Supplementary Materials

Using Imperfect Surrogates for Downstream Inference:
Design-based Supervised Learning for
Social Science Applications of Large Language Models

All data and source code for replication are available at https://osf.io/gjt87

Contents

1 Proof of Theoretical Results

We first provide a proof for a method of moment estimator (Proposition 2) and then apply it to a logistic regression problem (Proposition 1).

1.1 Method of Moment Estimator (Proposition 2)

Suppose researchers are interested in a method of moment estimator with a moment function $m(Y, Q, W, X; \beta, g)$ where (Y, Q, W, X) are the data, β are parameters of interest, and g is the supervised machine learning function. Then, the estimand of interest β^*_M can be written as the solution to the following moment equations.

$$\mathbb{E}(m(Y, Q, W, X; \beta, g^*)) = 0, \tag{1}$$

where g^* is the true conditional expectation $\mathbb{E}(Y \mid Q, W, X)$.

We define the moment function to be *design-based* when the moment function is insensitive to the first step machine learning function. That is, $\mathbb{E}(m(Y, Q, W, X; \beta, g)) = \mathbb{E}(m(Y, Q, W, X; \beta, g'))$ for any β and any machine learning functions g and g' that do not diverge.

In this general setup, the DSL estimator $\widehat{\beta}_M$ is a solution to the following moment equation.

$$\sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} m(Y_i, Q_i, W_i, X_i; \beta, \widehat{g}_k) = 0. \tag{2}$$

where we employ a K-fold cross-fitting procedure (Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins, 2018). We first partition the observation indices $i = 1, \dots, n$ into K groups $\mathcal{D}_k$ where $k = 1, \dots, K$. We then learn the supervised machine learning model $\widehat{g}_k$ by predicting Y using (Q, W, X) using all hand-coded documents *not* in $\mathcal{D}_k$.

Proposition 2 *Under Assumption 1, when the DSL estimator with a design-based moment is fitted with the cross-fitting approach, $\widehat{\beta}_M$ is consistent and asymptotically normal as sample size n goes to infinity.*

$$\sqrt{n}(\widehat{\beta}_M - \beta^*_M) \xrightarrow{d} \mathcal{N}(0, V_M). \tag{3}$$

where

$$V_M = \mathbb{E}\left(\frac{\partial m(L; \beta^*, \overline{g})}{\partial \beta}\right)^{-1} \mathbb{E}(m(L; \beta^*, \overline{g}) m(L; \beta^*, \overline{g})^\top) \mathbb{E}\left(\frac{\partial m(L; \beta^*, \overline{g})}{\partial \beta}\right)^{-1}.$$

Here we define $\overline{g}$ to be the probability limit of the estimated supervised machine learning function $\widehat{g}_k$ in the sense that $||\widehat{g}_k - \overline{g}||_2 = o_p(1)$ and $\mathbb{E}_k(||m(L; \beta^*, \widehat{g}_k) - m(L; \beta^*, \overline{g})||_2^2) = o_p(1)$. This probability limit does not need to be equal to the true conditional expectation g^*. Thus, we do not assume the correct specification of the estimated supervised machine learning function.

Note that this asymptotic regime is based on sample size n, and under Assumption 1, the probability of hand-coding is bounded away from zero, so the number of hand-coded documents n_R also goes to infinity. Despite this asymptotic regime, we will show strong finite-sample performance of the proposed DSL estimator with relatively small n_R (in the order of $50 \sim 200$) in Section 5 of the main paper and Section 2 of this supplement.

Proof. For notational simplicity, we use $L_i := (Y_i, Q_i, W_i, X_i)$ to denote observed data for document i. We also use $\widehat{\beta}$ and β^* to denote $\widehat{\beta}_M$ and β^*_M.

Using the mean value theorem, we can expand the moment equation around β^*_M and obtain

$$\frac{1}{n}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} m(L_i; \widehat{\beta}, \widehat{g}_k) \quad = \quad \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} m(L_i; \beta^*, \widehat{g}_k) + (\widehat{\beta} - \beta^*)\frac{1}{n}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} \frac{\partial m(L_i; \widetilde{\beta}, \widehat{g}_k)}{\partial \beta}$$

where $\widetilde{\beta}$ is a mean value, located between $\widehat{\beta}$ and β^*. Thus,

$$\sqrt{n}(\widehat{\beta} - \beta^*) \quad = \quad \underbrace{\left(-\frac{1}{n}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} \frac{\partial m(L_i; \widetilde{\beta}, \widehat{g}_k)}{\partial \beta}\right)^{-1}}_{(a)} \times \underbrace{\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} m(L_i; \beta^*, \widehat{g}_k).}_{(b)}$$

We will consider terms (a) and (b) in order.

We begin with the main term (b), which can be decomposed into three terms.

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} m(L_i; \beta^*, \widehat{g}_k) \quad = \quad R_1 + R_2 + R_3$$

where

$$\begin{aligned}
R_1 &= \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} (m(L_i; \beta^*, \widehat{g}_k) - \mathbb{E}_k(m(L; \beta^*, \widehat{g}_k)) - (m(L_i; \beta^*, \overline{g}) - \mathbb{E}_k(m(L; \beta^*, \overline{g})) \\
R_2 &= \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i \in \mathcal{D}_k} (m(L_i; \beta^*, \overline{g}) - \mathbb{E}_k(m(L; \beta^*, \overline{g})) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(m(L_i; \beta^*, \overline{g}) - \mathbb{E}(m(L; \beta^*, \overline{g})) \\
R_3 &= \frac{1}{\sqrt{n}}\sum_{k=1}^{K} n_k \times \mathbb{E}_k(m(L; \beta^*, \widehat{g}_k)).
\end{aligned}$$

Here we use $\mathbb{E}_k$ to denote the expectation over $\mathcal{D}_k$, which is independent of data used to learn $\widehat{g}_k$ in cross-fitting. Remember that we define $\overline{g}$ to be the probability limit of the estimated supervised machine learning function $\widehat{g}_k$ in a sense that $||\widehat{g}_k - \overline{g}||_2 = o_p(1)$ and $\mathbb{E}_k(||m(L; \beta^*, \widehat{g}_k) - m(L; \beta^*, \overline{g})||_2^2) = o_p(1)$. This probability limit does not need to be equal to the true conditional expectation g^*. Thus, we do not assume the correct specification of the estimated supervised machine learning function.

R_1 is known as the empirical process term. Given that we use cross-fitting and $\mathbb{E}_k(||m(L; \beta^*, \widehat{g}_k) - m(L; \beta^*, \overline{g})||_2^2) = o_p(1)$, we obtain $R_1 = o_p(1)$ by Lemma 1 of Kennedy, Balakrishnan and G'Sell (2020).

As for R_2, we can use the central limit theorem to show that

$$R_2 \xrightarrow{d} \mathcal{N}(0, \mathbb{E}(m(L; \beta^*, \overline{g})m(L; \beta^*, \overline{g})^\top)) \tag{4}$$

because $\mathbb{E}(m(L; \beta^*, \overline{g})) = 0$ for the design-based moment.

Finally, as for R_3, $\mathbb{E}_k(m(L; \beta^*, \widehat{g}_k)) = 0$ for the design-based moment, and thus, $R_3 = 0$.

Taken together, for the design-based moment, we obtain

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k} m(L_i;\beta^*,\widehat{g}_k) \quad \xrightarrow{d} \quad \mathcal{N}(0,\mathbb{E}(m(L;\beta^*,\overline{g})m(L;\beta^*,\overline{g})^\top)).$$

We now consider the term (a), and we need to show that

$$\left(\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}\frac{\partial m(L_i;\widetilde{\beta},\widehat{g}_k)}{\partial\beta}\right)^{-1} \xrightarrow{p} \mathbb{E}\left(\frac{\partial m(L_i;\beta^*,\overline{g})}{\partial\beta}\right)^{-1} \tag{5}$$

We require the standard regularity condition for GMM that requires the smoothness of the derivative of the moment, which holds true for most common method of moment estimators like the estimation of class prevalence, linear regression, and logistic regression problems.

Assumption 5 from Chernozhukov et al. (2022). $\mathbb{E}\left(\partial m(L_i;\beta^*,\overline{g})/\partial\beta\right)$ *exists and there is a neighborhood* $\mathcal{N}_\beta$ *of* β^* *such that: (i) for each* k, $||\widehat{g}_k-\overline{g}||_2 = o_p(1)$*; (ii) for all* $||g-\overline{g}||_2$ *small enough,* $m(L;\beta,g)$ *is differentiable in* β *on* $\mathcal{N}_\beta$ *with probability approaching one, and there are* $C>0$ *and* $d(L;g)$ *such that, for* $\beta\in\mathcal{N}_\beta$ *and* $||g-\overline{g}||_2$ *small enough,*

$$\left|\left|\frac{\partial m(L;\beta,g)}{\partial\beta}-\frac{\partial m(L;\beta^*,g)}{\partial\beta}\right|\right|_2 \leq d(L,g)||\beta-\beta^*||_2^{1/C};\quad \mathbb{E}(d(L,g))<C.$$

(iii) For each k *and* p *and* q, $\mathbb{E}(\partial m_p(L;\beta^*,\widehat{g}_k)/\partial\beta_q - \partial m_p(L;\beta^*,\overline{g})/\partial\beta_q) = o_p(1)$.

These regularity conditions are standard (Newey and McFadden, 1994). Among this regularity condition, the main requirement is that, for each k, $||\widehat{g}_k-\overline{g}||_2 = o_p(1)$. However, we define $\overline{g}$ to be the probability limit of $\widehat{g}_k$, and thus, this automatically holds. Therefore, under this assumption and $\widehat{\beta}-\beta^* = o_p(1)$, we obtain equation (5).

Combining terms (a) and (b), we have

$$\sqrt{n}(\widehat{\beta}-\beta^*) \quad \xrightarrow{d} \quad \mathcal{N}(0,V_M)$$

where

$$V_M = \mathbb{E}\left(\frac{\partial m(L;\beta^*,\overline{g})}{\partial\beta}\right)^{-1}\mathbb{E}(m(L;\beta^*,\overline{g})m(L;\beta^*,\overline{g})^\top)\mathbb{E}\left(\frac{\partial m(L;\beta^*,\overline{g})}{\partial\beta}\right)^{-1}, \tag{6}$$

which completes the proof. □

1.2 Logistic Regression (Proposition 1)

The logistic regression (Proposition 1) is a special case of Proposition 2.

For the logistic regression, the estimand β^* is defined as a solution to the following moment equation.

$$\mathbb{E}\{(Y-\text{expit}(X^\top\beta^*))X\} = 0, \tag{7}$$

where expit() is the inverse of the logit function. And, the DSL estimator uses the following moment equation.

$$m_{DSL}(L;\beta,g) \coloneqq \left(\left(\frac{R}{\pi(Q,W,X)}(Y-g(Q,W,X))+g(Q,W,X)\right)-\text{expit}(X^\top\beta)\right)X \tag{8}$$

To use Proposition 2, we just need to verify that $m(L; \beta, g)$ is a design-based moment.

$$\begin{aligned}
& \mathbb{E}\left\{\left(\left(\frac{R}{\pi(Q,W,X)}(Y - g(Q,W,X)) + g(Q,W,X)\right) - \operatorname{expit}(X^\top \beta)\right) X\right\} \\
= \;& \mathbb{E}\left\{\mathbb{E}\left\{\left(\left(\frac{R}{\pi(Q,W,X)}(Y - g(Q,W,X)) + g(Q,W,X)\right) - \operatorname{expit}(X^\top \beta)\right) X \,\middle|\, Q,W,X\right\}\right\} \\
= \;& \mathbb{E}\left\{\left(\mathbb{E}\left(\frac{R}{\pi(Q,W,X)}(Y - g(Q,W,X)) + g(Q,W,X)\,\middle|\, Q,W,X\right) - \operatorname{expit}(X^\top \beta)\right) X\right\}
\end{aligned}$$

where the first equality follows from the rule of total expectaton and the second from the rearrangement of terms.

