; just as many papers in causal inference clip the weights in practice [50]. Our near closed-form solution reveals that it's not necessary to estimate propensity scores for the final ATE estimation on the full dataset (though it is needed to estimate π^*). At π^* , observe that¹ $(e_z(x)\pi^*(z, x))^{-1} \propto \sqrt{\sigma_z^2(x)}^{-1}$ and is independent of the propensity score $e_z(x)$, so estimating it directly can directly exploit its *lower* statistical complexity. In causal inference and covariate shift, many methods (such as balancing weights) avoid the plug-in approach for inverse propensity methods in favor of direct estimation of the inverse propensity score [45, 60, 26, 30, 31, 13, 5]. We recommend estimation on the final dataset with such approaches or other types of direct estimation. For example, even estimation of $P(Z = z, R = 1 | X)$ directly helps:

$$\psi_z(e, \pi^*) = \frac{\mathbb{I}[Z=z, R=1]}{P(Z=z, R=1|X)}(Y - \mu_z(X)) + \mu_z(X). \quad (RZ\text{-plug-in.})$$

Extension to continuous treatments. Our analysis applies readily to other static causal inference estimands, such as those for continuous treatments. We introduce the analogous estimator and the optimal sampling probabilities. Let $e(z, X) = P(Z = z | X)$ be the generalized propensity score and $\mu(Z, X) = \mathbb{E}[Y | Z, X]$. The estimator for continuous treatments replaces the indicator function $\mathbb{I}[Z = z]$ with a local kernel function smoother localizing around z , $K_h(Z - z)$.²The following estimator for continuous treatments with missing outcomes is a direct extension of [33, 14]:

$$\psi(e, \mu) = \mu(z, X_i) + \frac{K_h(Z_i - z)\mathbb{I}[R=1]}{e(z, X_i)\pi(z, X_i)}(Y_i - \mu(z, X_i)); \quad \mathbb{E}[Y(z)] = E[\psi(e, \mu)]$$

We consider the same assumptions required as in [14], standard in kernel density estimation analysis. The optimal sampling probabilities minimize the part of the asymptotic variance of $\mathbb{E}[Y(z)]$ depending on π , subject to a budget constraint:

$$\pi^*(z, x) \in \arg \min_{\pi(z, x)} \left\{ \mathbb{E} \left[\frac{\sigma^2(z, X)}{e(z, X)\pi(z, X)} \right] : \mathbb{E}[\pi(z, X)K_h(Z - z)] \leq B \right\},$$

Theorem 3. Define the kernel localization of the generalized propensity score $e(z, x)$ around z under the kernel function $K_h(z' - z)$: $\tilde{e}_h(z, x) = \int K_h(z' - z)e(z', x)dz'$. Then

$$\pi^*(z, x) \propto \frac{\sqrt{\sigma^2(z, x)}}{e(z, x)} \sqrt{\frac{e(z, x)}{\tilde{e}_h(z, x)}}$$

The optimal sampling probabilities are quite similar, with the appropriate analogous conditional variance and generalized propensity score, up to a factor $(e(z, x)/\tilde{e}_h(z, x))^{1/2}$ from the implications of kernel-smoothing treatment for sampling budget. Consider a box kernel for simplicity, then $\tilde{e}_h(z, x)$ is the average of $e(z, x)$ over the interval $e(z - h, x), e(z + h, x)$.

6 Experiments

We evaluate our batch adaptive allocation protocol on synthetic and ground-truthed real-world datasets. We compare our method to a baseline (uniform random sampling in the same doubly-robust estimator) and a skyline (running the estimator on the complete dataset, which is generally infeasible). The baseline does not use our adaptively learned $\hat{\pi}(z, x)$, but instead uses uniform random sampling at different budget values. The skyline that we compare against is the standard AIPW estimator with fully observed outcomes, that is when the budget equals 1 or $R = 1$ for all data points.

A note on active learning. We also run pool-based active learning baselines. However, there are key differences that lead to poor performance of these baselines. In our setting, random sampling is the

¹This depends on some joint properties of κ, p_1 , whether it is feasible to find second-stage batch sampling probabilities π_2 so that $\kappa p_1 + (1 - \kappa)\pi_2(x) = \pi^*(x)$

²The kernel function $K_h(u)$, used in kernel density estimation, satisfies $\int_{-\infty}^{\infty} K(u)du = 1$ (normalizes to a probability density) and $K(-u) = K(u)$, for all u (symmetry), such as the Gaussian kernel with $K(u) = (2\pi)^{-\frac{1}{2}}e^{-u^2/2}$, or uniform $K(u) = 1/2\mathbb{I}[|u| \leq 1]$. We generally consider $K_h(u) = h^{-d}K(u/h)$.

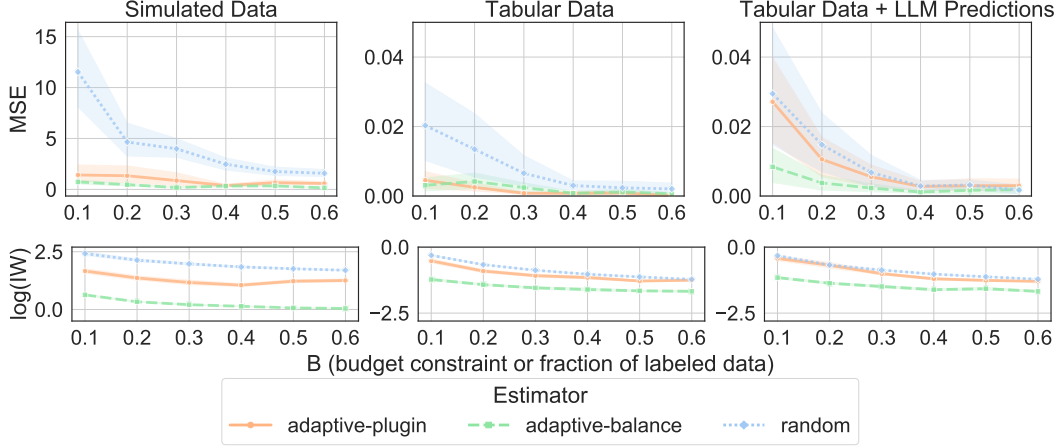


Figure 1: **Synthetic and Semi-Synthetic (Retail Hero) Data.** Mean squared error (top) and 95% confidence interval width on the log scale (bottom) averaged over 20 and 100 (for simulated data) trials across budget percentages of the data. Left and Center: Random forest prediction on tabular data. Right: including LLM predictions on text and serialized features.

strong baseline, because the sampling probabilities are much simpler. Any other sampling strategy in our 2-stage framework with AIPW performs suboptimally (since we’ve proved that ours is optimal). Additionally, the baselines used in related papers are either random sampling or the exclusion of model-based predictions (i.e. $\hat{\mu}$ or $\tilde{Y}$). However, because our task is inherently causal, our AIPW estimator relies on μ . We provide more details about active learning and baseline experiment results that we run in Appendix I.7

Retail Hero Data. We study a semi-synthetic dataset, RetailHero [52], augmented by Dhawan et al. [17] to include outcomes recorded in text. The dataset contains background customer information X , treatment Z as a text message ad sent to the customer, and outcomes Y of whether the customer made a purchase or not. Dhawan et al. [17] sampled datapoints according to an artificial propensity score and generated text from the binary outcomes prompting LLMs to generate social media posts following personas (given covariates) (details in Appendix I). These text posts are $\tilde{Y}$. The goal is to estimate the causal effect of SMS communication on purchase. This is an example of our contextual setting, where plentiful social media posts can offer insights into customer behaviors but companies may only be able to allocate a fixed amount of resources for ground-truth validation.

We implement our proposed methods using 1) random forest models to estimate the outcome model $\hat{\mu} = \mathbb{E}[Y|X]$ on tabular data only or 2) sampling from a set of five LLM predictions of purchase from social media posts $\tilde{Y}$ and then using them as predictors in a random forest to estimate $f(X, \tilde{Y})$ (We run the LLM predictions offline in batch to save cost and time). Then we estimate the outcome model $\hat{\mu}$ by ensembling, taking a weighted average between $\mathbb{E}[Y|X]$ (random forest) and $\mathbb{E}[Y|X, f(X, \tilde{Y})]$ (support vector machine), choosing the best models and weights to minimize the MSE of predicting Y on 20% of the full data. We average the results over 20 random data splits. We compute the AIPW estimator on all available data as a stand-in for ground-truth. (The dataset was too small for a separate held-out validation set). We have further experiments with simulated data to validate these results (more details on the data-generating process are given in Appendix I.)

Figure 1 shows the performance of our adaptive estimator either with 1) a direct estimation of $(e\pi^*)^{-1}$ using logistic regression that we plug-in (following Equation (RZ-plug-in.)) or 2) a random forest-based estimator of $(e\pi^*)^{-1}$ extracted from ForestRiesz [11], a random forest-based method to learn balancing weights, compared to a uniform baseline. Across different values of the budget, B , our batch adaptive procedure reduces the MSE by almost double and reduces the confidence interval width by almost one-unit in the interval width on the log scale. In Figure 10, we see the impact of our approach most clearly when we compute the percentage of the budget saved to reach the same interval width. We observe a minimum budget saved of 10% with the adaptive plug-in estimator and 45% with the adaptive balance estimator on tabular data. The LLM prediction we generate is based on simple zero-shot learning and direct serialization of the tabular data; further fine-tuning

could improve performance. Nonetheless, our method can provide robust valid guardrails around these black-box predictions.

Street Outreach Data. Next, we demonstrate our method on street outreach casenote data collected by a partnering nonprofit providing homelessness services. This analysis, which uses proprietary sensitive data, was approved by the Institutional Review Boards at [blinded for review].

The covariate data X consists of baseline characteristics on each client as tabular data (left, Figure 2), such as the number of previous outreach engagements, and (right, Figure 2) LLM generated summaries of case notes recorded before treatment. We construct the cohort in our dataset to include clients who are seen consistently at least once per month from 2019-2021. The binary treatment Z was based on the number of outreach engagements within the first 6 months of 2019. Clients with 1-2 engagements were assigned $Z = 0$ (131 clients), and those with 3-15 where assigned $Z = 1$ (355 clients). The outcome Y is the highest housing placement reached by 2021. Our final data set contained 471 clients. More information on the data can be found in Appendix I. We seek to estimate the causal effect of street outreach on housing placement. We use housing placement as an illustrative example because it is well-recorded ground truth data in our dataset. However, it could also be plausibly missing, in which case nonprofits have to decide how to expend their limited resources to obtain more information (i.e., caseworker follow-up calls or analyzing more recent casenotes $\tilde{Y}$).

Similar to Retail Hero, we demonstrate the utility of our approach by using a random forest model to estimate the outcome model on tabular data alone, $\hat{\mu} = \mathbb{E}[Y|X]$, and we incorporate LLM predictions $f(X)$ by including them as predictors in a random forest model to get $\hat{\mu} = \mathbb{E}[Y|X, f(X)]$.

In Figure 2 we see that overall our adaptive approach shows improvements over uniform random sampling. The MSE is doubled when going from both adaptive estimators to random sampling in the tabular data setting and tripled with LLM predictions from the adaptive estimator with balancing weights to random sampling. In Figure 11, we see that we can save between 43 – 75% of the budget using the plugin-in estimator on tabular data alone and by incorporating LLM predictions, and between 53 – 91% using the balance estimator over the random sampling baseline.

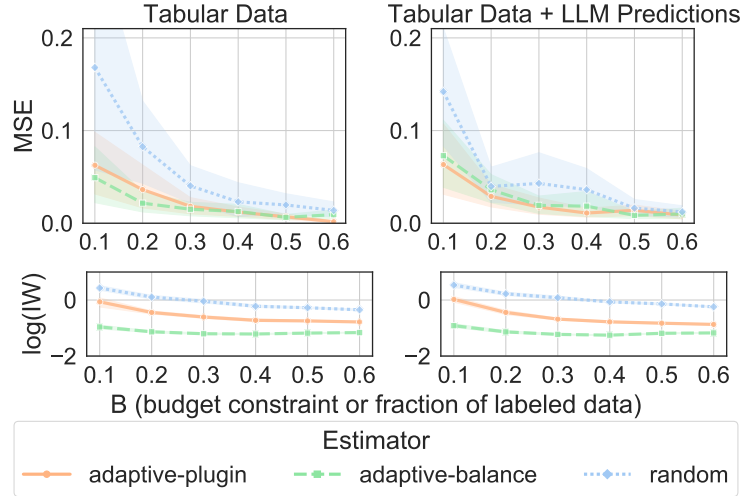


Figure 2: **Street Outreach Data.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

Conclusion, limitations, and future work. We have introduced a batch-adaptive annotation procedure and estimators that provides a framework for efficient data labeling and incorporating complex embedded outcomes into causal estimation. This work is not without limitations. We assume that annotations reveal ground truth, but there could be disagreement between expert annotators. Additionally, LLMs are still a black box and our theory requires them to be consistent to satisfy product error rate assumptions. In future work, we plan to explore other causal estimators.

