Ensembled Prediction Intervals for Causal Outcomes Under Hidden Confounding

Myrl G. Marmarelis MYRLM@ISI.EDU
Greg Ver Steeg GREGV@ISI.EDU
Aram Galstyan GALSTYAN@ISI.EDU
Fred Morstatter FREDMORS@ISI.EDU
USC Information Sciences Institute
4676 Admiralty Way
Marina del Rey, CA 90292

Editors: Francesco Locatello and Vanessa Didelez

Abstract

Causal inference of exact individual treatment outcomes in the presence of hidden confounders is rarely possible. Recent work has extended prediction intervals with finite-sample guarantees to partially identifiable causal outcomes, by means of a sensitivity model for hidden confounding. In deep learning, predictors can exploit their inductive biases for better generalization out of sample. We argue that the structure inherent to a deep ensemble should inform a tighter partial identification of the causal outcomes that they predict. We therefore introduce an approach termed **Caus-Modens**, for characterizing **caus**al outcome intervals by **mod**ulated **ens**embles. We present a simple approach to partial identification using existing causal sensitivity models and show empirically that Caus-Modens gives tighter outcome intervals, as measured by the necessary interval size to achieve sufficient coverage. The last of our three diverse benchmarks is a novel usage of GPT-4 for observational experiments with unknown but probeable ground truth.

Keywords: hidden confounding, sensitivity analysis, prediction intervals, deep ensembles

1. Introduction

In order for a regression model to make *causal* predictions, the effect of confounders must be disentangled from the effect of the treatment. For this reason, causal inference is closely related to the problem of domain shift, since the outcome predictor may be learned on observational data while being expected to perform well on the hypothetical domain with fully randomized treatments. More often than not, the available covariates are imperfect proxies for all the confounders in the causal system. This further compounds the task of causal inference, as the hidden confounders must somehow be taken into account. The best hope in these cases is to produce “ignorance intervals” that *partially* identify the causal estimands. The tighter the intervals, the more useful the partial identification, which depends on what can be said about the hidden confounders.

A sensitivity model (Rosenbaum and Rubin, 1983) in causal inference is a structural assumption (Manski, 2003) about the possible behavior of hidden confounders. It allows causal estimands to be partially identified as long as the extent of hidden confounding is consistent with the sensitivity model. The dependence of the treatment assignment on confounders, i.e. the conditional treatment *propensity*, is what makes a study observational rather than a fully randomized experiment. We consider sensitivity models that bound the *complete* propensity (colloquially, the true propensity of treatment assignments for an individual, taking into account all relevant variables, observed or not) in terms of the *nominal* propensity (based just on observed covariates, allowing it to be estimated by regression.) Sensitivity models of this kind were first introduced by Tan (2006) and have become popular due to their generality and simplicity. The most common setting for these models, in line with Tan's initial formulation, is of binary treatments (Jesson et al., 2021a; Kallus et al., 2019; Dorn et al., 2021), in which the Marginal Sensitivity Model (MSM) bounds the ratio of nominal-propensity odds to complete-propensity odds. When that ratio is unit, and the complete propensity equals the nominal propensity at all points, then the covariates are adequate to explain all the confounders. It is worthwhile to broaden the notion of the MSM in light of recent developments with MSM-like sensitivity models for continuous treatments (Jesson et al., 2022; Marmarelis et al., 2023) and other nonbinary domains. To accommodate these settings, we consider a general form of sensitivity model.

In this paper, we explore prediction intervals of causal outcomes due to interventions on the treatment variable, termed *outcome intervals*, that incorporate empirical uncertainties (Jesson et al., 2020) in addition to the orthogonal concept of hidden-confounding uncertainty. Outcome intervals predict *individual outcomes* of treatments disentangled from confounders, relying on a sensitivity model to guide partial identification in the presence of hidden confounders.

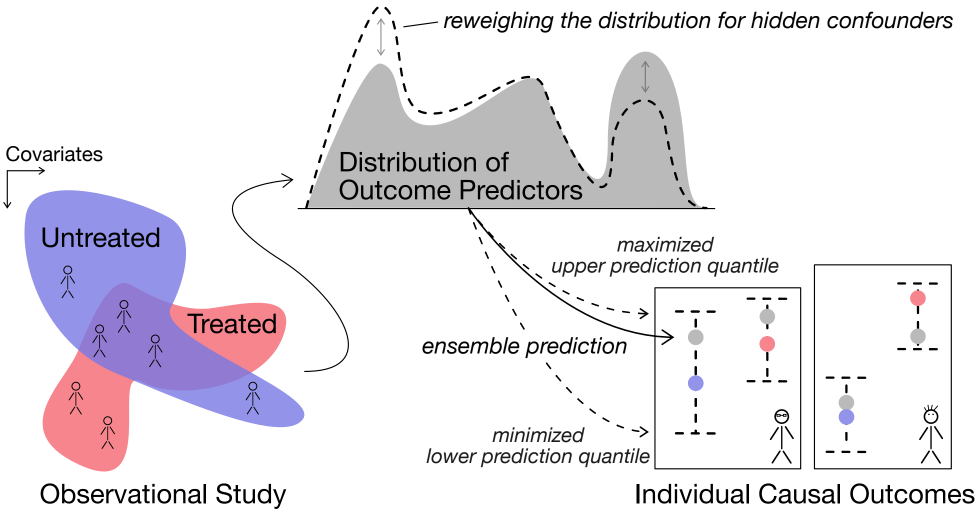

reweighing the distribution for hidden confounders
Covariates
Distribution of Outcome Predictors
Untreated
Treated
maximized upper prediction quantile
ensemble prediction
minimized lower prediction quantile
Observational Study
Individual Causal Outcomes

Figure 1. An illustration of the proposed method for causal outcome intervals. First, one samples predictors from a Bayesian posterior or otherwise learns an ensemble to approximate the distribution of outcome predictors that agree with the observational data. The ensemble average (grey dot) could be used to predict actual causal outcomes (red/blue dots). With hidden confounding, the learned ensemble might diverge substantially from the best predictor distribution to model causal outcomes. One cannot identify the correct distribution from observational data alone. Instead, a sensitivity model says how wrong this learned ensemble could be, and one optimizes with respect to weights on the ensemble elements for each individual and treatment in order to upper-bound the $(1-\alpha/2)$ quantile and lower-bound the $(\alpha/2)$ quantile of the ensemble prediction. These intervals incorporate both empirical uncertainties from prediction quantiles and hidden-confounding uncertainties from the ensemble modulation. They are evaluated against ground-truth causal outcomes by removing confounding through interventions on test-set individuals, using (semi-)synthetic data.

1.1. Related Work

Our goal diverges from the great strides that have been made in the realm of multiply debiased and robust estimators for *average* outcomes of populations or subpopulations (Athey et al., 2019; Chernozhukov et al., 2018). Largely in the binary-treatment context, these estimators have been augmented with *sharp* partial identification methods (Dorn and Guo, 2022) that are guaranteed to be valid while not overly conservative. Dorn et al. (2021) accomplishes this partial identification at the cost of having to re-estimate outcome regressions every time the sensitivity model changes. Partial identification is less explored for nonbinary treatments, which are receiving increased attention (Nie et al., 2021; Kaddour et al., 2021; Bica et al., 2020; Wang et al., 2022). Separately, outcome statistics that are more complex than expectations are also of key interest in machine learning (Kallus and Oprescu, 2023), with diverse purposes like fairness-oriented measures (Kallus, 2022). Tight partial identification of these other statistics requires novel methodology. Partially identified *outcome quantiles* would be a step in that direction. We solve that problem in this paper for the purpose bounding above and below the individual outcome intervals—our current focus.

The state of the art for partially identified outcome intervals from binary treatments is conformalized (Yin et al., 2022; Jin et al., 2023), building on domain shift (Lei and Candès, 2021). Conformal inference looks at the empirical performance of a model to decide how to size its prediction set (interval). The simplicity of this approach coupled with its finite-sample statistical guarantees makes it widely applicable. However, conformalized intervals even in the causal setting are based on the behavior of the outcome predictor *on the observed distribution.* We hypothesize that heavily confounded observational data might make it difficult for causal conformal prediction to extrapolate to the unconfounded domain. In this paper, we offer an alternative approach that can benefit from a predictor ensemble's inductive biases when constructing the causal outcome intervals.

