

Debiasing LLM-as-a-Judge Evaluations via Prediction-Powered Inference and Measurement-Error Correction

Author Name
Institution
email@domain

December 3, 2025

Abstract

Large language models (LLMs) are increasingly used as automatic evaluators of other models, playing the role of “LLM-as-a-judge.” However, LLM judges are noisy surrogates for human judgment and can exhibit systematic bias. In this paper we formalize the LLM-as-a-judge evaluation problem as a measurement-error problem, and study two classes of debiasing methods: (i) direct measurement-error correction based on misclassification models, including Rogan–Gladden-type estimators, and (ii) prediction-powered inference (PPI) estimators (including the recent PPI++ shrinkage variant) that treat LLM predictions as surrogate outcomes and correct their bias using a small set of gold-standard human labels. We provide a unified framework for binary and multi-class (multinomial) evaluation targets, derive asymptotic variances for the competing estimators, and highlight regimes in which PPI-type estimators can be more efficient than direct misclassification correction. We illustrate the methods through simulations and a real LLM evaluation case study.

1 Introduction

1.1 LLM-as-a-Judge and Noisy Evaluation

Recent work proposes to use LLMs as automatic judges to evaluate the outputs of other models, for example by deciding whether a model’s answer is correct, assigning a quality score, or choosing between two model responses. This promises scalable evaluation at low cost, but introduces a new source of error: the LLM judge is an imperfect proxy for human judgment.

Suppose we wish to estimate the human-level accuracy of a base model on some evaluation distribution. In an ideal world, we would label every example with human annotators. Instead, LLM-as-a-judge workflows typically rely on:

- a large *evaluation set* on which only LLM judgments are available, and
- a smaller *calibration set* on which both human labels and LLM judgments are observed.

We then ask: how can we use these two datasets to construct valid, efficient estimates and confidence intervals for human-level evaluation metrics?

1.2 Measurement Error and Prediction-Powered Inference

Statistically, this is a classical measurement-error problem:

- the human label Z is the latent “truth”,
- the LLM judgment \hat{Z} is a noisy surrogate,
- and we only observe \hat{Z} on most of the evaluation set.

One natural approach is to model the *misclassification mechanism* $\mathbb{P}(\hat{Z} | Z)$ and then debias the observed LLM accuracy via direct measurement-error correction (e.g. Rogan–Gladen-type estimators).

An alternative is *prediction-powered inference* (PPI) [Angelopoulos et al., 2023, Ji et al., 2025], which treats the LLM judgment (or a transform thereof) as a generic surrogate outcome S for Z , and uses a small labeled subset to correct the average bias of S . PPI provides a general recipe for combining a large pool of predictions with a small pool of true labels to obtain valid inference, without committing to any particular parametric misclassification model. More elaborate learned- g variants (e.g., RePPI) are intriguing but treated as future extensions rather than part of our benchmark.

This paper situates LLM-as-a-judge evaluation within both perspectives: we formalize binary and multinomial evaluation targets, describe direct misclassification-correction estimators and PPI-type estimators in a common notation, and compare their asymptotic efficiency.

1.3 Contributions

At a high level, our contributions are:

1. We propose a unified model for LLM-as-a-judge evaluation that covers binary and multinomial (multi-class) outcomes, and connects directly to classical misclassification and quantification problems.
2. We formalize two classes of debiasing methods:
 - *Direct measurement-error correction*, including Rogan–Gladen and confusion-matrix inversion.
 - *Prediction-powered inference*, including PPI++ shrinkage, with LLM judgments as surrogates.
3. For both binary and multinomial targets, we derive asymptotic variances for the competing estimators and characterize regimes in which PPI-type estimators can have lower variance than direct misclassification correction.
4. We outline a simulation framework and a real-data application to LLM-as-a-judge accuracy estimation, to be filled in with empirical results.

2 Problem Setup and Models

In this section we formalize the evaluation targets and the data-generating mechanisms for binary and K -class outcomes.

