Using Benchmark Datasets to Correct for Measurement Error in Privacy-Preserving Classifiers of AI Conversations

Jacob Carlson
Harvard University*

July 2026

1 Introduction

Consider the recent work of Johnston et al. [2026]. As part of their study of the adoption dynamics of agentic AI, they use an automated, *privacy-preserving* classifier of a Codex conversation to predict how "complex" a user request submitted to Codex is. The prompt for the classifier is in Appendix F of their paper; the prompt solicits the "active work time in minutes an experienced human would need" to complete the task, as a specific quantitative definition of task complexity.

Of course, what the classifier outputs is a prediction, not the "ground truth" for the how long the task would take to perform. As such, there is a measurement error question at the heart of analyses that are downstream of classifier-measured task complexity: if for some reason the automated `gpt-5-mini` classifier used by Johnston et al. [2026] is systematically biased at making predictions (even if it is good in a mean square error sense), it will bias downstream estimates of economic quantities of interest.

Johnston et al. [2026] actually do check the performance of their automated classifier with some *benchmark data*. They use a benchmark dataset of 1,000 random Codeforces problems, which contains pairs of coding tasks and (fastest) completion times for competitive coders. They find a correlation of 0.65 between the model-predicted times and the average completion time among the five fastest human solvers of the coding task. They also find that "in general, the model [classifier] tends to predict slower completion times than the scores derived from the metadata." Johnston et al. [2026] rightly point out that this is somewhat expected: "this is in part because we are measuring our modeled prediction against the speed of a nonrandomly selected group, which is largely made up of competitive programmers." Even so, this exercise is illustrative of the possibility for systematic measurement error raised earlier. (For a very recent example of frontier LLM classifiers systematically failing to replicate expert human judgment, see Su et al. [2026].)

*Email: `jacob_carlson@g.harvard.edu`

However, we do have access to some actual labels—some real ground truth—right? They just come from the benchmark sample. The question is then: might we be able to use information from this benchmark dataset to correct for the estimation-relevant consequences of measurement error in conversation classifiers? The intuition for how we might do this is somewhat straightforward: we can get some sense of how systematically wrong the classifier is on the benchmark dataset, and perhaps use that knowledge of wrongness to adjust our estimator based on the target conversation dataset, so long as we take into account that the types of errors made in the benchmark dataset may not be totally representative of those in the conversation dataset due to *distribution shift*. The purpose of this note will be to show that we can do exactly this wrt almost any downstream estimation goal of economic interest (for certain benchmark datasets, under certain assumptions).

Before moving on to specifics, it's worth commenting on the fact that a natural alternative idea here would be to subsample from the AI conversation dataset itself, and invoke some expensive process to label that subsample with ground truth values—which some literatures would call a *validation dataset*. That way, we ensure the bias we are measuring is representative in the way we want. Indeed, when one can do this, there are many statistical benefits to using a validation sample to correct for measurement error in downstream estimation [Angelopoulos et al., 2023, Egami et al., 2024, Carlson and Dell, 2025]. However, specifically in this setting where *privacy-preservation* is required, and no human should be looking at conversations, this route may not be feasible. Moreover, we might simply find the ground truth signal in a given benchmark to be most meaningful to match, e.g., humans carefully guessing how long it would take to complete a task versus actual task completion timestamps. That said, to the extent that automated, privacy-preserving means exist to collect meaningful ground truth for a subset of the conversation dataset, such a strategy could be preferred.

2 Warm-up

Let's add some notation here. Say that we have a dataset of unique AI users (for simplicity) indexed by i and their conversations, X_i, at a given point in time. Say these users are sampled iid. In principle, there exists some most meaningful ground truth measurement Y_i^* for each i, e.g., the "experienced" human task completion time associated with the task described in X_i. The problem is that we don't get to see Y_i^* for any of the individuals in the AI conversation dataset. This is a *missing data* problem.

However, we do have access to a benchmark dataset with some ground truth, i.e., we have some (Y_i^*, X_i) pairs, just drawn from a different distribution. Let's join our conversation and benchmark datasets, and for record keeping purposes we'll introduce the random variable D_i, which indicates whether that observation i is from the target AI conversation dataset of interest ($D_i = 1$) or from the benchmark dataset ($D_i = 0$). With this combined dataset convention, we can write the estimand (parameter) of interest—say, the average human task completion time in the target AI user population—as

$$\theta := E[Y^* \mid D = 1].$$

Recall we only get to see Y^* when $D = 0$, under this bookkeeping convention. To formally encode this missingness, we write that we get to see the random variable Y_i for *every* i in

the combined dataset, where

$$Y_i = (1 - D_i)Y_i^*.$$

That is, when $D_i = 0$ the $Y_i = Y_i^*$, but otherwise when $D_i = 1$ the $Y_i = 0$, where zero is just a convention for a missing value. (Note we can always differentiate a meaningful zero from a missingness zero by looking at D_i.)