Importantly, we have

$$\begin{aligned}
& \mathbb{E}\left(\frac{R}{\pi(Q,W,X)}(Y - g(Q,W,X)) + g(Q,W,X)\,\middle|\, Q,W,X\right) \\
= \;& \frac{\mathbb{E}\left(RY \mid Q,W,X\right)}{\pi(Q,W,X)} + \left(1 - \frac{\mathbb{E}(R \mid Q,W,X)}{\pi(Q,W,X)}\right) g(Q,W,X) \\
= \;& \frac{\mathbb{E}\left(R \mid Q,W,X\right)\mathbb{E}\left(Y \mid Q,W,X\right)}{\pi(Q,W,X)} + \left(1 - \frac{\mathbb{E}(R \mid Q,W,X)}{\pi(Q,W,X)}\right) g(Q,W,X) \\
= \;& \frac{\pi(Q,W,X)\mathbb{E}\left(Y \mid Q,W,X\right)}{\pi(Q,W,X)} + \left(1 - \frac{\pi(Q,W,X)}{\pi(Q,W,X)}\right) g(Q,W,X) \\
= \;& \mathbb{E}\left(Y \mid Q,W,X\right)
\end{aligned}$$

where the first equality follows from the rearrangement of terms. The second equality follows because $\mathbb{E}\left(RY \mid Q,W,X\right) = \mathbb{E}\left(R \mid Q,W,X\right)\mathbb{E}\left(Y \mid Q,W,X\right)$ based on Assumption 1, and the third from $\mathbb{E}\left(R \mid Q,W,X\right) = \pi(Q,W,X)$ by definition. Importantly, this equality does not require any assumption about the supervised machine learning method $g(Q,W,X)$.

Therefore,

$$\mathbb{E}\left(m_{DSL}(L; \beta, g)\right) = \mathbb{E}\left\{\left(\mathbb{E}\left(Y \mid Q,W,X\right) - \operatorname{expit}(X^\top \beta)\right) X\right\}.$$

This implies that $m_{DSL}(L; \beta, g)$ is a design-based moment, i.e., $\mathbb{E}\left(m_{DSL}(L; \beta, g)\right) = \mathbb{E}\left(m_{DSL}(L; \beta, g')\right)$ for any β and any machine learning functions g and g' that do not diverge.

Thus, using Proposition 2, under Assumption 1, when the DSL estimator is fitted with the cross-fitting approach (Algorithm 1 in the paper), estimated coefficients $\widehat{\beta}$ are consistent and asymptotically normal.

$$\sqrt{n}(\widehat{\beta} - \beta^*) \xrightarrow{d} \mathcal{N}(0, V),$$

where

$$V := \mathbb{E}\left(\frac{\partial m_{DSL}(L; \beta^*, \overline{g})}{\partial \beta}\right)^{-1} \mathbb{E}(m_{DSL}(L; \beta^*, \overline{g}) m_{DSL}(L; \beta^*, \overline{g})^\top) \mathbb{E}\left(\frac{\partial m_{DSL}(L; \beta^*, \overline{g})}{\partial \beta}\right)^{-1}.$$

Finally, we obtain the consistent estimator for variance V. For the logistic regression moment, we have

$$\mathbb{E}(m_{DSL}(L; \beta^*, \overline{g}) m_{DSL}(L; \beta^*, \overline{g})^\top) \;\; = \;\; \mathbb{E}\left((\widetilde{Y} - \operatorname{expit}(X^\top \beta))^2 X X^\top\right),$$

$$\mathbb{E}\left(\frac{\partial m_{DSL}(L;\beta,\overline{g})}{\partial \beta}\right) \;\; = \;\; \mathbb{E}\left(\mathrm{expit}(X^\top \beta)(1-\mathrm{expit}(X^\top\beta))XX^\top\right).$$

Therefore, using the standard theory of GMM, we have

$$\begin{aligned}
\widehat{\Omega} \;\; &:= \;\; \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in\mathcal{D}_k}(\widetilde{Y}_i^k - \mathrm{expit}(X_i^\top\widehat{\beta}))^2 X_i X_i^\top \xrightarrow{p} \mathbb{E}(m_{DSL}(L;\beta^*,\overline{g})m_{DSL}(L;\beta^*,\overline{g})^\top) \\
\widehat{\mathbf{M}} \;\; &:= \;\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{expit}(X_i^\top\widehat{\beta})(1-\mathrm{expit}(X_i^\top\widehat{\beta})))X_iX_i^\top \xrightarrow{p} \mathbb{E}\left(\frac{\partial m(L_i;\beta^*,\overline{g})}{\partial\beta}\right).
\end{aligned}$$

Taken together,

$$\widehat{V} := \widehat{\mathbf{M}}^{-1}\widehat{\Omega}\widehat{\mathbf{M}}^{-1},$$

which gives a consistent estimator for the variance. □

2 Experiment

Using simulations and 18 real-world datasets, we investigate the statistical properties of the proposed DSL estimator and compare them against those of existing alternatives.

2.1 Datasets

This section contains details on the 18 real-world datasets we use in our experiments in Section 5 of the main paper. We explain the choice of dataset, how it was collected, how the gold-standard labels are created, how we filter and subset the data, and how we create the variables that are used in subsequent analyses. We also explain how we generate the surrogate labels, including the prompts and model parameters we use.

2.1.1 Congressional Bills Project

For the logistic regression task in Section 5.1, we use data from the Congressional Bills Project (CBP, Adler and Wilkerson, 2006), a database of 400K public and private bills introduced in the U.S. House and Senate since 1947. This dataset of congressional bills provides human-coded labels of the topic of the proposed legislation and an array of covariates about the sponsor, and has been used widely in political science analyses of legislative behavior and lawmaking.

The topics are based on the hierarchical coding scheme created by the Comparative Agendas Project (CAP), which includes 20 major topics and 224 minor subtopics (see Table S1). Each bill is assigned a single major topic and minor subtopic. Regarding the coding process, the authors note:

> Trained human coders assign a primary topic (one of 19 major and one of 224 subtopics) to each bill based on their readings of either the short description or the title of the bill. Intercoder reliability across the 225 subtopics is very high. Intercoder disagreements can indicate coding errors, but we have found that most of them reflect legitimate disagreements about a bill's primary topic. For example, a bill proposing a health care program for children of illegal immigrants might be arguably coded as an immigration issue (530) or as a child health issue (332).[1]

We choose this dataset for our experiment because it reflects a plausible research scenario for social scientists: given a large collection of legislative documents, we want to estimate: a) how many pertain to a topic of interest, and b) how the likelihood of a document pertaining to a topic relates to the features of the legislator that proposed it.

Two versions of this dataset are available. One is from the original authors' project website (`http://congressionalbills.org/download.html`), and a second from the CAP (`https://www.comparativeagendas.net/datasets_codebooks`. We use a combination of the two datasets. The bill texts and topic labels are more standardized in the CAP version, while the authors' version has a greater number of covariates. The exact method of how these two datasets are downloaded and the additional covariates are merged can be found in the replication code in `01_download_data.py` and `03_additional_covariates.py`.

As noted in the main body of our paper, we consider a binary task in our experimental analysis: distinguishing `Macroeconomy` (the positive class) from `Law and Crime`, `Defense` and

[1] Source: `http://congressionalbills.org/about.html`. Retrieved 17 May 2023.

Code	Topic	Code	Topic
1	Macroeconomics	12	Law and Crime
2	Civil Rights	13	Social Welfare
3	Health	14	Housing
4	Agriculture	15	Domestic Commerce
5	Labor	16	Defense
6	Education	17	Technology
7	Environment	18	Foreign Trade
8	Energy	19	International Affairs
9	Immigration	20	Government Operations
10	Transportation	21	Public Lands

Table S1: Major CAP Codes

`International Affairs` (the negative class). The specific choice of these topics reflects three factors: the relative prevalence of each topic dataset after the preprocessing steps detailed below, the requirement for a diverse negative class (to reflect that in many social science analyses, we target a particular phenomenon out of a highly heterogeneous collection of events), and finally the condition that the positive class be sufficiently distinct from the negative classes to avoid an ambiguous task for the LLM. For instance, had we used `Labor`, `Domestic Commerce` or `Foreign Trade` in the negative class, then the correct answer for classifying whether a bill about wage negotiations is about the macroeconomy is ambiguous.[2]

From this task, we consider a balanced and an imbalanced condition. The imbalanced condition is particularly representative of the real-world corpora and tasks in social science applications, such as event detection in news or hate speech detection in social media posts. In the Balanced condition, there are 5K positive observations and 5K negative observations. In the Imbalanced condition, there are 1K positives and 9K negatives. All observations are drawn from the same pool of 14K observations (5K `Macroeconomy`, 3K `Law and Crime`, 3K `Defense`, 3K `International Affairs`) in order to reduce the cost of generating surrogate labels.

These 14K observations were randomly sampled (with each topic sampled separately) from the subset of the CBP dataset remaining after applying the following filtering conditions:[3]

1. All rows containing NA in the topic, bill description, chamber, party, passage, bill ID, year, or Congressional session columns were dropped.

2. Where there were duplicate bill descriptions, the first observation was retained and all subsequent ones removed.

[2] An plausible solution is to provide the LLM with all possible categories and then ask it to choose the one that best matches the text. Given that the focus of this experiment was to test the improvements from our estimator instead of designing the optimal GPT-3 prompt for CAP coding, we opted to create a task with fewer issues of ambiguity that still represented a plausible social science research scenario.

[3] These steps are implemented in `01_download_data.py` in the replication code.

2.1.2 Ziems et al. (2023) Datasets

Ziems et al. (2023) evaluate zero-shot performance of 13 LLMs on a diverse collection of CSS benchmark tasks including emotion, hate-speech, ideology and misinformation detection. We evaluate the performance of our approach versus the surrogate-only (SO) results for the 17 tasks reported in their Table 3. As noted in our paper, because many of these datasets do not have consistent covariates, we focus on a class prevalence estimation task (estimating the proportion of classes). This is representative of plausible descriptive quantities of interest for social science research questions such as measuring issue attention (Grimmer, 2010) or the prevalence of negative campaigning.

We use the datasets and classifications that are available in the replication materials for Ziems et al. (2023)[4]. The datasets are summarized in Table S2. Note that Ziems et al. (2023) only evaluate on a maximum of 500 observations drawn by stratifying on the true labels.

Dataset	Obs.	Surr.	Cls.	Citation
Dialect	266	13	23	Demszky et al. (2019)
Tempowic	344	12	2	Pilehvar and Camacho-Collados (2019)
RAOP	399	12	7	Althoff, Danescu-Niculescu-Mizil and Jurafsky (2014); Yang et al. (2019)
Persuasion	434	13	2	Wang et al. (2019)
Semeval Stance	435	12	3	Mohammad et al. (2016)
Discourse	497	13	7	Zhang, Culbertson and Paritosh (2017)
Politeness	498	13	3	Danescu-Niculescu-Mizil et al. (2013)
Emotion	498	11	6	Saravia et al. (2018)
Hate	498	13	6	ElSherief et al. (2021)
IBC	498	12	3	Gross et al. (2013); Iyyer et al. (2014)
Talklife	498	12	3	Sharma et al. (2020)
Media Ideology	498	13	3	Baly et al. (2020)
Humor	500	11	2	Weller and Seppi (2019)
Power	500	11	2	Danescu-Niculescu-Mizil et al. (2012)
Toxicity	500	12	2	Zhang et al. (2018)
Misinformation	500	10	2	Gabriel et al. (2022)
Figurative	500	13	4	Chakrabarty et al. (2022)

Table S2: Ziems et al. (2023) datasets used. Ordered by number of observations in dataset. Surr. indicates the number of different types of LLM labels available in the replication material. Cls. indicates the number of classes in the dataset/task.