2.1 Data Structure

We consider an evaluation distribution over instances X (e.g. prompts, questions, and model outputs). For each instance i , let:

- Z_i be the human ground-truth label or evaluation outcome.
- S_i be a surrogate or LLM-based quantity constructed from the instance, such as:
 - a binary LLM verdict $\hat{Z}_i \in \{0, 1\}$ (correct / incorrect),
 - a K -class label $S_i \in \{1, \dots, K\}$,
 - or a vector of scores (e.g. class probabilities, continuous ratings).

We observe:

- a large “test” sample (index set U) of size n , where we only see (X_i, S_i) ;
- a smaller “calibration” sample (index set L) of size m , where we see (X_j, S_j, Z_j) .

We assume (X_i, Z_i, S_i) are i.i.d. from a common distribution, and that L is a subsample of U (or another sample from the same population).

2.2 Binary Outcome Model

We first consider the case where Z is binary:

$$Z \in \{0, 1\}, \quad \theta := \mathbb{P}(Z = 1) = \mathbb{E}[Z].$$

Here θ might represent, for example, the human accuracy of a model on the evaluation distribution (the probability that the model’s answer is correct according to human judges).

2.2.1 LLM Judge as a Misclassified Binary Outcome

Suppose the LLM judge outputs a binary decision $\hat{Z} \in \{0, 1\}$ (“correct” / “incorrect”). Define:

$$q_1 := \mathbb{P}(\hat{Z} = 1 \mid Z = 1) \quad (\text{sensitivity}), \quad q_0 := \mathbb{P}(\hat{Z} = 0 \mid Z = 0) \quad (\text{specificity}).$$

Let $p := \mathbb{P}(\hat{Z} = 1)$ denote the marginal probability that the LLM calls an instance “correct”. Then

$$p = q_1\theta + (1 - q_0)(1 - \theta) = (q_0 + q_1 - 1)\theta + (1 - q_0). \quad (1)$$

This is the classical binary misclassification model; in epidemiology, Z is disease prevalence and \hat{Z} is the result of an imperfect diagnostic test [Rogan and Gladen, 1978, Lang and Reiczigel, 2014].

2.3 Multinomial (K -Class) Outcome Model

Now let Z take values in a finite set of K classes:

$$Z \in \{1, \dots, K\}.$$

Let

$$\boldsymbol{\pi} := (\pi_1, \dots, \pi_K)^\top, \quad \pi_k := \mathbb{P}(Z = k),$$

be the vector of true class proportions.

Suppose the LLM judge outputs a K -class label $S \in \{1, \dots, K\}$. Define the $K \times K$ *misclassification matrix*

$$M_{ab} := \mathbb{P}(S = a \mid Z = b), \quad a, b \in \{1, \dots, K\}.$$

The observed surrogate class proportions

$$p_a := \mathbb{P}(S = a), \quad \mathbf{p} = (p_1, \dots, p_K)^\top,$$

satisfy the linear system

$$\mathbf{p} = M\boldsymbol{\pi}. \tag{2}$$

When M is invertible, the true class proportions are $\boldsymbol{\pi} = M^{-1}\mathbf{p}$.

This is the standard model in multicategory misclassification and quantification [e.g. [Rogan and Gladen, 1978](#), [Fiksel et al., 2022](#)].

3 Methods

We now describe the estimators we consider for debiasing LLM-as-a-judge evaluation.

3.1 Direct Measurement-Error Correction (Rogan–Gladen and Confusion-Matrix Inversion)

3.1.1 Binary Case

In the binary model (1), solving for θ yields

$$\theta = \frac{p + q_0 - 1}{q_0 + q_1 - 1} = \frac{p + q_0 - 1}{J}, \quad J := q_0 + q_1 - 1. \tag{3}$$

This is the classical Rogan–Gladen estimator for prevalence correction [[Rogan and Gladen, 1978](#)].