References

- [1] Timothy B Armstrong. Asymptotic efficiency bounds for a class of experimental designs. *arXiv preprint arXiv:2205.02726*, 2022.
- [2] Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): pp. 133–161, 2021. ISSN 00129682, 14680262. URL <https://www.jstor.org/stable/48628848>.
- [3] Susan Athey, Raj Chetty, Guido W Imbens, and Hyunseung Kang. The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical report, National Bureau of Economic Research, 2019.
- [4] Michela Bia, Martin Huber, and Lukäs Laffärs. Double machine learning for sample selection models. *arXiv preprint arXiv:2012.00745*, 2021.
- [5] David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf019, 2025.
- [6] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, 2013. doi: 10.1109/ICDM.2013.104.
- [7] Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/ca9c267dad0305d1a6308d2a0cf1c39c-Paper.pdf.
- [8] Kamalika Chaudhuri, Prateek Jain, and Nagarajan Natarajan. Active heteroscedastic regression. In *International Conference on Machine Learning*, pages 694–702. PMLR, 2017.
- [9] Jacob M Chen, Rohit Bhattacharya, and Katherine A Keith. Proximal causal inference with text data. *arXiv preprint arXiv:2401.06687*, 2024.
- [10] 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.
- [11] Victor Chernozhukov, Whitney Newey, Victor M Quintas-Martinez, and Vasilis Syrgkanis. Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning*, pages 3901–3914. PMLR, 2022.
- [12] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *J. Artif. Int. Res.*, 4(1):129–145, March 1996. ISSN 1076-9757.
- [13] Eric R Cohn, Eli Ben-Michael, Avi Feller, and José R Zubizarreta. Balancing weights for causal inference. In *Handbook of Matching and Weighting Adjustments for Causal Inference*, pages 293–312. Chapman and Hall/CRC, 2023.
- [14] Kyle Colangelo and Ying-Ying Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.
- [15] Thomas Cook, Alan Mishler, and Aaditya Ramdas. Semiparametric efficient inference in adaptive experiments. In *Causal Learning and Reasoning*, pages 1033–1064. PMLR, 2024.
- [16] 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.
- [17] Nikita Dhawan, Leonardo Cotta, Karen Ullrich, Rahul Krishnan, and Chris J Maddison. End-to-end causal effect estimation from unstructured natural language data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

- [18] Maria Dimakopoulou, Zhimei Ren, and Zhengyuan Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34:1939–1951, 2021.
- [19] Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652, 2022.
- [20] 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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 68589–68601. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d862f7f5445255090de13b825b880d59-Paper-Conference.pdf.
- [21] Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. *Advances in Neural Information Processing Systems*, 32, 2019.
- [22] Claudio Gentile, Zhilei Wang, and Tong Zhang. Fast rates in pool-based batch active learning. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.
- [23] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66:315–331, 1998.
- [24] Jinyong Hahn, Keisuke Hirano, and Dean Karlan. Adaptive experimental design using the propensity score. *Journal of Business & Economic Statistics*, 29(1):96–108, 2011.
- [25] M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2025. ISBN 9781420076165. URL https://books.google.com/books?id=_KnHIAAACAAJ.
- [26] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.
- [27] Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004. ISSN 0034-6535. doi: 10.1162/003465304323023651. URL <https://doi.org/10.1162/003465304323023651>.
- [28] Diaz Ivan and van der Laan Mark J. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. *The International Journal of Biostatistics*, 9(2): 149–160, November 2013. doi: 10.1515/ijb-2013-0004. URL <https://ideas.repec.org/a/bpj/ijbist/v9y2013i2p149-160n8.html>.
- [29] Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schölkopf. Causal direction of data collection matters: Implications of causal and anticausal learning for nlp. *arXiv preprint arXiv:2110.03618*, 2021.
- [30] Nathan Kallus. Balanced policy evaluation and learning. *Advances in neural information processing systems*, 31, 2018.
- [31] Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(1):85–112, 2018.
- [32] Nathan Kallus and Xiaojie Mao. On the role of surrogates in the efficient estimation of treatment effects with limited outcome data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae099, 2024.
- [33] Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.
- [34] Nathan Kallus and Angela Zhou. Minimax-optimal policy learning under unobserved confounding. *Management Science*, 67(5):2870–2890, 2021.

- [35] Edward H. Kennedy. Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics*, 2020. URL <https://doi.org/10.1515/ijb-2019-0087>.
- [36] Harrison H Li and Art B Owen. Double machine learning and design in batch adaptive experiments. *Journal of Causal Inference*, 12(1):20230068, 2024.
- [37] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. ISSN 00063444, 14643510. URL <http://www.jstor.org/stable/2335739>.
- [38] Susanne M Schennach. Recent advances in the measurement error literature. *Annual Review of Economics*, 8(1):341–377, 2016.
- [39] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471*, 2012.
- [40] Burr Settles. Active learning literature survey. 2009. URL <https://api.semanticscholar.org/CorpusID:324600>.
- [41] Lei Shi, Waverly Wei, and Jingshen Wang. Using surrogates in covariate-adjusted response-adaptive randomization experiments with delayed outcomes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [42] Di Shu and Grace Y Yi. Causal inference with measurement error in outcomes: Bias analysis and estimation methods. *Statistical Methods in Medical Research*, 28(7):2049–2068, 2019. doi: 10.1177/0962280217743777. URL <https://doi.org/10.1177/0962280217743777>. PMID: 29241426.
- [43] Dhanya Sridhar and David M Blei. Causal inference from text: A commentary. *Science Advances*, 8(42):eade6585, 2022.
- [44] Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.
- [45] Yuta Tsuboi, Hisashi Kashima, Shohei Hido, Steffen Bickel, and Masashi Sugiyama. Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, 17:138–155, 2009.
- [46] Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS ’20*, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.
- [47] Victor Veitch, Dhanya Sridhar, and David Blei. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR, 2020.
- [48] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [49] Stefan Wager. Causal inference: A statistical learning approach, 2024. URL https://web.stanford.edu/~swager/causal_inf_book.pdf.
- [50] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.
- [51] Dongrui Wu, Chin-Teng Lin, and Jian Huang. Active learning for regression using greedy sampling. *Information Sciences*, 474:90–105, 2019. ISSN 0020-0255. doi: <https://doi.org/10.1016/j.ins.2018.09.060>. URL <https://www.sciencedirect.com/science/article/pii/S0020025518307680>.
- [52] X5. X5 retail hero: Uplift modeling for promotional campaign. *Challenge Dataset*, 2019.
- [53] Shu Yang and Peng Ding. Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 2020.

- 481 [54] Jinglong Zhao. Adaptive neyman allocation. *arXiv preprint arXiv:2309.08808*, 2023.
- 482 [55] Qingyuan Zhao, Dylan S Small, and Bhaswar B Bhattacharya. Sensitivity analysis for inverse
483 probability weighting estimators via the percentile bootstrap. *Journal of the Royal Statistical
484 Society Series B: Statistical Methodology*, 81(4):735–761, 2019.
- 485 [56] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
486 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and
487 chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.
- 488 [57] Yinglun Zhu and Robert Nowak. Active learning with neural networks: Insights from nonpara-
489 metric statistics. *Advances in Neural Information Processing Systems*, 35:142–155, 2022.
- 490 [58] Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Forty-first International
491 Conference on Machine Learning*, 2024.
- 492 [59] Tijana Zrnic and Emmanuel J Candès. Active statistical inference. *arXiv preprint
493 arXiv:2403.03208*, 2024.
- 494 [60] José R Zubizarreta. Stable weights that balance covariates for estimation with incomplete
495 outcome data. *Journal of the American Statistical Association*, 110(511):910–922, 2015.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: Yes, we provide our main results in Sections 4 and 5, along with any necessary assumptions. We include the proofs for all of our theorem statements in Appendix E. We also substantiate our claims by providing empirical evidence using synthetic and real world data.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Yes, we discuss some limitations of our work in the last subsection of Section 6. We also state the assumptions made by our framework in Sections 3, 4 and 5.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, we state all main theorems and the full set of assumptions that accompany them in Sections 3, 4, and 5. We include any additional results, that are not essential to the main argument, but still interesting in Appendix X. We provide the full proof derivations and any additional lemmas used in proofs in Appendix F and G.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We include the experimental details needed to reproduce the synthetic datasets. We provide the empirical results in Section 6, and include further details such as a description of the data generating process for the simulated data, description of the datasets, prompts used to query language models, and computer specifications to run experiments on in Appendix H. The data set provided by nonprofit collaborators cannot be released for privacy reasons, so instead we describe the features of the data and detailed instructions on how each variable was constructed, but we cannot release this dataset publicly. To compensate for this, we run experiments on simulated data and a second real-world dataset that is publicly available.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We plan to release the code for reproducing all the experimental results on synthetic and semi-synthetic data, along with scripts for reproducing the simulate data and running our algorithm. The street outreach data is private data that was released to us under a data use agreement, but it cannot be released publicly for privacy reasons. However, we will release our other experimental results and code with the final version of the paper.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

547 Justification: Yes, we describe the data details and construction in the experimental results
548 Sections 6 of the main paper. We go into much more detail about the data generation process
549 and model tuning details in Appendix H.

550 7. Experiment statistical significance

551 Question: Does the paper report error bars suitably and correctly defined or other appropriate
552 information about the statistical significance of the experiments?

553 Answer: [Yes]

554 Justification: We include error bands in all of our main results in Figures 2 and 3.

555 8. Experiments compute resources

556 Question: For each experiment, does the paper provide sufficient information on the com-
557 puter resources (type of compute workers, memory, time of execution) needed to reproduce
558 the experiments?

559 Answer: [Yes]

560 Justification: In Section 6, we provide sufficient information on what is needed to reproduce
561 the experiments, such as running LLM predictions offline and in batch and reference the
562 models used to run each experiment, such as random forest. We specify the type of compute
563 resources in more detail in Appendix H.

564 9. Code of ethics

565 Question: Does the research conducted in the paper conform, in every respect, with the
566 NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

567 Answer: [Yes]

568 Justification: We conform to all aspects of the NeurIPS Code of Ethics. We ensure author
569 anonymity by also removing identifying information about our nonprofit collaboration.

570 10. Broader impacts

571 Question: Does the paper discuss both potential positive societal impacts and negative
572 societal impacts of the work performed?

573 Answer: [Yes]

574 Justification: Throughout the paper, we highlight the potential positive impacts especially
575 in the introduction and motivation of this work. We discuss limitations of our method that
576 could potentially have negative societal impacts in the limitations section of the paper. We
577 go into more detail about the potential negative effects in an Impact Statement and the steps
578 that we take to mitigate these impacts in Appendix A.

579 11. Safeguards

580 Question: Does the paper describe safeguards that have been put in place for responsible
581 release of data or models that have a high risk for misuse (e.g., pretrained language models,
582 image generators, or scraped datasets)?

583 Answer: [Yes]

584 Justification: Yes, we mention how the data was released to our research team through a
585 collaboration with a nonprofit reviewed by Institutional Review Boards at author universities.
586 We discuss more about the steps taken to preserve the data privacy when training models in
587 the Impact Statement in Appendix A. We do not release any large models, but we do plan to
588 release scripts that reproduce our results on the sythethic data and publicly available data.

589 12. Licenses for existing assets

590 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
591 the paper, properly credited and are the license and terms of use explicitly mentioned and
592 properly respected?

593 Answer: [Yes]

594 Justification: Yes, we cite all the original owners of the code, data, and models in the main
595 text and in Appendix H.

596 13. New assets

597 Question: Are new assets introduced in the paper well documented and is the documentation
598 provided alongside the assets?

599 Answer: [Yes]

600 Justification: Yes, we include well documented code to run our algorithm and reproduce our
601 experimental results in an anonymized zip file included in supplementary material.

602 **14. Crowdsourcing and research with human subjects**

603 Question: For crowdsourcing experiments and research with human subjects, does the paper
604 include the full text of instructions given to participants and screenshots, if applicable, as
605 well as details about compensation (if any)?

606 Answer: [NA]

607 Justification: While one of our experiments includes real data about a vulnerable community,
608 this data was historical data collected by a nonprofit organization. None of this research
609 included crowdsourcing or directly involved human subjects.

610 **15. Institutional review board (IRB) approvals or equivalent for research with human**
611 **subjects**

612 Question: Does the paper describe potential risks incurred by study participants, whether
613 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)
614 approvals (or an equivalent approval/review based on the requirements of your country or
615 institution) were obtained?

616 Answer: [Yes]

617 Justification: Yes, we mention that IRB approval was obtained for the use of the street
618 outreach data in Section 6.

619 **16. Declaration of LLM usage**

620 Question: Does the paper describe the usage of LLMs if it is an important, original, or
621 non-standard component of the core methods in this research? Note that if the LLM is used
622 only for writing, editing, or formatting purposes and does not impact the core methodology,
623 scientific rigor, or originality of the research, declaration is not required.

624 Answer: [Yes]

625 Justification: Yes, we describe our use of LLMs to get predictions from complex embedded
626 outcome data, i.e. text data in Section 6. We explain in detail how we run the experiments
627 on a secure HIPAA compliant cloud platform in Appendix H.

A Impact Statement

Our work deals with sensitive information about a vulnerable community so care must be taken when deploying our methods. The case notes are redacted by the organization, and any sensitive information is removed from the notes. Furthermore, we use local LLMs accessed through a HIPAA-compliant fire-walled cloud instance to mitigate ensure the privacy of clients. We work in collaboration with a nonprofit to ensure that the necessary guardrails are in place and that their data is used responsibly and in line with their mission.

B Notation

Y_i	Ground truth outcomes, observed when label is provided by experts
$\tilde{Y}_i$	Complex embedded outcomes, such as raw text
X_i	Covariates included in estimation
Z_i	Treatment assignment indicator
R_i	Missingness indicator, indicates whether i is expertly labeled
$e_z(X_i)$	Propensity score, probability of being assigned treatment $Z = z$
$\pi(Z_i, X_i)$	Annotation probability, probability of sampling unit i for expert annotation
$f(\tilde{Y}_i)$	Estimated function of complex embedded outcomes, e.g. zero-shot LLM prediction from raw text
$\mu_z(X_i, f(\tilde{Y}_i))$	Estimated model predicting Y as function of $f(\tilde{Y})$ alone or $(X, f(\tilde{Y}))$

C Additional discussion on related work

Additional discussion on surrogate estimation In much of the surrogate literature, surrogates measure an outcome that is impossible to measure at the time of analysis. The canonical example in [3] studies the long-term intervention effects of job training on lifetime earnings, by using only short-term outcomes (surrogates) such as yearly earnings. In this regime, the ground truth cannot be obtained at the time of analysis. In this paper, we focus a different regime where obtaining the ground truth from expert data annotators is feasible but budget-binding.