2.1.3 Constructing Surrogates and Covariates

Surrogate labels for the CBP dataset are generated with zero-shot and five-shot classification using the `text-davinci-003` model from OpenAI (GPT-3, Brown et al., 2020). We opt for GPT-3 because it is widely available, used in several recent computational social science (CSS)

[4] `https://github.com/SALT-NLP/LLMs_for_CSS`

analyses (e.g. Ornstein, Blasingame and Truscott, 2022) and it demonstrates consistently strong performance in Ziems et al. (2023). Although existing CSS works use zero-shot classification, results in Brown et al. (2020) and subsequent analyses indicate the classifier performance should improve with in-context examples. Therefore we include both zero-shot and five-shot, with the rationale that it is a trivial amount of additional researcher labor to code five examples and include these in the prompts.

The prompts are designed based on the recommendations on the OpenAI website at the time of running the experiment. We test the performance of the prompts on a subset of 100 documents in order to ensure that the surrogate labels were of reasonable accuracy. The only change to the prompts we made as a result of this procedure was to change the word `macroeconomy` to `economy`, as the former produced an extremely high proportion of false negatives. The five exemplars in the five-shot prompt are cherry-picked from thirty randomly sampled positive and negative observations from the CBP data that were not included in the 14K randomly sampled observations used for the main analysis. We prioritized bills with longer and less generic descriptions because we hypothesized that having greater variance in the exemplars may be conducive to better performance across a wider distribution of texts.

CBP Zero-shot Prompt:

```
Does the following text relate to the economy? (True/False)

text: """
{content}
"""

label:
```

CBP Five-shot Prompt:

```
Does the following text relate to the economy? (True/False)

text: """
To provide that Federal expenditures shall not exceed Federal revenues, except in time of
    war or grave emergency declared by the Congress, and to provide for systematic
    reduction of the public debt
"""

label: True

text: """
To amend the Internal Revenue Act of 1954 to increase from $600 to $1,200 the personal
    income tax exemptions of a taxpayer (including the exemtion for a spouse, the
    exemption for a dependent, and the additional exemptions for old age and blindness)
"""

label: True
```

```
text: """
A bill to amend the Internal Revenue Code of 1954 to provide special loss carryover rules
    for insurance companies.
"""

label: False

text: """
To provide individuals with access to health information of which they are a subject, to
    ensure personal privacy, security, and confidentiality with respect to health related
    information in promoting the development of a nationwide interoperable health
    information infrastructure, to impose criminal and civil penalties for unauthorized
    use of personal health information, to provide for the strong enforcement of these
    rights, to protect States' rights, and for other purposes.
"""

label: True

text: """
A bill to amend title XVI of the Social Security Act to reduce from 21 to 18 the age at
    which a disabled child need no longer include his parents' income in determining his
    eligibility for supplemental security income benefits or the amount of such benefits.
"""

label: False

text: """
{content}
"""

label:
```

In addition to the zero- and five-shot surrogate labels, we prepare five covariates for use in our analysis. Four of these are features of the bill sponsor: `senate`, a binary variable indicating whether they are a senator; `democrat`, a binary variable indicating whether they are a Democrat; `dw1`, a continuous variable indicating the first DW-Nominate score of the sponsor; and `Postal`, a categorical variable indicating the state of the bill sponsor. The latter two variables are only available in the authors' original dataset. We mean impute missing values of `dw1`.

The fifth covariate, `dist_macro`, is a similarity score between each observation and the positive class description using the cosine distance of sentence transformer embeddings. The positive class description is constructed from the codebook description of the `Macroeconomy` topic:

> Issues related to general domestic macroeconomic policy inflation, cost of living, prices, interest rates, the unemployment rate, impact of unemployment the monetary policy, central bank, the treasury, public debt, budgeting, efforts to reduce deficits tax policy, the impact of taxes, tax enforcement manufacturing policy, industrial revitalization growth wage or price control, emergency price controls or other macroeconomics subtopics.

Embeddings are generated using the `all-mpnet-base-v2` sentence transformer model (Reimers and Gurevych, 2019; Song et al., 2020).[5]

The Ziems et al. (2023) datasets have between 10 and 13 surrogate labels available each. For our final analysis, we use the labels produced by `flan-ul2` (Tay et al., 2022). In Section 2.4 of this supplement, we also report results using the average of all available surrogates. We transform datasets and labels in Ziems et al. (2023) into binary tasks. This means that for a dataset with $c > 2$ classes, we convert them into an c datasets where in each dataset a single class is coded as positive and the remainder as negative. In each binary dataset, we use the sentence embedding cosine distance procedure described in Section 2.1.3 (below) to generate a document covariate to be used as inputs to the supervised learning (SL) and design-based supervised learning (DSL) estimators. The positive class embedding is generated from texts based on the class descriptions given in the zero-shot prompts in Ziems et al. (2023).[6]

2.2 Estimator Implementation

In this section, we detail the implementation of each of the four estimators described in Sections 3 and 4 of the main paper: Surrogate Only Estimation (SO), Gold-Standard Only Estimation (GSO), Supervised Learning (SL) and Design-based Supervised Learning (DSL). All implementations are done in the `R` programming language.

For clarity, we first restate the notation introduced in Section 2. Additional notation will be introduced as we explain the implementation of the estimators. The researcher begins with the following data:

- n documents indexed i
- $Y \in \{0, 1\}$: the outcome. This is the gold-standard label.
- $X \in \mathbb{R}^{d_X}$: the explanatory variables. These are the covariates of interest for the downstream regression model.
- $Q \in \mathbb{R}^{d_Q}$: the surrogate labels. These are the labels generated by the LLM. When there are multiple surrogates, these are indexed by j.
- $W \in \mathbb{R}^{d_W}$: optional document-level metadata. In our experiments, these are the similarity scores generated using the method detailed in the previous section, but in practice, they can be any data that helps predict Y.

The researcher then samples a subset of documents to annotate and obtain outcomes Y.

- $R_i \in \{0, 1\}$: the missing indicator. This is a binary indicator of whether document i has been hand coded. We use the $\{i : R_i = 1\}$ to denote labeled documents and $\{i : R_i = 0\}$ to denote unlabeled ones.

[5] Total computation time for 14K observations was 42 seconds using a laptop with a NVIDIA RTX 3080Ti Laptop GPU.

[6] For brevity, we do not list these all out here. See the variable `embed_text_map` in `css_mappings.py` in the replication code to see the exact class descriptions. Note that in general they are considerably shorter than the macroeconomy topic description above.

- $\pi(Q_i, W_i, X_i) := \Pr(R_i = 1 \mid Q_i, W_i, X_i)$: the probability of gold-standard labeling. This denotes the probability that researchers will obtain a gold-standard label for document i. Note that by design, the researcher knows $\pi(Q_i, W_i, X_i)$ because they can choose which document to hand-code. Moreover, they can ensure the probability is bounded away from zero, i.e., $\pi(Q_i, W_i, X_i) > 0$ for all i.
- $n_R = \sum_{i=1}^{n} R_i$: the number of labeled documents.

In the experiments, we test our theoretical expectations about the four estimators by comparing their performance. In the logistic regression case, our quantities of interest are the coefficients of the oracle logistic regression β^* in the model regressing Y on X. In the class prevalence case, the parameter of interest is the true proportion of positives, which we denote $\mu = \mathbb{E}(Y)$.

Performance is evaluated using three metrics: *bias*, *coverage* and *root mean square error* (RMSE). These are calculated for each data-subset and design (detailed in the subsequent section) over $S = 500$ simulations. The simulation iteration is indexed by $s \in S$.

2.2.1 Surrogate Only Estimation

The SO estimate $\widehat{\beta}^{SO}$ is calculated by taking the average of labels produced by different prompts $\overline{Q}_i = \frac{\sum_{j=1}^{d_Q} Q_{ij}}{d_Q}$ and then fitting a logistic regression of $\overline{Q}_i$ on X_i. We implement this using the standard `R` logistic regression implementation `glm`. Details can be found on lines 248-252 of `experiment_logit.R` in the replication code.

The SO estimate for class prevalence is simply the average of the surrogate labels: $\widehat{\mu}^{SO} = \frac{\sum_{i=1}^{n} \overline{Q}_i}{n}$). Details can be found on lines 170-173 of `experiment_measure.R` in the replication code.

2.2.2 Gold-Standard Only Estimation

$\widehat{\beta}^{GSO}$ is calculated by regressing Y on X in the labeled subset of the data, weighting observations by the inverse propensity of being labeled $\pi(Q_i, W_i, X_i)^{-1}$. We use the implementation of the weighted logistic regression function in the `survey` library, `svyglm`. Details can be found on lines 257-266 of `experiment_logit.R` in the replication code.

$\widehat{\mu}^{GSO}$ is the expectation over the labeled observations: $\mu^{GSO} = \frac{\sum_{i:R_i=1} Y}{n_R}$. Details can be found on lines 178-181 of `experiment_measure.R` in the replication code.

2.2.3 Supervised Learning

Our implementation of $\widehat{\beta}^{SL}$ uses bootstrap. Over $b \in B$ bootstrap iterations, we:

1. Draw bootstrapped sample $\mathcal{D}_b = \{Y_b, Q_b, W_b, X_b\}$ from $\mathcal{D}$.
2. Fit the supervised model on the labeled data: $g(Q, W, X)$ to predict Y with (Q, W, X). This yields the fitted model $\widehat{g}_b(\cdot)$.
3. Predict the outcome for the entire dataset: $\widehat{Y}_b \coloneqq \widehat{g}(Q_b, W_b, X_b)$.
4. Fit logistic regression of $\widehat{Y}_b$ on X_b, yielding estimated model coefficients $\widehat{\beta}_b$.

The SL estimate of $\widehat{\beta}^{SL}$ is the average over bootstrapping iterations: $\frac{\sum_{b=1}^{B}(\widehat{\beta}_b)}{B}$, and standard error is the standard deviation of $\widehat{\beta}_b$.

We use the `regression_forest` function from the `grf` library for $g(\cdot)$ with the default model parameters. We pass the inverse propensity weights $1/\pi$ as sample weights to the model. The implementation of all the above can be found in `impute_logit.R` in the replication code.

In estimating the logistic regression of $\widehat{Y}$ on X, we allow the outcome to be a real-valued number not between 0 and 1, i.e. $\widehat{Y} \notin [0,1]$. Because the `glm` implementation of logistic regression does not permit this, we directly solve logistic regression's moment equation as in equation (3) of the main paper. This implementation can be found on lines 144-199 of `debias_logit.R` in the replication code.

$\widehat{\mu}^{GSO}$ is estimated using the same method as above, except that in step 4 we calculate $\widehat{\mu}_b = \sum_{i=1}^{n} \widehat{Y}_{ib}/n$. The point estimate is $\widehat{\mu} = \sum_{b=1}^{B} \widehat{\mu}_b/B$ and the standard error is given by the standard deviation of $\widehat{\mu}_b$.

2.2.4 Design-based Supervised Learning

As our implementation of $\widehat{\beta}^{DSL}$ closely follows Algorithm 1 of the main paper, we do not restate here. Sample-splitting iteration estimates are aggregated using the median approach described in Chernozhukov, Demirer, Duflo and Fernandez-Val (2018). Specific details can be found in `debias_logit.R` of the replication code.

$\widehat{\mu}^{DSL}$ is fitted analogously. For each sample-splitting iteration r, we calculate the point estimate of class prevalence as $\widehat{\mu}_r = \sum_{k=1}^{K} \sum_{i \in \mathcal{D}_k} \widetilde{Y}_i^{kr}$ (with the standard error computed correspondingly). The final point estimate $\widehat{\mu}$ is the median of $\widehat{\mu}_r$, and its standard error given by the formula $\sqrt{\text{median}(\widehat{\text{se}}_r^2 + (\widehat{\mu}_r - \widehat{\mu})^2)}$. Specific details can be found in `debias_measure.R` of the replication code.