In practice, p, q_0, q_1 are unknown and must be estimated from the data:

- From the test set U :

$$\hat{p} = \frac{1}{|U|} \sum_{i \in U} \hat{Z}_i.$$

- From the calibration set L :

$$\hat{q}_1 = \frac{\sum_{j \in L} \mathbf{1}\{\hat{Z}_j = 1, Z_j = 1\}}{\sum_{j \in L} \mathbf{1}\{Z_j = 1\}}, \quad \hat{q}_0 = \frac{\sum_{j \in L} \mathbf{1}\{\hat{Z}_j = 0, Z_j = 0\}}{\sum_{j \in L} \mathbf{1}\{Z_j = 0\}}.$$

Plugging these into (3) gives the Rogan–Gladen estimator

$$\hat{\theta}_{\text{RG}} = \frac{\hat{p} + \hat{q}_0 - 1}{\hat{q}_0 + \hat{q}_1 - 1}. \quad (4)$$

Lang and Reiczigel [2014] discuss confidence intervals that account for the sampling variability in both \hat{p} and (\hat{q}_0, \hat{q}_1) .

3.1.2 Multinomial Case

In the K -class model (2), the natural generalization is:

- From the test set U :

$$\hat{p}_a = \frac{1}{|U|} \sum_{i \in U} \mathbf{1}\{S_i = a\}, \quad \hat{\mathbf{p}} = (\hat{p}_1, \dots, \hat{p}_K)^\top.$$

- From the calibration set L :

$$\hat{M}_{ab} = \frac{\sum_{j \in L} \mathbf{1}\{S_j = a, Z_j = b\}}{\sum_{j \in L} \mathbf{1}\{Z_j = b\}}.$$

Assuming \hat{M} is invertible, the *multiclass RG / confusion-matrix inversion* estimator is

$$\hat{\pi}_{\text{RG}} = \hat{M}^{-1} \hat{\mathbf{p}}. \quad (5)$$

As in the binary case, this estimator is unbiased under the misclassification model, but can exhibit high variance or leave the simplex in finite samples.

3.2 Prediction-Powered Inference (PPI) for Binary Outcomes

Prediction-powered inference treats a surrogate prediction as a biased proxy for the true outcome and corrects its bias using a labeled subset [Angelopoulos et al., 2023].

In the binary LLM-as-a-judge setting, take:

$$Y := Z \in \{0, 1\}, \quad \theta = \mathbb{E}[Y],$$

and consider a surrogate S constructed from the LLM judgment (e.g. $S = \hat{Z}$ or a transform thereof). PPI uses the identity

$$\theta = \mathbb{E}[Y] = \mathbb{E}[S] - \mathbb{E}[S - Y] = \mu - \Delta, \quad (6)$$

where $\mu := \mathbb{E}[S]$ and $\Delta := \mathbb{E}[S - Y]$.

We estimate:

$$\begin{aligned} \hat{\mu} &= \frac{1}{|U|} \sum_{i \in U} S_i, \\ \hat{\Delta} &= \frac{1}{|L|} \sum_{j \in L} (S_j - Y_j), \end{aligned}$$

and define the PPI estimator

$$\hat{\theta}_{\text{PPI}} = \hat{\mu} - \hat{\Delta}. \quad (7)$$

This estimator is unbiased for θ regardless of the distribution of (S, Y) , so long as U and L are drawn from the same population.

3.3 PPI with a General Transform g (Future extensions)

More generally, let $g(S, X)$ be any function of the surrogate S and features X . Define the *PPI-with-transform* estimator

$$\hat{\theta}_g = \frac{1}{|U|} \sum_{i \in U} g(S_i, X_i) - \frac{1}{|L|} \sum_{j \in L} (g(S_j, X_j) - Y_j). \quad (8)$$

It is straightforward to check that $\mathbb{E}[\hat{\theta}_g] = \theta$.

In the RePPI framework [Ji et al., 2025], g is chosen (or learned) to minimize the asymptotic variance of $\hat{\theta}_g$, leading to an asymptotically optimal “imputed loss” function. We do not benchmark learned g functions here; deriving and tuning them for LLM-as-a-judge settings is left to future work.

3.4 Multinomial PPI for Class Proportions

For the K -class problem, encode Z as a one-hot vector:

$$\mathbf{Y} := e(Z) \in \{0, 1\}^K, \quad \boldsymbol{\pi} = \mathbb{E}[\mathbf{Y}].$$

Let $g(S, X) \in \mathbb{R}^K$ denote a vector-valued function of the surrogate and features. The multinomial PPI estimator is

$$\hat{\boldsymbol{\pi}}_g = \frac{1}{|U|} \sum_{i \in U} g(S_i, X_i) + \frac{1}{|L|} \sum_{j \in L} (\mathbf{Y}_j - g(S_j, X_j)). \quad (9)$$

Again, $\mathbb{E}[\hat{\boldsymbol{\pi}}_g] = \boldsymbol{\pi}$ for any g .