1.2. Motivation

Instead of partial identification of (conditional) average treatment effect, (C)ATE, the conformal sensitivity analysis (CSA) (Yin et al., 2022; Jin et al., 2023) produces rigorous intervals for the individual treatment effect (ITE), in other words the outcome realization rather than expectation. CSA considers a predictor's performance in a calibration set as a guide for determining prediction intervals out of sample. Partial identification tends to be formulated adversarially, in terms of minimizing/maximizing a causal estimand that is admitted by the problem's constraints. CSA involves an optimization problem over the rebalancing weights applied to the calibration sample (Tibshirani et al., 2019; Lei and Candès, 2021). As the conformal method requires quantile estimates, it is impacted by theoretical implications on weighted quantile estimators. The variance of the estimator scales with the variance of the weights (Glynn et al., 1996, Theorem 1). If the weights were not inverse-propensity adjusted, then the conformal guarantees would fail due to distribution mismatch, so a large variance from covariate shift, for instance, cannot be avoided.

To avoid the statistical challenges associated with reweighing an observational sample, we posit that an ensemble capturing empirical uncertainties from the observational data could harness its inductive biases to extrapolate to causal outcomes (Jesson et al., 2020; Rame et al., 2022). These elements exist in Bayesian reasoning (Jaynes, 2003), which is a sound and scientific way to reconcile models with data. It incorporates parametric, distributional, structural, and prior knowledge into a *posterior* distribution of learned models that agree with the data. Even with large, deep models that are commonly developed in machine learning, the structural elements of the model contribute to its performance in a general domain (e.g. Edelman et al., 2022). In the method presented below, we allow an ensemble's learned biases to aid in extrapolation of the partially identifiable causal estimands. The mathematical connections are clear when the ensemble is supposedly from a Bayesian posterior, but in practice it can be learned in any way that sufficiently captures empirical uncertainties (Wild et al., 2023). See Figure 1 on the ensemble reweighing.

2. Approach

We present a versatile, modular procedure for taking an ensemble of outcome predictors and, in coordination with some causal sensitivity model, producing tight causal outcome intervals. We term this approach for **caus**al outcome intervals via **mod**ulated **ens**embles *"Caus-Modens."* The idea is to min/max an ensemble's conditional quantiles by reweighing the predictors, yielding individual causal outcome intervals. First we list the fundamental assumptions for our causal inference.

Assumption 1 (Potential Outcomes) *We adopt Rubin (1974)'s first two assumptions for potential outcomes. First, observation tuples of (outcome, assigned treatment, covariates) denoted as* $\{(y^{(i)}, t^{(i)}, x^{(i)})\}_{i=1}^{n}$, *are* i.i.d *from a single joint distribution. This subsumes the Stable Unit Treatment Value Assumption (SUTVA), where units/individuals cannot depend on one another. Secondly, all treatment values have a nonzero chance of assignment for every individual in the data.*

For a family of outcome predictor models $\mathcal{M}$, we use $p_{\mathcal{M}}$ to denote probability density functions constrained by one or more models in $\mathcal{M}$—that is, $\mathcal{M}$ conveys the hard constraints implied by the choice of parametrization θ and any other structural assumption. These models predict an outcome Y due to treatment assignment T and covariate X. In Bayesian notation the posterior $P(\Theta \mid \mathcal{D})$, given a dataset $\mathcal{D}$, induces a *posterior predictive* outcome distribution, which is described by a conditional expectation that averages the individual model predictions $p_{\mathcal{M}}(y \mid t, x; \theta)$:

$$p_{\mathcal{M}}(y \mid t, x; \mathcal{D}) = \mathbb{E}_{\Theta}\big[p_{\mathcal{M}}(y \mid t, x; \Theta) \mid \mathcal{D}\big]. \tag{1}$$

In practice, this integration over viable parameters is simulated by Monte Carlo with an ensemble of learned models. For our purposes, $\{\theta^{(j)}\}_{j=1}^{m}$ is assumed to be i.i.d from an estimator distribution (in the frequentist case) or posterior (in the Bayesian case, Li et al., 2023) with a density denoted as $p(\theta|\mathcal{D})$ in either case. Our sensitivity analysis requires the estimation of a nominal propensity function as well, denoted by $e_t(x)$, which can be a discrete probability or a continuous density.

The third potential-outcomes assumption is ignorability: absence of hidden confounders, where $\{(Y_t)_{t\in\mathcal{T}} \perp\!\!\!\perp T\} \mid X$. It states that while the outcome would depend on the assigned treatment, a *potential outcome* for any treatment should not be affected by the treatment assignment, after conditioning on covariates. Our setting allows a bounded violation to the ignorability assumption.

Definition 1 (Hidden Confounding) $\{(Y_t)_{t\in\mathcal{T}} \not\perp\!\!\!\perp T\} \mid X$, *hence* $P(Y_t \mid T, X)$ *may differ from* $P(Y_t \mid X)$ *for the potential outcomes* Y_t *outside the assigned treatment* ($T \neq t$)*, and similarly the complete propensity* $P(T \mid X, Y_t)$ *is not the nominal propensity* $P(T \mid X)$ *for any* Y_t.

Whichever sensitivity model is invoked to bound the extent of hidden confounding, all that is required for Caus-Modens is a pair of weight-bounding functions $\underline{\omega}(t, x), \overline{\omega}(t, x)$ that are partial identifiers of the potential-outcome probability density function, $p(y_t|x)$. We introduce one layer of indirection by referring to potential *outcome models* θ_t, heterogeneous in treatment t and covariate x (conditioning on the latter,) that can only be partially identified by means of the learned outcome model θ. The real potential outcomes are therefore (partially) identified by marginalization over the potential models: $p(y_t|x) = \int p(y|t,x;\theta_t)\, p(\theta_t|x;\mathcal{D})\, \mathrm{d}\theta_t$, assuming integrability. The role of the weights is in the relation $p(\theta_t|x;\mathcal{D}) = \omega(\theta,t,x)p(\theta|t,x;\mathcal{D})$, where $p(\theta|t,x;\mathcal{D}) = p(\theta|\mathcal{D})$ because the learned model is invariant. As mentioned, the weights can only be partially identified by the given sensitivity model. The reason for pushing our causal sensitivity analysis to the level of the outcome *model* is that it can be empirically favorable while remaining largely intuitive.

Assumption 2 (Sensitivity Model as Weights) *The sensitivity model under consideration uses the propensity $e_t(x)$ to produce bounds $0 < \underline{\omega}(t,x) \leq 1 \leq \overline{\omega}(t,x) < +\infty$ on weights for partial identification of the outcome model, in the sense that there exists some $\theta \mapsto \omega(\theta,t,x) \in [\underline{\omega}(t,x), \overline{\omega}(t,x)]$ that recovers the true potential outcome model density function, $p(\theta_t|x;\mathcal{D}) = \omega(\theta,t,x)\, p(\theta|\mathcal{D})$.*

This formulation readily accommodates a variety of recently proposed sensitivity models once we pose the *complete propensity* in terms of potential outcome models rather than direct potential outcomes. Instead of $P(T \mid Y_t, X)$, we consider $P(T \mid \Theta_t, X; \mathcal{D})$.

Example 1 *For binary treatments, an MSM with violation-of-ignorability $\Gamma > 1$ bounds the ratio of complete-propensity odds and nominal-propensity odds to $[1/\Gamma, \Gamma]$, implying (Kallus et al., 2019)*

$$\underline{\omega}(t,x) = e_t(x) + \Gamma[1 - e_t(x)], \quad \overline{\omega}(t,x) = e_t(x) + (1/\Gamma)[1 - e_t(x)], \tag{2}$$

where $e_t(x)$ is the nominal propensity of binary treatment $t \in \{0,1\}$. Suppose the complete propensity is denoted as $e_t(x,\theta)$ and gives the propensity of (binary) treatment given that the potential outcome model Θ_t is known to be θ. The MSM can help partially identify this quantity, which on its own would make it possible to identify the potential outcome model. By the MSM,

$$\left[\frac{e_t(x)}{1 - e_t(x)}\right]^{-1} \frac{e_t(x,\theta)}{1 - e_t(x,\theta)} \in [\Gamma^{-1}, \Gamma^{+1}].$$

The bounded ratio of odds permits characterization of the counterfactual in the following equation:

$$\begin{aligned} p(\theta_t|x) &= \underbrace{p(\theta_t \mid T = t, X = x)}_{\textit{(factual)}} e_t(x) + \underbrace{p(\theta_t \mid T = 1-t, X = x)}_{\textit{(counterfactual)}} [1 - e_t(x)] \\ &= p(\theta|t,x)e_t(x) + [1 - e_t(x,\theta)]p(\theta_t|x), \qquad \therefore \quad p(\theta_t|x) = \frac{e_t(x)}{e_t(x,\theta)} p(\theta|t,x), \end{aligned}$$

and $\omega(\theta,t,x) \triangleq e_t(x)/e_t(x,\theta)$ according to Assumption 2, partially identifiable by Equation 2.