2.3 Conducting the Experiment

In this section, we specify the exact experimental setups we executed and the resources and hardware required.

2.3.1 Logistic Regression

For the logistic regression simulation, we compute the performance of our estimators for the following designs:

- Class Balance: Balanced versus Imbalanced
- Surrogate: Zero-shot versus Five-shot label from `text-davinci-003`
- Number of "Labeled" Documents: $n_R = \{50, 100, 250, 500, 1000\}$

This yields a total of $2 \times 2 \times 5 = 20$ designs. For each design, we run $S = 500$ simulations, where we resample $\mathcal{D}_s$ from the original dataset $\mathcal{D}$. In each simulation we select n_R documents, stratifying on the surrogate outcome Q_i, to have $R_i = 1$ and treat the remainder as unlabeled. We fit each of the four estimators above to calculate point estimates and standard errors for each of the coefficients `senate`, `democrat` and `dw1` (described in the section above). We denote the estimate of coefficient $j \in \{\text{senate, democrat, dw1}\}$ in simulation $s \in S$ as $\widehat{\beta}_{j,s}$. For all estimators we compare to β^*, the coefficients of the oracle logistic regression on the original dataset $\mathcal{D}$.

We evaluate these four estimators across three metrics: *Bias*, *Coverage*, and *RMSE*. For clarity we state how we aggregate these metrics across coefficients. *Bias* is calculated as the normalized root mean squared difference between the average point estimate across simulations $\widehat{\beta}_j = \frac{\sum_{s=1}^{S} \widehat{\beta}_{j,s}}{S}$ and the oracle parameter β_j^* across all coefficients. *Coverage* is calculated as the average over the proportion of simulations in which the oracle lay within the confidence interval of the estimated coefficient. *RMSE* is calculated as the average root mean squared error across coefficients.

$$\begin{aligned}
\text{Bias} &= \sqrt{\mathbb{E}_j\left[(\widehat{\beta}_j - \beta_j^*)^2\right]}/\sqrt{\mathbb{E}_j[(\beta_j^*)^2]} \\
\text{Coverage} &= \mathbb{E}_j\left[\mathbb{E}_s(\mathbf{1}\{\widehat{\beta}_{j,s} - 1.96 \times \widehat{\text{se}}_{j,s} \leq \beta_j^*,\ \widehat{\beta}_{j,s} + 1.96 \times \widehat{\text{se}}_{j,s} \geq \beta_j^*\})\right] \\
\text{RMSE} &= \mathbb{E}_j\left[\sqrt{(\widehat{\beta}_j - \beta_j^*)^2 + \text{Var}_s(\widehat{\beta}_{j,s})}\right]
\end{aligned}$$

where $\mathbb{E}_j$ stands for the empirical average across coefficients and $\mathbb{E}_s$ for the empirical average over simulations. Var_s is the empirical variance over simulations.

2.3.2 Class Prevalence

For each of the 17 datasets from Ziems et al. (2023) outlined above, we compute the performance of our four estimators for the following designs:

- Surrogate: All available surrogates (`all`) and `flan-ul2`
- $n_R = \{25, 100, 200\}$

As noted earlier, for datasets with more than two classes, we compute separate estimates of $\widehat{\mu}_j$ for each class and aggregate these at the end. The only difference is for how we calculate bias, where we instead report the mean absolute error across classes: $\text{Bias} = \mathbb{E}_j[\text{abs}(\widehat{\beta}_j - \beta_j^*)]$.

2.3.3 Resources and Computation

All experiments were run in parallel on approximately 400 CPUs of various architectures on the HPC cluster at Princeton University. The total compute time for the logistic regression simulation was 28.7 CPU-hours. The total compute time for the class prevalence simulation was approximately 11K CPU-hours. Most computing time comes from the fact that we considered a large number of simulation designs for comprehensiveness, and each simulation only takes about 30 seconds. Thus, in practical applications in social sciences, researchers would typically need much less than 1 min (most often a matter of seconds) to implement our proposed methods. These calculations are based on average run times in the job logs, and are detailed in `02_experiment/README.md` of the code supplement.

2.4 Extended Results

In this section we provide the full set of results for our experiments (including results for values of n_R and surrogates not detailed in the main paper).

2.4.1 Complete Logistic Regression Results

The complete results of our logistic regression experiments on the CBP dataset are reported in Table S3 of this supplement. All reported values are based on the 500 simulations, and simulation standard errors are computed based on 1000 bootstraps and reported in parentheses.

In Table S3: *Bal* indicates the ratio of positives to negatives in the dataset; `shot` indicates the surrogate labels are zero- or five-shot; n_R indicates the number of gold-standard labels. Bias is the standardized root mean square bias averaged over four coefficients (including the intercept). Coverage is the average proportion of 95% confidence intervals containing the oracle logistic regression coefficient across four coefficients. RMSE is the average RMSE of the coefficients, with the fifth column $\frac{DSL}{GSO}$ included to show the improvement of DSL over GSO in RMSE.

Numbers in green indicate any estimator within 0.1 of the lowest bias for the row. Blue indicate any estimator achieving above 94.5% coverage. Orange indicates an estimator is within the average standard error of the DSL RMSE column of the best RMSE in the row. In this case, the average standard error of the DSL RMSE column is 0.00955 (3 s.f.).

Table S3 confirms the main results reported in the main paper. First, Surrogate-Only estimation has the largest bias and poorest coverage. Second, only GSO and DSL achieve low bias and proper coverage, both of which are fundamental to social science downstream statistical analyses. Looking at the last column, we see that the proposed DSL uniformly outperforms GSO in terms of RMSE. Importantly, SL achieves the lowest RMSE as expected, because it focuses on prediction. However, SL has a larger bias and poor coverage (as low as 60 %), which makes it unsuitable for social science downstream statistical analyses.

2.4.2 Complete Class Prevalence Results

The complete results of our class prevalence experiments with the Ziems et al. (2023) datasets are reported in Tables 4a through 4f. As with above, all reported values are based on the 500 simulations, and simulation standard errors are computed based on 1000 bootstrap and reported in parentheses.

Tables are ordered by the accuracy of the `flan-ul2` surrogate, indicated in *Acc.*. *LLM* is an abbreviated name of the surrogates used, with `ALL` indicating an average of all surrogates, and `UL2` indicating the `flan-ul2` surrogate. Bias is the mean absolute error across classes (where the task is binary we just report the absolute difference between the estimator and the oracle). Coverage and RMSE are teh same as with the CBP table. Likewise, $\frac{DSL}{GSO}$ shows the improvement of DSL over GSO in RMSE (2 d.p.).

Numbers in green indicate any estimator within 0.1pp of the lowest bias for the row. Blue indicate any estimator achieving above 94.5% coverage. Orange indicates an estimator is within the average standard error of the DSL RMSE column for all datasets and designs of the best RMSE in the row. In this case, the average standard error of the DSL RMSE is 0.00119 (3 s.f., not scaled by 100).

Table 4 confirms the main results reported in the main paper. First, Surrogate-Only estimation has the largest bias and poorest coverage. Second, GSO and DSL achieve low bias and proper coverage, both of which are fundamental to social science downstream statistical analyses. Looking at the last column, we see that the proposed DSL almost always outperforms GSO in terms of RMSE, and the gain is largest when the accuracy of LLMs-based surrogates is high. Given that LLMs-based surrogates will improve over time, we expect that the gain from DSL over GSO will increase in future applications where researchers engage with more prompt engineering and the performance of LLMs goes up. Importantly, SL also achieves relatively low bias and reasonable coverage, even though it does not have explicit theoretical guarantees for these properties. In terms of RMSE, DSL and SL are almost always one of the best performing methods, and their differences are often negligible.

In sum, we show that the DSL always has low bias and proper coverage, while achieving RMSE comparable to SL, which fails to provide valid inference in the logistic regression problem. Thus, the DSL can be seen as a safe strategy for using LLMs-based surrogates efficiently.

				Bias				Coverage (×100)				RMSE				
Bal	Acc.	shot	n_R	SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
1:1	68	0	100	3.46 (0.00)	0.13 (0.06)	0.82 (0.01)	0.11 (0.05)	23.3 (0.32)	94.9 (0.59)	65.5 (0.81)	96.8 (0.44)	0.64 (0.00)	0.66 (0.02)	0.22 (0.00)	0.58 (0.01)	0.88
			250	3.45 (0.00)	0.11 (0.04)	0.73 (0.01)	0.10 (0.03)	23.1 (0.32)	94.1 (0.67)	64.8 (0.84)	95.1 (0.63)	0.64 (0.00)	0.41 (0.01)	0.20 (0.00)	0.36 (0.01)	0.86
			500	3.46 (0.00)	0.05 (0.02)	0.68 (0.01)	0.04 (0.02)	22.4 (0.34)	95.4 (0.59)	65.2 (0.84)	95.6 (0.59)	0.64 (0.00)	0.28 (0.01)	0.17 (0.00)	0.23 (0.01)	0.84
			1K	3.45 (0.00)	0.04 (0.02)	0.60 (0.01)	0.03 (0.01)	23 (0.32)	94.7 (0.65)	62.9 (0.89)	94.9 (0.62)	0.64 (0.00)	0.20 (0.00)	0.15 (0.00)	0.17 (0.00)	0.85
	84	5	100	0.57 (0.00)	0.18 (0.09)	0.69 (0.01)	0.11 (0.06)	51.3 (0.85)	95.3 (0.58)	78 (0.83)	96.5 (0.47)	0.14 (0.00)	0.70 (0.02)	0.17 (0.00)	0.48 (0.01)	0.69
			250	0.56 (0.00)	0.14 (0.06)	0.47 (0.01)	0.12 (0.04)	52.5 (0.84)	95.3 (0.63)	82.4 (0.73)	95.6 (0.61)	0.13 (0.00)	0.41 (0.01)	0.13 (0.00)	0.28 (0.01)	0.69
			500	0.57 (0.00)	0.09 (0.04)	0.40 (0.01)	0.05 (0.02)	50.7 (0.83)	94.5 (0.67)	84 (0.73)	95.7 (0.54)	0.14 (0.00)	0.28 (0.01)	0.11 (0.00)	0.19 (0.00)	0.68
			1K	0.56 (0.00)	0.08 (0.03)	0.35 (0.01)	0.04 (0.02)	53.4 (0.78)	94.3 (0.65)	83.7 (0.83)	94.6 (0.69)	0.13 (0.00)	0.20 (0.00)	0.10 (0.00)	0.14 (0.00)	0.71
1:9	90	0	100	0.39 (0.00)	0.12 (0.02)	0.24 (0.00)	0.13 (0.02)	35.1 (0.54)	92.1 (0.69)	60 (0.75)	94.5 (0.6)	0.40 (0.00)	1.20 (0.14)	0.28 (0.00)	0.95 (0.02)	0.79
			250	0.39 (0.00)	0.04 (0.01)	0.19 (0.00)	0.05 (0.01)	36.2 (0.57)	93.7 (0.68)	64.1 (0.64)	94.2 (0.64)	0.40 (0.00)	0.59 (0.01)	0.23 (0.00)	0.54 (0.01)	0.93
			500	0.39 (0.00)	0.03 (0.01)	0.17 (0.00)	0.03 (0.01)	34.6 (0.57)	93.9 (0.64)	65.8 (0.62)	94.1 (0.61)	0.40 (0.00)	0.40 (0.01)	0.20 (0.00)	0.38 (0.01)	0.94
			1K	0.39 (0.00)	0.02 (0.00)	0.17 (0.00)	0.02 (0.01)	35.8 (0.57)	95 (0.62)	66.6 (0.6)	94.7 (0.66)	0.40 (0.00)	0.28 (0.01)	0.19 (0.00)	0.26 (0.01)	0.94
	88	5	100	0.33 (0.00)	0.40 (0.07)	0.18 (0.00)	0.11 (0.02)	31.4 (0.64)	92.9 (0.77)	62.4 (0.74)	94.6 (0.64)	0.31 (0.00)	1.69 (0.10)	0.24 (0.00)	0.87 (0.02)	0.52
			250	0.33 (0.00)	0.04 (0.01)	0.15 (0.00)	0.04 (0.01)	31.4 (0.63)	94.4 (0.66)	68.5 (0.88)	94.8 (0.59)	0.31 (0.00)	0.55 (0.01)	0.19 (0.00)	0.47 (0.01)	0.84
			500	0.33 (0.00)	0.01 (0.01)	0.13 (0.00)	0.02 (0.01)	31.7 (0.62)	95.3 (0.57)	71.4 (0.93)	95.3 (0.53)	0.31 (0.00)	0.36 (0.01)	0.16 (0.00)	0.31 (0.01)	0.86
			1K	0.33 (0.00)	0.02 (0.01)	0.11 (0.00)	0.02 (0.00)	30.9 (0.67)	95.6 (0.56)	75.3 (0.95)	96.2 (0.54)	0.31 (0.00)	0.26 (0.01)	0.14 (0.00)	0.22 (0.00)	0.85

Table S3: **Complete logistic regression results.** See text.

