BACON: BUDGETED HUMAN CALIBRATION FOR MODELING AND EVALUATION WITH MULTIPLE AI JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

AI judges are increasingly used to reduce the cost of human evaluation. While they provide a scalable and inexpensive alternative, their outputs can be biased relative to human preferences and highly item-dependent, with substantial variation across judges, tasks, and domains. When practitioners rely on uncalibrated AI evaluations for model ranking, item scoring, or quality reporting, systematic bias can propagate directly into downstream decisions and make the conclusions unreliable. We propose BACON, a four-stage pipeline that combines budgeted human calibration with multiple AI-judge outputs to obtain more accurate annotations. BACON first constructs full-coverage auxiliary features for every item in the evaluation pool, including multi-judge scores, token-level uncertainty statistics, and contextual embeddings. It then obtains human labels for a small sampled subset and trains a cross-fitted outcome model to produce calibrated item-level surrogate predictions. These predictions support two downstream modes: the first is population-level estimation of summary metrics, such as means, quantiles, or other estimands, using an augmented estimating-equation estimator with valid confidence intervals; and the second is individual-level surrogate scoring for item scoring and ranking. Throughout, BACON treats AI judges as auxiliary measurements rather than ground truth: human labels provide the calibration anchor, while AI-derived signals serve as surrogate predictions and improve efficiency. We validate BACON on a variety of tasks and domains. Across settings and labeling budgets, the cross-fitted outcome model improves predictive accuracy and ranking consistency, while the estimating-equation estimator reduces bias and variance relative to raw AI outputs and purely human-label-based methods. These results suggest that BACON provides a practical framework for scalable, statistically grounded evaluation with limited human annotation.

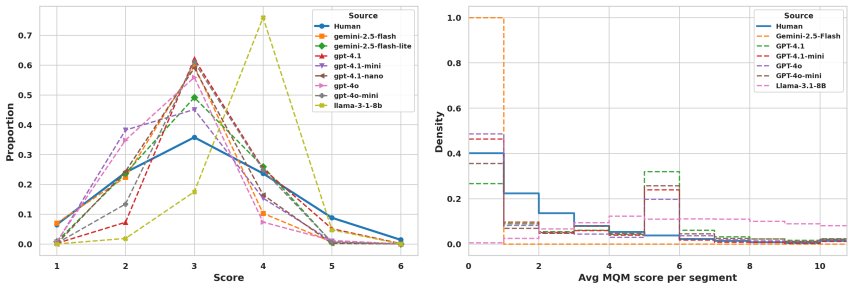
1 INTRODUCTION

1.1 BACKGROUND: BIAS IN AI-ASSISTED EVALUATIONS

Many important decisions rely on large-scale human evaluations: grading student writing, scoring persuasive essays, and rating the aesthetics, trustworthiness, and usability of websites, among others. Human labels are expensive, especially when the goal is not only to score individual items but also to estimate population-level summary statistics, such as the mean quality of a corpus, under a fixed budget. Using generative AI as scalable judges has become a popular practice, providing low-cost scores for large item pools. However, AI-judge outputs can be systematically biased relative to human preferences, and their errors are often item-dependent and model-specific. Below, we give three examples showing that this mismatch appears across tasks and modalities.

Case Study 1: Essay scoring with PERSUADE Dataset. The PERSUADE dataset (Crossley et al., 2024) contains over 25,000 argumentative essays from U.S. students in grades 6–12, each assigned a holistic human score from 1 to 6. We prompt several LLMs to score the same essays and compare their outputs to the human labels (prompt in Appendix B). Figure 1a compares human and LLM score distributions on PERSUADE. Human scores are more heavy-tailed than the AI scores, and some AI models (e.g. LLAMA-3.1-8b) concentrate around an entirely different peak.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: Score distribution comparisons between human raters and AI judges for PERSUADE and MQM. Panel (a): Grade-10 PERSUADE score proportions for human labels and all 8 LLM judges. Human shown as solid line; LLMs as dashed lines. Panel (b): MQM segment-level score distributions for human raters and six LLM judges on WMT 2020 en–de. Both datasets show systematic disagreement between human labels and AI judges.

Case Study 2: Machine translation evaluation. We also study the WMT 2020 English–German MQM dataset (Freitag et al., 2021), where expert raters annotate translation errors by category and severity and each segment receives a non-negative MQM score (lower is better). We use six LLM judges (listed in Figure 1b) to produce MQM-style annotations across ten MT systems, including three human translations. Figure 1b shows systematic miscalibration: some LLM judges largely underestimate or overestimate error rates (Gemini-2.5-Flash and Llama-3.3-70B), while others demonstrate a two peak distribution that is not observed in human labels.

Case Study 3: Web-design scoring. Finally, we consider visual evaluation of website screenshots, where models must judge perceptual qualities such as aesthetics, trustworthiness, typicality, and usability (Miniukovich & Figl, 2023). Figure 2 shows the score distributions for the Universities subset of the WebDesign dataset and reveals substantial disagreement between human labels and VLM judges.