Example 2 *For continuous-valued treatments, a δMSM (Marmarelis et al., 2023) with parameter $\Gamma > 1$ likewise defines $[\underline{\omega}, \overline{\omega}]$ in terms of analytically tractable integrals over the propensity density.*

Formally, Assumption 2 can be derived as a consequence of using a sensitivity model on the Radon-Nikodym derivative of the potential model Θ_t with respect to the learned model Θ, while assuming absolute continuity between the two distributions:

$$\omega(\theta,t,x) \triangleq \frac{\mathrm{d}P(\Theta_t = \theta \mid X = x; \mathcal{D})}{\mathrm{d}P(\Theta = \theta \mid T = t, X = x; \mathcal{D})} = \frac{p(\theta_t \mid x; \mathcal{D})}{p(\theta \mid \mathcal{D})}.$$

Moving forward, we denote empirical propensity estimates as $\tilde{e}_t(x)$ and hence the empirical weights as $\tilde{\omega}(\theta, t, x)$ with sensitivity bounds $[\underline{\tilde{\omega}}, \overline{\tilde{\omega}}]$. These are learned from the training data with sample size n. On the other hand, posterior quantities approximated by a finite ensemble of size m shall use the hat symbol. The finite-sample version of Equation 3 below, for instance, is denoted as $\hat{p}_{\mathcal{M}}(y_t \mid x; \mathcal{D})$.

2.1. Sensitivity Analysis on Quantiles via Ensemble

Our proposal is to apply the weights from Assumption 2 directly over the outcome models. Concretely, the potential outcomes are described as a weight-modulated version of Equation 1:

$$p_{\mathcal{M}}(y_t \mid x; \mathcal{D}) = \mathbb{E}_{\Theta}\big[\omega(\Theta, t, x)\, p_{\mathcal{M}}(y \mid t, x; \Theta) \mid \mathcal{D}\big], \quad \text{where } \omega(\cdot | t, x) \in [\underline{\omega}(t, x), \overline{\omega}(t, x)]. \tag{3}$$

Prediction intervals for individual outcomes would take the form $[F_{\omega}^{-1}(\alpha/2),\ F_{\omega}^{-1}(1-\alpha/2)]$ with expected miscoverage α (mirroring the conformal usage,) where $F_{\omega}(y)$ is the cumulative density of $p_{\mathcal{M}}(y_t \mid x; \mathcal{D})$, given in Equation 3. Partial identification entails an optimization over the outcome interval for maximal ignorance as admitted by the sensitivity model. We construct the program

$$\begin{aligned} Y_t \mid X \ &\in \Big[\inf F_{\omega_1}^{-1}(\alpha/2),\ \ \sup F_{\omega_2}^{-1}(1-\alpha/2)\Big], \\ \text{subject to} \quad &\omega_1(y),\ \omega_2(y) \in \big[\underline{\omega}(t, x),\ \overline{\omega}(t, x)\big], \\ &F_{\omega_1}(y) \text{ and } F_{\omega_2}(y) \text{ are probability distributions.} \end{aligned} \tag{4}$$

A globally optimal greedy solution to the finite-sample problem with an ensemble is presented in Supplementary Algorithm 2, with the optimality criterion addressed in Theorem 4. A general procedure for maximizing (& minimizing) the ensemble quantile is provided in Algorithm 1.

```
Algorithm 1: General Quantile Maximizer
Input: Quantile rank $\beta$, weight bounds $(\underline{\omega}, \overline{\omega})$ like those described in Assumption 2, and
       invertible cumulative density functions $F_1(y), F_2(y), \ldots, F_n(y)$, which can be
       considered the conditional prediction distributions from the ensemble.
Output: Ensemble's $\beta$-quantile, $q \coloneqq \sup_w F^{-1}(\beta)$.
Initialize $w_i \leftarrow 1$ for all $i = 1, 2, \ldots n$;
while global optimality is not met, according to Theorem 4, do
    Compute $\beta$-quantile of current $F(y) \coloneqq n^{-1} \sum_i w_i F_i(y)$;
    Find pair(s) of weight indices $(r, s)$ that violate the optimality condition per Theorem 4;
    Transfer weight between pair(s) $(w_r, w_s)$ such that the condition is satisfied for the pair(s);
end
```

3. Estimation Properties

Our main assumption beyond Assumptions 1 & 2 that enables a simple coverage guarantee of causal outcomes Y_t is that they are independently generated by some unobserved $\Theta_t \sim P(\Theta_t | \mathcal{D})$. This requirement, marked by the subscript $\mathcal{M}$, aligns with our parametric setting and a Bayesian perspective. However, we acknowledge that this result is not as general as the conformal alternatives. We note, additionally, that the empirical evaluations of §4 do not necessarily enforce these conditions.

Lemma 2 (Empirical Coverage) *For fixed values t, x, and $\alpha \in (0, 1)$, consider empirical weights $\tilde{\omega}(\theta, t, x)$. Let $\hat{F}_{\tilde{\omega}}$ be the cumulative distribution of the empirical, finite-ensemble estimate for the potential outcome of Equation 3, i.e. $\hat{p}_{\mathcal{M}}(y_t \mid x; \mathcal{D}) = \hat{\mathbb{E}}_m[\, \tilde{\omega}(\Theta, t, x) \times p_{\mathcal{M}}(y \mid t, x; \Theta) \mid \mathcal{D}\,]$. Then for any $\varepsilon > 0$ and $\beta = \alpha + \varepsilon + 2\,\mathbb{E}\,|\tilde{\omega} - \omega|$, it holds with probability at least $1 - \beta$ that*

$$\mathbb{P}_{\mathcal{M}}\big[Y_t \in \big(\hat{F}_{\tilde{\omega}}^{-1}(\alpha/2),\ \hat{F}_{\tilde{\omega}}^{-1}(1 - \alpha/2)\big) \bigm| X = x\big] \ > \ 1 - 2\exp\big\{-m\varepsilon^2/2\big\}.$$

We blend the finite-sample coverage result in Lemma 2 with partial identification. Theorem 3 characterizes the validity of the causal-outcome intervals from a finite ensemble of size m.

Theorem 3 (Valid Partial Identification) *For fixed values t, x, and $\alpha \in (0, 1)$, consider weight boundary estimates $[\underline{\tilde{\omega}}, \overline{\tilde{\omega}}]$ yielded from a sensitivity model according to Assumption 2. Estimating a solution to the program in Equation 4 produces outcome intervals with hidden-confounding ignorance. Assume for the admitted extrema $\inf \hat{F}_{\tilde{\omega}}^{-1}(\alpha/2)$ and $\sup \hat{F}_{\tilde{\omega}}^{-1}(1 - \alpha/2)$, that $\tilde{\omega}(\Theta) \in \{\underline{\tilde{\omega}}, \overline{\tilde{\omega}}\}$ almost surely. Now let $\beta = \alpha + \varepsilon + 2\,\mathbb{E}[\,|\underline{\tilde{\omega}} - \underline{\omega}| \vee \big|\overline{\tilde{\omega}} - \overline{\omega}\big|\,]$ for any margin constant $\varepsilon > 0$. In this case, with probability at least $1 - \beta$,*

$$\mathbb{P}_{\mathcal{M}}\big[Y_t \in \big(\inf \hat{F}_{\tilde{\omega}}^{-1}(\alpha/2),\ \sup \hat{F}_{\tilde{\omega}}^{-1}(1 - \alpha/2)\big) \bigm| X = x\big] \ > \ 1 - 2\exp\big\{-m\varepsilon^2/2\big\}. \tag{5}$$

Next, we justify our Supplementary Algorithm 2 by revealing a global optimality condition that can be reached greedily. Theorem 4 suggests a simple, monotonically nondecreasing update rule for an optimization algorithm: find pairs of ensemble components that disprove the optimality condition, and transfer weight between them.

Theorem 4 (Global Optimality Condition) *The weight assignments $\omega(\theta^{(i)}) \in [\underline{\omega}, \overline{\omega}]$ for a predictor ensemble $\{\theta^{(i)}\}_{i=1}^m$ maximize the β-quantile of the finite weighted mixture in the space of all admissible weight assignments if and only if there exists no pair of mixture components $(\theta^{(j)}, \theta^{(k)})$ such that weight can be transferred from j to k, i.e. $\omega(\theta^{(j)}) > \underline{\omega}$ and $\omega(\theta^{(k)}) < \overline{\omega}$, and j has more leftward mass than k, i.e. $F(q; \theta^{(j)}) > F(q; \theta^{(k)})$ where q is the current β-quantile: $\beta = m^{-1} \sum_i \omega(\theta^{(i)}) F(q; \theta^{(i)})$.*

Some of the empirical tightness of the outcome intervals might stem from the preservation of continuity in the partially identified densities; by the definition of Lipschitz continuity on the reals,

Proposition 5 (Continuity of Outcome Density) *If the predictor densities $p_{\mathcal{M}}(y \mid t, x; \theta_i)$ are C-Lipschitz continuous, then the posterior outcome density $p_{\mathcal{M}}(y_t \mid x; \mathcal{D})$ is $\overline{\omega}C$-Lipschitz.*

On the other hand, a sample-based reweighing scheme like from Kallus et al. (2019); Jesson et al. (2021a) does not preserve any implied continuity of the partially identified probability density.