As such, θ is not "identified" under our missing dataset setup (i.e., it is not computable using our observed data, even if we had as much of it as we want) *unless* we make an assumption that looks like the following:

$$[Y^* \perp\!\!\!\perp D] \mid X.$$

This resembles an "unconfoundedness" assumption in the causal inference literature (another quintessential missing data problem) but can also be seen substantively as a *covariate shift* assumption in the present context, which is that for all possible $x \in \mathcal{X}$

$$Y^* \mid (X = x, D = 1) \stackrel{L}{=} Y^* \mid (X = x, D = 0),$$

where $\stackrel{L}{=}$ denotes equality in law (distribution).[1] That is, the conditional distribution of $Y^* \mid X$ is "stable" across the AI conversation and benchmark datasets, though the marginal distributions of X are able to vary. Concretely, for our running example, this would mean that conditional on the coding task, the distribution of human completion times is the same in the conversation and benchmark datasets, but the marginal distribution of coding tasks may differ across the conversation and benchmark datasets.

If we can reasonably assume this condition, then we can leverage the stability of the relationship $Y^* \mid X$ across the conversation and benchmark distributions in order to debias our estimates of θ wrt systematic measurement error in the conversation classifier. Doing this well will come down to how well we can effectively model the covariate shift, the change in the marginal distribution of X—though, fortunately, in practice, this will just mean we need to build one more classifier, this time of whether a given conversation/task belongs in the conversation or benchmark dataset. Finally, note that what we can debias effectively then is a Y^* that is conditionally stable, i.e., here, completion times for competition coders, given the nature of our benchmark dataset.

3 Framework

Now that we've established some basic notation and intuition for the strategy of handling systematic measurement error, we turn to an overview of a fully general and rigorous framework for accomplishing our goals.

In particular, we will adapt the framework of Chen et al. [2008], in the spirit of Carlson and Dell [2025]. Specifically, we can look to the "verify-out-of-sample" results of Chen et al. [2008]. These results will actually allow us to correct for measurement error based on privacy-preserving classifier predictions used to estimate a parameter defined by *any* GMM

[1] This is an if and only if relationship (see Proposition 1 in the appendix).

identification strategy (so almost anything of interest: means, linear regressions/models, structural models, and more).

To start, we use the same essential setup as Chen et al. [2008], generalizing what we did in our warm-up. We start with an iid sample of data $\{(X_i, Y_i, Z_i, D_i)\}_{i=1}^n$, where X_i are conversations (say, that contain a coding task within them), Z_i are other covariates for the observation (say, other structured metadata about the conversation), and D_i keeps track of where the observation came from: $D_i = 1$ if the conversation came from the target dataset, and $D_i = 0$ if it came from a benchmark dataset. We again define $Y_i = (1 - D_i)Y_i^*$. To simplify some notation later on, we'll also introduce the sets $\mathcal{C} := \{i : D_i = 1\}$ and $\mathcal{B} := \{i : D_i = 0\}$, the indices of observations in the target conversation and benchmark datasets, respectively.

Say we are interested in estimating a parameter $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ given by the moment condition

$$E[m((Y^*, Z); \vartheta) \mid D = 1] = 0 \quad \text{iff} \quad \vartheta = \theta$$

where $m(\cdot; \vartheta)$ is a set of functions with dimension $d_m \geq d_\theta$.