				Bias (×100)				Coverage (×100)				RMSE (×100)				
Dataset	Acc.	LLM	n_R	SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
Misinfo.	78	ALL	25	3.9 (0.0)	0.4 (0.3)	1.0 (0.5)	0.4 (0.3)	4.0 (0.9)	92.0 (1.2)	91.2 (1.3)	91.2 (1.3)	4.06 (0.04)	10.7 (0.34)	10.3 (0.33)	10.7 (0.34)	1.00
			100	3.8 (0.0)	0.3 (0.2)	0.2 (0.2)	0.2 (0.1)	3.0 (0.8)	92.9 (1.2)	90.8 (1.3)	93.3 (1.1)	3.97 (0.04)	5.3 (0.16)	4.5 (0.14)	4.6 (0.14)	0.86
			200	3.8 (0.0)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	2.0 (0.6)	95.0 (1.0)	93.0 (1.1)	95.0 (1.0)	4.0 (0.04)	3.5 (0.11)	3.03 (0.09)	3.05 (0.09)	0.87
		UL2	25	8.7 (0.1)	0.4 (0.3)	0.9 (0.4)	0.4 (0.3)	3.2 (0.8)	93.1 (1.1)	92.9 (1.2)	93.9 (1.1)	9.01 (0.10)	9.9 (0.32)	9.6 (0.32)	9.9 (0.32)	1.00
			100	8.7 (0.1)	0.3 (0.2)	0.2 (0.1)	0.3 (0.2)	1.0 (0.5)	94.8 (1.0)	95.6 (0.9)	96.0 (0.9)	9.0 (0.10)	4.9 (0.15)	4.31 (0.13)	4.33 (0.13)	0.87
			200	8.7 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	1.6 (0.6)	93.8 (1.1)	94.2 (1.0)	94.9 (1.0)	8.9 (0.10)	3.6 (0.13)	3.13 (0.11)	3.13 (0.11)	0.87
Emotion	70	ALL	25	8.6 (0.0)	0.4 (0.1)	1.9 (0.1)	0.4 (0.1)	19.6 (0.4)	90.3 (0.5)	87.1 (0.6)	90.5 (0.5)	8.8 (0.01)	7.6 (0.10)	6.99 (0.10)	7.6 (0.10)	1.00
			100	8.6 (0.0)	0.1 (0.0)	0.5 (0.1)	0.1 (0.0)	19.2 (0.4)	94.5 (0.4)	92.3 (0.5)	94.6 (0.4)	8.8 (0.01)	3.7 (0.05)	2.96 (0.04)	3.1 (0.04)	0.85
			200	8.7 (0.0)	0.2 (0.0)	0.2 (0.0)	0.1 (0.0)	19.3 (0.4)	94.3 (0.4)	91.8 (0.5)	95.1 (0.4)	8.8 (0.01)	2.6 (0.03)	2.16 (0.03)	2.20 (0.03)	0.85
		UL2	25	2.1 (0.0)	0.4 (0.1)	1.6 (0.1)	0.4 (0.1)	69.4 (0.6)	91.2 (0.5)	87.7 (0.6)	91.8 (0.5)	2.90 (0.02)	7.4 (0.10)	6.8 (0.09)	7.4 (0.10)	1.00
			100	2.1 (0.0)	0.2 (0.0)	0.6 (0.1)	0.2 (0.0)	68.8 (0.6)	94.4 (0.4)	91.3 (0.5)	94.0 (0.4)	2.92 (0.02)	3.7 (0.05)	3.2 (0.04)	3.2 (0.04)	0.87
			200	2.1 (0.0)	0.1 (0.0)	0.2 (0.0)	0.1 (0.0)	69.9 (0.6)	95.2 (0.4)	92.5 (0.5)	95.4 (0.4)	2.9 (0.03)	2.6 (0.04)	2.22 (0.03)	2.24 (0.03)	0.87
Figur.	64	ALL	25	7.6 (0.0)	0.4 (0.1)	1.6 (0.2)	0.4 (0.1)	3.8 (0.4)	92.2 (0.6)	90.3 (0.7)	91.7 (0.6)	7.61 (0.01)	8.9 (0.14)	8.4 (0.14)	9.0 (0.15)	1.00
			100	7.6 (0.0)	0.2 (0.1)	0.5 (0.1)	0.2 (0.1)	4.0 (0.4)	93.9 (0.5)	91.8 (0.6)	94.4 (0.5)	7.6 (0.01)	4.4 (0.07)	3.56 (0.06)	3.7 (0.06)	0.85
			200	7.6 (0.0)	0.1 (0.0)	0.3 (0.1)	0.1 (0.0)	3.8 (0.4)	94.2 (0.5)	92.2 (0.6)	94.5 (0.5)	7.6 (0.01)	3.1 (0.05)	2.57 (0.04)	2.63 (0.04)	0.85
		UL2	25	5.6 (0.0)	0.4 (0.1)	1.3 (0.2)	0.4 (0.1)	39.8 (0.8)	92.6 (0.6)	91.9 (0.6)	93.3 (0.6)	6.03 (0.04)	8.6 (0.14)	8.0 (0.14)	8.6 (0.14)	1.00
			100	5.5 (0.0)	0.3 (0.1)	0.4 (0.1)	0.3 (0.1)	40.3 (0.8)	93.9 (0.5)	91.9 (0.6)	94.3 (0.5)	6.0 (0.04)	4.4 (0.07)	3.84 (0.06)	3.87 (0.06)	0.89
			200	5.6 (0.0)	0.1 (0.1)	0.2 (0.1)	0.1 (0.0)	41.5 (0.8)	95.2 (0.5)	93.3 (0.6)	95.1 (0.5)	6.1 (0.04)	3.1 (0.05)	2.69 (0.04)	2.69 (0.04)	0.88

Table S4a: **Class prevalence estimation** for Misinfo., Emotion, Figur..

				Bias (×100)				Coverage (×100)				RMSE (×100)				
Dataset	Acc.	LLM	n_R	SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
Power	61	ALL	25	17.2 (0.0)	0.4 (0.3)	0.4 (0.3)	0.4 (0.3)	0.0 (0.0)	92.0 (1.2)	92.0 (1.2)	91.2 (1.3)	17.2 (0.03)	10.70 (0.34)	11.0 (0.35)	10.73 (0.34)	1.00
			100	17.2 (0.0)	0.3 (0.2)	0.6 (0.2)	0.2 (0.2)	0.0 (0.0)	92.9 (1.2)	90.9 (1.3)	92.7 (1.2)	17.2 (0.03)	5.32 (0.16)	5.24 (0.16)	5.22 (0.16)	0.98
			200	17.2 (0.0)	0.1 (0.1)	0.3 (0.1)	0.1 (0.1)	0.0 (0.0)	95.0 (1.0)	93.8 (1.1)	94.2 (1.1)	17.2 (0.03)	3.5 (0.11)	3.38 (0.11)	3.40 (0.10)	0.97
		UL2	25	26.9 (0.1)	0.4 (0.3)	0.5 (0.4)	0.4 (0.3)	0.0 (0.0)	93.1 (1.1)	92.7 (1.2)	93.9 (1.1)	27.0 (0.09)	9.94 (0.32)	10.2 (0.33)	9.94 (0.32)	1.00
			100	26.8 (0.1)	0.3 (0.2)	0.4 (0.2)	0.3 (0.2)	0.0 (0.0)	94.8 (1.0)	94.0 (1.1)	94.8 (1.0)	26.9 (0.09)	4.95 (0.15)	4.98 (0.16)	4.90 (0.16)	0.99
			200	26.8 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	93.8 (1.1)	93.8 (1.1)	94.4 (1.0)	26.8 (0.08)	3.61 (0.13)	3.61 (0.13)	3.55 (0.13)	0.98
Humor	59	ALL	25	22.9 (0.0)	0.6 (0.4)	0.9 (0.4)	0.5 (0.4)	0.0 (0.0)	94.1 (1.0)	94.3 (1.0)	92.3 (1.2)	22.9 (0.03)	10.18 (0.31)	10.3 (0.31)	10.20 (0.31)	1.00
			100	22.9 (0.0)	0.2 (0.2)	0.3 (0.2)	0.2 (0.2)	0.0 (0.0)	95.2 (0.9)	93.6 (1.1)	95.4 (0.9)	22.9 (0.03)	4.92 (0.14)	4.92 (0.14)	4.84 (0.13)	0.98
			200	22.9 (0.0)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.0 (0.0)	93.6 (1.2)	92.3 (1.2)	93.6 (1.1)	22.9 (0.03)	3.59 (0.11)	3.54 (0.11)	3.51 (0.11)	0.98
		UL2	25	22.8 (0.1)	0.4 (0.3)	0.4 (0.3)	0.5 (0.3)	0.0 (0.0)	94.8 (1.0)	94.8 (1.0)	94.4 (1.0)	22.9 (0.09)	9.76 (0.35)	9.83 (0.36)	9.83 (0.35)	1.01
			100	22.7 (0.1)	0.2 (0.1)	0.2 (0.2)	0.2 (0.1)	0.0 (0.0)	94.6 (1.0)	93.2 (1.1)	94.8 (1.0)	22.7 (0.08)	4.97 (0.15)	4.96 (0.15)	4.95 (0.15)	1.00
			200	23.0 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	95.4 (0.9)	93.3 (1.1)	94.6 (1.0)	23.1 (0.09)	3.56 (0.11)	3.55 (0.11)	3.52 (0.11)	0.99
Toxic.	57	ALL	25	26.5 (0.0)	0.4 (0.3)	0.5 (0.4)	0.4 (0.3)	0.0 (0.0)	92.0 (1.2)	91.6 (1.3)	91.2 (1.3)	26.5 (0.02)	10.70 (0.34)	11.0 (0.35)	10.73 (0.34)	1.00
			100	26.5 (0.0)	0.3 (0.2)	0.6 (0.2)	0.3 (0.2)	0.0 (0.0)	92.9 (1.2)	91.1 (1.3)	93.5 (1.1)	26.5 (0.02)	5.32 (0.16)	5.32 (0.16)	5.28 (0.16)	0.99
			200	26.5 (0.0)	0.1 (0.1)	0.3 (0.1)	0.1 (0.1)	0.0 (0.0)	95.0 (1.0)	93.6 (1.1)	95.0 (1.0)	26.5 (0.02)	3.50 (0.11)	3.48 (0.11)	3.46 (0.11)	0.99
		UL2	25	34.9 (0.1)	0.4 (0.3)	0.6 (0.4)	0.4 (0.3)	0.0 (0.0)	93.1 (1.1)	93.5 (1.1)	93.9 (1.1)	34.9 (0.07)	9.94 (0.32)	10.02 (0.32)	9.94 (0.32)	1.00
			100	35.1 (0.1)	0.3 (0.2)	0.6 (0.2)	0.3 (0.2)	0.0 (0.0)	94.8 (1.0)	93.4 (1.1)	95.4 (0.9)	35.1 (0.07)	4.95 (0.15)	5.1 (0.16)	4.94 (0.15)	1.00
			200	35.1 (0.1)	0.2 (0.1)	0.3 (0.2)	0.2 (0.1)	0.0 (0.0)	93.8 (1.1)	93.6 (1.1)	93.4 (1.1)	35.2 (0.07)	3.61 (0.13)	3.59 (0.13)	3.54 (0.13)	0.98

Table S4b: **Class prevalence estimation** for Power, Humor, Toxic..