Additional discussion on more adaptive allocation methods beyond batch. We outline how our approach is a good fit for our motivating data annotation setting. Full-adaptivity is less relevant in our setting with ground-truth annotation from human experts, due to distributed-computing-type issues with random times of annotation completion. But standard tools such as the martingale CLT can be applied to extend our theoretical results to full adaptivity. Additionally, many recent works primarily focus on the different problem of treatment allocation for ATE estimation. In-sample regret is less relevant for our setting of data annotation, which is a pure-exploration problem.

Optimizing asymptotic variance of the ATE vs. active learning. An extensive literature in machine learning studies where to sample data to improve machine learning predictors, in the subfield of active learning. The biggest difference is that we target functional estimation, aka improving estimation and inference on the average treatment effect, rather than improving estimation of the black-box nuisance predictors, so our approach is complementary to other approaches for active learning. Approaches for active learning with nonparametric regression include Zhu and Nowak [57], Chaudhuri et al. [8]. Active learning generally requires additional structural conditions, such as margin or low-noise conditions, in order to show improvements. Our work highlights optimality leveraging the structure of our final treatment effect inferential goal.

Relationship to causal inference and NLP There is a large and rapidly growing literature on causal inference with text data [19, 43, 47]. Throughout, we have deliberately used the terminology of measurement error to characterize our approach: that text measures outcomes of interest. [17] also adopt this stance towards text and note that it differs from prior works on causal inference and NLP, which focuses on questions of substantive interest related to the text itself.

Although we can define a potential outcome $\tilde{Y}(Z)$, we are generally uninterested in causal inference in the ambient high-dimensional space of $\tilde{Y}(Z)$ itself - corresponding to, in our examples, the

effect of the presence of a tumor on the pixel image, the effect of street outreach on the linguistic characteristics of casenotes written for documentation, etc — $\tilde{Y}(Z)$ is relevant to causal estimation insofar as it is informative of latent outcomes $Y(Z)$.

This is consistent with viewing certain types of NLP tasks as “anti-causal learning” [39], wherein outcomes cause measurements thereof, in analogy to anti-causal learning in supervised classification where a label of “cat” or “dog” causes the classification covariates (e.g. image) [29]. Analogously, we view the underlying ground-truth outcomes Y as causing the measurement thereof, $\tilde{Y}$.

D Diagram of Cross-fitting Procedure

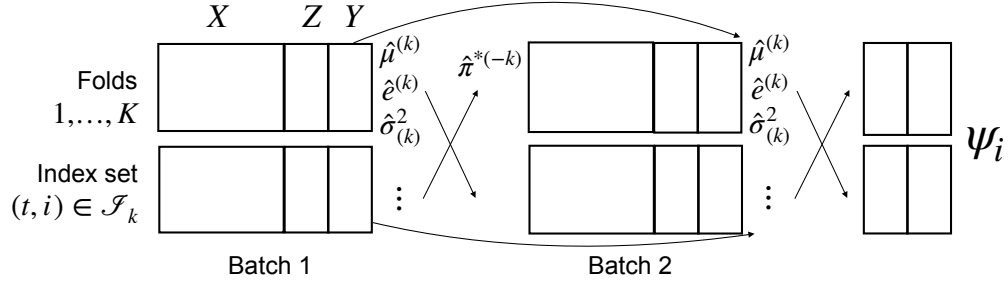


Figure 3: Illustration of cross-fitting (K folds within batches)

674 E Algorithm

Algorithm 2 (Full Algorithm) Batch Adaptive Causal Estimation With Complex Embedded Outcomes

Input: Data $\mathcal{D} = \{(X_i, Z_i, Y_i, \tilde{Y}_i)\}_{i=1}^n$, sampling budget B_z for $z \in \{0, 1\}$
Output: ATE estimator $\hat{\tau}_{AIPW}$
 Partition $\mathcal{D}$ into 2 batches and K folds $\mathcal{D}_1^{(k)}, \mathcal{D}_2^{(k)}$ for $k = 1, \dots, K$
Batch 1:
for $k = 1, \dots, K$ **do**
 On $\mathcal{D}_1^{(k)}$: Sample $R_1 \sim \text{Bern}(\pi_1(Z, X))$, where $\pi_1(z, x) = B_z$.
 Estimate nuisance models: Where $R = 1$, estimate $\hat{\mu}_z^{(k)}$ by regressing Y on $X, \tilde{Y}$, and $\hat{\sigma}_z^{2(k)}$ by regressing $(Y - \hat{\mu}_z)^2$ on X . Estimate $\hat{e}_z^{(k)}$ by regressing Z on X .
end for
Batch 2:
for $k = 1, \dots, K$ **do**
 On $\mathcal{D}_2^{(k)}$: Obtain π^* by optimizing eq. (OPT (global budget)), plugging in $\hat{\mu}_z^{(-k)}, \hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$.
 Solve for $\hat{\pi}_2^{(k)}(X_i) = \frac{1}{1-\kappa}(\pi^*(X_i) - \kappa\pi_1)$
 Sample $R_2 \sim \text{Bern}(\hat{\pi}_2^{(k)}(X_i))$
end for
 Obtain $\mathcal{D}^{(k)}$ for $k = 1, \dots, K$ by pooling across batches $\mathcal{D}_1^{(k)}$ and $\mathcal{D}_2^{(k)}$
 On $\mathcal{D}^{(k)}$, re-estimate $\hat{\mu}_z^{(k)}, \hat{\sigma}_z^{2(k)}$, and $\hat{e}_z^{(k)}$ on observed outcomes RY for $k = 1, \dots, K$
 On $\mathcal{D}^{(k)}$, run optimization procedure to get $\pi^{*(-k)}$ with out of fold nuisances $\hat{\mu}_z^{(-k)}, \hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$.
 On full data $\mathcal{D}$, estimate ATE by using AIPW estimator in eq. (1) and out of fold nuisances $\pi^{*(-k)}, \hat{\mu}_z^{(-k)}, \hat{\sigma}_z^{2(-k)}$, and $\hat{e}_z^{(-k)}$

675 F Additional Results

676 F.1 Treatment- z -specific budgets B_z

We also consider a setting with different a priori fixed budgets within each treatment group, where sampling budget proportion $B_z \in [0, 1]$

677 is the max percentage of the treated group $Z = z$ that can be annotated. Given that we are trying to
 678 choose the π that minimizes this variance bound, we only need to focus on the terms that depend
 679 on π and can drop the rest. Supposing oracle knowledge of propensities and outcome models, the
 680 optimization problem, for each $z \in \{0, 1\}$ is:

$$\min_{0 < \pi(z, x) \leq 1, \forall z, x} \left\{ \mathbb{E} \left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z, X)} \right] : \mathbb{E} [\pi(z, X) \mid Z = z] \leq B_z, z \in \{0, 1\} \right\} \quad (\text{z-budget})$$

Theorem 4. *The solution to the within- z -budget problem is:*

$$\pi^*(z, X) = \frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E} \left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z \right]} \cdot B_z$$

681 .

682 **G Proofs**

683 *Proof of Proposition 1.* We simplify the expression for the asymptotic variance of the ATE with
 684 missing outcomes to isolate the components affected by the data annotation probability.

685 First the variance of the ATE defined in terms of the efficient influence function ψ_z is

$$\begin{aligned} \text{Var}[\psi_z - \psi_{z'}] &= \text{Var} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X) - \frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X) \right] \\ &= \underbrace{\text{Var} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X) \right]}_{V_1} + \underbrace{\text{Var} \left[\frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X) \right]}_{V_2} \\ &\quad - \underbrace{2\text{Cov} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X), \frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X) \right]}_{V_3} \end{aligned}$$

686 **For V_3 :**

$$\begin{aligned} &2\text{Cov} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X), \frac{Z' \cdot R \cdot [Y - \mu_{z'}(X)]}{e_{z'}(X) \cdot \pi(z', X)} + \mu_{z'}(X) \right] \\ &= 2 \left[\mathbb{E} \left[\frac{Z \cdot R}{e_z(X) \cdot \pi(z, X)} \underbrace{[\mathbb{E}[Y|Z, R=1, X] - \mu_z(X)]}_{=0} \right] \right] \\ &\quad + \left[\mathbb{E} \left[\mu_z(X) \cdot \frac{Z' \cdot R}{e_{z'}(X) \cdot \pi(z', X)} \underbrace{[\mathbb{E}[Y|Z', R=1, X] - \mu_{z'}(X)]}_{=0} + \mu_{z'}(X) \right] \right] \\ &\quad - \mathbb{E} \left[\frac{Z \cdot R}{e_z(X) \cdot \pi(z, X)} \underbrace{[\mathbb{E}[Y|Z, R=1, X] - \mu_z(X)]}_{=0} + \mu_z(X) \right] \\ &\quad \times \mathbb{E} \left[\frac{Z' \cdot R}{e_{z'}(X) \cdot \pi(z', X)} \underbrace{[\mathbb{E}[Y|Z', R=1, X] - \mu_{z'}(X)]}_{=0} + \mu_{z'}(X) \right] \\ &= 2 \left[\mathbb{E}[\mu_z(X) \cdot \mu_{z'}(X)] - \mathbb{E}[\mu_z(X)\mu_{z'}(X)] \right] \end{aligned}$$

687 **For V_1 :**

$$\begin{aligned} &\text{Var} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} + \mu_z(X) \right] \\ &= \text{Var} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} \right] + \text{Var}[\mu_z(X)] + \underbrace{2 \text{Cov} \left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)}, \mu_z(X) \right]}_{=0} \\ &= \mathbb{E} \left[\left[\frac{Z \cdot R \cdot [Y - \mu_z(X)]}{e_z(X) \cdot \pi(z, X)} \right]^2 \right] - \left[\frac{Z \cdot R}{e_z(X) \cdot \pi(z, X)} \underbrace{[\mathbb{E}[Y|Z, R=1, X] - \mu_z(X)]}_{=0} \right]^2 \\ &\quad + \mathbb{E}[\mu_z(X)^2] - \mathbb{E}[\mu_z(X)]^2 \\ &= \mathbb{E} \left[\left[\frac{Z^2 \cdot R^2}{e_z^2(X) \cdot \pi^2(z, X)} \cdot [Y - \mu_z(X)]^2 \right] \right] + \mathbb{E}[\mu_z(X)^2] - \mathbb{E}[\mu_z(X)]^2 \\ &= \mathbb{E} \left[\frac{Z \cdot R}{e_z^2(X) \cdot \pi^2(z, X)} \cdot [Y - \mu_z(X)]^2 \right] + \mathbb{E}[\mu(z, 1, X)^2] - \mathbb{E}[\mu_z(X)]^2 \\ &= \mathbb{E} \left[\frac{1}{e_z(X) \cdot \pi(z, X)} \cdot [Y - \mu_z(X)]^2 \right] + \mathbb{E}[\mu_z(X)^2] - \mathbb{E}[\mu_z(X)]^2 \end{aligned}$$

688 Lastly, $V_1 = V_2$. So the full variance term is

$$\begin{aligned}\text{Var}[\psi_z - \psi_{z'}] &= \mathbb{E} \left[\frac{1}{e_z(X) \cdot \pi(z, X)} \cdot [Y - \mu_z(X)]^2 \right] + \mathbb{E} \left[\frac{1}{e_{z'}(X) \cdot \pi(z', X)} \cdot [Y - \mu_{z'}(X)]^2 \right] \\ &\quad + \mathbb{E} [(\mu_z(X) - \mu_{z'}(X))^2] - \mathbb{E} [\mu_z(X) - \mu_{z'}(X)]^2 \\ &= \mathbb{E} \left[\frac{1}{e_z(X) \cdot \pi(z, X)} \cdot [Y - \mu_z(X)]^2 \right] + \mathbb{E} \left[\frac{1}{e_{z'}(X) \cdot \pi(z', X)} \cdot [Y - \mu_{z'}(X)]^2 \right] \\ &\quad + \text{Var} [\mu_z(X) - \mu_{z'}(X)]\end{aligned}$$

689 Rewriting the bound from Hahn (1998), we get

$$\begin{aligned}V &\geq \mathbb{E} \left[\frac{1}{e_z(X) \cdot \pi(z, X)} \cdot [Y - \mu_z(X)]^2 \right] + \mathbb{E} \left[\frac{1}{e_{z'}(X) \cdot \pi(z', X)} \cdot [Y - \mu_{z'}(X)]^2 \right] \\ &\quad + \text{Var} [\mu_z(X) - \mu_{z'}(X)]\end{aligned}$$

690 □

691 *Proof of Theorem 4.* Finding the optimal π can be separated into sub-problems for each treatment
692 $z \in \{0, 1\}$, since the objective and dual variables are separable across z . We first look at a solution
693 for $\pi(z, X)$ for a given z :

$$\begin{aligned}\min_{\pi(z, x)} \mathbb{E} \left[\frac{\sigma_z^2(X)}{e_z(X) \pi(z, X)} \right] & \quad (\text{z-budget}) \\ \text{s.t. } \mathbb{E} [\pi(z, X) \mid Z = z] &\leq B_z, \\ 0 < \pi(z, x) &\leq 1, \forall x\end{aligned}$$