Going forward, let's label $\tilde{X} := (X, Z)$. Then the more general assumption we make to identify θ from observed data is just

$$[Y^* \perp\!\!\!\perp D] \mid \tilde{X}.$$

Under this assumption, by the law of total expectation

$$\begin{aligned} E[m((Y^*, Z); \vartheta) \mid D = 1] &= E[E[m((Y^*, Z); \vartheta) \mid D = 1, (X, Z)] \mid D = 1] \\ &= E[E[m((Y^*, Z); \vartheta) \mid D = 0, \tilde{X}] \mid D = 1] \\ &= E[E[m((Y, Z); \vartheta) \mid D = 0, \tilde{X}] \mid D = 1], \end{aligned}$$

which is a moment condition solely in terms of the observed data. We also need natural positivity assumptions here so that this expectation exists, sufficiently that

$$p := P(D = 1) \in (0, 1), \quad p(\tilde{X}) := P(D = 1 \mid \tilde{X}) \in (0, 1) \text{ a.s.}$$

Because $\tilde{X}$ is plausibly quite high-dimensional (it is unstructured conversation data), this condition—sometimes called "overlap"—may plausibly fail to hold [D'Amour et al., 2021], or fail in practice (causing high variance/finite sample instability). That said, the condition suggests that we want to *choose* a benchmark dataset that has high covariate overlap with our target conversation dataset, all else equal.[2] Otherwise, we want to deploy strategies at our disposal (mostly borrowed from the causal inference literature) to ameliorate this problem, e.g., change the parameter to one that is conditioned on a high(er) overlap region of the covariate space; "trim" the $p(\tilde{X})$; or invoke representation learning strategies to lower the dimensionality of $\tilde{X}$ without introducing more bias [Clivio et al., 2026].

We know from Chen et al. [2008] what the "efficient influence function" (EIF) for any θ in this framework is:

$$\Psi(Y, \tilde{X}, D; \theta) = -\mathcal{J}_\theta^{-1} \times F_\theta(Y, \tilde{X}, D)$$

[2]Technically, a minimal viable condition here is just that $P_{\tilde{X}|D=1} \ll P_{\tilde{X}|D=0}$. That is, we should choose a benchmark that *covers* the target conversations, but does not necessarily need to *match* them.

where

$$F_\theta(Y, \tilde{X}, D) = \frac{1-D}{p}\frac{p(\tilde{X})}{1-p(\tilde{X})}[m((Y,Z);\theta) - \mathcal{E}(\tilde{X};\theta)] + \frac{\mathcal{E}(\tilde{X};\theta)}{p}D$$

for

$$\mathcal{E}(\tilde{X};\theta) = E[m((Y^*,Z);\theta) \mid \tilde{X}], \quad \mathcal{J}_\theta = \frac{\partial}{\partial\theta}E[m((Y^*,Z);\theta) \mid D=1].$$

What is this EIF thing, and why do we care about it? We care about the EIF in this setting primarily because it can be used as a "doubly robust," Neyman orthogonal estimating equation [Chen et al., 2026], which is more permissive of using black-box AI/ML models to estimate "nuisance" functions, i.e., the $p(\tilde{X}), \mathcal{E}(\tilde{X};\theta)$ here.[3] Concretely, it is also true that

$$E[\Psi(Y, \tilde{X}, D; \vartheta)] = 0 \quad \text{iff} \quad \vartheta = \theta,$$

and so Ψ can play the same role as m, but, when used as an estimating equation, yields a more robust estimator. Moreover, estimators derived from EIF-based estimating equations will typically have the lowest possible variance of any nice, "regular" estimator of our parameter of interest. As such, there is a very real sense that, if we want to use benchmark data to debias our estimator of a quantity of interest θ, we can do no better from a robustness and efficiency standpoint than an estimator derived using the above framework.

4 Revisiting Johnston et al. [2026]

With a general framework in place, let's go back to the simple motivating example, estimating the mean human completion time in some period of AI use. Then the relevant moment condition is simply

$$m(Y^*;\vartheta) = Y^* - \vartheta$$

which implies that

$$\mathcal{E}(X;\theta) = E[Y^* \mid X] - \theta = E[Y \mid D=0, X] - \theta, \quad \mathcal{J}_\theta = -1,$$

as here $\tilde{X} = X$. Thus we can compute that in this setting

$$\begin{aligned}
F_\theta(Y, \tilde{X}, D) &= \frac{1-D}{p}\frac{p(X)}{1-p(X)}[m(Y;\theta) - \mathcal{E}(X;\theta)] + \frac{\mathcal{E}(X;\theta)}{p}D \\
&= \frac{1-D}{p}\frac{p(X)}{1-p(X)}[(Y-\theta) - (E[Y \mid D=0, X] - \theta)] + \frac{E[Y \mid D=0, X] - \theta}{p}D \\
&= \frac{1-D}{p}\frac{p(X)}{1-p(X)}(Y - E[Y \mid D=0, X]) + \frac{E[Y \mid D=0, X] - \theta}{p}D
\end{aligned}$$

and so

$$\Psi(Y, \tilde{X}, D; \theta) = \frac{1-D}{p}\frac{p(X)}{1-p(X)}(Y - E[Y \mid D=0, X]) + \frac{D}{p}E[Y \mid D=0, X] - \frac{D}{p}\theta.$$

[3] These are "nuisance" functions in the sense that we don't care about learning them for their own sake, per se, but just towards the ends of learning θ.