				Bias (×100)				Coverage (×100)				RMSE (×100)				
Dataset	Acc.	LLM	n_R	SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
Stance	55	`ALL`	25	13.3 (0.0)	0.5 (0.2)	1.2 (0.2)	0.5 (0.2)	0.0 (0.0)	93.3 (0.7)	92.1 (0.7)	92.4 (0.7)	13.3 (0.03)	9.7 (0.17)	9.23 (0.17)	9.7 (0.17)	1.00
			100	13.3 (0.0)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.0 (0.0)	94.8 (0.6)	92.3 (0.7)	95.0 (0.6)	13.4 (0.03)	4.6 (0.08)	3.97 (0.07)	4.02 (0.07)	0.87
			200	13.3 (0.0)	0.1 (0.1)	0.2 (0.1)	0.1 (0.0)	0.0 (0.0)	95.4 (0.6)	93.0 (0.7)	95.0 (0.6)	13.3 (0.03)	3.3 (0.06)	2.85 (0.05)	2.88 (0.06)	0.88
		`UL2`	25	20.7 (0.1)	0.4 (0.2)	0.7 (0.2)	0.4 (0.2)	2.8 (0.4)	93.2 (0.6)	92.8 (0.7)	92.7 (0.7)	20.8 (0.05)	9.7 (0.17)	9.32 (0.17)	9.7 (0.17)	1.00
			100	20.7 (0.1)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	3.5 (0.5)	94.6 (0.6)	91.9 (0.7)	93.4 (0.6)	20.9 (0.05)	4.8 (0.09)	4.32 (0.08)	4.34 (0.08)	0.91
			200	20.7 (0.0)	0.1 (0.1)	0.2 (0.1)	0.1 (0.0)	2.1 (0.4)	95.7 (0.5)	94.6 (0.6)	96.6 (0.5)	20.8 (0.05)	3.2 (0.06)	2.94 (0.05)	2.93 (0.05)	0.90
Seman.	54	`ALL`	25	16.6 (0.0)	0.4 (0.3)	0.4 (0.3)	0.4 (0.3)	0.0 (0.0)	94.5 (1.0)	94.1 (1.0)	94.1 (1.1)	16.7 (0.03)	9.59 (0.30)	9.62 (0.31)	9.58 (0.30)	1.00
			100	16.7 (0.0)	0.4 (0.2)	0.4 (0.2)	0.4 (0.2)	0.0 (0.0)	93.8 (1.1)	92.4 (1.1)	93.4 (1.1)	16.7 (0.02)	5.2 (0.16)	4.91 (0.15)	4.98 (0.15)	0.96
			200	16.7 (0.0)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	95.6 (0.9)	94.6 (1.0)	95.0 (1.0)	16.7 (0.02)	3.45 (0.11)	3.35 (0.11)	3.34 (0.11)	0.97
		`UL2`	25	45.6 (0.0)	0.5 (0.3)	0.4 (0.3)	0.5 (0.3)	0.0 (0.0)	95.0 (1.0)	93.2 (1.1)	93.6 (1.1)	45.6 (0.05)	9.69 (0.30)	9.9 (0.30)	9.69 (0.30)	1.00
			100	45.6 (0.1)	0.5 (0.2)	0.3 (0.2)	0.4 (0.2)	0.0 (0.0)	94.0 (1.1)	92.2 (1.2)	93.2 (1.1)	45.6 (0.05)	5.2 (0.18)	5.14 (0.18)	5.10 (0.17)	0.98
			200	45.7 (0.1)	0.2 (0.1)	0.1 (0.1)	0.1 (0.1)	0.0 (0.0)	95.6 (0.9)	95.6 (0.9)	96.8 (0.8)	45.7 (0.05)	3.3 (0.11)	3.23 (0.10)	3.22 (0.10)	0.96
Pers. I	53	`ALL`	25	30.6 (0.0)	0.4 (0.3)	0.4 (0.3)	0.4 (0.3)	0.0 (0.0)	94.1 (1.1)	94.3 (1.1)	92.2 (1.3)	30.6 (0.03)	10.44 (0.38)	10.7 (0.38)	10.44 (0.38)	1.00
			100	30.6 (0.0)	0.2 (0.1)	0.2 (0.2)	0.2 (0.1)	0.0 (0.0)	96.8 (0.8)	95.2 (0.9)	96.4 (0.8)	30.6 (0.02)	4.64 (0.16)	4.74 (0.16)	4.62 (0.16)	1.00
			200	30.7 (0.0)	0.2 (0.1)	0.4 (0.2)	0.2 (0.1)	0.0 (0.0)	93.8 (1.1)	91.8 (1.3)	94.0 (1.1)	30.7 (0.03)	3.60 (0.12)	3.61 (0.12)	3.56 (0.12)	0.99
		`UL2`	25	42.2 (0.1)	0.4 (0.3)	0.4 (0.3)	0.4 (0.3)	0.0 (0.0)	92.2 (1.2)	90.8 (1.3)	93.4 (1.1)	42.2 (0.06)	10.14 (0.34)	10.4 (0.35)	10.11 (0.33)	1.00
			100	42.1 (0.1)	0.2 (0.2)	0.4 (0.2)	0.2 (0.2)	0.0 (0.0)	96.0 (0.9)	95.8 (0.9)	95.8 (0.9)	42.1 (0.06)	4.68 (0.16)	4.78 (0.17)	4.69 (0.16)	1.00
			200	42.2 (0.1)	0.1 (0.1)	0.2 (0.1)	0.1 (0.1)	0.0 (0.0)	95.5 (0.9)	94.8 (1.0)	95.1 (0.9)	42.2 (0.06)	3.51 (0.11)	3.56 (0.11)	3.50 (0.11)	1.00

Table S4c: **Class prevalence estimation** for Stance, Seman., Pers. I..

				Bias (×100)				Coverage (×100)				RMSE (×100)				
Dataset	Acc.	LLM	n_R	SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
Polite.	53	ALL	25	11.7 (0.0)	0.6 (0.2)	1.1 (0.2)	0.6 (0.2)	0.0 (0.0)	94.2 (0.6)	93.2 (0.6)	92.7 (0.7)	11.7 (0.02)	9.4 (0.16)	9.30 (0.16)	9.4 (0.16)	1.00
			100	11.7 (0.0)	0.2 (0.1)	0.5 (0.1)	0.2 (0.1)	0.0 (0.0)	94.1 (0.6)	92.7 (0.7)	94.0 (0.6)	11.7 (0.02)	4.8 (0.09)	4.23 (0.08)	4.33 (0.08)	0.90
			200	11.7 (0.0)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.0 (0.0)	95.0 (0.6)	92.6 (0.7)	94.9 (0.6)	11.7 (0.02)	3.3 (0.06)	3.00 (0.06)	3.02 (0.06)	0.91
		UL2	25	16.9 (0.1)	0.4 (0.2)	0.7 (0.2)	0.4 (0.2)	20.2 (0.7)	93.4 (0.6)	92.0 (0.7)	93.1 (0.6)	17.1 (0.05)	9.41 (0.17)	9.40 (0.16)	9.43 (0.17)	1.00
			100	16.8 (0.0)	0.2 (0.1)	0.4 (0.1)	0.2 (0.1)	21.7 (0.7)	95.0 (0.6)	94.0 (0.6)	95.1 (0.6)	17.1 (0.05)	4.7 (0.08)	4.59 (0.08)	4.50 (0.08)	0.97
			200	16.9 (0.1)	0.1 (0.1)	0.3 (0.1)	0.1 (0.0)	21.9 (0.7)	95.5 (0.5)	94.6 (0.6)	95.9 (0.5)	17.1 (0.05)	3.3 (0.06)	3.16 (0.06)	3.14 (0.06)	0.96
Pers. II	49	ALL	25	6.9 (0.0)	0.3 (0.1)	1.4 (0.1)	0.4 (0.1)	17.7 (0.3)	89.8 (0.5)	87.4 (0.6)	87.3 (0.6)	7.1 (0.01)	7.1 (0.09)	6.64 (0.09)	7.1 (0.09)	1.00
			100	6.9 (0.0)	0.2 (0.0)	0.6 (0.1)	0.2 (0.0)	17.1 (0.3)	93.7 (0.4)	91.1 (0.5)	93.7 (0.4)	7.1 (0.01)	3.5 (0.04)	3.12 (0.04)	3.19 (0.04)	0.91
			200	6.9 (0.0)	0.1 (0.0)	0.3 (0.0)	0.1 (0.0)	17.6 (0.3)	94.2 (0.4)	92.1 (0.5)	94.4 (0.4)	7.1 (0.01)	2.5 (0.03)	2.25 (0.03)	2.27 (0.03)	0.91
		UL2	25	7.9 (0.0)	0.3 (0.1)	1.2 (0.1)	0.3 (0.1)	10.6 (0.5)	89.5 (0.5)	86.9 (0.5)	87.3 (0.5)	8.1 (0.03)	7.1 (0.08)	6.59 (0.08)	7.1 (0.08)	1.00
			100	7.9 (0.0)	0.2 (0.0)	0.6 (0.1)	0.2 (0.0)	10.7 (0.5)	94.0 (0.4)	91.9 (0.5)	93.7 (0.4)	8.1 (0.03)	3.5 (0.04)	3.20 (0.04)	3.26 (0.04)	0.94
			200	8.0 (0.0)	0.1 (0.0)	0.3 (0.0)	0.1 (0.0)	10.8 (0.5)	94.7 (0.4)	92.4 (0.5)	94.8 (0.4)	8.2 (0.03)	2.5 (0.03)	2.34 (0.03)	2.35 (0.03)	0.94
Books	48	ALL	25	4.6 (0.0)	0.6 (0.2)	0.9 (0.2)	0.6 (0.2)	9.6 (0.7)	94.2 (0.6)	93.3 (0.6)	92.6 (0.7)	4.67 (0.02)	9.4 (0.16)	9.4 (0.16)	9.5 (0.16)	1.00
			100	4.6 (0.0)	0.2 (0.1)	0.4 (0.1)	0.2 (0.1)	10.7 (0.7)	94.1 (0.6)	92.5 (0.7)	93.5 (0.6)	4.7 (0.02)	4.8 (0.09)	4.36 (0.08)	4.44 (0.08)	0.93
			200	4.6 (0.0)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	10.2 (0.7)	95.0 (0.6)	92.9 (0.7)	94.6 (0.6)	4.7 (0.02)	3.3 (0.06)	3.09 (0.06)	3.12 (0.06)	0.94
		UL2	25	22.6 (0.0)	0.4 (0.2)	0.9 (0.2)	0.4 (0.2)	0.0 (0.0)	93.4 (0.6)	92.2 (0.7)	93.0 (0.7)	22.6 (0.05)	9.41 (0.16)	9.51 (0.16)	9.44 (0.16)	1.00
			100	22.4 (0.0)	0.2 (0.1)	0.4 (0.1)	0.2 (0.1)	0.0 (0.0)	95.0 (0.6)	93.8 (0.6)	95.2 (0.6)	22.5 (0.05)	4.67 (0.09)	4.66 (0.09)	4.58 (0.09)	0.98
			200	22.5 (0.0)	0.1 (0.1)	0.2 (0.1)	0.1 (0.1)	0.0 (0.0)	95.6 (0.5)	94.5 (0.6)	95.6 (0.5)	22.5 (0.05)	3.26 (0.06)	3.18 (0.06)	3.16 (0.06)	0.97

Table S4d: **Class prevalence estimation** for Polite., Pers. II, Books..