We define the Lagrangian of the optimization problem and introduce dual variables λ for the budget constraint and η and ν for the constraint that $0 < \pi(z, X) \leq 1$:

$$\mathcal{L} = \mathbb{E} \left[\frac{(Y - \mu_z(X))^2}{e_z(X) \pi(z, X)} \right] + \lambda_z (\mathbb{E} [\pi(z, X) \mid Z = z] - B_z) + \sum_{x \in \mathcal{X}} (\nu_x^z (\pi(z, x) - 1) - \eta_x^z \pi(z, x))$$

Define the conditional outcome variance $\sigma^2(X) = \mathbb{E} [(Y - \mu(z, 1, X))^2 \mid X]$. Note that by iterated expectations,

$$\mathcal{L} = \mathbb{E} \left[\frac{\sigma_z^2(X)}{e_z(X) \pi(z, X)} \right] + \lambda_z (\mathbb{E} [\pi(z, X) \mid Z = z] - B_z) + \sum_{x \in \mathcal{X}} (\nu_x^z (\pi(z, x) - 1) - \eta_x^z \pi(z, x))$$

694 We can find the optimal solution by setting the derivative equal to 0. Since $p(X = x \mid Z = z) =$
695 $\frac{e_z(x)p(x)}{p(Z=z)}$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi(z, X)} &= -\frac{\sigma^2(X)}{e_z(X)(\pi^2(z, X))} p(x) + \lambda_z \frac{e_z(x)p(x)}{p(Z=z)} + \nu_x - \eta_x = 0, \text{ where } p(x) > 0 \\ &= -\frac{\sigma^2(X)}{e_z^2(X)\pi^2(z, X)} + \frac{\lambda_z}{p(Z=z)} + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)} = 0\end{aligned}$$

Therefore

$$\pi(z, x) = \sqrt{\frac{\sigma^2(x)}{e_z^2(x) \left(\frac{\lambda_z}{p(Z=z)} + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)} \right)}}$$

696 Next we give a choice of λ that results in an interior solution with $0 \leq \pi(z, x) \leq 1$, so that ν_x^z, η_x^z
 697 can be set to 0 without loss of generality to satisfy complementary slackness.

We posit a closed form solution

$$\pi^*(z, X) = \frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E} \left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z \right]} \cdot B_z$$

698 .

699 Note that this solution is self-normalized to satisfy the budget constraint such that

$$\mathbb{E} [\pi^*(z, X) \mathbb{I}[Z = z]] = \mathbb{E} \left[\frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E} \left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z \right]} B_z \mid Z = z \right] = B_z$$

700 This solution corresponds to a choice of $\lambda_z^* = p(Z=z) \mathbb{E} \left[\sqrt{\mathbb{I}[Z=z] \sigma^2(X)/e_z^2(X)} \right]^2 / B_z^2$ in the prior
 701 parametrized expression.

$$\begin{aligned} \pi_\lambda(z, X) &= \pi^*(z, X) \\ \sqrt{\frac{\sigma_z^2(X)}{e_z^2(X) \frac{\lambda}{p(Z=z)}}} &= \frac{\sqrt{\sigma_z^2(X)/e_z^2(X)}}{\mathbb{E} \left[\sqrt{\sigma_z^2(X)/e_z^2(X)} \mid Z = z \right]} \cdot B_z \end{aligned}$$

702 We can check that the KKT conditions are satisfied at $\pi^*(z, X)$ and λ^* . We note that since $\pi^*(z, X)$
 703 is an interior solution then w.l.o.g we can fix $\nu_x, \eta_x = 0$ to satisfy complementary slackness.

704 It remains to check that $\frac{\partial \mathcal{L}}{\partial \pi^*(z, X)} = 0$, we have that:

$$\frac{\partial \mathcal{L}}{\partial \pi(z, X)} = -\frac{\sigma_z^2(X)}{e_z(X)} \cdot \frac{e_z^2(X) \mathbb{E} \left[\sqrt{\sigma_z^2(X)/e_z(X)} \mid Z = z \right]^2}{\sigma_z^2(X) \cdot B_z^2} + \frac{\mathbb{E} \left[\sqrt{\sigma^2(X)/e_z(X)} \mid Z = z \right]^2 \sigma_z^2(X) e_z(X)}{\sigma_z^2(X) \cdot B_z^2} + 0 = 0.$$

705 Thus we have shown that $\pi^*(z, X)$ is optimal.

706

□

707 *Proof of Theorem 1* . Proceed as in the proof of Theorem 4.

708 The Lagrangian of the optimization problem (with a single global budget constraint) is:

$$\begin{aligned} \mathcal{L} = & \sum_{z \in \{0,1\}} \mathbb{E} \left[\frac{(Y - \mu_z(X))^2}{e_z(X) \pi(z, X)} \right] + \sum_{x \in \mathcal{X}} (\nu_x^z (\pi(z, x) - 1) - \eta_x^z \pi(z, x)) \\ & + \lambda (\mathbb{E} [\pi(1, X) \mathbb{I}[Z = 1] + \pi(0, X) \mathbb{I}[Z = 0]] - B) \end{aligned}$$

Again by iterated expectations,

$$\mathcal{L} = \mathbb{E} \left[\frac{\sigma_z^2(X)}{e_z(X) \pi(z, X)} \right] + \lambda (\mathbb{E} [\pi(1, X) e_1(X) + \pi(0, X) e_0(X)] - B_z) + \sum_{x \in \mathcal{X}} (\nu_x^z (\pi(z, x) - 1) - \eta_x^z \pi(z, x))$$

709 We can find the optimal solution by setting the derivative equal to 0.

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi(z, X)} &= -\frac{\sigma^2(X)}{e_z(X) (\pi^2(z, X))} p(x) + \lambda p(x) e_z(x) + \nu_x^z - \eta_x^z = 0, \text{ where } p(x) > 0 \\ &= -\frac{\sigma^2(X)}{e_z^2(X) \pi^2(z, X)} + \lambda + \frac{(\nu_x^z - \eta_x^z)}{p(x) e_z(x)} = 0 \end{aligned}$$

Therefore we obtain a similar expression parametrized in λ , but this parameter is the same across both groups under a global budget.

$$\pi(z, x) = \sqrt{\frac{\sigma^2(x)}{e_z^2(x)(\lambda + \frac{(\nu_x^z - \eta_x^z)}{p(x)e_z(x)})}}$$

We can similarly give a closed-form expression for a different choice of λ yielding an interior solution, so that we can set $\nu_x^z, \eta_x^z = 0$ without loss of generality.

$$\lambda = \frac{\mathbb{E} \left[\mathbb{I}[Z = 1] \sqrt{\sigma_1^2(X)/e_1^2(X)} + \mathbb{I}[Z = 0] \sqrt{\sigma_0^2(X)/e_0^2(X)} \right]^2}{B^2}$$

710 Notice that this satisfies the normalization requirement that $\mathbb{E}[\pi^\lambda(1, X)\mathbb{I}[Z = 1] + \pi^\lambda(0, X)\mathbb{I}[Z =$
711 $0]] \leq B$, and similarly note that the partial derivatives with respect to $\pi(z, x)$ are 0. $\square$

712 *Proof of Theorem 3 .* The objective function arises from the asymptotic variance expression in [14,
713 Thm. 3]; it follows readily from following their proof of Thm. 3 with our analysis of the asymptotic
714 variance as in Proposition 1. The proof of the optimal solution follows our analysis in Theorem 1
715 with a few slightly different expressions, discussed as follows.

Then the Lagrangian is

$$\int \frac{\sigma^2(z | x)}{e(z, x)\pi(z, x)} f(x) dx + \lambda \left(\int \int \pi(z, x) K_h(z' - z) e(z', x) dz' f(x) dx \right)$$

Define the kernel localization of $e(z, x)$ around z under the kernel function $K_h(z' - z)$:

$$\tilde{e}_h(z, x) = \int K_h(z' - z) e(z', x) dz'$$

Taking derivatives with respect to $\pi(z, x)$, we obtain the FOC

$$\nabla_{\pi(t|x)} \mathcal{L} = \frac{-\sigma^2(z, x)}{e(z, x)\pi(z, x)^2} f(x) + \lambda \tilde{e}_h(z, x) f(x) = 0$$

Solving the FOC, we obtain

$$\frac{-\sigma^2(z, x)}{e(z, x)\pi(z, x)^2} + \lambda \tilde{e}_h(z, x) = 0 \implies \pi^*(z, x) = \frac{1}{\lambda} \frac{\sqrt{\sigma^2(z, x)}}{e(z, x)} \sqrt{\frac{e(z, x)}{\tilde{e}_h(z, x)}}$$

We conclude that

$$\pi^*(z, x) \propto \frac{\sqrt{\sigma^2(z, x)}}{e(z, x)} \sqrt{\frac{e(z, x)}{\tilde{e}_h(z, x)}}$$

716 $\square$

717 *Proof of Theorem 2 . Proof sketch.*

718 The proof proceeds in two steps. The first establishes that the feasible AIPW estimator converges to
719 the AIPW estimator with oracle nuisances. It follows from standard analysis with cross-fitting, in
720 particular the variant used across batches.

Preliminaries In the analysis, we write the score function as a function of R in addition to other nuisance functions:

$$\psi_{z,i}(R_i, e, \pi, \mu) = \frac{\mathbb{I}[Z_i = z] R_i (Y_i - \mu_z(X_i))}{e_z(X_i) \pi(z, X_i)} + \mu_z(X_i)$$

The AIPW estimator can be rewritten as a sum over estimators within batch- t , fold- k , $\hat{\tau}_{AIPW}^{(t,k)}$, as follows:

$$\hat{\tau}_{AIPW} = \sum_{t=1}^2 \sum_{k=1}^K \frac{n_{t,k}}{n} \sum_{(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \{ \hat{\psi}_{1,i}(R, \hat{e}, \hat{\pi}, \hat{\mu}) - \hat{\psi}_{0,i}(R, \hat{e}, \hat{\pi}, \hat{\mu}) \} = \sum_{t=1}^2 \sum_{k=1}^K \frac{n_{t,k}}{n} \hat{\tau}_{AIPW}^{(t,k)}$$

We introduce an intermediate quantity. The realized treatments are sampled with probability $\hat{\pi}(X_i)$, $R_i \sim \text{Bern}(\hat{\pi}(Z_i, X_i))$. In the asymptotic framework, we study treatments sampled from a mixture distribution over the two batches, $\tilde{R}_i \sim \text{Bern}(\pi^*(Z_i, X_i))$.

$$\tilde{\tau}_{AIPW} = \sum_{t=1}^2 \sum_{k=1}^K \frac{n_{t,k}}{n} \sum_{(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \{ \hat{\psi}_{1,i}(\tilde{R}, \hat{e}, \hat{\pi}, \hat{\mu}) - \hat{\psi}_{0,i}(\tilde{R}, \hat{e}, \hat{\pi}, \hat{\mu}) \}$$

721 We also denote the AIPW estimator with oracle nuisances, $\hat{\tau}_{AIPW}^*$, as

$$\hat{\tau}_{AIPW}^* = \sum_{t=1}^2 \sum_{k=1}^K \frac{n_{t,k}}{n} \sum_{(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \{ \psi_{1,i}(\tilde{R}_i, e, \pi, \mu) - \psi_{0,i}(\tilde{R}_i, e, \pi, \mu) \}$$

722 We study convergence within a batch- t , fold- k subset; the decompositions above give that conver-
723 gence also holds for the original estimators.

The first step studies the limiting mixture distribution propensity arising from the two-batch process and shows that the use of the double-machine learning estimator (AIPW), under the weaker product error assumptions, gives that the oracle estimator is asymptotically equivalent to the oracle estimator where missingness follows the limiting mixture missingness probability. The latter of these is a sample average of iid terms and follows a standard central limit theorem. Recalling that $\tilde{R}_i = \mathbb{I}[U_i \geq \pi^*(X_i)]$, we wish to show:

$$\sum_z \mathbb{E}_n[\psi_{z,i}(R, \hat{e}, \hat{\pi}, \hat{\mu})] - \mathbb{E}_n[\psi_{z,i}(\tilde{R}, e, \pi, \mu)] = o_p(n^{-\frac{1}{2}}).$$

Next we show that the estimator with feasible nuisance estimators converges to the estimator with oracle knowledge of the nuisance functions

$$\sqrt{n}(\tilde{\tau}_{AIPW}^{(t,k)} - \hat{\tau}_{AIPW}^{(t,k)}) \rightarrow_p 0.$$

724 The result follows by the standard limit theorem applied to the estimator with oracle nuisance
725 functions.

726 Step 1

727 Let $\tilde{R}_i = \mathbb{I}[U_i \geq \pi^*(Z_i, X_i)]$. Restricting attention to a single treatment value $z \in \{0, 1\}$, we want
728 to show that:

$$\begin{aligned} & \sum_{t=1}^2 \sum_{k=1}^K \frac{n_{t,k}}{n} \sum_{(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \{ \hat{\psi}_{1,i}(\tilde{R}, \hat{e}, \hat{\pi}, \hat{\mu}) - \hat{\psi}_{1,i}(R, \hat{e}, \hat{\pi}, \hat{\mu}) \} \\ &= \sum_{t=1}^2 \sum_{k=1}^K \frac{n_{t,k}}{n} \sum_{(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \left\{ \frac{\mathbb{I}[Z_i = z] \tilde{R}_i (Y_i - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i) \hat{\pi}(z, X_i)} - \frac{\mathbb{I}[Z_i = z] R_i (Y_i - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i) \hat{\pi}(z, X_i)} \right\} = o_p(n^{-1/2}). \end{aligned}$$

729 Without loss of generality we further consider one summand on batch- t , fold- k data, the same
730 argument will apply to the other summands and the final estimator.