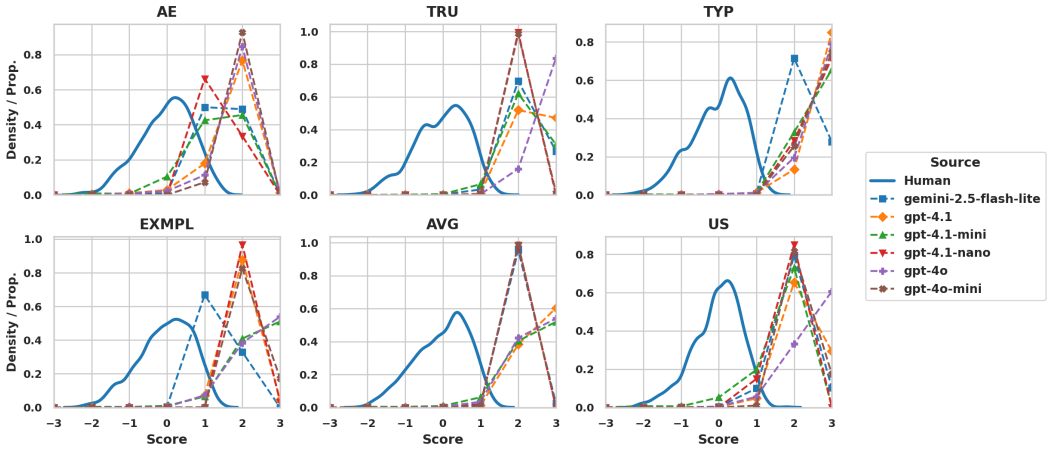


Figure 2: WebDesign Universities score distributions for six perceptual outcomes (scores -3 to 3). Human scores are shown as KDE curve; VLM judges as proportion lines.

Together, these examples show a common problem: AI scores usually don't target the same measurement as the human labels. If practitioners simply average AI scores or use them to rank items or systems, they are making an implicit and usually unvalidated measurement assumption. We propose BACON to address this problem.

1.2 BACON FRAMEWORK OVERVIEW AND OUR CONTRIBUTIONS

BACON is an evaluation protocol for making human-aligned claims under limited labels. For a target item population, the evaluator specifies the human outcome and the evaluation report of interest, such as a population mean, system ranking, or item-level score. BACON collects full-coverage auxiliary measurements from multiple AI judges, including scores, uncertainty statistics, and item embeddings. A small portion of samples is then labeled by humans and used to fit a cross-fitted calibration model from AI measurements to the target human outcome. BACON reports calibrated item-level surrogate scores and aggregate human-aligned estimates with uncertainty intervals. Throughout the protocol, AI judges are treated as auxiliary measurements rather than replacements for human labels: human calibration anchors the validity of the final claim, while AI measurements improve efficiency when they are predictive.

To summarize, we make the following contribution: (1) **Human-calibrated AI-judge evaluation protocol.** We formulate AI-assisted evaluation as a budgeted human-calibration problem in which multiple AI judges are treated as auxiliary measurements, not ground truth. AI judge scores provide full-coverage outcome predictions, and a small uniform human sample supplies a bias correction, decoupling estimation accuracy from the quality of the outcome model. (2) **Multi-judge calibration recipes.** BACON distinguishes two evaluation tasks: empirically calibrated item-level surrogate scoring and statistically valid summary metric estimation. For item-level evaluation, we propose building outcome models from multiple judge scores, uncertainty summaries, and context embeddings, and discuss practical model choices for different type of outcomes such as ordinal, zero-inflated, continuous, and categorical evaluation outcomes. For summary statistics report, BACON follows an estimating equation-based statistical framework to deliver estimators that are consistent for human assessment and provide valid uncertainty quantification through confidence intervals. (3) **Cross-domain empirical study of AI-judge calibration.** We validate BACON across three real-world datasets spanning multiple tasks and modalities, including PERSUADE essay scoring, MQM machine translation evaluation, and WebDesign web-design scoring. The results demonstrate substantial improvements in judge-model predictive accuracy for individual item quality and the estimation efficiency for summary metrics over several baselines.

1.3 RELATED WORKS

LLM-as-a-judge. Zheng et al. (2023) establish the canonical LLM-as-judge paradigm via MT-Bench and Chatbot Arena; Liu et al. (2023) extend it with chain-of-thought prompting for text-generation evaluation; and Dubois et al. (2024) develop AlpacaEval for instruction-following comparison. Despite their utility, LLM judges exhibit systematic biases in terms of position, verbosity, and self-enhancement (Wang et al., 2024a), that distort rankings; surveys (Li et al., 2024a; Gu et al., 2026) catalog these failure modes. BACON highlights such biases as a starting point and provides a principled calibration framework to correct them using a small sample of human labeled data.

Multi-judge aggregation. Self-consistency (Wang et al., 2023) and multi-agent debate (Du et al., 2024; Liang et al., 2024) use ensembles of reasoning paths or models to improve reliability. In the evaluation setting, ChatEval (Chan et al., 2023), jury-style panels (Verga et al., 2024), and structured multi-agent collaboration (Qian et al., 2026) show that aggregating diverse judges reduces individual biases. BACON differs: rather than debate or ad-hoc ensembling, we treat judge scores as features in a supervised outcome model fit against human labels, yielding statistically grounded estimates.