Using the EIF as an estimating equation immediately suggests the estimator (based on a few lines of algebra only):

$$\hat{\theta} = \left(\frac{1}{n} \sum_i \frac{D_i}{\hat{p}} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i}{\hat{p}} \texttt{Cls}(X_i) + \frac{1 - D_i}{\hat{p}} \frac{\hat{p}(X_i)}{1 - \hat{p}(X_i)} (Y_i - \texttt{Cls}(X_i)) \right),$$

where $\hat{p} = \frac{1}{n} \sum_i D_i$, and $\texttt{Cls}(X_i) = \hat{E}[Y \mid D = 0, X_i]$ and $\hat{p}(X_i) = \hat{P}(D = 1 \mid X_i)$, our best feasible approximations of $E[Y \mid D = 0, X_i], P(D = 1 \mid X_i)$, respectively. Simplifying,

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i}{\hat{p}} \texttt{Cls}(X_i) + \frac{1 - D_i}{\hat{p}} \frac{\hat{p}(X_i)}{1 - \hat{p}(X_i)} (Y_i - \texttt{Cls}(X_i)) \right)$$

or equivalently, writing out terms as averages from the conversation and benchmark datasets, respectively,

$$\hat{\theta} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \texttt{Cls}(X_i) + \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{|\mathcal{B}|}{|\mathcal{C}|} \frac{\hat{p}\,(X_i)}{1 - \hat{p}\,(X_i)} \left\{ Y_i - \texttt{Cls}(X_i) \right\}.$$

Notice that the role of $\texttt{Cls}(x)$ is usually played by the privacy-preserving classifier, which in the status quo is implemented with a frontier LLM; the efficiency theory here tells us what the classifier should approximate is a particular conditional expectation function, $x \mapsto E[Y^* \mid X = x]$. The $\hat{p}(x)$ is another classifier we should build—a "discriminating" binary classifier of the probability a conversation/task came from the target conversation data versus the benchmark data (perhaps approximated using the softmax distribution over the final token of a calibrated LLM, or some small neural network trained on an independently held-out dataset). That is, there is *one* additional prediction problem we must do for appropriate handling of measurement error, which we can still plausibly do in privacy-preserving way. Further notice that the issue of poor "overlap" directly manifests in the properties of the weights $\frac{\hat{p}(x)}{1-\hat{p}(x)}$.

For comparison, the (possibly biased) estimator implicitly being used in the status quo, which we might call $\tilde{\theta}$, is just the first term in $\hat{\theta}$, the

$$\tilde{\theta} := \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \texttt{Cls}(X_i).$$

The work above informs us of the appropriate debiasing correction for $\tilde{\theta}$ using some benchmark data, assuming that we can model well the probability a given conversation/task comes from the benchmark or comes from the AI usage dataset, the $\hat{p}(x)$.

Recall, this machinery is general to any θ of economic interest defined by a GMM identification condition, though we have fully worked out the case for a simple mean estimation question.

For the sake of this note, we will not derive the large-sample properties of this estimator, but doing something formal about inference would be the natural next step in the project. That said, this is almost certainly the actual estimator one would implement, given how it was derived—there are simply some additional questions about how well one needs to approximate $E[Y^* \mid X = x]$ and $p(x)$ that formal asymptotic analysis would clarify. If we knew $E[Y^* \mid X = x]$ and $p(x)$ with certainty, the inference results here would be immediate.

5 More Intuition

For more intuition, we can observe that the estimator $\hat{\theta}$ also naturally corresponds to a sensible estimator that measures the systematic mistakes in the benchmark predictions and then applies that as a bias correction to predictions on the target conversation dataset, appropriately adjusting for covariate shift across benchmark and target with a density ratio (Radon-Nikodym derivative).