Note that by consistency of potential outcomes, for any data point we have that

$$\frac{\mathbb{I}[Z_i = z] \tilde{R}_i (Y_i - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i) \hat{\pi}(z, X_i)} - \frac{\mathbb{I}[Z_i = z] R_i (Y_i - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i) \hat{\pi}(z, X_i)} = \frac{\mathbb{I}[Z_i = z] (\tilde{R}_i - R_i) (Y_i(z) - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i) \hat{\pi}(z, X_i)}$$

731 For each batch $t = 1, \dots, T$ and fold $k = 1, \dots, K$, according to the CSBAE crossfitting procedure,
732 we observe that conditional on $\mathcal{I}_{(-k)}$ for a given batch and the observed covariates, the summands

(namely $R_i = \mathbb{I}[U_i \leq \hat{\pi}^{(-k)}(X_i)]$) are independent mean-zero. The final estimator will consist of the sum over batches and folds. We start by looking at the estimator over one batch t and one fold k and the rest follows for the other batches and folds.

$$\begin{aligned}
& \frac{1}{n_{t,k}} \sum_{(t,i) \in \mathcal{I}_k} \frac{\mathbb{I}[Z_i = z](\tilde{R}_i - R_i)(Y_i(z) - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z, X_i)} \\
&= \frac{1}{n_{t,k}} \sum_{(t,i) \in \mathcal{I}_k} \frac{\mathbb{I}[Z_i = z] \left((\tilde{R}_i - \pi^*(z, X_i)) + (\pi^*(z, X_i) - \hat{\pi}(z, X_i)) + (\hat{\pi}(z, X_i) - R_i) \right) (Y_i(z) - \hat{\mu}_z(X_i))}{\hat{e}_z(X_i)\hat{\pi}(z, X_i)} \\
&\leq \nu_e \gamma \sigma^2 \frac{1}{n_{t,k}} \sum_{(t,i) \in \mathcal{I}_k} \mathbb{I}[Z_i = z] \left((\tilde{R}_i - \pi^*(z, X_i)) + (\pi^*(z, X_i) - \hat{\pi}(z, X_i)) + (\hat{\pi}(z, X_i) - R_i) \right) (Y_i(z) - \hat{\mu}_z(X_i))
\end{aligned}$$

Applying Cauchy-Schwarz to each of these terms, we obtain product error rate terms. For the second term, we obtain that

$$\begin{aligned}
& \nu_e \gamma \sigma^2 \frac{1}{n_{t,k}} \sum_{(t,i) \in \mathcal{I}_k^z} (\pi^*(X_i) - \hat{\pi}(X_i))(Y_i(z) - \hat{\mu}_z(X_i)) \\
&\leq \nu_e \gamma \sigma^2 \sqrt{\frac{1}{n_{t,k}} \sum_{(t,i) \in \mathcal{I}_k^z} (\pi^*(X_i) - \hat{\pi}(X_i))^2} \sqrt{\frac{1}{n_{t,k}} \sum_{(t,i) \in \mathcal{I}_k^z} (Y_i(z) - \hat{\mu}_z(X_i))^2} \\
&= \nu_e \gamma \sigma^2 \|\pi^*(X_i) - \hat{\pi}(X_i)\|_{2,n} \|Y_i(z) - \hat{\mu}_z(X_i)\|_{2,n} \\
&= o_p(n^{-\frac{1}{2}}) \quad (\text{Assumption 7})
\end{aligned}$$

Analogously, we conclude that the first and third terms are $o_p(n^{-\frac{1}{2}})$, applying Cauchy-Schwarz to each of them in turn.

Step 2 (feasible estimator converges to oracle)

If we look at one term for one treatment and datapoint in the above (the rest follows for the others), we obtain the following decomposition into error and product-error terms:

$$\begin{aligned}
& \frac{Z_i \tilde{R}_i (Y_i - \hat{\mu}_1(X_i))}{\hat{e}_1(X_i) \hat{\pi}(1, X_i)} - \frac{Z_i \tilde{R}_i (Y_i - \mu_1(X_i))}{e_1(X_i) \pi(1, X_i)} + (\hat{\mu}_1(X_i) - \mu_1(X_i)) \\
&= (\mu_1(X_i) - \hat{\mu}_1(X_i)) \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right) + Z_i \tilde{R}_i (Y_i - \hat{\mu}_1(X_i)) \left(\frac{1}{\hat{e}_1(X_i) \hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i) \pi(1, X_i)} \right) \\
&\quad (\text{by } \pm \frac{Z_i \tilde{R}_i (Y_i - \hat{\mu}_1(X_i))}{e_1(X_i) \pi(1, X_i)}) \\
&= (\mu_1(X_i) - \hat{\mu}_1(X_i)) \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right) + Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \left(\frac{1}{\hat{e}_1(X_i) \hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i) \pi(1, X_i)} \right) \\
&\quad + Z_i \tilde{R}_i (\mu_1(X_i) - \hat{\mu}_1(X_i)) \left(\frac{1}{\hat{e}_1(X_i) \hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i) \pi(1, X_i)} \right) \\
&\quad (\text{by } \pm Z_i \tilde{R}_i \mu_1(X_i) \left(\frac{1}{\hat{e}_1(X_i) \hat{\pi}(1, X_i)} - \frac{1}{e_1(X_i) \pi(1, X_i)} \right)) \\
&= (\mu_1(X_i) - \hat{\mu}_1(X_i)) \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right) \\
&\quad + Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \left(\hat{\pi}(1, X_i)^{-1} (\hat{e}_1(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1} (\hat{\pi}(1, X_i)^{-1} - \pi(1, X_i)^{-1}) \right) \\
&\quad + Z_i \tilde{R}_i (\mu_1(X_i) - \hat{\mu}_1(X_i)) \left(\hat{\pi}(1, X_i)^{-1} (\hat{e}_1(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1} (\hat{\pi}(1, X_i)^{-1} - \pi(1, X_i)^{-1}) \right) \\
&\quad (\text{by } \pm \frac{1}{e\hat{\pi}})
\end{aligned}$$

We want to show that

$$\sqrt{n_{t,k}} (\hat{\tau}_{AIPW}^{(t,k)} - \hat{\tau}_{AIPW}^{*,(t,k)}) \rightarrow_p 0$$

743 Now that we have written out this expansion for one datapoints, we can write out this expansion
 744 within a batch- t , fold- k subset, and write out the cross-fitting terms for reference:

$$\begin{aligned}
 & \sqrt{n_{t,k}} \left(\hat{\tau}_{AIPW}^{(t,k)} - \hat{\tau}_{AIPW}^{*,(t,k)} \right) \\
 &= \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(1, X_i)) \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right) \\
 &+ \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \times \\
 &\quad \left(\hat{\pi}^{(-k)}(1, X_i)^{-1} (\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1} (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1}) \right) \\
 &+ \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} Z_i \tilde{R}_i (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(1, X_i)) \times \\
 &\quad \left(\hat{\pi}^{(-k)}(1, X_i)^{-1} (\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) + e_1(X_i)^{-1} (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1}) \right)
 \end{aligned}$$

745 **Bound for third term:**

$$\begin{aligned}
 & \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} Z_i \tilde{R}_i (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) (\hat{\pi}^{(-k)}(1, X_i)^{-1} (\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) \\
 &\quad + e_1(X_i)^{-1} (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1})) \\
 &= \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} Z_i \tilde{R}_i \hat{\pi}^{(-k)}(1, X_i)^{-1} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) (\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) \\
 &\quad + Z_i \tilde{R}_i e_1(X_i)^{-1} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1}) \\
 &\leq (\lambda_\pi + \nu_e) \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) (\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1}) \\
 &\quad + (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1}) \\
 &\leq (\lambda_\pi + \nu_e) \delta_n n^{-1/2}
 \end{aligned}$$

746 where the last inequality makes use of product error rate assumptions 5-6 and nuisance function
 747 convergence rates from Lemma 4. Thus, we find that this term is $o_p(1/\sqrt{n})$

748 **Bound for the first term:**

749 The key to bounding the first term is that cross-fitting allows us to treat this term as the average
 750 of independent mean-zero random variables. We will bound it with Chebyshev's inequality, which
 751 requires a bound on the second moment on the summands in the first term.

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} \left((\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right) \right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&= \text{Var} \left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i)) \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right) \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \mathbb{E} \left[(\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))^2 \left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&\quad \text{(expectation of } (\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1)^2) \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \frac{1 - e_1(X_i) \pi(z, X_i)}{e_1(X_i) \pi(1, X_i)} (\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))^2 \\
&\leq \frac{1 - \nu_e \lambda_\pi}{\nu_e \lambda_\pi} \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} ((\mu_1(X_i) - \hat{\mu}_1^{(-k)}(X_i))^2) = o_p\left(\frac{1}{n^{1+2r_\mu}}\right)
\end{aligned}$$

752 where for the third equality, we use the fact that

$$753 \quad \mathbb{E} \left[\left(\frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} - 1 \right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] = \mathbb{E} \left[\left(\frac{Z_i^2 \tilde{R}_i^2}{e_1^2(X_i) \pi^2(1, X_i)} - 2 \frac{Z_i \tilde{R}_i}{e_1(X_i) \pi(1, X_i)} + 1 \right) \mid \mathcal{I}_{(-k)}, \{X_i\} \right] = \frac{1}{e_1(X_i) \pi(1, X_i)} - 1$$

754 Since $r_\mu \geq 0$, we can conclude by Chebyshev's inequality that the first term is $o_p(n^{-1/2})$.

755 **Bound for the second term:** We bound the second term following a similar argument as above.

$$\begin{aligned}
& \mathbb{E} \left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} \left(Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \left(\hat{\pi}^{(-k)}(1, X_i)^{-1} (\hat{e}_1^{(-k)}(X_i))^{-1} - e_1(X_i)^{-1} \right) \right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&+ \mathbb{E} \left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} \left(Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \left(e_1(X_i)^{-1} (\hat{\pi}^{(-k)}(1, X_i))^{-1} - \pi(1, X_i)^{-1} \right) \right)^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&= \text{Var} \left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} \left(Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \left(\hat{\pi}^{(-k)}(1, X_i)^{-1} (\hat{e}_1^{(-k)}(X_i))^{-1} - e_1(X_i)^{-1} \right) \right) \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&+ \text{Var} \left[\frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} \left(Z_i \tilde{R}_i (Y_i - \mu_1(X_i)) \left(e_1(X_i)^{-1} (\hat{\pi}^{(-k)}(1, X_i))^{-1} - \pi(1, X_i)^{-1} \right) \right) \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \mathbb{E} \left[\left(\hat{\pi}^{(-k)}(1, X_i)^{-1} (\hat{e}_1^{(-k)}(X_i))^{-1} - e_1(X_i)^{-1} \right)^2 \frac{Z_i^2 \tilde{R}_i^2}{(\hat{\pi}^{(-k)}(1, X_i))^2} (Y_i - \mu_1(X_i))^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&+ \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \mathbb{E} \left[\left(e_1(X_i)^{-1} (\hat{\pi}^{(-k)}(1, X_i))^{-1} - \pi(1, X_i)^{-1} \right)^2 \frac{Z_i^2 \tilde{R}_i^2}{(\hat{\pi}^{(-k)}(1, X_i))^2} (Y_i - \mu_1(X_i))^2 \mid \mathcal{I}_{(-k)}, \{X_i\} \right] \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \frac{e_1^2(X_i) \pi^2(z, X_i)}{(\hat{\pi}^{(-k)}(1, X_i))^2} \mathbb{E}[\sigma^2(X_i) \mid \mathcal{I}_{(-k)}, \{X_i\}] \hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1})^2 \\
&+ \frac{e_1^2(X_i) (\pi^{(-k)}(z, X_i))^2}{e_1(X_i)} \mathbb{E}[\sigma^2(X_i) \mid \mathcal{I}_{(-k)}, \{X_i\}] (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1})^2 \\
756 &\leq \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \frac{\nu_e^2 \lambda_\pi^2}{(\hat{\pi}^{(-k)}(1, X_i))^2} B_{\sigma^2} (\hat{e}_1^{(-k)}(X_i)^{-1} - e_1(X_i)^{-1})^2 + \frac{\nu_e^2 \lambda_\pi^2}{\nu_e^2} B_{\sigma^2} (\hat{\pi}^{(-k)}(1, X_i)^{-1} - \pi(1, X_i)^{-1})^2 \\
&= o_p\left(\frac{1}{n^{1+2r_e+2r_\pi}}\right)
\end{aligned}$$

757 where the last inequality is because $\sigma^2(X)$ is bounded above, $\sigma^2(X) \leq B_{\sigma^2}$, by Lemma 4. Thus, by
758 similar argument to the first term, since this term is a sum of zero-mean random variables and since

759 $r_\pi, r_e \geq 0$, we can apply Chebyshev's inequality and get that this term is also $o_p(1/\sqrt{n})$. This holds
 760 for both treatments. Therefore,

$$\sqrt{n_{t,k}}(\hat{\tau}_{AIPW}^{(t,k)} - \hat{\tau}_{AIPW}^{*,(t,k)}) \rightarrow_p 0.$$

761 Putting these results from Step 1 and Step 2 together, along with the fact that $\frac{n_{t,k}}{n} \rightarrow \frac{1}{K}$, gives the
 762 theorem. $\square$