Prediction-powered inference (PPI) and semi-supervised estimation. Angelopoulos et al. (2023a) introduce PPI, using model predictions plus a small labeled set to form valid confidence intervals; Angelopoulos et al. (2023b) extend this to PPI++ with adaptive weighting; Zrnic & Candès (2024) remove the held-out requirement via cross-fitting. These ideas root in classical survey regression estimators (Särndal et al., 1992; Cassel et al., 1976) and semi-supervised inference (Zhang et al., 2019; Cai & Guo, 2020). In the application of summary metrics estimation, BACON follows directly this line of work. For the special case of estimating a population mean, our estimator is mathematically equivalent to a PPI / regression-adjusted estimator with a learned prediction function. The contribution of BACON is therefore not a new mean estimator. Instead, BACON addresses a problem left open by generic PPI formulations: how to construct reliable auxiliary predictions in AI-assisted evaluation when the available predictors are multiple biased AI judges, uncertainty statistics, as well as item embeddings, and how to report both aggregate inferential quantities and item-level

surrogate scores without treating AI judgments as ground truth, which serves as the protocol layer between LLM-as-judge and PPI.

Training and fine-tuning LLM judges. JudgeLM (Zhu et al., 2025), Prometheus (Kim et al., 2024a;b), and PandaLM (Wang et al., 2024b) fine-tune open-source models as scalable evaluators, while Li et al. (2024b) produce evaluative rationales rather than scalar scores. However, Huang et al. (2025) find that fine-tuned judges often underperform GPT-4 out-of-domain. BACON is training-free: off-the-shelf judge outputs feed a lightweight calibration layer, requiring no judge-specific training data and supporting closed-source models.

2 BACON Framework

2.1 Overview and notation setup

We present a training-free pipeline for AI-assisted evaluation. BACON integrates three components to mitigate bias in AI judge outputs: (1) aggregating predictions from multiple judges to capture cross-judge variability; (2) leveraging contextual features to debias judge predictions; and (3) calibrating AI evaluations using a small sample of human-labeled data.

We study AI-assisted modeling and evaluation under limited human labeling budgets. Suppose we have a pool of items, each associated with a scalar or vector-valued human outcome. Our goal is to estimate population-level performance and improve item-level prediction by combining AI judge outputs in a principled way, with a small amount of human labels for calibration.

Formally, for item $i \in \{1, \ldots, N\}$, let y_i denote the human score (or one outcome component), and let $\ell_i \in \mathbb{R}^M$ collect M model-judge scores. We also construct contextual embedding features e_i (e.g., text embeddings for essays and image embeddings for screenshots). We then fit predictive models for y_i using judge scores, uncertainty features, and embedding features to capture their relationship with human evaluations.

2.2 Budgeted Evaluation Framework

The proposed evaluation framework is as follows; Figure 3 gives an overview.

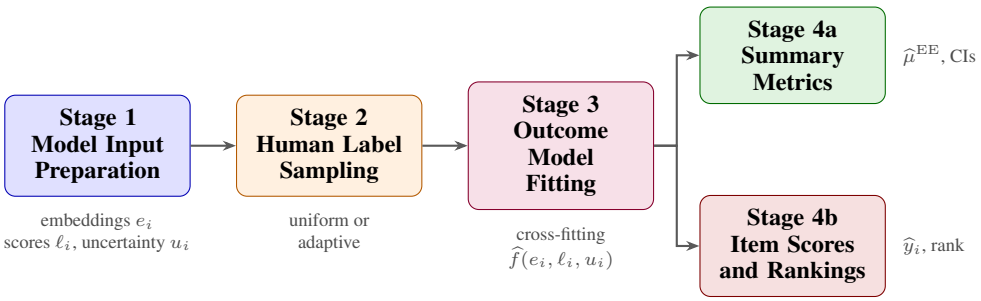


Figure 3: BACON evaluation pipeline. Stage 1 prepares inputs (context features and AI jury scores) for all items; Stage 2 collects a small human label sample; Stage 3 fits the outcome model with cross-fitting; Stage 4 routes to summary-metric estimation or individual scoring and ranking.

(1) Stage 1: Model input preparation. (i) Context features extraction. For each item in the evaluation pool, we extract some context features. This includes the content embeddings, summary statistics, emotion and psychological attributes, among others (Mozer & Miratrix, 2025). (ii) AI Jury Scoring. We use multiple AI judges to score all the items in the pool. We can record auxiliary information such as the token entropy and perplexity. We can also run the AI judges for multiple rounds.

(2) Stage 2: Human Label Sampling. We sample a small number of items to get human labels. The basic approach is to sample items uniformly at random, given a fixed budget. Also, we can apply adaptive sampling strategies to improve the sampling efficiency. We leave an detailed discussion of adaptive sampling strategies to Appendix E.

(3) Stage 3: Outcome Model Fitting. We fit an outcome model to the human labels and the AI scores. For this part, we use cross-fitting to obtain out-of-fold predictions to avoid overfitting.

(4) Stage 4: Route the model for downstream tasks. Here we list two representative tasks to use the scores: (i) Estimate summary metrics. We build regression-adjusted estimators to estimate aggregated metrics such as the population average of human scores, as well as providing uncertainty quantifications through confidence intervals. (ii) Get individual item predictions and ranks. We can use the model to predict surrogate human scores and rank the items. The surrogate scores can be used to flag low quality items.

Below are some implementation details.

Joint modeling of context features and multi-judge evaluations. Overall, we try to fit a predictive model for y_i from the context features e_i and the judge evaluations ℓ_i. Moreover, we can incorporate additional features u_i which encodes the uncertainty of the judge evaluations. We propose to fit a joint model for y_i from the context features e_i, the judge evaluations ℓ_i with uncertainty features u_i: $\widehat{y}_i = f(e_i, \ell_i, u_i)$, where f is a function that we need to learn from data. We will discuss the choice of f later. Intuitively, we hope the function f can serve as a meta-judge that aggregates the single-judge evaluations and steer the prediction towards the human preferences.