4. Empirical Evaluations

We present three benchmarks comparing the tightness of the outcome intervals produced by Caus-Modens and the prevailing conformalized causal sensitivity analyses. As discussed in §1, these conformal approaches encompass the state of the art in partial idenfiticaiton of individual outcomes and not outcome expectations. We first list the baselines and then detail the evaluation procedure.

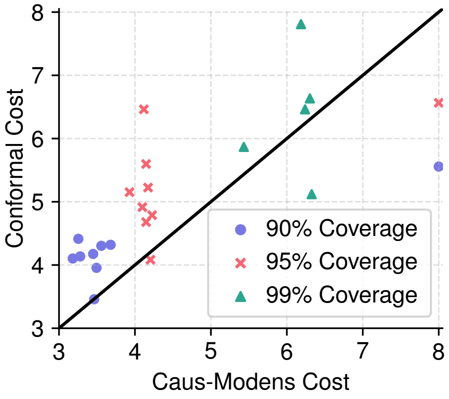
The baseline methods. We consider various combinations and ablations of conformal sensitivity analysis (CSA) (Yin et al., 2022; Jin et al., 2023) with the two state-of-the-art conformal backbones: distributional conformal prediction (DCP) (Chernozhukov et al., 2021) and conformalized quantile regression (CQR) (Romano et al., 2019). The CSA studies relied on CQR for their implementations. In the meanwhile an even more adaptive procedure, DCP, was proposed. For completeness in our analysis we constructed a "supercharged" baseline that combined CSA with DCP.

Since we learned an entire ensemble for each benchmark, we usually allowed the conformal alternatives to also harness the empirical uncertainties captured by the ensemble. Again, this was done in an attempt to be as favorable to the conformal alternatives as possible. Ensembled baselines were marked by the "Ens-" prefix, and the one baseline without that simply used a single model drawn from the ensemble. The predictive modeling foundations for all methods were kept the same so that there was no question about differential modeling performance leading to different results between Caus-Modens and baselines. This ensured the benchmarks were commensurate. For instance, whereas CQR normally calls for quantile regression, we fed it quantiles of the ensemble-marginalized distributional prediction. We list the actual baselines implemented:

- Ens-CSA-DCP — the main conformal baseline with all the beneficial components;
- Ens-CSA-CQR — similar to the above, but with the more standard CQR;
- CSA-DCP — the non-ensembled ablation;
- Ens-DCP — the non-CSA (non-causal) ablation.

How tightness is measured. Most of our results were reported with a concept of coverage efficiency. As is customary with studies on conformal inference, we set a target coverage level. Then we observed the size of the intervals required to achieve that level of coverage on a causal test set where treatments were de-confounded (the smaller the intervals, the more *efficient* the coverage.) The logic of this strategy is that a tighter partial identification should require less implied hidden confounding (via the sensivitiy model) to cover the causal outcomes, relative to a more conservative method that does not utilize the sample statistics or problem assumptions effectively. Each result section (§4.1, §4.2, §4.3) defined a domain-specific cost function to measure the size of the outcome intervals. Tighter intervals had lower cost. We always used Tan (2006)'s sensitivity model for binary treatments (MSM) and varied its single parameter Γ for the extent of violation to the ignorability assumption. We explored the landscape between and including $\Gamma = 1$, where ignorability holds, and $\Gamma = 50$, where all methods would plateau. We used binary search to identify the smallest $\Gamma := \Gamma^* \in [1, 50]$ that achieves a target coverage, like 95% of the test set. That point could be $\Gamma = 1$, the no-hidden-confounding condition. On the other hand, an experiment was classified as a failure if the method never reached the target coverage. The cost function evaluated at a successful Γ^* was termed the *coverage cost*.

Training the predictors. Whereas Caus-Modens was conceived in a Bayesian framework, in practice, deep ensembles tend to achieve better accuracy than Bayesian neural networks and similarly quantify empirical uncertainty (Lakshminarayanan et al., 2017; Pearce et al., 2018; Fort et al., 2019; Rahaman et al., 2021). Caus-Modens ultimately requires a sample of predictive models, whether from a posterior or an estimator distribution. The focus of our evaluations is the sensitivity analyses once models have been learned. Hence, we use deep ensembles of cardinality 16 in our reported benchmarks. We trained fully connected feedforward neural networks with sigmoid activations for both the outcome and the propensity predictors. Hyperparameter and architecture selection was done



# Failures	Caus-Modens	Baseline
90% Coverage	**0** / 10	1 / 10
95% Coverage	**0** / 10	1 / 10
99% Coverage	**0** / 10	**5** / 10

Figure 2. Coverage costs of Caus-Modens versus the main ensemble-conformal baseline, Ens-CSA-DCP, for the ten IHDP realizations used by Louizos et al. (2017). In the scatter plot, we only display the cost pairs, clipped at 8.0, where both methods achieved the target coverage. The table shows that all the failures to reach adequate coverage occurred on the conformal method.

by grid search. Ensembles were trained by maximum likelihood on bootstrap-resampled training sets and randomly initialized weights. Caus-Modens and all the baselines relied on this set of predictors, either in whole or by randomly drawing individual models in the case of non-ensemble ablations.

4.1. Classical Benchmark (`IHDP`)

The `IHDP` dataset in causal literature is a semi-synthetic classical benchmark for CATE estimation (e.g. Louizos et al., 2017; Jesson et al., 2021b; Samothrakis et al., 2022). It contains binary treatments with covariate shift, and simulated real-valued outcomes, for 747 individuals. The original covariates are eight real and nineteen binary attributes. To induce hidden confounding, we obscured the binary covariates. The benchmark task was to predict the $T = 1$ (potential) outcomes of the test set. Due to the smallness of the sample, we randomly allocated 10% of the data to the validation set and 20% to the test set. In addition to hyperparameter selection, the validation set served for calibration in the conformal baselines for maximal resourcefulness. The whole calibration set consisted of 3/7$^{\text{ths}}$ of the otherwise-labeled training set for an ultimate 50-50 estimation-calibration split, as is recommended with split conformal prediction (SCP) (Papadopoulos et al., 2002). In other words, Caus-Modens utilized the entire training set, and the conformal baselines used 4/7$^{\text{ths}}$ of the training set for estimation and the rest for calibration. We also tested a 75-25 SCP split rather than 50-50, to similar effect.

Cost function & results. The cost function was the absolute length of the interval scaled to the empirical standard deviation of the outcomes. We first tested Caus-Modens against Ens-CSA-DCP for three target coverages shown in Figure 2. Caus-Modens produced tighter intervals with Wilcoxon signed-rank test $p < 0.05$. This dataset was noteworthy for the occurrence of failures in the conformal approach and complete success in Caus-Modens for achieving the target coverage. Supplemental Table 3 shows how other conformal configurations induced more failures. For Caus-Modens we found that the size of the ensemble beyond 16 predictors ceased to impact the coverage cost.

4.2. Novel Semi-synthetic Benchmark (PBMC)

Recent widely celebrated single-cell RNA sequencing (scRNAseq) modalities have enabled an unprecedented view into human physiology (Jovic et al., 2022). The complex relations between the expressions of roughly 20,000 genes makes it a good source for benchmark datasets with unintuitive statistics. We obtained a relatively clean dataset of well-characterized peripheral blood mononuclear cells (PBMCs) (Kang et al., 2018) and randomly projected the gene expressions into 32 observed and 32 unobserved confounders, as well as a treatment variable that was discretized to binary values. The simulated outcome was a completely random quadratic form of all these 32+32+1 variables, ensuring arbitrary relations between treatment assignment, confounders, and outcome.

A current shortcoming of partial identifiers for expectations of causal quantities, like ATE and CATE, is that they were not designed for heavy-tailed outcomes. We showcase Caus-Modens in this light, using the Cauchy distribution for simulated PBMC outcomes. The Cauchy distribution has several scientific uses, including the modeling of physical & financial phenomena (e.g. Kagan, 1994) and specifying priors for variance (Gelman, 2006). It is considered "pathological" because it has no mean or higher moments. The sample mean is also Cauchy distributed—for it is a stable distribution (Nolan, 2020)—and diverges in large samples. However, a viable alternative is to estimate the tail parameters by maximum likelihood (e.g. Huisman et al., 2001; Taleb, 2020). Parametric approaches are paramount to characterizing pathological distributions like the Cauchy, which are punctuated by extreme rare events. This simple benchmark highlights the value of inductive bias.