763 H Additional Lemmas

764 H.1 Results appearing in other works, stated for completeness.

765 **Lemma 1** (Conditional convergence implies unconditional convergence, from [10]). *Lemma 6.1.*
 766 *(Conditional Convergence implies unconditional) Let $\{X_m\}$ and $\{Y_m\}$ be sequences of random*
 767 *vectors. (a) If, for $\epsilon_m \rightarrow 0$, $\Pr(\|X_m\| > \epsilon_m \mid Y_m) \rightarrow_{Pr} 0$, then $\Pr(\|X_m\| > \epsilon_m) \rightarrow 0$. In*
 768 *particular, this occurs if $E[\|X_m\|^q / \epsilon_m^q \mid Y_m] \rightarrow_{Pr} 0$ for some $q \geq 1$, by Markov's inequality. (b)*
 769 *Let $\{A_m\}$ be a sequence of positive constants. If $\|X_m\| = O_P(A_m)$ conditional on Y_m , namely,*
 770 *that for any $\ell_m \rightarrow \infty$, $\Pr(\|X_m\| > \ell_m A_m \mid Y_m) \rightarrow_{Pr} 0$, then $\|X_m\| = O_P(A_m)$ unconditionally,*
 771 *namely, that for any $\ell_m \rightarrow \infty$, $\Pr(\|X_m\| > \ell_m A_m) \rightarrow 0$.*

772 **Lemma 2** (Chebyshev's inequality). *Let X be a random variable with mean μ and variance σ^2 .*
 773 *Then, for any $t > 0$, we have*

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

774 **Lemma 3** (Theorem 8.3.23 (Empirical processes via VC dimension), [48]). *Let $\mathcal{F}$ be a class of*
 775 *Boolean functions on a probability space (Ω, Σ, μ) with finite VC dimension $\text{vc}(\mathcal{F}) \geq 1$. Let*
 776 *$X, X_1, X_2, \dots, X_n$ be independent random points in Ω distributed according to the law μ . Then*

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right| \leq C \sqrt{\frac{\text{vc}(\mathcal{F})}{n}}$$

777 H.2 Lemmas

778 **Lemma 4** (Convergence of $\hat{\pi}$). *Assume that with high probability, for some large constant K ,*
 779 *$\|\hat{e}(X) - e(X)\|_2 \leq K n^{-r_e}$, $\|\hat{\sigma}^2(X) - \sigma^2(X)\|_2 \leq K n^{-r_\sigma}$. Assume Assumption 8. Assume that*
 780 *$\sigma^2(X) > 0$ so that its inverse is bounded $1/\sigma^2(X) \leq \gamma_\sigma$. Recall that Theorem 1 gives that*

$$\pi^*(z, X) = \sqrt{\frac{\sigma_z^2(X)}{e_z^2(X)}} B \left(\mathbb{E} \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right] \right)^{-1}$$

Define $\hat{\pi}^*(z, x)$ to be a plug-in version of the above (with $\hat{\sigma}^2, \hat{e}$, and $\mathbb{E}_n[\cdot]$). Then

$$\|\hat{\pi}^*(z, X) - \pi^*(z, X)\|_2 = o_p(n^{-\min(r_e, r_\sigma, 1/2)}).$$

781 *Proof.* Let $a = \frac{\sigma_z^2(X)}{e_z^2(X)}$, $b = \mathbb{E} \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right]$.

782 Let $c = \frac{\hat{\sigma}_z^2(X)}{\hat{e}_z^2(X)}$, $d = \mathbb{E}_n \left[\mathbb{I}[Z=1] \sqrt{\frac{\hat{\sigma}_1^2(X)}{\hat{e}_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\hat{\sigma}_0^2(X)}{\hat{e}_0^2(X)}} \right]$.

783 Then $\|\pi^*(z, X) - \hat{\pi}^*(z, X)\|_2 = \|a/b - c/d\|_2$.

784 Positivity of $\sigma_z^2(X)$ gives the elementary equality that $\frac{a}{b} - \frac{c}{d} = \left(\frac{a-b}{b}\right) + \left(\frac{d-c}{d}\right)$.

785 Therefore, by triangle inequality and boundedness,

$$\begin{aligned} \|\pi^*(z, X) - \hat{\pi}^*(z, X)\|_2 &\leq \gamma_\sigma \left\| \sqrt{\sigma^2(X)/e^2(X)} - \sqrt{\hat{\sigma}^2(X)/\hat{e}^2(X)} \right\|_2 \\ &+ \gamma_\sigma \left| \mathbb{E}_n \left[\mathbb{I}[Z=1] \sqrt{\frac{\hat{\sigma}_1^2(X)}{\hat{e}_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\hat{\sigma}_0^2(X)}{\hat{e}_0^2(X)}} \right] - \mathbb{E} \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right] \right| \end{aligned} \quad (2)$$

786 Next we show that for $z \in \{0, 1\}$,

$$\left\| \sqrt{\hat{\sigma}_z^2(X)/\hat{e}_z^2(X)} - \sqrt{\sigma_z^2(X)/e_z^2(X)} \right\|_2 \leq \nu_e B_{\sigma^2} \left(\left\| \sqrt{\hat{\sigma}_z^2(X)} - \sqrt{\sigma_z^2(X)} \right\|_2 + \|e_z(X) - \hat{e}_z(X)\|_2 \right) \quad (3)$$

787 In the below, we drop the z argument.

788 By the triangle inequality, boundedness of $1/\hat{e}(X) \leq \nu_e$, and of $\sigma^2(X) \leq B_{\sigma^2}$:

$$\begin{aligned} &\left\| \sqrt{\hat{\sigma}^2(X)/\hat{e}^2(X)} - \sqrt{\sigma^2(X)/e^2(X)} \right\|_2 \\ &= \left\| \sqrt{\hat{\sigma}^2(X)/\hat{e}^2(X)} \pm \sqrt{\sigma^2(X)/\hat{e}^2(X)} - \sqrt{\sigma^2(X)/e^2(X)} \right\|_2 \\ &\leq \nu_e \left\| \sqrt{\hat{\sigma}^2(X)} - \sqrt{\sigma^2(X)} \right\|_2 + B_{\sigma^2} \left\| \frac{1}{e(X)} - \frac{1}{\hat{e}(X)} \right\|_2 \end{aligned}$$

789 For the second term:

$$B_{\sigma^2} \left\| \frac{1}{e(X)} - \frac{1}{\hat{e}(X)} \right\|_2 \leq B_{\sigma^2} \left\| \frac{1}{e(X)} - \frac{1}{\hat{e}(X)} \right\|_2 \leq B_{\sigma^2} \nu_e \|e(X) - \hat{e}(X)\|_2$$

790 since $1/e(X)$ is Lipschitz on the assumed bounded domain (overlap assumption).

For the first term:

$$\nu \left\| \sqrt{\hat{\sigma}^2(X)} - \sqrt{\sigma^2(X)} \right\|_2 \leq \nu_e B_{\sigma^2} \|\hat{\sigma}^2(X) - \sigma^2(X)\|_2$$

791 since $\sigma^2(X)$ is bounded away from 0, then $\sqrt{\sigma^2(X)}$ is Lipschitz.

This proves Equation (3), which bounds the first term of Equation (2). For the second term, denote for brevity

$$\hat{\beta}(\sigma, e) = \mathbb{E}_n \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right],$$

792 and $\beta(\sigma, e)$ to be the above with $\mathbb{E}[\cdot]$ instead of $\mathbb{E}_n[\cdot]$. Then the second term of Equation (2) is

793 $\hat{\beta}(\hat{\sigma}, \hat{e}) - \beta(\sigma, e)$, and decomposing further, that

$$\hat{\beta}(\hat{\sigma}, \hat{e}) - \beta(\sigma, e) = \hat{\beta}(\hat{\sigma}, \hat{e}) - \hat{\beta}(\sigma, e) + \hat{\beta}(\sigma, e) - \beta(\sigma, e).$$

Note that by Cauchy-Schwarz inequality, and Lemma 3 (chaining with VC-dimension),

$$\hat{\beta}(\hat{\sigma}, \hat{e}) - \hat{\beta}(\sigma, e) \leq 2\nu_e B_{\sigma^2} \left(\left\| \sqrt{\hat{\sigma}_z^2(X)} - \sqrt{\sigma_z^2(X)} \right\|_2 + \|e_z(X) - \hat{e}_z(X)\|_2 \right) + 2C \sqrt{\frac{\text{vc}(\mathcal{F} \sqrt{\frac{\sigma^2}{e}})}{n}}$$

794 And another application of Lemma 3 gives that

$$\hat{\beta}(\sigma, e) - \beta(\sigma, e) = (\mathbb{E}_n - \mathbb{E}) \left[\mathbb{I}[Z=1] \sqrt{\frac{\sigma_1^2(X)}{e_1^2(X)}} + \mathbb{I}[Z=0] \sqrt{\frac{\sigma_0^2(X)}{e_0^2(X)}} \right] \leq 2C \sqrt{\frac{\text{vc}(\mathcal{F} \sqrt{\frac{\sigma^2}{e}})}{n}}.$$

795 Combining the above bounds with Equation (2), we conclude that $\|\pi^*(z, X) - \hat{\pi}^*(z, X)\|_2 =$
796 $O_p(n^{-\min(r_e, r_\sigma, 1/2)})$. $\square$

I Additional Experiment, Details and Discussion

I.1 Additional details

All experiments using our full algorithm 2 were conducted on a 2021 13-inch MacBook Pro equipped with a 2.3 GHz Quad-Core Intel Core i7 processor and 32 GB of memory. This setup was used to train standard nuisance models using machine learning, evaluated our algorithm, and conduct the analysis tasks reported in this paper. The average compute time for the experiments on real world data with 20 trials was less than 30 minutes, while the simulated data with 100 trials took less than 60 minutes. Additionally, for all experiments, we allocate 55% of the data to batch 1 and 45% to batch 2.

We run the ML nuisance models, logistic regression, random forest and support vectors machines, using popular Python packages (i.e. sklearn and scipy). We use logistic regression to estimate the propensity scores. For the outcome and variance models, we use random forest with the following hyperparameters:

- max_depth: None
- min_samples_leaf: 4
- min_samples_split: 10
- n_estimators: 100
- random_state: 42

We also use SVM model for the outcome models incorporating LLM predictions, and we use the following hyperparameters:

- kernel: 'rbf'
- C: 1

We chose these hyperparameters by doing a grid search over hyperparameters and chose the ones that performed the best.

We run LLM calls on Together.AI since they provide enterprise-secure deployments of local models, which is required for sensitive data. Because we need to use local LLMs for the real-world street outreach data, we also use the same local LLMs for the other experiments. We use "Llama-3.3-70B-Instruct-Turbo" for all experiments using LLMs. (Larger models provide effectively similar performance).

To solve our optimization problem, we used the python package CVXPY and we specifically used the Splitting Conic Solver (SCS) solver.

Once the experiments are run, we display the means and 95% confidence interval bands, obtained through bootstrapping, in each of our figures.

I.2 Synthetic Data

Before running our batch adaptive algorithm, we split the data into a validation set (35% of data) in which we estimate the ATE on. Then we use the remainder of the data to run our algorithm, which splits that data into the two batches in the way we described previously.

Data Generating Process. We generate a dataset $\mathcal{D} = \{X, Z, Y, Y(1), Y(0)\}$, of size 1000 and where the true ATE $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = 3$. We sample each covariate $X \in \mathbb{R}^5$ from a standard normal distribution, $X \sim \mathcal{N}(0, I_5)$. Treatment Z is drawn with logistic probability $\gamma_z(X) = (1 + e^{X_2 + X_3 + 0.5})$. We define $\sigma_z^2(X)$ as follows:

$$\begin{aligned}\sigma_1^2(X) &:= \max[1.3 + 0.4\sin(X_1), 0] \\ \sigma_0^2(X) &:= \max[3.5 + 0.3\cos(X_3), 0].\end{aligned}$$

Finally, the outcome models are defined as:

$$\begin{aligned}Y(0) &= 5 + X_1 - 2X_2 + \epsilon_0 \\ Y(1) &= Y(0) + \theta_0 + \epsilon_1,\end{aligned}$$

838 where $\epsilon_0 \sim \mathcal{N}(0, \sigma_0(X))$ and $\epsilon_1 \sim \mathcal{N}(0, \sigma_1(X))$. The observed outcomes are $Y = Z \cdot Y(1) + (1 -$
839 $Z) \cdot Y(0)$.

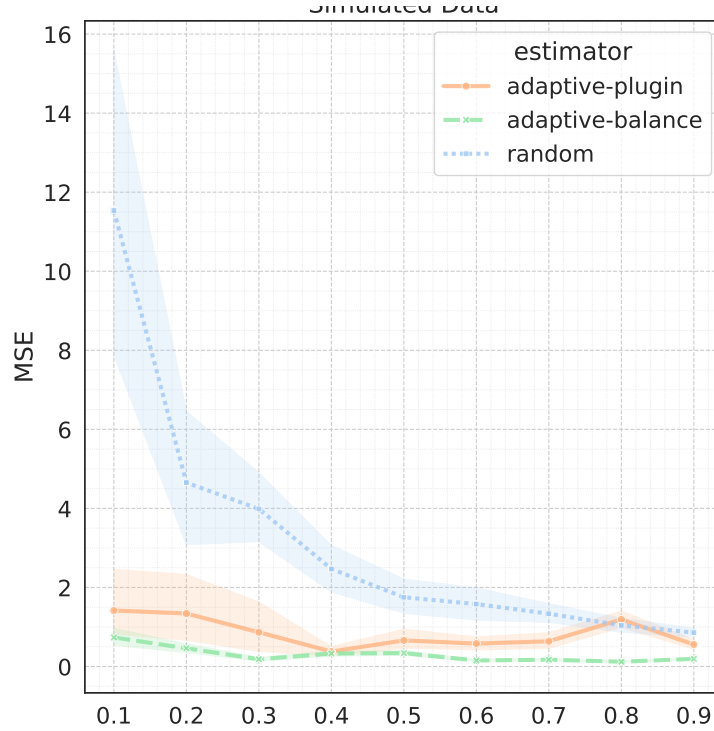


Figure 4: Mean squared error between estimated ATE and true ATE averaged over 100 trials across varying budgets.