To see this, we start with another identification for a θ which is the expected completion time (generically, a mean of missing data). In particular, say P is the target conversation distribution of interest, and Q is the distribution of the benchmark data. We make a covariate shift assumption as before, now denoting it as $P_{Y|X} = Q_{Y|X}$, but again where P_X and Q_X are able to differ. Then we have that

$$\theta := E_P[Y^*] = \underbrace{E_P[\texttt{Cls}(X)]}_{\text{Expected classifier prediction}} + \underbrace{E_Q\left[w(X)(Y^* - \texttt{Cls}(X))\right]}_{\text{Bias in benchmark, reweighted for cov. shift}}$$

where $w(x) = \frac{dP_X}{dQ_X}(x)$, the Radon-Nikodym derivative.

The natural plug-in estimator implied here, call it $\check{\theta}$, is just (assuming we knew the density ratio and p momentarily)

$$\check{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{p}\texttt{Cls}(X_i) + \frac{1-D_i}{1-p}w(X_i)(Y_i^* - \texttt{Cls}(X_i))\right).$$

By Bayes, however,

$$w(x) = \frac{dP_X}{dQ_X}(x) = \frac{f(x \mid D=1)}{f(x \mid D=0)} = \frac{P(D=1 \mid x)f(x)/p}{P(D=0 \mid x)f(x)/(1-p)} = \frac{p(x)(1-p)}{(1-p(x))p}$$

and so if we estimate p with $\hat{p} = n^{-1}\sum_i D_i$ and $p(x)$ with $\hat{p}(x)$

$$\check{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\hat{p}}\texttt{Cls}(X_i) + \frac{1-D_i}{\hat{p}}\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}(Y_i^* - \texttt{Cls}(X_i))\right).$$

Notice that $\hat{\theta} = \check{\theta}$; it is equivalent to the estimator we found earlier with the verify-out-of-sample approach. Thus the intuition that allowed us to build $\check{\theta}$ is the right intuition for $\hat{\theta}$ and the verify-out-of-sample approach too.

6 Takeaways

What are the overall takeaways from this note?

1. To estimate something like the average complexity of tasks for some target AI user population, researchers are currently using a privacy-preserving automated classifier in an estimator that looks like:

$$\tilde{\theta} := \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\texttt{Cls}(X_i).$$

This estimator may suffer from bias due to systematic measurement error in `Cls`. To purge this bias, we could leverage a benchmark dataset, so long as it only differs from the target population of interest by a covariate shift (change in marginal distribution of the conversations being had); a highly general framework based on this assumption would tell us to use the debiased (and efficient) estimator:

$$\hat{\theta} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \texttt{Cls}(X_i) + \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{|\mathcal{B}|}{|\mathcal{C}|} \frac{\hat{p}\left(X_i\right)}{1 - \hat{p}\left(X_i\right)} \left\{Y_i - \texttt{Cls}(X_i)\right\}.$$

Naturally, if there is no measurement error, $Y_i - \texttt{Cls}(X_i) = 0$ almost surely, and we recover the unadjusted estimator.

2. The theory that underlies this general framework would make it easy to debias any θ of interest fairly straightforwardly, be that a θ from a linear regression, structural model, or the like.

3. The theory underlying the framework reveals that we should ideally be choosing a benchmark dataset that has conversations that are as similar as possible (in a distributional sense) to the ones in the target population.

4. The framework tells us we only need to do one more piece of engineering to debias our estimates: fit a discriminating classifier of whether an observation came from the target or benchmark dataset. Plausibly, this can also be done in a privacy-preserving way.

5. The benchmark outcome (the "ground truth" it contains) naturally dictates which attribute of the target conversation dataset we can most credibly leverage for estimation, e.g., if the outcome in the benchmark is competition coder completion times, that's the outcome we can most credibly debias wrt (as opposed to completion times for average-skill coders), and the choice of privacy-preserving classifier prompt should reflect this. By the same logic, then, in the setting of Johnston et al. [2026], curating a benchmark of real-world reminiscent tasks and timing their completion by non-competitive coders would be preferred.

6. If we could annotate at least some of our AI conversations of interest with ground truth in privacy-preserving way, that would be the most promising thing to do. However, the present approach provides a naturally privacy-preserving alternative, involving no human looking at user conversations.

7 Appendix

Proposition 1. *Assume that* $P(D = 1) \in (0, 1)$, $P(D = 1 \mid X) \in (0, 1)$ *a.s. Then* $[Y^* \perp\!\!\!\perp D] \mid X$ *iff* $Y^* \mid (X, D = 1) \stackrel{L}{=} Y^* \mid (X, D = 0)$.