Cost function & results. For an interpretable measure than can be aggregated across multiple experiments, the cost function for Cauchy-outcome intervals was the interval length scaled to the smallest achieved length in each setting. Table 1 displays these relative costs for a high coverage target of 99%, evaluating each method's ability to characterize Cauchy tails. Caus-Modens achieved *significantly* lower costs than the CSA baselines while meeting target coverage on average, and the non-causal Ens-DCP had similar cost for greater miscoverage, with failure on average.

Method	Achieved Coverage ↑	Coverage Cost ↓	Avg. Coverage Loss ↓
Caus-Modens	**99.15** (0.20) %	**0.28** (0.15)	0.028 % pts
Ens-CSA-DCP	**99.58** (0.32) %	1.51 (1.94)	0.002 % pts
Ens-CSA-CQR	**99.57** (0.32) %	1.51 (1.84)	0.003 % pts
CSA-DCP	**99.60** (0.32) %	1.50 (1.82)	0.002 % pts
Ens-DCP	98.95 (0.45) %	0.30 (0.20)	0.206 % pts

Table 1. PBMC results from 16 independent dataset generations and inferences. We set the random seed to the predetermined value 0 prior to generation for reproducibility and transparency. We present average achieved coverages and standard deviations for a target of 99%, accompanied by relative coverage costs for the trials that met the target, and the average nonnegative loss in coverage percentage points, which was positive for trials with coverage below 99%.

4.3. Novel Benchmark via GPT-4 (AITA)

Semi-synthetic causal benchmarks like PBMC can be designed to harness the arbitrary statistical relations in real data. Still, the outcome must have pre-specified functional relations with the treatment and confounders. With the proliferation of causal-infernece studies proposing new methods for various settings, there is a need for flexible yet realistic benchmarks. In this result section we took a step in building a new kind of observational dataset that includes intervention results without the challenges of actually bringing in a randomized control experiment for testing the causal inference. We used the celebrated large language model (LLM) GPT-4 (OpenAI, 2023) that has demonstrated remarkable capabilities in emulating human text (Brown et al., 2020). One can use an LLM to sample complex outcomes from observational datasets and also query textual *interventions*. We seek to promote this usage of large generative models for benchmarking causal inference (Curth et al., 2021).

We framed the novel inference task in the format of the r/AmITheAsshole subreddit (hence the name of this benchmark, `AITA`.) The subreddit is a scientifically attractive setting because the rules and structure of the forum are clean: users post personal stories of conflict, and comments offer opinions on whether the author was at fault in the way the story transpired. A verdict is determined by the upvote mechanism on comments. Data from this subreddit have recently served as a vessel for human perspectives (Botzer et al., 2022) and moral judgment (Plepi et al., 2022) in the computational social sciences. For the sake of a causal benchmark, we asked GPT-4 to act as moral arbiter on real posts from the subreddit (O'Brien, 2020). That way there would be no doubt about the real-world salience of the data, while permitting interventions via the LLM. The treatment variable was the customary self-identified gender, which is binary between 'F' and 'M' (a limiting and problematic format.) Nevertheless, this variable allowed us to assess a bias in GPT-4's verdicts.

The mechanics of GPT-4's "intuitive process" are so complex that it would be difficult to predict its moral judgments through a much simpler outcome predictor that would necessarily be trained on a relatively small sample of text embeddings: the `AITA` posts that have discernible gender markers. We changed this benchmark in order to simplify the prediction problem. Concretely, the outcome predictor was tasked with denoising an artificially noised GPT-4 verdict, utilizing the gender (treatment) and topic (covariates) information of the post. See Supplemental Figure 4 for a diagram. Topics were represented by five-dimensional embeddings like in BERTopic (Grootendorst, 2022). To strengthen the bias in GPT-4, we coupled the real indicated gender with synthetic ages so that authors were either 30-year old men ($T = 0$) or 70-year old women ($T = 1$). Caus-Modens and baselines were tasked with predicting the causal effect of T on the denoised verdict.

Cost function & results. The GPT-4 verdict was a number from 1–99, afterwards rescaled to the unit interval, and then logit-transformed to be modeled by normal distributions. These outcomes had a standard deviation of 0.62 and the artificial noise standard deviation was selected to be 0.5. After learning the outcome (and propensity,) we focused on the woman gender arm of potential outcomes for the intervention testing set. Verdict-noise variance out of sample was reduced by 40% with the predictors. We chose two cost functions for evaluating coverage efficiency: the absolute units of outcome interval size, and *mass* units computed as the integral of the empirical marginal outcome distribution along the interval. Figure 3 shows the rate of tightest intervals for all the evaluated methods that reached sufficient coverage, in both units and at a wide array of coverage targets.

At 70% coverage, the verdict overlap between genders is smaller with Caus-Modens outcome intervals than with Ens-CSA-DCP for $20.0\,(0.8)\,\%$ of the posts, with the rest being equal in the empirical probability mass units, and none in the other direction. This suggests that Caus-Modens identifies more gender bias in GPT-4's moral judgments. The doubly robust and debiased ATE estimator (Yao et al., 2021) for gender effect suggests that ATE $= -0.07$ with Student t-test $p < 0.05$, likewise revealing a gender bias towards young men being more wrong than old women according to

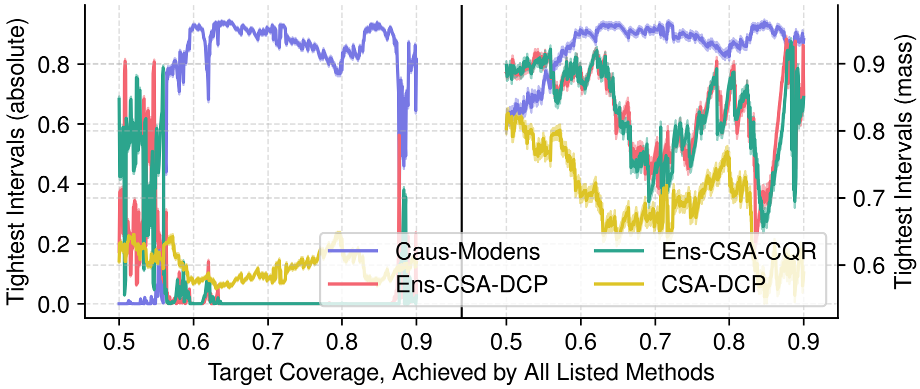


Figure 3. Share of the test set that each method produced the tightest intervals, shown in absolute (left) and mass (right) units, at a wide range of target coverages. Lines are widened by standard errors. The four listed methods always achieved the target, whereas the unlisted Ens-DCP failed frequently. The presence of ties, particularly in the right subplot, allows the shares to add up to more than unit.

GPT-4. Table 2 compares GPT-4 verdicts to the original Reddit verdicts and displays coverage costs for different judgment regimes. Caus-Modens consistently outperformed Ens-CSA-DCP, which was the best of the baselines according to Figure 3, even in empirical mass units.

Counts	Reddit right	Reddit wrong
GPT-4 right	1303 [58%]	276 [12%]
GPT-4 wrong	430 [19%]	250 [11%]

Costs	Reddit right	Reddit wrong
GPT-4 right	**0.36*** / 0.38	**0.38** / 0.39
GPT-4 wrong	**0.39*** / 0.41	0.35 / 0.37

Table 2. Confusion matrix of Reddit versus GPT-4 verdicts (**left**) and the 70%-coverage costs of Caus-Modens against Ens-CSA-DCP for those stratified posts (**right**). Bold: $p < 0.05$; asterisk: $p < 0.01$. As the GPT-4 verdict was given on a sliding scale, we chose a right/wrong threshold by equalizing its marginal rate with Reddit's. If we chose the midpoint of the verdict spectrum, there would have been many more wrong than right verdicts from GPT-4. This can be explained by a higher threshold for the designation of "asshole" to the author of a post, as is the Reddit protocol.

5. Discussion

We present three benchmarks, IHDP, PBMC, and AITA, on which our proposed Caus-Modens yields tighter coverage than the state of the art in adaptive conformal prediction with causal sensitivity analysis. In IHDP (§4.1), the conformal baselines failed at least once out of ten trials to achieve coverage by reweighing the calibration set. The costs of non-failures tended to be larger than Caus-Modens' at the same target coverage. Caus-Modens did not fail at all for IHDP. The failure rate of the conformalized sensitivity analyses highlights potential pitfalls of reweighing a finite observational sample to achieve coverage of causal outcomes. Valid coverage for CSA is contingent on correct propensity specification (see Lemma 3, Yin et al., 2022), which may be a challenge in certain applications with strong, high-dimensional covariate shift.