Embedding features e_i capture item-level effects, such as topic, style, visual composition, that are not explained by judge scores alone, acting as fixed-effect controls. Multiple judge scores ℓ_i are combined via model fitting rather than ad-hoc majority voting or averaging, allowing the outcome model to down-weight biased or unreliable judges. Uncertainty features u_i encode judge confidence: token-level entropy (available when log-probabilities are accessible) or sample-level variance from multiple judge runs (applicable to any model).

Example models. The choice of f depends on the score pattern of the outcome. For example, if the outcome is real-valued, we can use ridge regression; if it is non-negative and zero-inflated, we can use a hurdle model; for ordinal outcomes, a proportional-odds model is appropriate; and for categorical outcomes, a multinomial logistic regression can be used. See Table 16 in Appendix C for a summary of several instantiations; all share the same feature inputs (e_i, ℓ_i, u_i).

Cross-fitting for outcome model fitting. A key requirement is that the outcome model $\widehat{f}$ must not be fitted on the same sampled observations whose residuals are used for bias correction. Cross-fitting (Chernozhukov et al., 2018) resolves this by partitioning the sampled set $\mathcal{S} = \{i : S_i = 1\}$ into K folds $\mathcal{S}_1, \ldots, \mathcal{S}_K$. For each fold k, the outcome model $\widehat{f}^{(-k)}$ is trained on all sampled observations *except* fold k, and its predictions on fold k are used as the out-of-fold (OOF) predictions. Formally, for $i \in \mathcal{S}_k$, $\widehat{y}_i = \widehat{f}^{(-k)}(e_i, \ell_i, u_i)$, $\widehat{f}^{(-k)} = \mathrm{fit}(\{(e_j, \ell_j, u_j, Y_j) : j \in \mathcal{S} \setminus \mathcal{S}_k\})$. Then we fit the an aggregated model $\widehat{f}$ on the full sample S and use it to predict the labels for items that are not in S. Through this formulation, the predictions $\widehat{y}_i$'s cover all N items. Because $\widehat{f}^{(-k)}$ is fitted on a held-out split, the residuals $Y_i - \widehat{y}_i$ are honest as they reflect genuine generalisation error rather than in-sample fit.

Summary metrics estimation with uncertainty quantification. Summary metrics include population mean, median/quantiles, variance, etc. We can unify the estimation of common metrics using an estimating equation framework (Hardin & Hilbe, 2002). We save the details in Appendix D.2.

3 Real-World Results

3.1 Setup

Datasets and Outcomes. (i) *PERSUADE.* We use the PERSUADE training corpus with essay-level holistic labels (Crossley et al., 2024). In particular, main experiments target grade-10 essays. (ii) *Machine translation.* We use the WMT 2020 English–German MQM (Multidimensional Quality Metrics) dataset (Freitag et al., 2021), where expert human raters annotate translation errors by category and severity. Each segment receives a non-negative MQM score (lower is better; zero means no errors detected). (iii) *WebDesign.* We use screenshot datasets from multiple universities (Miniukovich & Figl, 2023). For each screenshot, we model six human-rated outcomes: Aesthetics,

Trustworthiness, Typicality, Exemplar Goodness, Family Resemblance, and Usability. Due to space limit, we defer the WebDesign results to Appendix A.3.

Embedding Features. (i) *Text embeddings.* Essays/Machine translation segments are embedded using `text-embedding-3-large`[1] from OpenAI. (ii) *Image embeddings.* Screenshots are embedded with SigLIP (Zhai et al., 2023). To stabilize downstream linear models, we apply PCA to high-dimensional embeddings, retaining a tunable number of principal components.

Baselines and general metrics. We compare BACON against both uncalibrated AI-judge baselines and calibrated outputs. The uncalibrated baselines use raw judge scores directly, including single-judge scores and averages across judges. For calibrated models, to isolate the value of different signals, we evaluate calibrated models using only judge scores and uncertainty features, only embedding features, or the full joint feature set. For item-level prediction, we report R^2, root mean squared error (RMSE), and Spearman correlation. For summary-metric estimation, we report bias, standard deviation, and RMSE across repeated sampling trials. Monte Carlo experiments are run on an internal research cluster with nodes equipped with Intel Xeon Gold 6338 CPUs.

3.2 PERSUADE: HOW WELL CAN WE FIT AN OUTCOME MODEL?

Oracle performance. Table 1 reports out-of-fold judge performance when fitting the outcome model on grade-10 essays under six configurations with all human labels revealed, in order to evaluate the oracle performance of the model. We use an ordinal regression model. The full 8-judge model combined with embeddings achieves the best performance. Single-judge results reveal a slight performance drop with embeddings, and a large performance drop without embeddings, indicating that embedding features carry substantial predictive signal. As uncalibrated baselines, simply averaging all judges' raw scores (no outcome model) yields R^2=0.361, and averaging per-judge raw metrics yields R^2=0.150—substantially below all calibrated configurations.

Table 1: PERSUADE judge-model performance (5-fold OOF CV, ordinal regression). Lower RMSE is better; higher R^2 and Spearman ρ are better.

