2025-10-09

Benchmarking Debiasing Methods for LLM-based Parameter Estimates

Nicolas Audinet de Pieuchon, Adel Daoud, Connor T. Jerzak, Moa Johansson, Richard Johansson

Chalmers University of Technology

GÖTEBORGS UNIVERSITET

WASP—HS

Motivation

Statistical analysis:

$$P(Y = 1 \mid X) = \sigma(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)$$

Data is often only available as free-form text

Expert Annotations

Generated Annotations

Debiasing methods:

Use small subset of expert annotations to bias-correct model from generated annotations.

Prediction-Powered Inference (PPI)*

Correct imputation estimate with rectifier:

$$\hat{\theta} = \tilde{\theta} - r_\theta$$

Compute rectifier from gradients:

$$r_\theta = \mathbb{E}[\nabla_\theta \ell_\theta(x_i, y_i) - \nabla_\theta \ell_\theta(x_i, \hat{y}_i)]$$

Design-based Supervised Learning (DSL)†

Assume known expert labelling distribution:

$$\pi(x_i, \hat{y}_i) = P(R_i = 1 \mid x_i, \hat{y}_i)$$

Double-robust estimate:

$$\tilde{y}_i = \hat{y}_i - \frac{R_i}{\pi_i}(\hat{y}_i - y_i)$$

$$\mathbb{E}[\tilde{y}_i - y_i \mid x_i] = 0$$

*Angelopoulos, Anastasios N., et al. "Prediction-powered inference." Science 382.6671 (2023): 669-674.

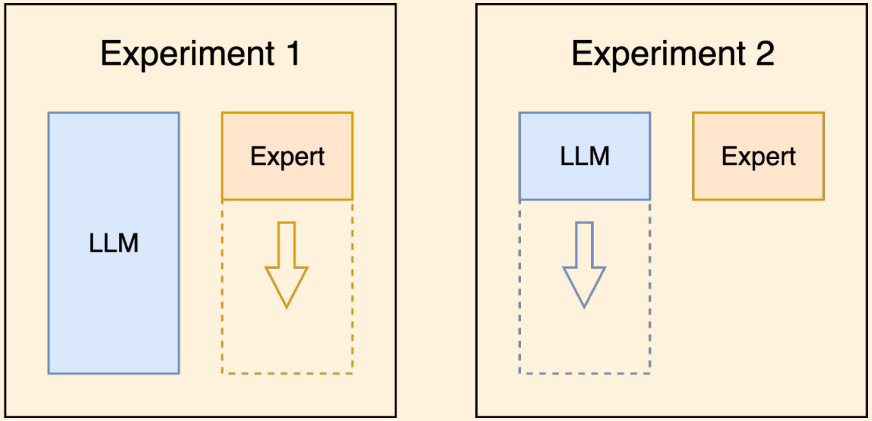
†Egami, Naoki, et al. "Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses." Preprint from November 17 (2024): 2024.

Research Questions

RQ1: When is it preferable to use debiasing methods over just the expert annotations?

RQ2: What are the performance differences between debiasing methods?