In the much larger PBMC (§4.2) benchmark, Caus-Modens achieved the tightest coverage for the majority of the trials. PBMC leveraged the relations between gene expressions in human cells by randomly projecting them into confounders and treatment. Further, it enabled the demonstration of Caus-Modens quantile predictions for extremely heavy tails. In the novel AITA benchmark (§4.3), which used an LLM to generate extremely complex causal outcomes with access to interventions, we also demonstrated that the outcome intervals produced by Caus-Modens were consistently more informative, in terms of share of tightest coverage, than the intervals produced by the baselines.

Empirical uncertainty via ensembles. Ensemble-based uncertainty quantification in deep learning remains an active field of study (Shen and Cremers, 2022; Ashukha et al., 2020; Wimmer et al., 2023; Theisen et al., 2024). We employed the classic deep ensemble (Lakshminarayanan et al., 2017) for our results. However, our approach is flexible and works with many kinds of "ensembles," even those that are implicitly defined, as with MC dropout (Gal and Ghahramani, 2016), or the inducible model distribution in Bayesian neural networks (Wenzel et al., 2020).

Using a sensitivity model. As mentioned in Examples 1 & 2, popular sensitivity models for hidden confounding like the MSM are parametrized by a single $\Gamma \geq 1$, with $\Gamma = 1$ signifying no hidden confounders. In past studies, Γ has been informed by domain knowledge (Cornfield et al., 1959) or data-driven heuristics (Hsu and Small, 2013). Without domain knowledge, a sensitivity model can help rank heterogeneous causal effects (across observational units, or even types of treatment) by apparent robustness to hidden confounding. Our empirical evaluations tested the alignment of the ensembled sensitivity model with the ground truth of semi-synthetic hidden confounders.

Societal concerns. We wish to emphasize that GPT-4's verdicts on morality in the AITA scheme should not be considered as an approximation of human moral judgements. GPT-4 is biased in poorly understood ways by its architecture and training data (Yu et al., 2024; Gallegos et al., 2023; Ferrara, 2023). Further, the biases isolated in the circumscribed scope of these experiments do not necessarily even reflect the general biases present in GPT-4.

Limitations. Causal inference methods always depend on the appropriateness of the assumptions. In our case, this includes assumptions about the data-generating process (namely constraints on the hidden confounders, via a sensitivity model) and the quality of the estimates. Our method elevates the role of empirical uncertainty relative to many prior works. However, in contrast with recent alternatives, Caus-Modens may be less conservative by relying more on parametric, structural, and inductive constraints. Hence, it could also be more vulnerable to model misspecification.

6. Conclusion

Our simple ensemble-based partial identification of outcome quantiles is a promising approach to prediction intervals that leverages the inductive biases of deep models. In addition to coverage efficiency, it accommodates various sensitivity models and adapts them to a novel formulation of potential posteriors that justifies the weight-modulation of an ensemble. Future work could explore how to regularize Caus-Modens to be sharper in its partial identification, especially in general (non-binary) treatment domains that are less explored.

Acknowledgments

This work was funded in part by Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002.

References

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

Nicholas Botzer, Shawn Gu, and Tim Weninger. Analysis of moral judgment on reddit. *IEEE Transactions on Computational Social Systems*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)*, 2021.

Jacob Dorn and Kevin Guo. Sharp sensitivity analysis for inverse propensity weighting via quantile balancing. *Journal of the American Statistical Association*, pages 1–13, 2022.

Jacob Dorn, Kevin Guo, and Nathan Kallus. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, 29, 2016.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

A Gelman. Prior distributions for variance parameters in hierarchical models (comment on an article by browne and draper). *Bayesian Analysis*, 1:515–533, 2006.

Peter W Glynn et al. Importance sampling for monte carlo estimation of quantiles. In *Mathematical Methods in Stochastic Simulation and Experimental Design: Proceedings of the 2nd St. Petersburg Workshop on Simulation*, pages 180–185. Citeseer, 1996.

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jesse Y Hsu and Dylan S Small. Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811, 2013.

Ronald Huisman, Kees G Koedijk, Clemens J M Kool, and Franz Palm. Tail-index estimates in small samples. *Journal of Business & Economic Statistics*, 19(2):208–216, 2001.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33: 11637–11649, 2020.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. *ICML*, 2021a.

Andrew Jesson, Panagiotis Tigas, Joost van Amersfoort, Andreas Kirsch, Uri Shalit, and Yarin Gal. Causal-bald: Deep bayesian active learning of outcomes to infer treatment-effects from observational data. *Advances in Neural Information Processing Systems*, 34:30465–30478, 2021b.

Andrew Jesson, Alyson Rose Douglas, Peter Manshausen, Maëlys Solal, Nicolai Meinshausen, Philip Stier, Yarin Gal, and Uri Shalit. Scalable sensitivity and uncertainty analyses for causal-effect estimates of continuous-valued interventions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Ying Jin, Zhimei Ren, and Emmanuel J Candès. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *Proceedings of the National Academy of Sciences*, 120(6): e2214889120, 2023.

Dragomirka Jovic, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12(3):e694, 2022.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854, 2021.

Yan Y Kagan. Observational evidence for earthquakes as a nonlinear dynamic process. *Physica D: Nonlinear Phenomena*, 77(1-3):160–192, 1994.

Nathan Kallus. Treatment effect risk: Bounds and inference. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 213–213, 2022.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 6037–6060. PMLR, 2023.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pages 2281–2290. PMLR, 2019.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1): 89–94, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

Fan Li, Peng Ding, and Fabrizia Mealli. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247):20220153, 2023.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Charles F Manski. *Partial identification of probability distributions*, volume 5. Springer, 2003.

Myrl G Marmarelis, Elizabeth Haddad, Andrew Jesson, Neda Jahanshad, Aram Galstyan, and Greg Ver Steeg. Partial identification of dose responses with hidden confounders. In *Uncertainty in Artificial Intelligence*, pages 1368–1379. PMLR, 2023.

Lizhen Nie, Mao Ye, qiang liu, and Dan Nicolae. {VCN}et and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021.

John P Nolan. Univariate stable distributions. *Springer Series in Operations Research and Financial Engineering, DOI*, 10:978–3, 2020.

Elle O'Brien. iterative/aita_dataset: Praw rescrape of entire dataset, February 2020. URL `https://doi.org/10.5281/zenodo.3677563`.

OpenAI. Gpt-4 technical report, 2023.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR, 2018.

Joan Plepi, Béla Neuendorf, Lucie Flek, and Charles Welch. Unifying data perspectivism and personalization: An application to social norms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7391–7402, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.500`.

Rahul Rahaman et al. Uncertainty quantification and deep ensembles. *Advances in Neural Information Processing Systems*, 34:20063–20075, 2021.

Alexandre Rame, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, patrick gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

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

Spyridon Samothrakis, Ana Matran-Fernandez, Umar Abdullahi, Michael Fairbank, and Maria Fasli. Grokking-like effects in counterfactual inference. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

Yuesong Shen and Daniel Cremers. Deep combinatorial aggregation. *Advances in Neural Information Processing Systems*, 35:32299–32310, 2022.

Nassim Nicholas Taleb. Statistical consequences of fat tails: Real world preasymptotics, epistemology, and applications. *arXiv preprint arXiv:2001.10488*, 2020.

Zhiqiang Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637, 2006.

Ryan Theisen, Hyunsuk Kim, Yaoqing Yang, Liam Hodgkinson, and Michael W Mahoney. When are ensembles really effective? *Advances in Neural Information Processing Systems*, 36, 2024.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Xin Wang, Shengfei Lyu, Xingyu Wu, Tianhao Wu, and Huanhuan Chen. Generalization bounds for estimating causal effects of continuous treatments. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pages 10248–10259. PMLR, 2020.

Veit David Wild, Sahra Ghalebikesabi, Dino Sejdinovic, and Jeremias Knoblauch. A rigorous link between deep ensembles and (variational) bayesian methods. *arXiv preprint arXiv:2305.15027*, 2023.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Uncertainty in Artificial Intelligence*, pages 2282–2292. PMLR, 2023.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46, 2021.