Configuration	Type	R^2	RMSE	Spearman ρ
Eight judges + embedding	Calibrated	0.699	0.602	0.831
Eight judges (no embedding)	Calibrated	0.399	0.851	0.619
Single judge + embedding	Calibrated	0.683	0.618	0.822
Single judge avg. (no embedding)	Calibrated	0.269	0.938	0.460
Embedding-only (ordinal)	Calibrated	0.674	0.627	0.816
Intercept-only constant	Calibrated	−0.000	1.098	−0.020
Raw avg. of all judges	Uncalibrated	0.361	0.877	0.601
Raw single-judge avg.	Uncalibrated	0.150	1.007	0.502

Outcome model quality vs. budget. A key question is how well the outcome model itself can be estimated from the sampled essays. Figure 4 traces out-of-fold R^2, RMSE, and Spearman ρ of the judge model as η increases, using only the sampled subset for fitting. At the smallest budget (η=0.05, roughly 230 essays), the full model already achieves meaningful predictive performance; quality improves steadily as more essays are observed. This confirms that the joint model provides sufficient signal for reliable outcome modelling even at modest sampling rates, supporting the practical viability of BACON in annotation-scarce settings.

3.3 MACHINE TRANSLATION: HOW WELL CAN WE INFER SUMMARY METRICS?

We apply the BACON framework to machine translation quality evaluation using the WMT 2020 English–German MQM dataset.

Simulation on summary metrics. We evaluate each outcome model as a plug-in for the AEE estimator under uniform sampling budgets $\eta \in \{0.05, \ldots, 0.30\}$ across all ten MT systems. In

[1] `https://developers.openai.com/api/docs/models/text-embedding-3-large`

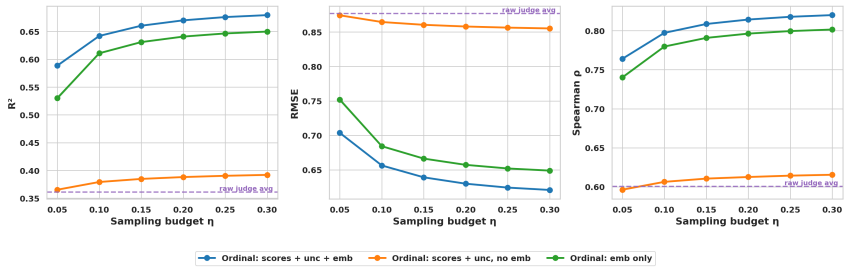


Figure 4: Outcome model quality vs. budget η.

each of 200 Monte Carlo trials, a fraction η of segments per system is sampled uniformly; 5-fold cross-fitting with winsorised LLM score features is used in the evaluation. Figure 5 reports estimator bias, SD, and RMSE across trials.

All methods remain approximately unbiased throughout (max $|\text{bias}| < 0.02$), confirming that the AEE estimator corrects for outcome-model misspecification. The hurdle models (both full and no-emb variants) reduce estimator SD relative to the intercept-only baseline at every budget. The embedding-only hurdle is slightly less efficient than the full model at larger budgets, suggesting that LLM judge scores contribute residual predictive signal beyond text embeddings alone. Besides, all calibrated methods maintain Bonferroni-corrected CI coverage at or above the nominal 95% level throughout, which suggests the validity of the uncertainty quantification.

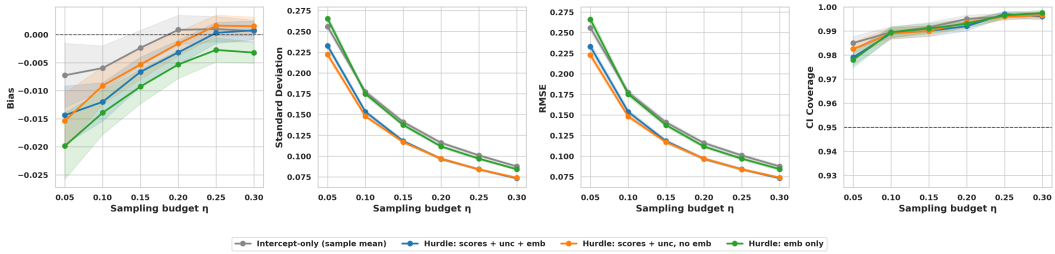


Figure 5: MQM AEE estimator quality vs. sampling budget η in terms of bias, SD, and RMSE and Bonferroni-corrected 95% CI coverage.

Translation system comparison under budgeted sampling. One key practical goal is correctly ranking the MT systems by quality given limited human labels. Figure 6 reports Spearman ρ, Kendall τ, and mean absolute rank error (MAE) of rankings derived from AEE estimates vs. the true system ranking under budgets. All methods produce increasingly accurate rankings as η grows. For this goal, hurdle models with both full and multiple-judge-only configurations achieve strong performance at only small budgets. This suggests that system ranking can be achieved with limited human label and proper calibration.

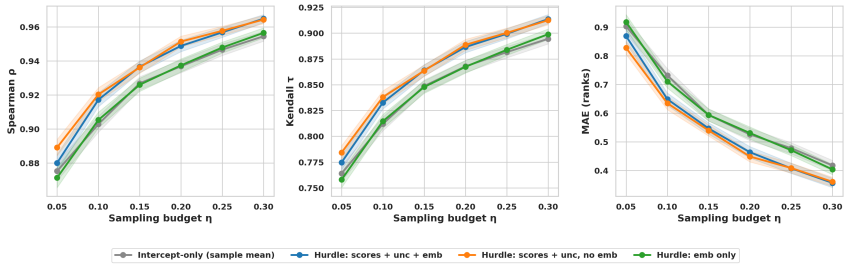


Figure 6: MQM system ranking quality vs. sampling budget η (200 MC trials, 5-fold cross-fitting). Spearman ρ, Kendall τ, and mean absolute rank error (MAE) comparing rankings estimated from AEE to the true system ranking. Hurdle models improve over the intercept baseline at all budgets.