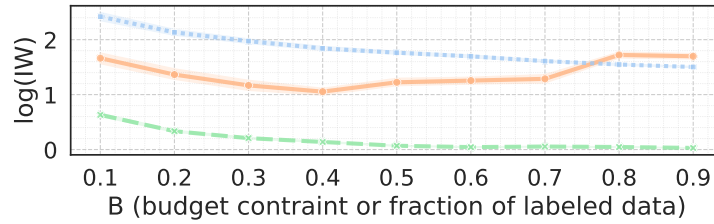


Figure 5: Average confidence interval width averaged over 100 trials across varying budgets.

840 **Results.** We see the greatest advantage with our adaptive estimation for budgets between 0.1 and
841 0.4. While for larger budgets, even as the MSE for both estimators converge, the interval width for
842 the adaptive estimator is still relatively small. Adaptive annotation with a larger budget introduces
843 additional variation in inverse annotation probabilities, as compared to uniform sampling, which is
844 equivalent to full-information estimation at a marginally smaller budget. This regime of improvement
845 for small budgets is nonetheless practically relevant and consistent with other works.

846 To stabilize the estimation of the inverse annotation probabilities, we use the plug-in estimator
847 following eq. (RZ-plug-in.) and the ForestReisz method to estimate the balancing weights [11]. This
848 approach provides an automatic machine learning debiasing procedure to learn the Reisz representer,
849 or unique weights that automatically balances functions between treated and control groups using a
850 random forest model.

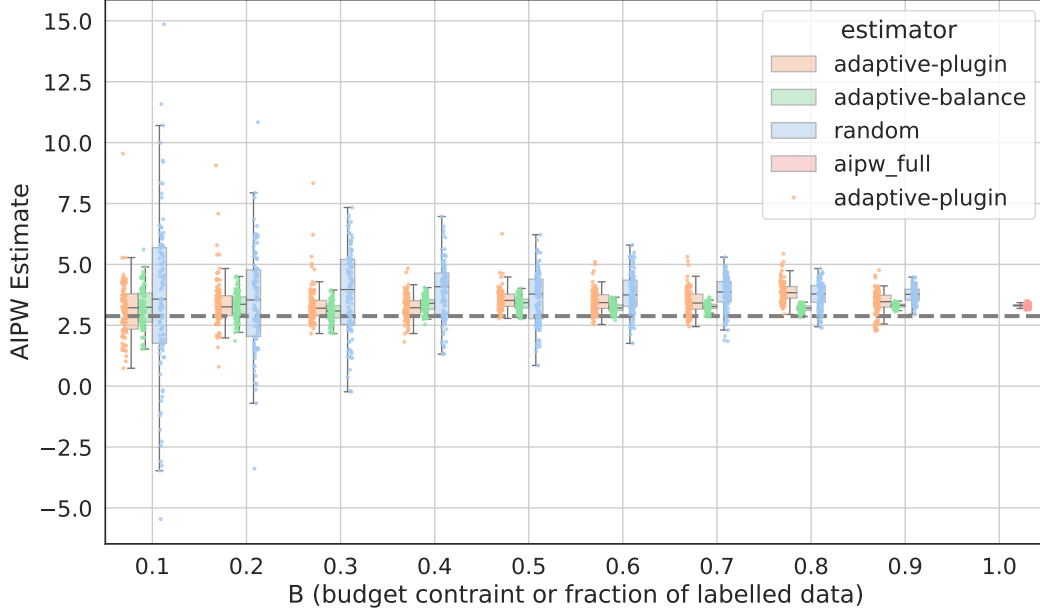


Figure 6: Boxplots of ATE estimates compared to skyline $\hat{\tau}_{AIPW}$ when the labeling budget is the entire dataset in red and the grey dotted line is τ .

I.3 Real-world Dataset Details

We provide further details about the treatment, covariates and outcomes for each dataset. Table 1 and table 2 describe the variables in the retail hero and outreach datasets, respectively. We refer the reader to [17] for further details about the dataset. For the outreach data, we constructed the binary treatment variable by binning the frequency of outreach engagements for each client within the first 6 months of the treatment period. We checked for overlap in propensity scores and decided to use treatments in the middle of the distribution as they had the most overlap. Additionally, by corollary 1, our method does well even when the propensity scores do not have good overlap.

Variable	Description	Discrete Category
Outcome		
Purchase	whether a customer purchased a product	[Yes,No]
Treatment		
SMS communication	whether a text was sent to encourage customer to continue shopping	[Yes, No]
Covariates		
avg. purchase	avg. purchase value per transaction	[1-263, 264-396, 397-611, > 612]
avg. product quantity	avg. number of products bought	[≤ 7 , > 7]
avg. points received	avg. number of points received	[≤ 5 , > 5]
num transactions	total number of transactions so far	[≤ 8 , 9 - 15, 16 - 27, > 28]
age	age of user	[≤ 45 , > 45]

Table 1: Covariate, treatment, and outcome descriptions and discrete category definitions for Retail-Hero dataset.

I.4 Additional Context on Street Outreach

In New York City alone, approximately \$80,000,000 per year is invested in homeless street outreach to an unclear effect. It is a time-consuming process, and it is unclear how the impacts of such intensive

Variable	Description	Discrete Category
Outcome		
Placement	The greatest housing placement attained by the client between 2019–2021	[3:permanent housing, 2: shelter/transitional housing, 1: other (e.g., hospital), 0: streets]
Treatment		
Street outreach	Binned frequency of outreach within the first three months of 2019	[More outreach (3–15), Less outreach (1–2)]
Covariates		
DateFirstSeen	Ordinal date when the client was first seen by the outreach team	NA
Program	Outreach or service program the client belonged to	[Brooklyn Library, Grand Central Partnership, Hospital to Home, K-Mart Alley, Macy's, MetLife, Penn Post Office, Pyramid Park, S2H Bronx, S2H Brooklyn, S2H Manhattan, S2H Queens, Starbucks, Superblock, Vornado, Williamsburg Stabilization Bed]
BelievedChronic	Perceived by outreach workers as chronically homeless individual	[Yes, No]
Gender	Perceived or disclosed gender of client	[Female, Male, Transgender]
Race	Perceived or disclosed race of client	[American Indian/Alaskan Native, Asian, Black/African American, Native Hawaiian/Pacific Islander, White/Caucasian]
Ethnicity	Perceived or disclosed ethnicity of client	[Hispanic/Latino, Non-hispanic/latino]
Age	Perceived or disclosed age range of client	[< 30 years old, 30–50 years old, > 50 years old]
Was311Call	Whether outreach workers were responding to a 311 city call	[Yes, No]
Was911Call	Whether 911 was called to the scene	[Yes, No]
Removal958	Whether outreach workers were responding to removal hotline call	[Yes, No]
Housing application	Whether any mention of the housing application was found in casenotes	[Yes, No]
Service refusal	Whether outreach worker documented that a client refused their services in casenotes	[Yes, No]
Important documents	Whether there was mention of any important documents (i.e. social security card, drivers license, etc.) in casenotes	[Yes, No]
Benefits	Whether there was any mention of social service benefits in the casenotes (i.e. foodstamps, SSI)	[Yes, No]
num contacts	number of engagements with an outreach worker prior to 2019	NA
max Placement	maximum housing placement reached before 2019	[3:permanent housing, 2: shelter/transitional housing, 1: other (e.g., hospital), 0: streets]

Table 2: Covariates, treatment, and outcome descriptions and discrete category definitions for the Street Outreach dataset.

862 individualized outreach might compare to other proposed approaches, such as those focusing on
863 placing entire networks of individuals together. While the nonprofit reports key metrics such as
864 number of completed placements in housing services, these can be somewhat rare due to length

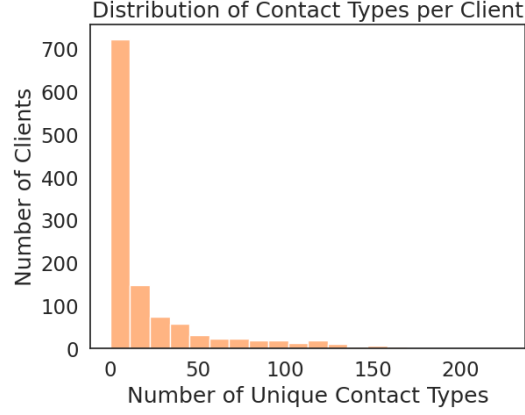


Figure 7: Distribution of street outreach engagements for client population.

865 of outreach, delays in waiting for housing, matching issues, etc; moreover, much of a successful
 866 placement is out of the control of outreach due to highly limited housing capacities. Measuring
 867 the impacts of street outreach on intermediate outcomes such as accessing benefits and services,
 868 completing required appointments and interviews, can better reflect the immediate impacts of street
 869 outreach.

870 I.5 Robustness Check on Street Outreach Data

871 To further demonstrate the utility of our approach, we run experiments on the Street Outreach data
 872 with $\tilde{Y}$. To recap, our setup consists of covariates X , which includes client characteristics at baseline
 873 and LLM-generated summaries of case notes recorded before the treatment period. In the main text,
 874 we used LLMs to summarize casenotes prior to outreach during the interventional period, and used
 875 them in zero-shot prediction of later placement outcomes. Here we also incorporate LLM-generated
 876 summaries of case notes recorded post-treatment. These represent $\tilde{Y}$ in our framework.

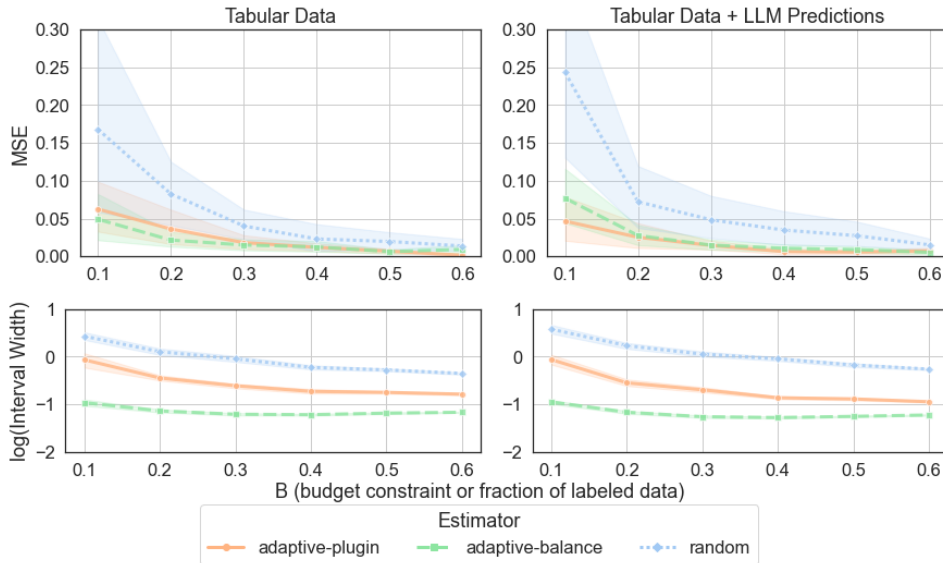


Figure 8: **Street outreach data with post-treatment summaries only.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

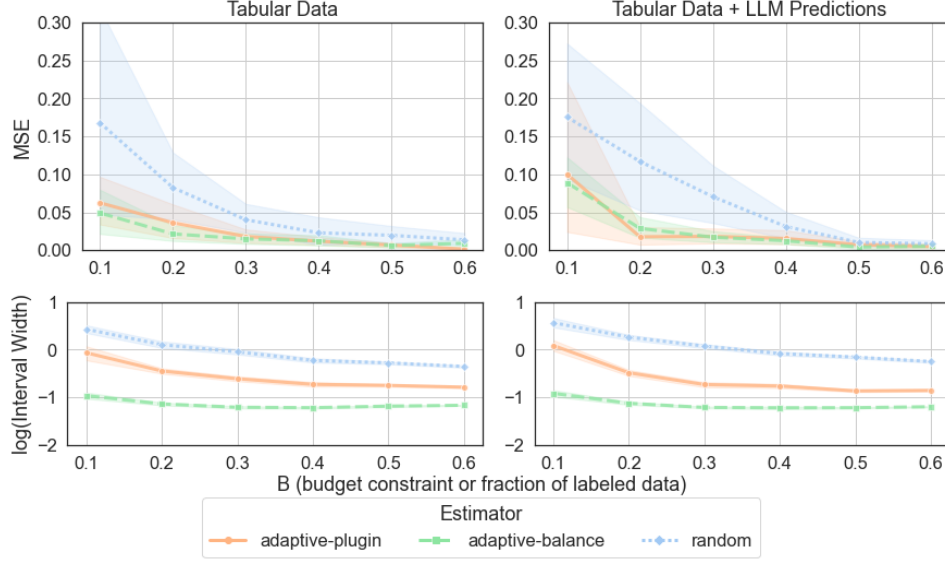


Figure 9: **Street outreach data with pre- and post-treatment summaries.** Mean squared error and 95% confidence interval width averaged over 20 trials across budget percentages of the data. This plot makes use of tabular data and the best-performing random forest outcome model (left) and text-encoded outcomes using LLMs (right).

In Figure 8 and Figure 9, we see that our results and analysis are preserved, and qualitatively similar. Our adaptive approach still shows improvements over uniform random sampling. The MSE is tripled when going from our adaptive estimators to random sampling in the tabular data. The MSE is five times higher when going from adaptive to random sampling in the setting where we have added LLM predictions using post-treatment summaries $\tilde{Y}$ only and it is nearly doubled when using both pre- and post-treatment summaries.