Proof. Our goal is to show that almost surely

$$[Y^* \perp\!\!\!\perp D] \mid X \iff Y^* \mid (X, D = 1) \stackrel{L}{=} Y^* \mid (X, D = 0).$$

The first direction then is

$$[Y^* \perp\!\!\!\perp D] \mid X \implies Y^* \mid (X, D = 1) \stackrel{L}{=} Y^* \mid (X, D = 0).$$

To show this, it is immediate that assuming $[Y^* \perp\!\!\!\perp D] \mid X$ implies

$$P(Y^* = y \mid X, D = 1) = P(Y^* = y \mid X) = P(Y^* = y \mid X, D = 0).$$

To show the other direction,

$$[Y^* \perp\!\!\!\perp D] \mid X \impliedby Y^* \mid (X, D = 1) \stackrel{L}{=} Y^* \mid (X, D = 0)$$

observe that if $Y^* \mid (X, D = 1) \stackrel{L}{=} Y^* \mid (X, D = 0)$, then

$$P(Y^* = y, D = d \mid X) = P(Y^* = y \mid X, D = d)P(D = d \mid X)$$

and then that by the law of total probability

$$P(Y^* = y \mid X) = P(Y^* = y \mid X, D = 1)P(D = 1 \mid X) + P(Y^* = y \mid X, D = 0)P(D = 0 \mid X).$$

But under the covariate shift assumption, then for all $d \in \{0, 1\}$

$$\begin{aligned} P(Y^* = y \mid X) &= P(Y^* = y \mid X, D = 1)P(D = 1 \mid X) + P(Y^* = y \mid X, D = 0)P(D = 0 \mid X) \\ &= P(Y^* = y \mid X, D = d)P(D = 1 \mid X) + P(Y^* = y \mid X, D = d)P(D = 0 \mid X) \\ &= P(Y^* = y \mid X, D = d)(P(D = 1 \mid X) + P(D = 0 \mid X)) \\ &= P(Y^* = y \mid X, D = d) \end{aligned}$$

and so

$$\begin{aligned} P(Y^* = y, D = d \mid X) &= P(Y^* = y \mid X, D = d)P(D = d \mid X) \\ &= P(Y^* = y \mid X)P(D = d \mid X) \end{aligned}$$

which is the definition of conditional independence, $Y^* \perp\!\!\!\perp D \mid X$. □

References

Drew Johnston, David Holtz, Alex Martin Richmond, Christopher Ong, Prasanna Tambe, and Aaron Chatterji. The Shift to Agentic AI: Evidence from Codex, June 2026. URL https://arxiv.org/abs/2606.26959. Version Number: 1.

Sarah Su, Kevin Zhu, Emily Xiao, Rohan Alur, and Daniel Kang. Learning to replicate expert judgment in financial tasks. *Thinking Machines Lab: News*, 2026. https://thinkingmachines.ai/news/learning-to-replicate-expert-judgment-in-financial-tasks/.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, November 2023. doi: 10.1126/science.adi6000. URL https://www.science.org/doi/10.1126/science.adi6000.

Naoki Egami, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses. June 2024. URL https://naokiegami.com/paper/dsl_ss.pdf.

Jacob Carlson and Melissa Dell. A Unifying Framework for Robust and Efficient Inference with Unstructured Data, 2025. URL https://arxiv.org/abs/2505.00282. Version Number: 2.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2), April 2008. ISSN 0090-5364. doi: 10.1214/009053607000000947. URL https://projecteuclid.org/journals/annals-of-statistics/volume-36/issue-2/Semiparametric-efficiency-in-GMM-models-with-auxiliary-data/10.1214/009053607000000947.full.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654, April 2021. ISSN 0304-4076. doi: 10.1016/j.jeconom.2019.10.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0304407620302694.

Oscar Clivio, Alexander D'Amour, Alexander Franks, David Bruns-Smith, Chris Holmes, and Avi Feller. Deconfounding Scores and Representation Learning for Causal Effect Estimation with Weak Overlap, 2026. URL https://arxiv.org/abs/2604.00811. Version Number: 1.

Yuxi Chen, Edward H. Kennedy, and Sivaraman Balakrishnan. On the Equivalence between Neyman Orthogonality and Pathwise Differentiability, 2026. URL https://arxiv.org/abs/2603.15817. Version Number: 1.