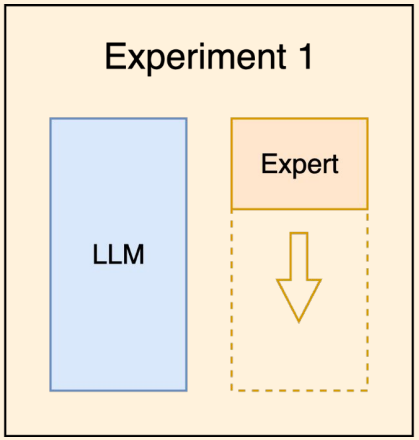
Experiments



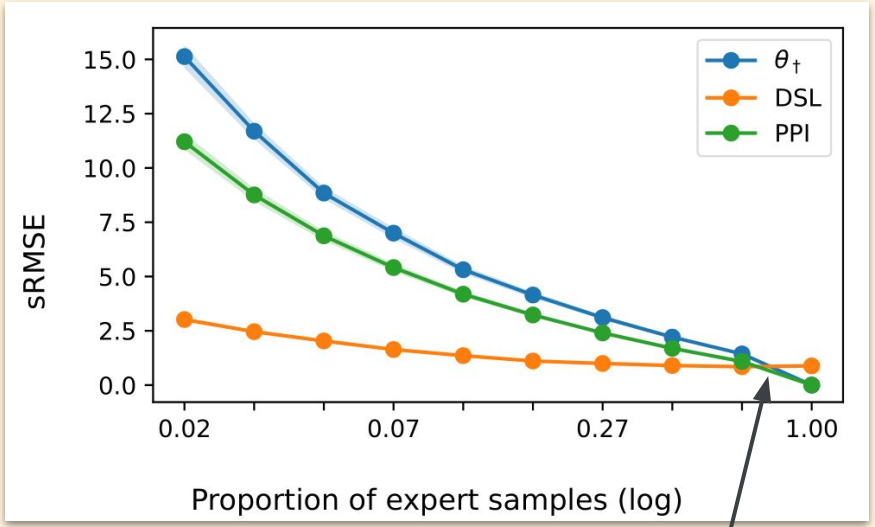
For both experiments:

- Logistic regression
  - 4 discrete input features
  - Binary output
- 4 different LLMs
- 4 different datasets
- Evaluation: standardised RMSE with a reference model

Results

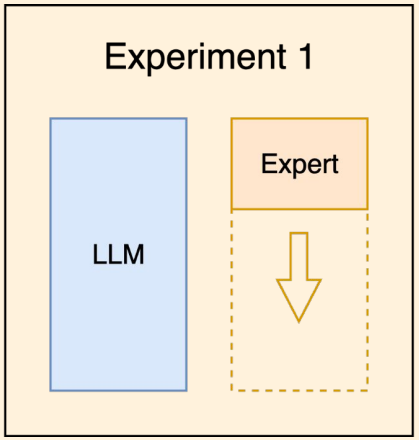


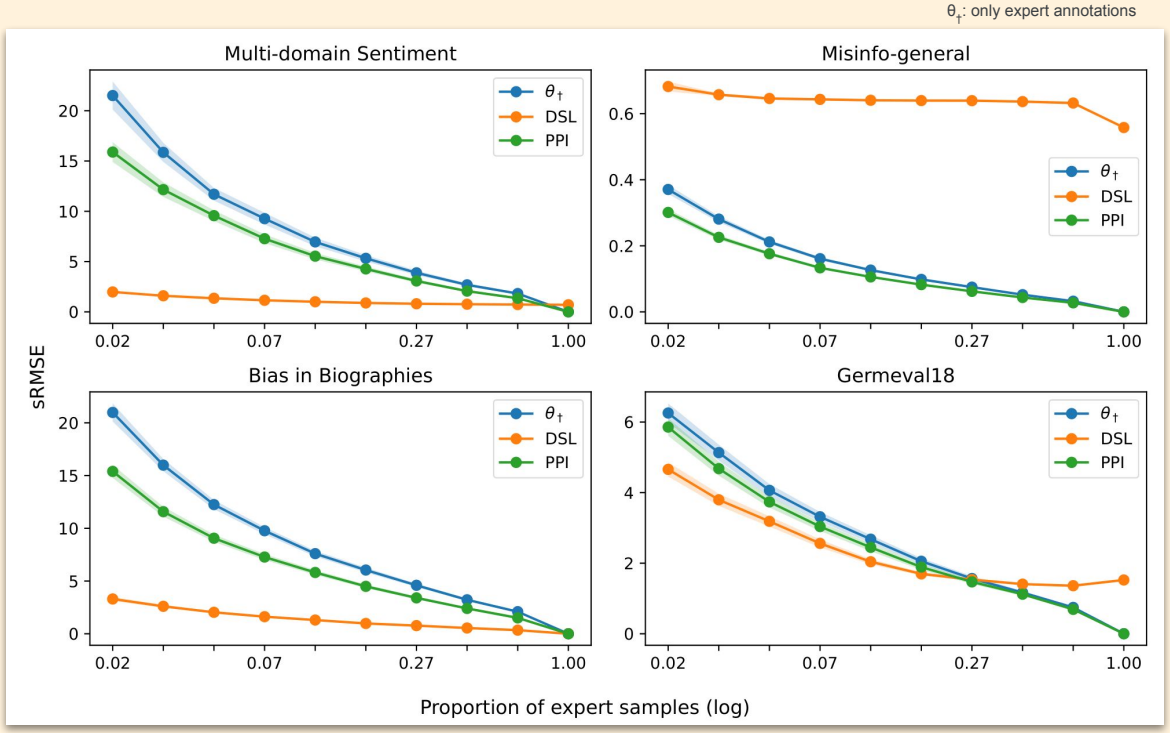
$\theta_\dagger$: only expert annotations