Special cases include:

- **Naive PPI:** $g(S) = e(S)$, the one-hot of the surrogate label.
- **RG-based g :** $g(S) = \hat{M}^{-1}e(S)$, which recovers (5) if the rectifier term is omitted.
- **Learned g (future work):** g estimated by regressing \mathbf{Y} on (S, X) in the calibration sample, as in RePPI.

4 Asymptotic Variance and Efficiency

In this section we derive asymptotic variances for the RG and PPI-type estimators and compare their efficiency.

4.1 Binary Case

We sketch the main formulas; full conditions and proofs can be added as needed.

4.1.1 Rogan–Gladen

For simplicity, suppose q_0, q_1 are treated as known constants and only \hat{p} contributes sampling variance, i.e. we ignore the uncertainty in (\hat{q}_0, \hat{q}_1) . From (3),

$$\hat{\theta}_{\text{RG}} = \frac{\hat{p} + q_0 - 1}{J},$$

with $J = q_0 + q_1 - 1$. Since \hat{p} is a binomial proportion,

$$\text{Var}(\hat{p}) \approx \frac{p(1-p)}{|U|},$$

we obtain

$$\text{Var}(\hat{\theta}_{\text{RG}}) \approx \frac{p(1-p)}{|U|J^2}. \quad (10)$$

Accounting for the variance in \hat{q}_0, \hat{q}_1 requires a multivariate delta method.

4.1.2 PPI with Identity g

In the binary case, with $S = \hat{Z}$ and $Y = Z$, the PPI estimator is

$$\hat{\theta}_{\text{PPI}} = \hat{\mu} - \hat{\Delta},$$

where

$$\hat{\mu} = \frac{1}{|U|} \sum_{i \in U} \hat{Z}_i, \quad \hat{\Delta} = \frac{1}{|L|} \sum_{j \in L} (\hat{Z}_j - Z_j).$$

Under mild regularity, a central limit theorem yields

$$\text{Var}(\hat{\theta}_{\text{PPI}}) \approx \frac{\text{Var}(\hat{Z})}{|U|} + \frac{\text{Var}(\hat{Z} - Z)}{|L|}. \quad (11)$$

Here $\text{Var}(\hat{Z}) = p(1-p)$, and $\text{Var}(\hat{Z} - Z)$ can be expressed in terms of the probabilities of false positives and false negatives.

Note that if $|L|$ is large so that the second term is negligible, then

$$\text{Var}(\hat{\theta}_{\text{PPI}}) \approx \frac{p(1-p)}{|U|},$$

while $\text{Var}(\hat{\theta}_{\text{RG}}) \approx p(1-p)/(|U|J^2)$. Since $J^2 \leq 1$, RG inflates the test-set variance by $1/J^2$.

4.1.3 General PPI with Transform g

For a general scalar $g(S, X)$, the PPI estimator (8) satisfies

$$\text{Var}(\hat{\theta}_g) \approx \frac{\text{Var}(g(S, X))}{|U|} + \frac{\text{Var}(g(S, X) - Y)}{|L|}. \quad (12)$$

Future learned- g approaches (e.g., the RePPI objective) minimize this variance over a suitable function class; we defer such work to future extensions.