In this experimental setup, we find that tabular estimation with ground-truth validated codes overall performs comparably as using more advanced LLM estimation. In this setup, we use placement outcomes as the measure of interest, in part because it is (nearly) fully recorded in our dataset, and hence we can consider it as having access to the “ground-truth” outcome in our methodological setup. On the other hand, we also expect that casenotes are weakly informative of placement, as compared with other outcomes we might seek to extract from casenotes (but do not have the ground-truth for). Nonetheless, this validates the usefulness of the method, and we leave further empirical developments for future work.

I.6 Budget Saved Plots

We compute the amount of budget saved due to our batch adaptive sampling approach. We find the sample size required to achieve the same confidence interval width with batch adaptive annotations using balancing weights (green) and RZ-plug-in (orange) compared to uniform random sampling.

I.7 Active Learning Baselines

Active learning is not a strong baseline and we argue this on theoretical and empirical fronts. Active learning for regression can’t improve statistical rates of convergence, while the doubly-robust AIPW estimator in causal inference can, so using AIPW is optimal. Additionally, using pool-based active learning algorithms in AIPW blows up variance due to near-deterministic annotation probabilities.

Active learning models only target μ_z , but the outcome model contributes $\frac{\sigma_z^2(x)}{e_z(x)\pi(z,x)}$ to the causal Avar, and our optimal annotation correctly balances the effect of all factors, but active learning only considers the first.

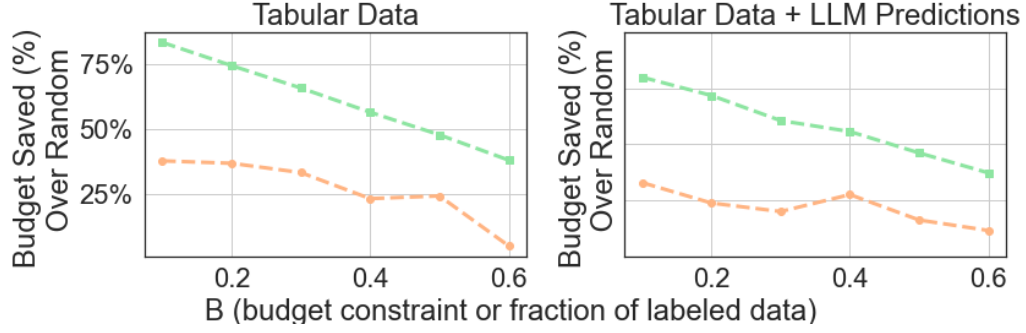


Figure 10: **RetailHero Data.** Budget saved due to batch adaptive annotation. The reduction in annotation sample size needed to achieve the same confidence interval width with batch adaptive annotation on tabular data (left) and on tabular data + complex embedded outcomes (right) compared to random sampling.

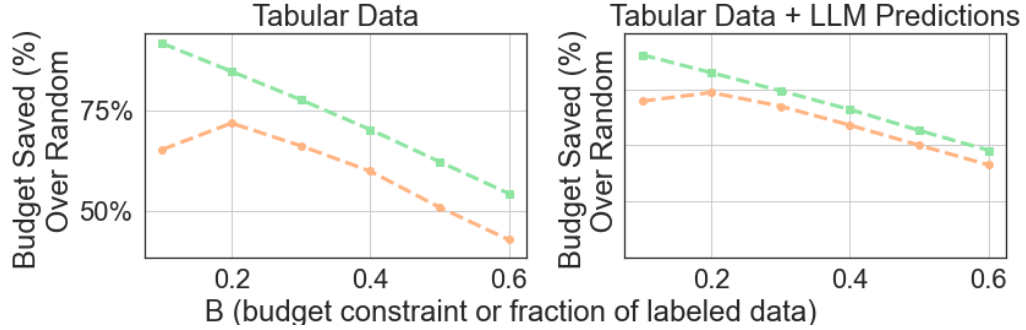


Figure 11: **Street Outreach Data.** Budget saved due to batch adaptive annotation. The reduction in annotation sample size needed to achieve the same confidence interval width with batch adaptive annotation on tabular data (left) and on tabular data + complex embedded outcomes (right) compared to random sampling.

903 In summary, active learning does something *completely different for prediction error, suboptimal for*
 904 *causal inference.*

905 Empirically, we run active learning algorithms to learn μ in AIPW and find that it *totally fails* for
 906 these reasons; if these objectives line up, it can do well, but in general, the prediction and causal error
 907 objectives are different.

908 **Theoretical comparison to active learning.** As a reminder, we optimize:

$$AVar_{ATE} = Var[CATE(X)] + \sum_{z \in \{0,1\}} E\left[\frac{\sigma_z^2(X)}{e_z(X)\pi(z, X)}\right]$$

909 (The first term is the variance of $CATE = E[Y(1) - Y(0)|X]$; it is never observed.)

910 To go more in detail on our experiments 1) we compare to theoretical results in batch *pool-based*
 911 *active learning*, Chaudhuri et al. [7] and Gentile et al. [22] (henceforth GWZ), which show that active
 912 learning doesn't improve convergence rates for regression, only multiplicative constants. Instead,
 913 the AIPW estimator is optimal for causal estimation: if the outcome and propensity scores can only
 914 achieve $n^{-1/4}$ convergence, the AIPW estimator is $O(n^{-1/2})$ -rate convergent, so AIPW can speed
 915 up outcome model convergence rates. Therefore using the AIPW estimator is best, and random
 916 sampling + AIPW is a stronger baseline than active learning.

917 To emphasize the different objectives, consider a simple example with two regions:

918 • Region 1 (Poor Overlap), $X > 0$: Propensity score $e(X) = 0.01$; outcome noise
919 $\sigma_1(X), \sigma_0(X)=1$.

920 • Region 2 (High Prediction Uncertainty), $X < 0$: Propensity score $e(X) = 0.5$; outcome
921 noise $\sigma_1(X), \sigma_0(X) = 10$ and the outcome model is complex.

922 Our method compares the ATE variance contribution in either region:

923 • Region 1: $\frac{\sqrt{1}}{0.01} = 100$

924 • Region 2: $\frac{\sqrt{100}}{0.5} = 10$

925 and samples in Region 1, where the causal variance is five times higher. Uncertainty-based active
926 learning samples in Region 2, to the detriment of causal variance.

927 **Active Learning Empirical Evaluations.** We evaluate our method against 2-3 active learning base-
928 lines for each experiment from two popular and well-established python packages (scikit-activeML
929 and modAL). Different active learning algorithms are appropriate for different outcome models, so
930 we choose the sampling strategy based on our modeling task, and we use pool-based active learning
931 matching our two-batch approach. (Note our approach is *model-agnostic*, while active learning
932 methods are not). For the classification tasks on our two real-world datasets (RetailHero/Street
933 Outreach), we use UncertaintySampling with margin sampling and least confident sampling as query
934 strategies, which both choose x with highest uncertainty measure based on classification probabilities
935 $P(\hat{Y} = 1 | x)$ [40]. For the regression tasks, we use Expected Model Variance Reduction [12],
936 Expected Model Change Maximization [6], and Improved Greedy Sampling [51]; these choose x
937 that maximizes greatest future variance reduction, maximally change the current model via the loss
938 gradient, and diversity in feature and output space, respectively.

939 We run each approach over 50 trials and take the average MSE. Across the board, we see that our
940 approach does better than the popular active learning strategies that are not optimized for causal
941 estimation.

942 Result Tables

Estimator	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
active-evar	0.313	17.3	85.1	579	1.31e+03	3.87e+03	1.27e+04	5.03e+04	8.93e+05
active-greedy	6.13	79.9	369	852	1.99e+03	5.06e+03	1.33e+04	5.09e+04	2.95e+05
active-mvar	10.6	94.3	314	883	2.17e+03	5.70e+03	1.21e+04	3.87e+04	2.99e+05
adaptive-balance	0.471	0.227	0.276	0.236	0.265	0.246	0.198	0.176	0.203
adaptive-plugin	1.7	1.17	0.831	0.196	0.83	0.449	0.507	0.93	0.481
random	8.99	4.56	2.19	1.54	1.7	1.61	1.46	0.956	0.987

Table 3: Averaged MSEs for Synthetic Data.

Estimator	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
active-margin	3.53e+03	0.047	0.087	12.5	8.38e+03	2.25e+06	1.49e+06	6.53e+05	1.43e+07
active-uncertain	16.1	38.9	70.4	75.9	115	112	168	250	402
adaptive-balance	0.004	0.002	0.002	0.001	0.001	0.001	0	0	0
adaptive-plugin	0.004	0.001	0.001	0.001	0.001	0	0	0	0
random	0.027	0.012	0.009	0.006	0.005	0.003	0.001	0.001	0

Table 4: Averaged MSEs for RetailHero Data.

Estimator	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
active-margin	0.009	28.5	4.47	0.501	0.449	0.044	0.099	0.412	0.209
active-uncertain	0.017	0.009	0.018	0.008	0.017	0.018	0.025	0.023	0.024
adaptive-balance	0.046	0.031	0.013	0.006	0.005	0.003	0.004	0.003	0.002
adaptive-plugin	0.045	0.025	0.027	0.012	0.006	0.004	0.004	0.006	0.001
random	0.113	0.061	0.037	0.045	0.014	0.012	0.011	0.003	0.001

Table 5: Averaged MSEs for Street Outreach Data.

Gentile et al. [22] chooses a point x maximizing a diversity measure, $D(x, S)$ that quantifies model uncertainty and is directly influenced by the observation noise, $\sigma_z^2(X)$. For general function approximation, they introduce a maximal disagreement measure over the regression function class $\mathcal{F}$ $\sup_{f, g \in \mathcal{F}} \frac{(f(x) - g(x))^2}{\sum_{z \in S} (f(z) - g(z))^2 + 1}$, where S is the set of already sampled points. If $\sigma^2(x)$ is large for some x , their disagreement measure is also large. Their diversity measure finds points where it is possible for two functions, f, g , to have similar predictions on the already-labeled data S (a small denominator) but different predictions for a new point x (a large numerator). When observation noise $\sigma^2(x)$ is larger, many different functions can be considered "plausible" fits and can agree on S but disagree elsewhere, leading to a high diversity score. In contrast, low noise tightly constrains all plausible functions, resulting in low disagreement.

I.8 LLM Prompts

Prompt 1 (Retail Hero):

You are a user who used a website for online purchases in the past one year and want to share your background and experience with the purchases on social media.

Attributes:

The following are attributes that you have, along with their descriptions.

$\{features\}$

Personality Traits The following dictionary describes your personality with levels (High or Low) of the Big Five personality traits.

$\{traits\}$

Your Instructions:

Write a social media post in first-person, accurately describing the information provided. Write this post in the tone and style of someone with the given personality traits, without simply listing them.

Only return the post that you can broadcast on social media and nothing more.

—
 $\{post\}$
—

Prompt 2 (Street Outreach Casenote Summaries) :

Objective: Your task is to summarize a trajectory of case notes of a client in street homelessness outreach, focusing on client interactions, the challenges they are facing, goals they are working towards, and progress towards housing placement. These are all from the same client. This summary is designed to help caseworkers and organizations assess client history at a glance, remind of prior personal information and important challenges mentioned (like veteran status or other information that is relevant for eligibility for housing, medical issues, and status of their support network), allocate resources effectively, and improve support for individuals experiencing chronic homelessness.

Context: *{task_context}*

The summary should be a concise overview of the client's situation, highlighting key points from the case notes. It should not include any personal opinions or assumptions about the client's future or potential outcomes. The goal is to provide a clear and informative summary that can be used by caseworkers and organizations to better understand the client's history and current status.

Here are the case notes for batch *{batch_num}* of *{total_batches}*:

— START NOTES —

{notes}

— END NOTES —

Based **only** on the notes provided above for this batch, generate a comprehensive summary focusing on key events, decisions, and progress during this specific period. The target length is approximately *{target_length}* words. Ensure the summary strictly reflects the content of these notes.

958

Prompt 2 (Street Outreach Classification) :

You are an expert analyst specializing in predicting long-term housing stability for individuals experiencing homelessness. Your task is to analyze client data, including demographic information, historical interactions, and case note summaries, to predict the **most stable housing placement level** the client is likely to achieve and maintain over the **next two years**.

Input Data:

You will be provided with the following information for each client:

Prediction Task:

Based **only** on the provided attributes and the case notes summary, predict the single most stable housing placement level the client is likely to maintain over the next two years.

Housing Placement Levels (Prediction Output):

Your prediction must be an integer between 0 and 3:

- **0**: No stable placement (remains on the street or in emergency shelters).
- **1**: Transitional Housing (temporary placement with support, aiming for longer-term housing).
- **2**: Rapid Re-housing (time-limited rental assistance and services).
- **3**: Permanent Supportive Housing (long-term housing with ongoing support services).

959

Reasoning Guidance (Internal Thought Process - Do Not Output This):

- Consider factors that promote stability: housing application progress, possession of documents, benefit acquisition, engagement with services (unless contacts are excessive without progress), prior successful placements (even if temporary), positive recent developments in the case notes.
- Consider factors that hinder stability: chronic homelessness indicators, frequent service refusals, mental health crises (Removal958), lack of documents/income, lack of prior placements, patterns of instability noted in the summary.
- Weigh the structured data against the nuances presented in the case note summary. The summary provides vital context.

Client Information:

Prediction:

Provide **only** the predicted number (0, 1, 2, or 3) as the output. Do not include any other text, explanation, or formatting.

Examples: {*examples*}