Mingzhang Yin, Claudia Shi, Yixin Wang, and David M Blei. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, pages 1–14, 2022.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix A. Algorithm

```
Algorithm 2: Greedy Quantile Maximizer                    (minimizer version is trivial)
Input: Quantile rank $\beta$, weight bounds $(\underline{\omega}, \overline{\omega})$ like those described in Assumption 2, and
       invertible cumulative density functions $F_1(y), F_2(y), \ldots, F_n(y)$, which can be
       considered the conditional prediction distributions from the ensemble.
Output: Ensemble's $\beta$-quantile, $q \coloneqq \sup_w F^{-1}(\beta)$.
Initialize $w_i \leftarrow 1$ for all $i = 1, 2, \ldots n$;
Compute initial search bounds $\underline{q} \leftarrow \min_i F_i^{-1}(\beta)$ and $\overline{q} \leftarrow \max_i F_i^{-1}(\beta)$;
while not converged do
    Binary-search for $q \leftarrow F^{-1}(\beta) \in (\underline{q}, \overline{q})$, where $F(y) \coloneqq n^{-1} \sum_i w_i F_i(y)$;
    Compute masses $\alpha_i \coloneqq F_i(q)$ for every $i$ and sort in ascending order (without relabeling);
    Find receiver $r \coloneqq \arg\min_i \alpha_i$ such that $w_i < \overline{\omega}$;
    Find sender   $s \coloneqq \arg\max_i \alpha_i$ such that $w_i > \underline{\omega}$;
    if $r \geq s$ then
        break;
    end
    Compute receivable $a \coloneqq \overline{\omega} - w_r$ and sendable $b \coloneqq w_s - \underline{\omega}$;
    if $a < b$ then
        Transfer $w_r \leftarrow \overline{\omega}$ and $w_s \leftarrow w_s - a$;
    else
        Transfer $w_s \leftarrow \underline{\omega}$ and $w_r \leftarrow w_r + b$;
    end
    Refine search bounds $\underline{q} \leftarrow q$;
end
```

This algorithm has quadratic asymptotic runtime in the ensemble size. As the bottleneck tends to be the quantile-search subroutine, one may benefit from implementing a *bulk* weight-transfer procedure using Algorithm 2 as a starting point.

Appendix B. Proofs

Proof [Lemma 2] We study estimation errors in the weights in a manner inspired by Theorem 3 (supplementary) of Lei and Candès (2021). In our case, we study the cumulative distribution function (CDF) estimate

$$\hat{F}_{\tilde{\omega}}(y) = m^{-1} \sum_{j=1}^{m} \tilde{\omega}_j F_j(y) = m^{-1} \sum_{j=1}^{m} \tilde{\omega}(\Theta^{(j)}) F(y;\, \Theta^{(j)})$$

for $\hat{p}_{\mathcal{M}}(y_t \mid x; \mathcal{D}) = \hat{\mathbb{E}}_m[\tilde{\omega}(\Theta, t, x) \times p_{\mathcal{M}}(y \mid t, x; \Theta) \mid \mathcal{D}]$ with t and x omitted for brevity. We wish to accurately predict $Y_t \sim p_{\mathcal{M}}(Y_t \mid X = x; \mathcal{D})$. Our main tool will be Hoeffding's inequality (Wainwright, 2019) in this endeavor. Namely, for any $u > 0$,

$$\mathbb{P}[m\hat{F}_{\tilde{\omega}}(y) - \mathbb{E}\, m\hat{F}_{\tilde{\omega}}(y) \geq u] \leq \exp\{-2u^2/m\}$$

because the individual CDFs are independent conditional on a fixed y, and take the range $[0, 1]$. We focus on the upper bound (with quantile $1 - \alpha/2$) of the prediction interval first, and extend that result to the lower bound by a symmetry argument.

Resolving the expectation and factoring out m, we find

$$\mathbb{P}[\hat{F}_{\tilde{\omega}}(y) - F_{\tilde{\omega}}(y) \geq u] \leq \exp\big\{-2mu^2\big\} \quad \text{where } F_{\tilde{\omega}}(y) = \mathbb{E}[\tilde{\omega} F(y;\, \Theta)]. \tag{6}$$

We observe that $F_{\tilde{\omega}}(y) = \mathbb{E}[\omega F(y;\, \Theta)] + \mathbb{E}[(\tilde{\omega} - \omega) F(y;\, \Theta)] = F_{\omega}(y) + F_{\tilde{\omega}-\omega}(y)$. Now let $u = 1 - F_{\tilde{\omega}}(y) + \alpha/2$ for the upper bound. This implies $\mathbb{P}[\hat{F}_{\tilde{\omega}}(y) \geq 1 - \frac{\alpha}{2}] \leq \exp\big\{-2m[1 - F_{\tilde{\omega}}(y) + \frac{\alpha}{2}]^2\big\}$. Now we introduce the margin $\varepsilon > 0$ and note that when $F_{\tilde{\omega}}(y) + \varepsilon \leq 1 - \frac{\alpha}{2}$, then $u > 0$ and we have

$$\mathbb{P}\Big[\hat{F}_{\tilde{\omega}}(y) \geq 1 - \frac{\alpha}{2}\Big] \leq \exp\big\{-2m\varepsilon^2\big\} \iff \mathbb{P}\Big[\hat{F}_{\tilde{\omega}}(y) < 1 - \frac{\alpha}{2}\Big] > 1 - \exp\big\{-2m\varepsilon^2\big\}.$$

Plugging in $y := Y_t$ from the test set defined above, we find that this law is satisfied when $F_{\omega}(Y_t) \leq 1 - (\frac{\alpha}{2} + \varepsilon + F_{\tilde{\omega}-\omega}(Y_t))$. The fact that $F_{\omega}(Y_t)$ is uniformly distributed (following Assumption 2) allows us to conclude that the condition is met with probability at least $1 - (\frac{\alpha}{2} + \varepsilon + \mathbb{E}\,|\tilde{\omega} - \omega|)$, observing the triangle inequality of the absolute norm.

Applying the same reasoning to the prediction interval's lower bound eventually yields

$$\mathbb{P}\Big[\frac{\alpha}{2} < \hat{F}_{\tilde{\omega}}(Y_t) < 1 - \frac{\alpha}{2}\Big] > 1 - 2\exp\big\{-2m\varepsilon^2\big\}.$$

with aggregate probability $1 - \beta'$ for $\beta' = \alpha + 2\varepsilon + 2\,\mathbb{E}\,|\tilde{\omega} - \omega|$. Applying the inverse CDF and substituting $\varepsilon := \varepsilon/2$ (without loss of generality), we obtain the final form $\beta = \alpha + \varepsilon + 2\,\mathbb{E}\,|\tilde{\omega} - \omega|$; hence with probability $1 - \beta$,

$$\mathbb{P}\Big[Y_t \in \Big(\hat{F}_{\tilde{\omega}}^{-1}\big(\alpha/2\big),\, \hat{F}_{\tilde{\omega}}^{-1}\big(1 - \alpha/2\big)\Big)\Big] > 1 - 2\exp\big\{-m\varepsilon^2/2\big\}.$$

■

Proof [Theorem 3] Here we extend Lemma 2 to partially identifiable weights that yield an admissible set of prediction intervals. The outcome interval under consideration is the union of all these admissible intervals. This is attained by a supremum over the upper bound and an infimum over the lower bound. In applying Hoeffding's bound on the upper bound, we can replace Equation 6 with

$$\mathbb{P}[\hat{F}_{\tilde{\omega}^+}(y) - F_{\tilde{\omega}^+}(y) \geq u] \leq \exp\big\{-2mu^2\big\} \quad \text{such that } \hat{F}_{\tilde{\omega}^+}(y) = \sup \hat{F}_{\tilde{\omega}}(y).$$

Additionally, we define ω^+ to satisfy $\hat{F}_{\omega^+}(y) = \sup \hat{F}_{\omega}(y)$. Let $\beta = \alpha + \varepsilon + \mathbb{E}\,|\tilde{\omega}^+ - \omega^+| + \mathbb{E}\,|\tilde{\omega}^- - \omega^-|$ where $(\tilde{\omega}^-, \omega^-)$ are analogously defined for their respective infima. By a straightforward extension of Lemma 2, we have Equation 5 with probability at least $1 - \beta$. By the assumption stated in this theorem, we know that $\tilde{\omega}^{\pm}(\Theta) \in \{\underline{\tilde{\omega}}, \overline{\tilde{\omega}}\}$ almost surely over the weight assignments. It is clear, then, that $\mathbb{E}\,|\tilde{\omega}^+ - \omega^+| + \mathbb{E}\,|\tilde{\omega}^- - \omega^-| \leq 2\,\mathbb{E}[\,|\underline{\tilde{\omega}} - \underline{\omega}| \vee |\overline{\tilde{\omega}} - \overline{\omega}|\,]$, completing the proof. ■

Proof [Theorem 4] The putative optimality condition for the maximization problem solved by Algorithm 2, restated, is for there to be no pair of mixture components $(\theta^{(j)}, \theta^{(k)})$ such that $\omega(\theta^{(j)}) > \underline{\omega}$ and $\omega(\theta^{(k)}) < \overline{\omega}$, as well as $F(q; \theta^{(j)}) > F(q; \theta^{(k)})$ where q is the current β-quantile:

$$\beta = F(q) \triangleq m^{-1} \sum_i \omega(\theta^{(i)}) F(q; \theta^{(i)}).$$

We will prove both directions of entailment to establish equivalence. First, we must show that if the quantile is maximized, then the condition holds. Suppose that q is the maximal quantile under the problem constraints and the condition is not satisfied, so there indeed is a pair $(\theta^{(j)}, \theta^{(k)})$ as described. This implies that there is weight that could be transferred, of the amount

$$\Delta\omega \triangleq \min\{\omega(\theta^{(j)}) - \underline{\omega},\ \overline{\omega} - \omega(\theta^{(k)})\} > 0.$$

Transferring that weight would yield a new mixture

$$G(\cdot) = \frac{1}{m}\sum_i \omega(\theta^{(i)})F(\cdot\,;\theta^{(i)}) + \frac{\Delta\omega}{m}[F(\cdot\,;\theta^{(k)}) - F(\cdot\,;\theta^{(j)})],$$

with the consequence of $G(q) < F(q)$ because $F(q;\theta^{(j)}) > F(q;\theta^{(k)})$. Therefore $G^{-1}(\beta) > q$ due to monotonicity and q is not the optimal quantile. By contraposition, optimality entails our stated optimality condition. Now for the converse.