3.4 WHETHER AN APPLICATION NEEDS MANY JUDGES?

Whether an application needs many judges or just one is an empirical question, and we answer it directly by measuring the panel frontier: how the accuracy of the calibrated outcome model grows as we enlarge the judge panel, over and above a full-coverage baseline that uses only item embeddings. Table 2 reports this frontier for PERSUADE and MQM. Each row is the mean over all sampled judge panels of the stated size.

Table 2: Marginal utility of panel size. Each row averages over all sampled judge panels of the given size, using the full hybrid feature set (scores + uncertainty + embeddings), 5-fold cross-validated. Higher R^2/Spearman and lower RMSE are better.

PERSUADE				**MQM**			
Judges	R^2	RMSE	Spearman	Judges	R^2	RMSE	Spearman
1	0.683	0.618	0.822	1	0.304	1.944	0.594
2	0.688	0.613	0.825	2	0.346	1.884	0.618
4	0.692	0.610	0.828	3	0.363	1.860	0.630
6	0.698	0.603	0.832	4	0.376	1.841	0.640
8	0.699	0.602	0.832	5	0.385	1.828	0.644
-	-	-	-	6	0.393	1.815	0.650

Table 2 reveals that, how much judges help at all is governed by how much the judge-free embedding baseline already explains. On PERSUADE, embeddings alone reach $R^2 = 0.673$, so judges have little left to add: the mean panel rises only from 0.683 (one judge) to 0.699 (eight), with clear diminishing returns after the first one or two judges. On MQM the embedding baseline is weak (0.170), judges carry most of the signal, and the gains accrue more gradually. Therefore, because the frontier flattens at a different point for every task, no universal judge count is warranted. In practice, one should test, for a given task and budget, how much each additional judge actually buys, so a practitioner can deploy the smallest panel rather than defaulting to the full set.

3.5 COST-MATCHED COMPARISON: DOES A CHEAPER PANEL BEAT A SINGLE STRONG JUDGE?

We directly test the allocation trade-off—one strong judge versus a broader panel of cheaper judges—with paired, cost-matched simulations. For each budget η and each of 100 paired draws, the two arms receive the same human calibration sample, embeddings, uncertainty features, and AEE estimator; they differ only in the judge-score panel. Costs are direct API list-price estimates for the completed runs, excluding the shared embedding pass. Table 3 presents the results for PERSUADE and MQM, and Table 9 in Appendix A.3 presents the results for WebDesign.

Table 3: Equal-cost allocation. PERSUADE: GPT-4.1 ($5.53) vs. a six-judge panel ($5.20: Gemini 2.5 Flash and Flash-Lite, GPT-4.1 mini and nano, GPT-4o mini, Llama 3.3 70B). MQM: GPT-4.1 ($92.02) vs. four judges ($91.12: Gemini 2.5 Flash, GPT-4.1 mini, GPT-4o mini, Llama 3.3 70B), where the Llama price uses the project's deployment ratio ($2.3\times$ Gemini Flash) so the arms sit within 1% on cost. RMSE is against the population mean; Cov. is empirical 95%-CI coverage; bold marks the lower RMSE; 100 paired trials per cell.

PERSUADE						**MQM**					
		GPT-4.1		Six-judge panel				GPT-4.1		Four-judge panel	
η	Labels	RMSE	Cov.	RMSE	Cov.	η	Labels	RMSE	Cov.	RMSE	Cov.
5%	233	**0.0440**	0.97	0.0475	0.96	5%	709	0.0658	0.96	**0.0635**	0.97
10%	466	0.0261	0.99	**0.0259**	0.99	10%	1418	0.0510	0.96	**0.0496**	0.96
20%	931	**0.0196**	0.95	0.0200	0.96	20%	2835	0.0302	0.97	**0.0290**	1.00
30%	1397	0.0116	1.00	**0.0111**	0.99	30%	4253	**0.0218**	1.00	0.0230	0.98

Overall, in these three tasks, the multi-judge panel performs competitively with or even much better than the single strong judge at equal or cheaper cost. On the two cost-matched tasks the multi-judge

Table 4: Illustrative computation for full-coverage scoring of the PERSUADE grade-10 slice. Token counts are realized from run logs; dollar amounts use public list prices (embeddings and judges) or a conservative crowd-labor estimate of 2–4 minutes per essay at \$15/hr plus a 20% MTurk requester fee.

Line item	Input tokens	Output tokens	Cost (USD)
Embeddings (one pass)	2.6M	—	\$0.34
Eight-judge panel (one pass)	22.2M	37K	\$16–\$21
One crowd holistic rating (all essays)	—	—	\$2,792–\$5,585
Three crowd ratings (all essays)	—	—	\$8,377–\$16,754
One rating for 10% calibration (465 essays)	—	—	\$279–\$559

panel wins two of four PERSUADE budgets and three of four MQM budgets. On WebDesign, the three-judge panel matches or beats the single strong judge in 21 of 24 outcome–budget cells while costing 48% as much. Therefore, a carefully chosen panel of cheaper judges can provide a cost-effective alternative to relying solely on a single strong judge, especially when human calibration is preserved and the panel size is optimized for the task at hand.