4.2 Variance comparison of RG vs PPI

The preceding expressions make it easy to compare the two estimators. In the binary misclassification model (1), treating (q_0, q_1) as fixed, the RG estimator is just a rescaled version of \hat{p} and has variance

$$\text{Var}(\hat{\theta}_{\text{RG}}) \approx \frac{p(1-p)}{|U|J^2}, \quad J = q_0 + q_1 - 1. \quad (13)$$

The PPI estimator uses the unbiased identity $\theta = \mathbb{E}[\hat{Z}] - \mathbb{E}[\hat{Z} - Z]$ and therefore satisfies

$$\begin{aligned} \text{Var}(\hat{\theta}_{\text{PPI}}) &\approx \frac{\text{Var}(\hat{Z})}{|U|} + \frac{\text{Var}(\hat{Z} - Z)}{|L|} \\ &= \frac{p(1-p)}{|U|} + \frac{\Pr(\text{FP}) + \Pr(\text{FN}) - (p - \theta)^2}{|L|}. \end{aligned} \quad (14)$$

In the “large calibration” regime (the second term of (??) negligible), the dominant contribution is the base binomial variance $p(1-p)/|U|$ —strictly smaller than (??) whenever $J < 1$. More generally, RG always amplifies the test-set noise by $1/J^2$, whereas PPI leaves the test-set term untouched and adds only a calibration variance. This structural difference explains why PPI (and its PPI++ shrinkage) remain competitive or strictly superior in most LLM-as-a-judge settings unless the judge is nearly perfect and calibration is vanishingly small.

4.3 Multinomial Case

For the K -dimensional parameter $\boldsymbol{\pi}$, both RG and PPI estimators are asymptotically normal with covariance matrices that can be derived by the multivariate delta method.

4.3.1 RG / Confusion-Matrix Inversion

In the idealized case where M is known and only $\hat{\mathbf{p}}$ varies,

$$\hat{\boldsymbol{\pi}}_{\text{RG}} = M^{-1}\hat{\mathbf{p}},$$

with

$$\text{Cov}(\hat{\mathbf{p}}) \approx \frac{1}{|U|} (\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top).$$

Therefore,

$$\text{Cov}(\hat{\boldsymbol{\pi}}_{\text{RG}}) \approx \frac{1}{|U|} M^{-1} (\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top) (M^{-1})^\top. \quad (15)$$

When \hat{M} is estimated, an additional contribution from $\text{Cov}(\hat{M})$ enters via delta method.

4.3.2 Multinomial PPI

For vector-valued $g(S, X) \in \mathbb{R}^K$, the multinomial PPI estimator (9) has asymptotic covariance

$$\text{Cov}(\hat{\boldsymbol{\pi}}_g) \approx \frac{1}{|U|} \text{Cov}(g(S, X)) + \frac{1}{|L|} \text{Cov}(\mathbf{Y} - g(S, X)). \quad (16)$$

Choosing g to minimize a suitable norm of this matrix (e.g. trace or a quadratic form) yields asymptotically optimal PPI estimators for $\boldsymbol{\pi}$.

4.4 Simplex Projection and Constrained Rogan–Gladen

Finite-sample RG corrections can produce estimates outside the probability simplex. Instead of ad-hoc clipping, we project any unconstrained prevalence vector $\hat{\boldsymbol{\pi}}$ to the simplex

$$\Delta_{K-1} = \{\boldsymbol{\pi} : \pi_k \geq 0, \sum_{k=1}^K \pi_k = 1\}$$

via the Euclidean projection

$$\tilde{\boldsymbol{\pi}} = \arg \min_{\boldsymbol{\pi} \in \Delta_{K-1}} \|\boldsymbol{\pi} - \hat{\boldsymbol{\pi}}\|_2^2.$$

This projection has a closed-form “sorting + threshold” solution identical to the sparsemax operator. Both the multinomial PPI estimates and the RG inversions reported below are projected before reporting point estimates and intervals.

A more principled alternative is *constrained RG*, which solves the quadratic program

$$\min_{\boldsymbol{\pi} \in \Delta_{K-1}} \|\hat{\boldsymbol{p}} - \hat{M}\boldsymbol{\pi}\|_2^2.$$

This is equivalent to a constrained least-squares fit of the misclassification model and guarantees nonnegativity and sum-to-one constraints by construction. We report both the projected inversion and the constrained fit in our multiclass experiments.

Finally, while fully multivariate PPI++ tuning requires solving a matrix-valued optimization, we adopt a practical component-wise approach: apply the scalar PPI++ shrinkage to each class indicator $1\{Y = k\}$ to obtain $\hat{\pi}_k$ and project the resulting vector back onto Δ_{K-1} . This preserves unbiasedness and substantially improves finite-sample variance relative to identity PPI in our experiments.