Dataset	Acc.	LLM	n_R	Bias (×100)				Coverage (×100)				RMSE (×100)				
				SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
Disc.	42	ALL	25	6.9 (0.0)	0.3 (0.1)	1.1 (0.1)	0.3 (0.1)	12.0 (0.3)	89.6 (0.5)	87.3 (0.5)	86.5 (0.6)	7.0 (0.01)	7.2 (0.09)	6.76 (0.08)	7.2 (0.09)	1.00
			100	6.9 (0.0)	0.2 (0.0)	0.5 (0.1)	0.1 (0.0)	12.6 (0.2)	93.2 (0.4)	91.1 (0.5)	93.4 (0.4)	6.9 (0.01)	3.6 (0.05)	3.20 (0.04)	3.3 (0.04)	0.93
			200	6.9 (0.0)	0.1 (0.0)	0.3 (0.0)	0.1 (0.0)	12.3 (0.3)	93.0 (0.4)	90.9 (0.5)	93.4 (0.4)	7.0 (0.01)	2.6 (0.03)	2.33 (0.03)	2.38 (0.03)	0.93
		UL2	25	5.5 (0.0)	0.3 (0.1)	0.8 (0.1)	0.3 (0.1)	38.0 (0.6)	89.0 (0.5)	88.6 (0.5)	87.2 (0.6)	5.86 (0.02)	7.1 (0.10)	6.8 (0.09)	7.1 (0.10)	1.00
			100	5.5 (0.0)	0.2 (0.0)	0.5 (0.1)	0.2 (0.0)	37.2 (0.6)	93.9 (0.4)	91.4 (0.5)	93.9 (0.4)	5.9 (0.02)	3.5 (0.04)	3.38 (0.04)	3.43 (0.04)	0.97
			200	5.5 (0.0)	0.1 (0.0)	0.3 (0.0)	0.1 (0.0)	38.0 (0.6)	94.4 (0.4)	92.6 (0.5)	94.4 (0.4)	5.9 (0.02)	2.45 (0.03)	2.37 (0.03)	2.36 (0.03)	0.96
News	40	ALL	25	8.3 (0.0)	0.6 (0.2)	1.1 (0.2)	0.6 (0.2)	0.1 (0.1)	94.2 (0.6)	93.2 (0.6)	92.7 (0.7)	8.35 (0.02)	9.4 (0.16)	9.4 (0.16)	9.4 (0.16)	1.00
			100	8.3 (0.0)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.0 (0.0)	94.1 (0.6)	92.3 (0.7)	93.7 (0.6)	8.4 (0.02)	4.8 (0.09)	4.36 (0.08)	4.45 (0.08)	0.93
			200	8.3 (0.0)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.0 (0.0)	95.0 (0.6)	92.1 (0.7)	94.3 (0.6)	8.4 (0.02)	3.3 (0.06)	3.09 (0.06)	3.10 (0.06)	0.93
		UL2	25	10.5 (0.1)	0.4 (0.2)	0.6 (0.2)	0.4 (0.2)	4.4 (0.5)	93.4 (0.6)	92.5 (0.7)	93.1 (0.7)	10.7 (0.05)	9.41 (0.17)	9.49 (0.17)	9.44 (0.17)	1.00
			100	10.6 (0.1)	0.2 (0.1)	0.4 (0.1)	0.2 (0.1)	4.8 (0.5)	95.0 (0.6)	93.9 (0.6)	95.4 (0.5)	10.8 (0.05)	4.7 (0.09)	4.62 (0.09)	4.52 (0.09)	0.97
			200	10.5 (0.1)	0.1 (0.1)	0.2 (0.1)	0.1 (0.1)	4.0 (0.5)	95.5 (0.5)	94.1 (0.6)	95.5 (0.5)	10.8 (0.05)	3.26 (0.06)	3.22 (0.06)	3.19 (0.06)	0.98
Emp.	40	ALL	25	16.7 (0.0)	0.6 (0.2)	0.6 (0.2)	0.6 (0.2)	0.0 (0.0)	94.2 (0.6)	93.7 (0.6)	92.7 (0.7)	16.7 (0.01)	9.45 (0.17)	9.6 (0.17)	9.45 (0.17)	1.00
			100	16.6 (0.0)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.0 (0.0)	94.1 (0.6)	93.0 (0.7)	93.9 (0.6)	16.6 (0.01)	4.79 (0.09)	4.84 (0.10)	4.77 (0.09)	1.00
			200	16.6 (0.0)	0.2 (0.1)	0.3 (0.1)	0.2 (0.1)	0.0 (0.0)	95.0 (0.5)	93.0 (0.7)	95.0 (0.6)	16.6 (0.01)	3.32 (0.06)	3.33 (0.06)	3.31 (0.06)	1.00
		UL2	25	24.6 (0.0)	0.4 (0.2)	0.4 (0.2)	0.4 (0.2)	0.0 (0.0)	93.4 (0.6)	92.9 (0.7)	93.1 (0.7)	24.7 (0.05)	9.41 (0.17)	9.6 (0.17)	9.44 (0.17)	1.00
			100	24.7 (0.0)	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	95.0 (0.5)	93.6 (0.6)	94.9 (0.6)	24.7 (0.05)	4.67 (0.08)	4.75 (0.08)	4.65 (0.08)	1.00
			200	24.7 (0.0)	0.1 (0.1)	0.3 (0.1)	0.1 (0.1)	0.0 (0.0)	95.5 (0.5)	94.9 (0.6)	95.3 (0.5)	24.8 (0.04)	3.26 (0.06)	3.28 (0.06)	3.25 (0.06)	1.00

Table S4e: **Class prevalence estimation** for Disc., News, Emp..

				Bias (×100)				Coverage (×100)				RMSE (×100)				
				SO	GSO	SL	DSL	SO	GSO	SL	DSL	SO	GSO	SL	DSL	$\frac{\text{DSL}}{\text{GSO}}$
Dataset	Acc.	LLM	n_R													
Hate	36	ALL	25	8.8 (0.0)	0.4 (0.1)	1.0 (0.1)	0.4 (0.1)	15.5 (0.2)	90.3 (0.6)	88.4 (0.6)	90.4 (0.6)	8.9 (0.01)	7.6 (0.11)	7.46 (0.11)	7.6 (0.11)	1.00
			100	8.8 (0.0)	0.2 (0.0)	0.6 (0.1)	0.1 (0.0)	15.8 (0.2)	94.5 (0.4)	92.3 (0.5)	94.6 (0.4)	8.9 (0.01)	3.7 (0.05)	3.46 (0.05)	3.55 (0.05)	0.97
			200	8.8 (0.0)	0.2 (0.0)	0.4 (0.0)	0.2 (0.0)	15.9 (0.2)	94.3 (0.4)	92.6 (0.5)	94.7 (0.4)	8.9 (0.01)	2.6 (0.03)	2.47 (0.03)	2.50 (0.03)	0.96
		UL2	25	9.7 (0.0)	0.4 (0.1)	0.7 (0.1)	0.4 (0.1)	17.0 (0.2)	91.2 (0.5)	89.8 (0.6)	91.8 (0.5)	10.1 (0.03)	7.38 (0.10)	7.31 (0.09)	7.40 (0.10)	1.00
			100	9.7 (0.0)	0.2 (0.0)	0.4 (0.1)	0.2 (0.0)	17.4 (0.3)	94.4 (0.4)	92.6 (0.5)	94.1 (0.4)	10.1 (0.03)	3.72 (0.05)	3.63 (0.05)	3.63 (0.05)	0.98
			200	9.7 (0.0)	0.1 (0.0)	0.3 (0.0)	0.1 (0.0)	17.0 (0.3)	95.2 (0.4)	93.1 (0.5)	95.2 (0.4)	10.1 (0.03)	2.59 (0.03)	2.54 (0.03)	2.53 (0.03)	0.98
Dialect	24	ALL	25	5.4 (0.0)	0.2 (0.0)	0.3 (0.0)	0.2 (0.0)	37.0 (0.3)	47.1 (0.5)	46.8 (0.5)	46.9 (0.5)	5.5 (0.00)	3.4 (0.03)	3.17 (0.03)	3.5 (0.03)	1.01
			100	5.4 (0.0)	0.1 (0.0)	0.2 (0.0)	0.1 (0.0)	36.9 (0.3)	86.2 (0.3)	84.8 (0.3)	86.2 (0.3)	5.5 (0.00)	1.7 (0.01)	1.59 (0.01)	1.7 (0.01)	0.99
			200	5.4 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	37.5 (0.3)	89.9 (0.3)	89.1 (0.3)	89.9 (0.3)	5.5 (0.00)	1.20 (0.01)	1.16 (0.01)	1.19 (0.01)	0.99
		UL2	25	5.2 (0.0)	0.2 (0.0)	0.3 (0.0)	0.2 (0.0)	48.1 (0.3)	47.9 (0.4)	47.7 (0.4)	47.8 (0.4)	5.5 (0.01)	3.4 (0.03)	3.17 (0.03)	3.4 (0.03)	1.01
			100	5.2 (0.0)	0.1 (0.0)	0.2 (0.0)	0.1 (0.0)	48.1 (0.3)	87.3 (0.3)	86.0 (0.3)	87.2 (0.3)	5.5 (0.01)	1.70 (0.01)	1.59 (0.01)	1.68 (0.01)	0.99
			200	5.2 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	47.6 (0.3)	90.2 (0.3)	89.6 (0.3)	90.3 (0.3)	5.5 (0.01)	1.19 (0.01)	1.15 (0.01)	1.18 (0.01)	0.99

Table S4f: **Class prevalence estimation** for Hate, Dialect..

2.5 Simulation (Figure 1)

To illustrate the main methodological challenges of the surrogate only estimation method, we use the following simple simulation for Figure 1. For full realistic experiments, see Section 5 of our main paper.

We generate $n = 5000$ i.i.d. observations ($i \in \{1, \ldots, 5000\}$) as follows.

- Covariates: $X_{ik} \sim \mathcal{N}(0, 1)$ where $k = 1, \ldots, 10$.
- Binary Outcome: $Y_i \sim \text{Bernoulli}(\text{expit}(W_i))$ where

  $$W_i = -1 + \frac{0.1}{1 + \exp(0.5X_{i3} - 0.5X_{i2})} + \frac{1.3X_{i4}}{1 + \exp(-0.1X_{i2})} + 1.5X_{i4}X_{i6} + 0.5X_{i1}X_{i2} + 0.3X_{i1} + 0.2X_{i2}$$

  This data-generating process is similar to the one in Vansteelandt and Dukes (2022). It contains various nonlinear transformation of X and it is difficult to correctly model the outcome function.
- Surrogate: $Q_i = P_i Y_i + (1 - P_i)(1 - Y_i)$ where $P_i \sim \text{Bernoulli}(P_q)$ and P_q controls the accuracy of the surrogate. When $P_q = 0.9$, $Q_i = Y_i$ with 90% and $Q_i = 1 - Y_i$ with 10%. We vary P_q in our simulation.
- Gold-standard Labeling: For simplicity, we use simple random sampling of 500 documents for gold-standard labeling. Thus, $\Pr(R_i = 1) = 0.1$.