3.6 How much are we saving compared with full human labeling?

In this section, we compare the cost of BACON with the cost of obtaining equivalent human ratings. Each judge request carries roughly 600 input tokens (prompt plus essay) and returns a single scalar, so one output token per essay. Summing the realized usage over the eight judges that returned valid scores gives 22.2M input and 37K output tokens for a full-coverage pass; a single `text-embedding-3-large` pass over the same essays adds about 2.6M input tokens.[2] Table 4 converts these counts into representative costs and places them next to what the same coverage would cost in human ratings. From Table 4, we see that the eight-judge panel plus embeddings costs on the order of \$16–\$21, or roughly 0.35–0.45 cents per essay, for a signal at every item in the corpus. Even rating only 10% of the essays once (\$279–\$559) costs more than running the entire panel. Hence, the comparison that matters is against human coverage, not human quality. The issue is not whether judges replace people, but whether a limited human-calibration budget, augmented by full-coverage judge and representation features, beats spending that same budget on human labels alone, which is the central question our earlier experiments address.

Remark. Appendix A provides the full set of supporting experiments. It includes additional budgeted-estimation results for PERSUADE; outcome-model and per-system ablations for MQM; and, for WebDesign, feature ablations, estimator and outcome-model performance across labeling budgets, panel-size effects, and the cheaper-panel comparison. We also stress-test BACON under noisy and disagreeing human labels and leave-one-group-out distribution shifts, examine sensitivity to the number of embedding principal components and the inclusion of uncertainty features, and extend the AEE estimator from means to quantiles.

4 Discussion: summary, limitations.

BACON is a pipeline that aims to provide better guidance for the use of AI judges in evaluation. It is closely related to LLM collaboration literature, as well as PPI-style estimators. The contribution is to bridge these practices in an unified framework and make them operational. Item-level BACON scores are only calibrated surrogate predictions: they are useful for triage, ranking, and auditing, but should not be interpreted as unbiased measurements for each item. Moreover, calibration may degrade under distribution shift in items, prompts, model versions, or rating rubrics; practical deployment requires periodic recalibration, version logging, and residual monitoring on fresh human labels.

[2]Token counts are read directly from the run logs (`judge_usage.csv`); dollar figures are computed from public list prices and are illustrative. Pricing references: OpenAI embeddings `https://openai.com/api/pricing/`; OpenAI GPT-4.1 `https://openai.com/api/pricing/`; Google Gemini API `https://ai.google.dev/gemini-api/docs/pricing`; Amazon MTurk requester fees `https://requester.mturk.com/pricing`.

AI USE STATEMENT

In this work, we used generative AI tools for:

> Design or provide feedback on research methodology or experiments; implement methods; clean and reformat dataset; support qualitative and thematic data analysis; assist in the writing of proofs; providing LLM judge labels as part of the core research idealization.

We have not used generative AI tools for:

> Generate synthetic data sets; help develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; propose or refine hypotheses; assist with translation; interpret results.

and the following is not applicable to this work:

> Assist with translation.

Additionally, we used generative AI tools for:

> Create or modify scientific figures or images, suggest experimental parameters, summarize or analyse existing literature, identify relevant literature.

We have reviewed all AI-assisted work. Specifically: methodology and experiment designs suggested with AI assistance were assessed by the authors for soundness and prior art through a manual literature survey before adoption; AI-assisted code implementing the estimators and experiment harness was read, run, and tested for correctness by at authors; dataset cleaning and reformatting scripts were checked against the source data to confirm no records were dropped or altered; qualitative and thematic analyses produced with AI assistance were verified against the underlying model outputs and human labels; and we use AI to help write latex formulas when drafting proofs but guide the process manually to ensure correctness. AI-generated figures, experimental parameters, and literature pointers were likewise inspected and confirmed by the authors before inclusion. Also our research target is LLM-as-judges, so we generated and evaluated many LLM judge labels as part of the core research idealization. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We are not aware of any significantethical concerns or potential conflicts of interest related to this work.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the experiments and results reported in this paper are reproducible. In the supplementary materials, we provide detailed descriptions of the experimental setup, data processing steps, and source code necessary to reproduce our results. For theoretical results, Appendix D provides clear statements of all results and proofs.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

T. Tony Cai and Zijian Guo. Semisupervised inference for explained variance in high dimensional linear regression and its applications. *Journal of the Royal Statistical Society Series B: Statistical*

Methodology, 82(2):391–419, 2020. doi: 10.1111/rssb.12357. URL https://doi.org/10.1111/rssb.12357.

Claes M. Cassel, Carl-Erik Särndal, and Jan H. Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

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

Scott A Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Meg Benner, and Ulrich Boser. A large-scale corpus for assessing written argumentation: Persuade 2.0. *Assessing Writing*, 61: 100865, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 2024.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-judge. *The Innovation*, 7(6): 101253, 2026. doi: 10.1016/j.xinn.2025.101253. URL https://doi.org/10.1016/j.xinn.2025.101253.

Max Hamilton, Jinlin Lai, Wenlong Zhao, Subhransu Maji, and Daniel R. Sheldon. Active measurement: Efficient estimation at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4878. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d6adf82fa531dcb8bfa53c224ce5e466-Abstract-Conference.html.

James W Hardin and Joseph M Hilbe. *Generalized Estimating Equations*. Chapman & Hall/CRC, 2002.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5880–5895, 2025.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=8euJaTveKw.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024a.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/747dc7c6566c74eb9a663bcd8d057c78-Abstract-Conference.html.