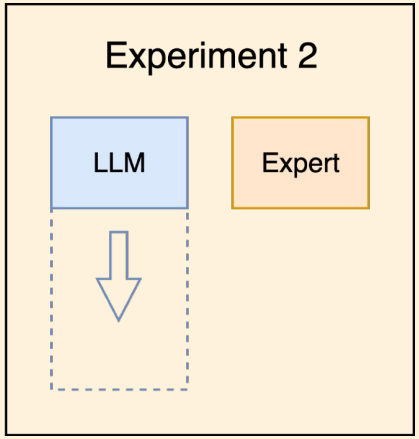
Cross-over effect for DSL!

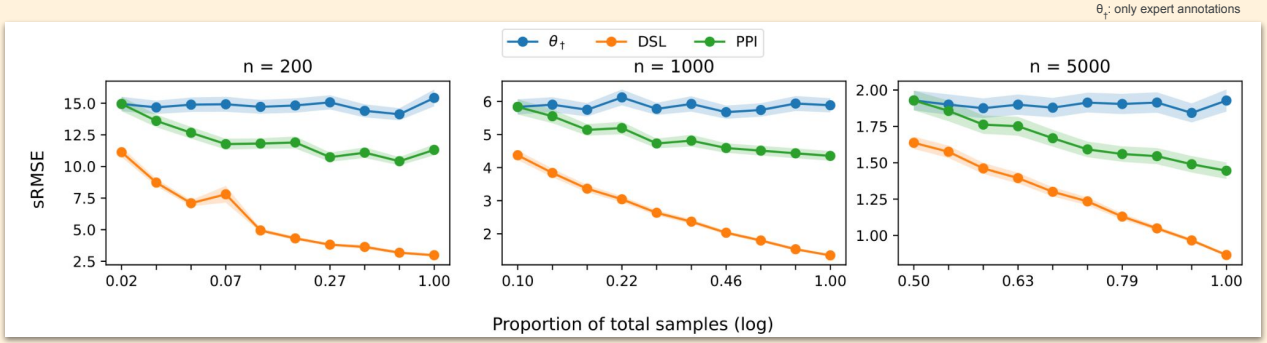
Results


Experiment 1
LLM
Expert


$\theta_\dagger$: only expert annotations
Multi-domain Sentiment
Misinfo-general
Bias in Biographies
Germeval18
$\theta_\dagger$
DSL
PPI
sRMSE
Proportion of expert samples (log)
0.02
0.07
0.27
1.00

Results


Experiment 2
LLM
Expert


θ↑: only expert annotations
θ↑
DSL
PPI
n = 200
n = 1000
n = 5000
sRMSE
15.0
12.5
10.0
7.5
5.0
2.5
6
5
4
3
2
2.00
1.75
1.50
1.25
1.00
0.02
0.07
0.27
1.00
0.10
0.22
0.46
1.00
0.50
0.63
0.79
1.00
Proportion of total samples (log)

Conclusions

RQ1: When is it preferable to use debiasing methods over just the expert annotations?

Both DSL and PPI strictly outperform using only expert annotations.

RQ2: What are the performance differences between debiasing methods?

DSL tends to outperform PPI, but performance is more variable.

Thank you for listening!

Chalmers University of Technology

GÖTEBORGS UNIVERSITET

WASP—HS