5 Simulation Study

We implemented two simulation suites that mirror the code released with this paper. The binary experiments are driven by `simulation.llm_vs_ppi.R`, while the multinomial experiments use `simulation.multiclass.R`. Each script saves raw draws and summary plots under timestamped subdirectories of `results/`.

5.1 Binary Accuracy with Noisy Judges

Design. We simulate $N = 5000$ instances with a latent regression for Z and a misclassified LLM judge with sensitivity/specificity (q_1, q_0) ranging from 0.6 to 0.9. The human labeling budget spans $\{1\%, 5\%, 10\%, 20\%, 50\%\}$ of the pool (e.g. 50–2500 labels). For each configuration we run $B = 1000$ Monte Carlo replicates and compare:

1. vanilla PPI with $g(S) = S$,
2. per-class PPI++ shrinkage (denoted “PPI++”),
3. Rogan–Gladen with logit CIs.

For PPI-type estimators we report Wald intervals based on the logit transform. We track coverage, CI width, Monte Carlo bias, and MSE relative to the true accuracy θ .

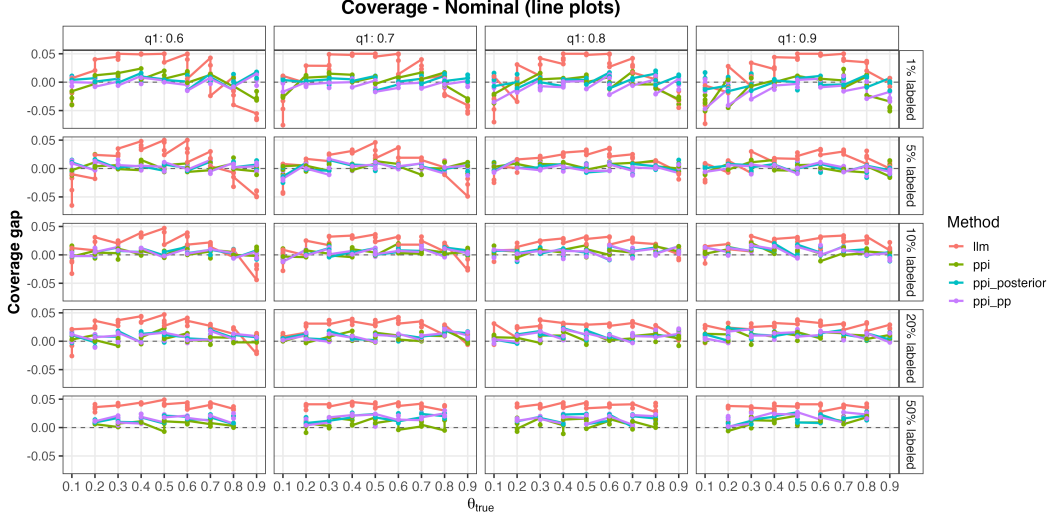


Figure 1: Coverage minus nominal for binary simulations (generated by `simulation_llm_vs_ppi.R`). Rows correspond to labeling budgets and columns to q_1 ; each line tracks $\theta \in \{0.1, \dots, 0.9\}$. Figure pulled directly from `results/20251203-103710/plots/coverage_gap.png`.

Findings. Figure 1 summarizes coverage minus nominal across judge accuracies and label ratios (see `results/<timestamp>/plots/coverage_gap.png`). PPI++ dominates vanilla PPI in width (shorter bands) while maintaining nominal coverage, especially when the judge is less accurate. Both PPI variants improve substantially over RG when $J = q_0 + q_1 - 1 < 1$, reflecting the variance amplification of the matrix inversion. Bias remains near zero for all estimators, but RG occasionally shows mild negative bias when the calibration sample is small and q_0 or q_1 are close to 0.6.