With similar notation as above, we have $F(q) = \beta$ but come into the posession of some feasible $G(\cdot)$ where $G(q^*) = \beta$ and $q^* > q$, so q is no longer optimal. Deconstruct $G(\cdot) = m^{-1}\sum_i \omega'(\theta^{(i)})F(\cdot\,;\theta^{(i)})$. By monotonicity, $G(q) < F(q)$. Hence

$$m^{-1}\sum_i [\omega'(\theta^{(i)}) - \omega(\theta^{(i)})]F(\cdot\,;\theta^{(i)}) < 0.$$

Ignoring the identical pairs of weights between F and G,

$$\sum_{i\in\mathcal{A}}[\omega(\theta^{(i)}) - \omega'(\theta^{(i)})]F(q;\theta^{(i)}) > \sum_{i\in\mathcal{B}}[\omega'(\theta^{(i)}) - \omega(\theta^{(i)})]F(q;\theta^{(i)}),$$

$$\mathcal{A} \triangleq \{i : \omega(\theta^{(i)}) > \omega'(\theta^{(i)})\}, \quad \mathcal{B} \triangleq \{i : \omega'(\theta^{(i)}) > \omega(\theta^{(i)})\}.$$

At the same time, $\sum_{i\in\mathcal{A}}[\omega(\theta^{(i)}) - \omega'(\theta^{(i)})] = \sum_{i\in\mathcal{B}}[\omega'(\theta^{(i)}) - \omega(\theta^{(i)})]$ because of the constraint on the probability simplex. For the above inequality to be valid alongside this equality, there *must* be at least one pair $(j \in \mathcal{A},\ k \in \mathcal{B})$ such that $F(q;\theta^{(j)}) > F(q;\theta^{(k)})$. Hence the negation of the optimality condition holds, and by contraposition we prove the other entailment direction. ■

Appendix C. Experimental Details on PBMC

The original PBMC dataset (Kang et al., 2018) had 14,039 cells that were used as data points for our benchmark. In each of the 16 trials, we randomly allocated 8,192 (2^{13}) training instances, 2,048 (2^{11}) validation instances, and 2,048 (2^{11}) test insances from the original sample. Then we designed the causal system by projecting the cells' 17,796 gene expressions into vectors

$$V \triangleq \langle 32 \text{ visible confounders}\ldots, 1 \text{ treatment}, 32 \text{ hidden confounders}\ldots\rangle \in \mathbb{R}^{65}$$

by drawing 65 i.i.d normal coefficients. The 64 confounding entries were rank-normalized to give them Uniform$[0, 1)$ marginals. The treatment entry was binarized by thresholding at $2/3$ so that the data were slightly unbalanced with more $(T = 0)$ observations. The outcome link was determined by a random matrix $M_{ij} \sim$ i.i.d Normal$(0, 1)$. The diagonal coefficient $M_{33,33}$ corresponding to the treatment entry was upscaled by a factor of 64 to keep the treatment effect discernible from the rest. The pre-noised outcome $U \triangleq V^{\mathsf{T}} MV$ endured strong quadratic treatment and confounding effects. Finally, the observed outcome was Cauchy$(\mu = U, \sigma = 1)$-distributed as motivated in §4.2. We aimed to keep this mathematical construction of the semi-synthetic benchmark *parsimonious* by introducing a minimal number of design choices and nontrivial default values.

Appendix D. Experimental Details on AITA

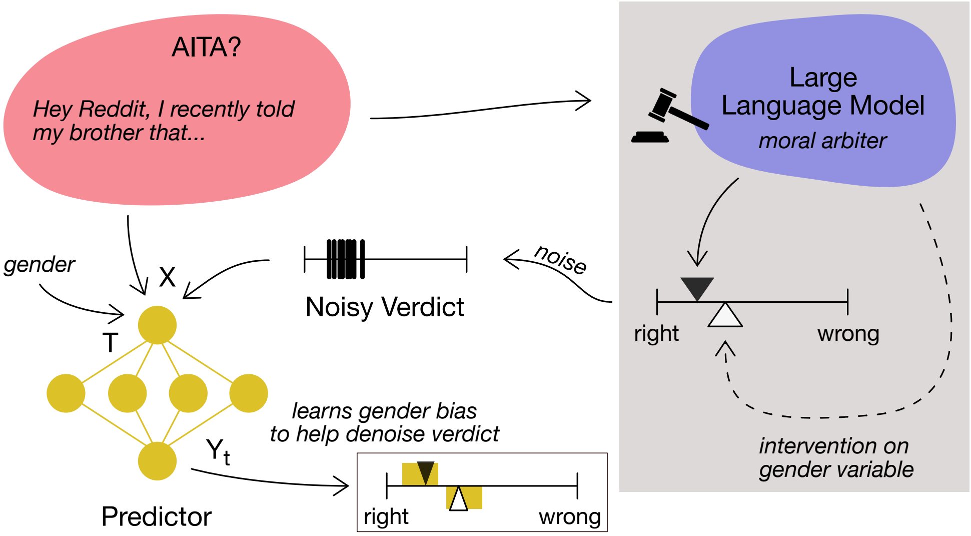

AITA?
Hey Reddit, I recently told my brother that...
Large Language Model
moral arbiter
gender
X
T
Noisy Verdict
noise
right
wrong
Yt
learns gender bias to help denoise verdict
Predictor
right
wrong
intervention on gender variable

Figure 4. Illustration for the AITA benchmark as a denoising task. The predictor receives a rough verdict through a noisy channel, relieving it from having to learn the moral judgment of GPT-4 and allowing it to focus on the biases due to gender and topics that help with denoising.

Figure 4 depicts the experimental setup. We used OpenAI's gpt-4-0314 API model. The O'Brien (2020) dataset contained 11,862 posts with detectable gender identifiers, of which 9,000 were reserved for the observational data. 8,000 of those were for training and the other 1,000 for validation. The 2,862 posts that were not included in the observational dataset were used as the causal outcome test set by intervening on the gender variable before feeding them into GPT-4. Please refer to `scripts/aita-preprocessing.py` in the supplementary source for all the querying and preprocessing details. Processed data are included as CSVs. View the chat-prompt prologue below.

System	You are an average person acting as a subjective moral arbiter on r/AmITheAsshole. Whenever someone presents a story, you respond with a number from 1 to 99 where 1 is noble, 49 is completely neutral (rare), and 99 is atrocious.
User	I was mean for no reason.
Assistant	90
User	I tipped the waiter more than usual.
Assistant	10

Unfortunately, we hit OpenAI's GPT-4 quota before we could evaluate all arms of the potential outcomes. It did not hinder the benchmark, but it did prevent thorough analysis of the causal system.

Appendix E. Additional Results for IHDP

Conformal Baseline	50 - 50 Split	75 - 25 Split	87.5 - 12.5 Split*
Ens-CSA-DCP	1 Failure / 10 Trials	2 Failures / 10 Trials	5 Failures / 10 Trials
Ens-CSA-CQR	2 Failures / 10 Trials	2 Failures / 10 Trials	6 Failures / 10 Trials

Table 3. Failure rates of the baseline methods applied to the IHDP benchmark with 95% target coverage. The results in Figure 2 use the 50-50 split that appears to work best. Asterisk marks the arrangement where the entire original training set is used for estimation, and the validation set for calibration. The other benchmarks (PBMC, AITA) have larger samples that obviate this issue.

Appendix F. Hardware Details

All results were obtained on an Intel Xeon server using a single NVIDIA GeForce RTX 2080 Ti.