Our estimand of interest is the coefficients of the Oracle logistic regression where we regress Y_i on $(X_{i1}, X_{i1}^2, X_{i2}, X_{i4})$.

We evaluated bias, coverage, RMSE of the SO and DSL in Figure 1. See Section 2.3.1 of this supplement for the definitions of bias, coverage, and RMSE, and See Sections 2.2.1 and 2.2.4 for the implementations of the SO and DSL.

We found that the SO has a large bias and invalid confidence intervals, which makes it unsuitable for social science downstream analyses. The DSL has low bias and proper coverage of confidence intervals regardless of the accuracy of the LLMs-based surrogates.

2.6 Further Comparison between SL and DSL

In this section, we briefly discuss further comparisons between SL and DSL, and include experimental results for an alternative state-of-the-art SL method (Wang, McCormick and Leek, 2020) as an additional baseline.

For reasons explained in the main body of the paper, our evaluations of the estimators follow the social science priority for bias and coverage. On these metrics SL performs poorly in contrast to DSL, being biased and having invalid confidence intervals. In contrast, we expect that SL methods, which are generally optimized on minimizing RMSE, should outperform DSL on RMSE. However, in our experiment we show that DSL, while maintaining unbiasedness and valid confidence intervals, can achieve RMSE comparable to SL (even though it is often higher than SL in a finite sample).

Moreover, as sample size increases, DSL provably dominates SL in terms of RMSE because bias of SL does not vanish with sample size, while variance of DSL will vanish. This is visible in the right-hand panel of figure S1, which shows that as we increase the size of the gold-standard data, DSL outperforms SL on RMSE as well as bias and coverage.

As noted in the main body of the paper, SL describes a broad collection of estimators, of which the one implemented in this paper is only one. We therefore also include bias, coverage and RMSE values for the state-of-the-art SL method presented in Wang, McCormick and Leek (2020) for the CBP data in Figure S1. We again find that DSL outperforms this SL method on bias and coverage, and that as sample size increases, DSL achieves parity and then outperforms SL on RMSE.

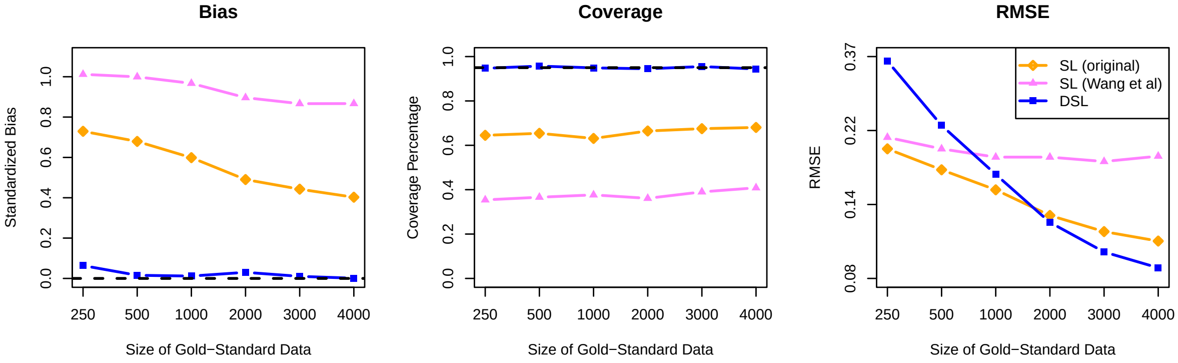

Figure S1: **Comparison of DSL and SL (including Wang, McCormick and Leek (2020)) with CBP data for larger sizes of gold-standard data.** We extend the comparison of the DSL and SL estimators for the CBP data for values of n_R up to 4000, and add an additional state-of-the-art SL method introduced in Wang, McCormick and Leek (2020). We see that the new SL is likewise biased and provides invalid confidence intervals. Moreover, the panels show that as n_R increases, DSL dominates SL in terms of RMSE as well.

3 Ethical Considerations

Given that the research did not involve human subjects, we do not discuss IRB or human subjects research. However we briefly address the impacts and implications of our research from an ethical perspective.

3.1 Social Impact

We conducted this work in a context where several social science works are already advocating for using LLMs as a cost-effective and sufficiently accurate alternative to crowd-sourced human coders (Gilardi, Alizadeh and Kubli, 2023; Ornstein, Blasingame and Truscott, 2022; Törnberg, 2023; Ziems et al., 2023). Although our method is applicable in a wider set of cases than LLMs as surrogates (or even text-as-data), the motivating use case is as a statistical correction to improve the validity of research utilizing LLM labeling in its pipeline. In of itself, we believe that widespread adoption of our method would have a generally positive impact by better aligning research using LLM labeling with the labels that an expert (or other gold standard) would assign.

However, we wish to emphasize to end-users the risk of conflating various forms of *bias*, of which statistical is only one. Our method does not guarantee that the downstream analysis is free of any form of social or psychological bias. Differential biases exhibited by LLMs may remain in finite samples, and our approach especially does not address differential biases in the gold-standard labeling procedure. Researchers should continue to engage critically with their questions, data, and operationalization of concepts and not rely on a statistical procedure to eliminate researcher bias.

3.2 Resource Usage

The experimental validation of the method presented in this paper used a considerable amount of compute (detailed above in Section 2.3.3). Although this is non-trivial, we note that end-users will only need to run the method once, which takes less than a minute on single CPU.

List of Tables

References

Adler, E Scott and John Wilkerson. 2006. "Congressional bills project." *NSF* 880066:00880061.

Althoff, Tim, Cristian Danescu-Niculescu-Mizil and Dan Jurafsky. 2014. "How to Ask for a Favor: A Case Study on the Success of Altruistic Requests." *Proceedings of the International AAAI Conference on Web and Social Media* 8(1):12–21.
URL: *https://ojs.aaai.org/index.php/ICWSM/article/view/14547*

Baly, Ramy, Giovanni Da San Martino, James Glass and Preslav Nakov. 2020. We Can Detect Your Bias: Predicting the Political Ideology of News Articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics pp. 4982–4991.
URL: *https://aclanthology.org/2020.emnlp-main.404*

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

Chakrabarty, Tuhin, Arkadiy Saakyan, Debanjan Ghosh and Smaranda Muresan. 2022. FLUTE: Figurative Language Understanding through Textual Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics pp. 7139–7159.
URL: *https://aclanthology.org/2022.emnlp-main.481*

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *Econometrics Journal* 21:C1 – C68.

Chernozhukov, Victor, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey and James M Robins. 2022. "Locally robust semiparametric estimation." *Econometrica* 90(4):1501–1535.

Chernozhukov, Victor, Mert Demirer, Esther Duflo and Ivan Fernandez-Val. 2018. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Technical report National Bureau of Economic Research.

Danescu-Niculescu-Mizil, Cristian, Lillian Lee, Bo Pang and Jon Kleinberg. 2012. Echoes of Power: Language Effects and Power Differences in Social Interaction. In *Proceedings of the 21st International Conference on World Wide Web*. WWW '12 New York, NY, USA: Association for Computing Machinery p. 699–708.
URL: *https://doi.org/10.1145/2187836.2187931*

Danescu-Niculescu-Mizil, Cristian, Moritz Sudhof, Dan Jurafsky, Jure Leskovec and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers). Sofia, Bulgaria: Association for Computational Linguistics pp. 250–259.
URL: *https://aclanthology.org/P13-1025*

Demszky, Dorottya, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics pp. 2970–3005.
URL: *https://aclanthology.org/N19-1304*

ElSherief, Mai, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury and Diyi Yang. 2021. "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech.".

Gabriel, Saadia, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi and Yejin Choi. 2022. Misinfo Reaction Frames: Reasoning about Readers' Reactions to News Headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics pp. 3108–3127.
URL: *https://aclanthology.org/2022.acl-long.222*

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." *Proceedings of the National Academy of Sciences* 120(30).
URL: *https://doi.org/10.1073/pnas.2305016120*

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1):1–35.

Gross, Justin H., Brice Acree, Yanchuan Sim and Noah A. Smith. 2013. Testing the Etch-a-Sketch Hypothesis: A Computational Analysis of Mitt Romney's Ideological Makeover During the 2012 Primary vs. General Elections. In *APSA 2013 Annual Meeting Paper, American Political Science Association 2013 Annual Meeting*.
URL: *Available at SSRN: https://ssrn.com/abstract=2299991*

Iyyer, Mohit, Peter Enns, Jordan Boyd-Graber and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics pp. 1113–1122.
URL: *https://aclanthology.org/P14-1105*

Kennedy, Edward H, Sivaraman Balakrishnan and Max G'Sell. 2020. "Sharp instruments for classifying compliers and generalizing causal effects." *Annals of Statistics* .

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for

Computational Linguistics pp. 31–41.
URL: *https://aclanthology.org/S16-1003*

Newey, Whitney K and Daniel McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." *Handbook of econometrics* 4:2111–2245.

Ornstein, Joseph T, Elise N Blasingame and Jake S Truscott. 2022. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." Working Paper.

Pilehvar, Mohammad Taher and Jose Camacho-Collados. 2019. WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Minneapolis, Minnesota: Association for Computational Linguistics pp. 1267–1273.
URL: *https://aclanthology.org/N19-1128*

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics.
URL: *http://arxiv.org/abs/1908.10084*

Saravia, Elvis, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu and Yi-Shin Chen. 2018. CARER: Contextualized Affect Representations for Emotion Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* Brussels, Belgium: Association for Computational Linguistics pp. 3687–3697.
URL: *https://aclanthology.org/D18-1404*

Sharma, Ashish, Adam Miner, David Atkins and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP).* Online: Association for Computational Linguistics pp. 5263–5276.
URL: *https://aclanthology.org/2020.emnlp-main.425*

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu and Tie-Yan Liu. 2020. "Mpnet: Masked and permuted pre-training for language understanding." *Advances in Neural Information Processing Systems* 33:16857–16867.

Tay, Yi, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby and Donald Metzler. 2022. "Unifying language learning paradigms." *arXiv preprint arXiv:2205.05131* .

Törnberg, Petter. 2023. "ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning." *arXiv preprint arXiv:2304.06588* .

Vansteelandt, Stijn and Oliver Dukes. 2022. "Assumption-lean Inference for Generalised Linear Model Parameters." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84(3):657–685.
URL: *https://doi.org/10.1111/rssb.12504*

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

Wang, Xuewei, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang and Zhou Yu. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics pp. 5635–5649.
URL: *https://aclanthology.org/P19-1566*

Weller, Orion and Kevin Seppi. 2019. Humor Detection: A Transformer Gets the Last Laugh. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics pp. 3621–3625.
URL: *https://aclanthology.org/D19-1372*

Yang, Diyi, Jiaao Chen, Zichao Yang, Dan Jurafsky and Eduard Hovy. 2019. Let's Make Your Request More Persuasive: Modeling Persuasive Strategies via Semi-Supervised Neural Nets on Crowdfunding Platforms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics pp. 3620–3630.
URL: *https://aclanthology.org/N19-1364*

Zhang, Amy, Bryan Culbertson and Praveen Paritosh. 2017. "Characterizing Online Discussion Using Coarse Discourse Sequences." *Proceedings of the International AAAI Conference on Web and Social Media* 11(1):357–366.
URL: *https://ojs.aaai.org/index.php/ICWSM/article/view/14886*

Zhang, Justine, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics pp. 1350–1361.
URL: *https://aclanthology.org/P18-1125*

Ziems, Caleb, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang and Diyi Yang. 2023. "Can Large Language Models Transform Computational Social Science?" *arXiv preprint arXiv:2305.03514* .