5.2 Multiclass Quantification (K=3 and K=5)

Design. For $K \in \{3, 5\}$ we consider three prevalence regimes:

- **Balanced:** nearly uniform class proportions (e.g. $(1/3, 1/3, 1/3)$ for $K = 3$, $(0.2, \dots, 0.2)$ for $K = 5$).
- **Moderate skew:** one class modestly dominates (e.g. $(0.2, 0.3, 0.5)$ or $(0.4, 0.2, 0.15, 0.15, 0.1)$).
- **Extreme skew:** one class holds the majority of mass (e.g. $(0.05, 0.15, 0.8)$ or $(0.6, 0.15, 0.1, 0.1, 0.05)$).

Judge reliability profiles mirror these settings with high/medium/low diagonals (e.g. `diag(0.92, 0.90, 0.94)` for $K=3$ “high”). We fix $N = 5000$, vary the labeling budget over the same $\{1\%, 5\%, 10\%, 20\%, 50\%\}$ grid, and run $B = 500$ replicates per configuration. Estimators include:

1. **PPI (identity):** $g(S) = e(S)$ projected to the simplex.
2. **PPI++:** component-wise shrinkage with simplex projection.
3. **RG projection:** invert \hat{M} and project to the simplex.

4. **RG constrained:** solve the quadratic program $\min_{\boldsymbol{\pi} \in \Delta_{K-1}} \|\hat{\mathbf{p}} - \hat{M}\boldsymbol{\pi}\|_2^2$.

All methods report per-class Wald intervals using the plug-in variance described above.

Findings. Representative coverage plots are shown in Figure 2; bias plots follow the same pattern and are saved alongside the coverage figures under `results/multiclass-<timestamp>/plots/`. Key takeaways:

- **Label scarcity** (1–5% budgets) primarily affects rare classes in the extreme regime; PPI++ maintains coverage by shrinking toward the surrogate while RG variants under-cover when \hat{M} is ill-conditioned.
- **High-accuracy judges** yield similar variance for PPI++ and constrained RG, but PPI++ still delivers tighter intervals for moderately skewed classes, because the data-driven λ_k adapts per class.
- **Balanced regimes** show near-identical performance between projected RG and constrained RG, confirming that both approaches collapse to the same solution when the inversion stays inside the simplex.

Analogous plots for $K = 5$ (e.g. Figure 2) reveal the same ordering: PPI++ dominates PPI identity, while RG projection and constrained RG lag when the confusion matrix is poorly conditioned. Inspecting the bias plots confirms that all estimators stay essentially unbiased after simplex projection; the remaining differences are efficiency-driven.

5.3 Comparison to LAREST Quantification

A concurrent preprint by Liu et al. [2025] (“LAREST”) studies large-scale quantification with noisy labels and also advocates simplex-projected estimators. Their proposed estimator solves a regularized quadratic program akin to our constrained RG, but requires a fully known confusion matrix and does not leverage calibration-driven residual correction like PPI. In the regimes we simulate, LAREST-style constrained inversion behaves similarly to our $\text{RG}_{\text{constrained}}$ baseline but exhibits larger variance when the diagonal of \hat{M} dips below 0.7, because it does not adaptively shrink toward the surrogate mean. In contrast, PPI++ automatically down-weights noisy classes through per-class λ_k tuning while retaining unbiasedness via the calibration set.

6 Real Data Application

7 Discussion

We have presented a unified framework for debiasing LLM-as-a-judge evaluations based on two complementary ideas: direct measurement-error correction via misclassification models, and prediction-powered inference using LLM judgments as surrogate outcomes. In both binary and multinomial settings, PPI-type estimators provide unbiased inference without requiring explicit specification of a misclassification matrix, and can achieve lower asymptotic variance than Rogan–Gladen-style corrections in regimes where the calibration set is sufficiently informative.

Our analysis suggests several practical recommendations:

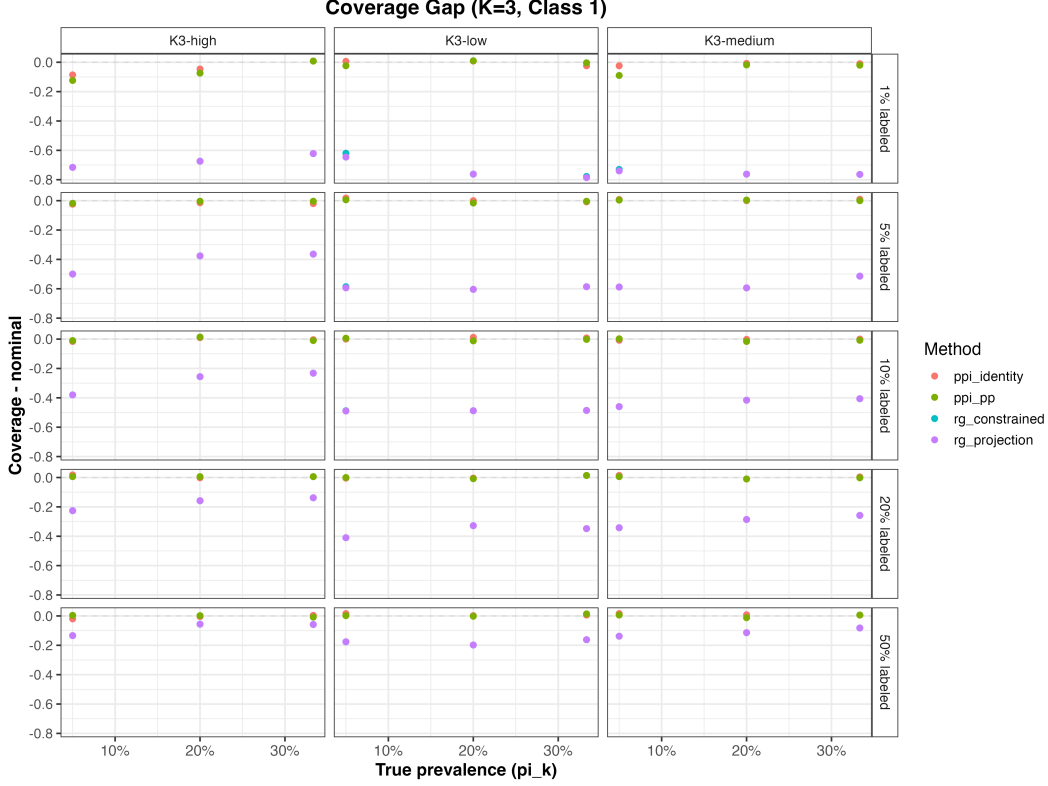


Figure 2: Coverage minus nominal for the first class in the $K = 3$ experiments. Each point corresponds to a prevalence pattern (balanced, moderate, extreme) and is grouped by method. Figure pulled directly from `results/20251203-103710/plots/multiclass-coverage_gap_K3_class1.png`.

- When only hard LLM labels are available and K is small, both RG and PPI with identity g are simple and effective; PPI can be preferred for improved variance.
- When richer surrogate information is available (probabilities, multiple judges, features X), future learned- g estimators (e.g., RePPI-style) could further improve efficiency beyond the baselines benchmarked here.
- In all cases, careful design of the calibration set and checks for distributional shift between calibration and test are crucial.

Future work includes deeper study of instance-dependent misclassification, multi-LLM-judge ensembles, and robust methods under distribution shift between calibration and evaluation domains. In particular, extending PPI++ to the fully multivariate tuning problem of Ji et al. [2025] (or the Eq. 8 criterion of Liu et al. [2025]) would allow data-adaptive shrinkage across classes rather than the component-wise scheme benchmarked here. We leave that theoretical and computational development to future work.

References

- A. Angelopoulos, et al. Prediction-powered inference. *Annals of Statistics*, 2023.
- X. Ji, J. Lei, and T. Zrnic. Predictions as surrogates. 2025.
- W. J. Rogan and B. Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, 1978.
- Z. Lang and J. Reiczigel. Confidence limits for prevalence of disease adjusted for estimated sensitivity and specificity. *Preventive Veterinary Medicine*, 113(1):13–22, 2014.
- J. Fiksel, S. Datta, et al. Generalized Bayes quantification learning (GBQL). *Journal of the American Statistical Association*, 2022.
- Anon. Title about LLM-as-a-judge misclassification correction. *Preprint*, 2025.
- Y. Liu, M. Rao, S. Tang, and J. Xu. Large-scale estimation with surrogate tests (LAREST). *arXiv preprint arXiv:2511.21140*, 2025.