Puheng Li, Tijana Zrnic, and Emmanuel Candès. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2312. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/6389470564214983604d1ac81631c2c5-Abstract-Conference.html.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Aliaksei Miniukovich and Kathrin Figl. The effect of prototypicality on webpage aesthetics, usability, and trustworthiness. *International Journal of Human-Computer Studies*, 179:103103, 2023. doi: 10.1016/j.ijhcs.2023.103103. URL https://doi.org/10.1016/j.ijhcs.2023.103103.

Reagan Mozer and Luke Miratrix. More power to you: Using machine learning to augment human coding for more efficient inference in text-based randomized trials. *The Annals of Applied Statistics*, 19(1):440–464, 2025.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=XccDXrDNLek.

Yiyue Qian, Shinan Zhang, Yun Zhou, Haibo Ding, Diego Socolinsky, and Yi Zhang. CollabEval: Enhancing LLM-as-a-judge via multi-agent collaboration, 2026. URL https://arxiv.org/abs/2603.00993.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer, New York, 1992.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *International Conference on Learning Representations*, 2024b.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Anru Zhang, Lawrence D. Brown, and T. Tony Cai. Semi-supervised inference: General theory and estimation of means. *The Annals of Statistics*, 47(5):2538–2566, 2019.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7f8f73134e253845a8f82983219a8452-Abstract-Conference.html.

Tijana Zrnic and Emmanuel J Candès. Cross-prediction-powered inference. *Proceedings of the National Academy of Sciences*, 121(15):e2322083121, 2024.

A Additional Evaluation Results

A.1 More on PERSUADE evaluation

Simulation on summary metrics. We evaluate how well several configurations support the bias and efficiency claim of the AEE estimator under a uniform sampling budget $\eta \in \{0.05, \ldots, 0.30\}$. In each Monte Carlo trial, a fraction η of grade-10 essays is sampled uniformly at random. We run 200 independent trials per budget and report estimator bias, standard deviation (SD), and RMSE across trials. Figure 7 shows the AEE estimator quality as a function of η. All calibrated methods are approximately unbiased when the sampling budget is appropriate, while the uncalibrated baselines are biased. The full hybrid model mostly achieves the lowest SD and RMSE across budgets, demonstrating that a better outcome model directly translates to a more efficient estimator. Across all budgets, the reduction in estimator variance from the full model over the intercept-only baseline reflects the predictive R^2 gains seen in Table 1.

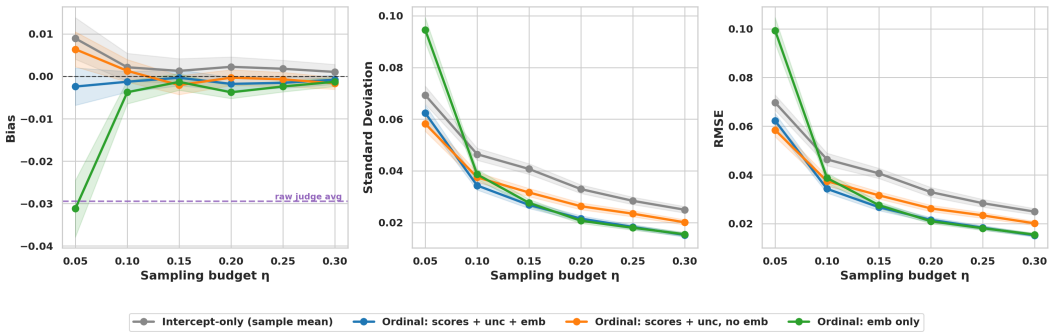


Figure 7: Estimating-equation estimator quality

A.2 More on MQM evaluation

Importance of fitting the appropriate outcome model. Table 5 reports pooled 5-fold OOF metrics for predicting human MQM scores. Features include per-LLM mean scores and variance-based uncertainty, with embeddings from a large text embedding model. Because MQM scores are non-negative with a substantial zero mass (∼13% pooled, up to 29% for human translations), we compare the standard Ridge model against a *Hurdle model* (logistic gate predicting $P(\text{MQM} > 0)$ combined with Ridge on the positive subset). We also compare two variance encodings: the original $\log(\text{Var}+\epsilon)$ transform versus $\sqrt{\text{Var}}$ with a binary has-variance indicator, since 57–67% of items have zero LLM variance (all raters agree), creating an artificial bimodal cliff at $\log(10^{-8}) \approx -18.4$ in the log encoding.

Table 5: MQM pooled judge-model ablation (5-fold OOF, all 10 systems, n=14,180). The Hurdle model with $\sqrt{\text{Var}}$ encoding achieves the best R^2 while eliminating negative predictions.

Configuration	R^2	Spearman
Full (sqrt, hurdle)	**0.369**	**0.657**
Full (sqrt, ridge)	0.355	0.646
Full (log, hurdle)	0.365	0.646
Full (log, ridge)	0.351	0.634
LLM scores only (sqrt, ridge)	0.287	0.557
Embedding-only ridge	0.167	0.497
Intercept-only baseline	0.000	0.000

The $\sqrt{\text{Var}}$ encoding resolves pathological failures on high-quality systems: for `Human-B.0` (29% zeros), Ridge with log-variance yields $R^2 = -0.14$ due to extreme predictions driven by the -18.4

Table 6: Per-system MQM judge-model performance (5-fold OOF, Hurdle model, $\sqrt{\text{Var}}$ encoding). R^2 and Spearman ρ are reported for each feature ablation.

	Full		LLM only		Embedding only		Intercept only	
System	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
Human-A.0	0.273	0.554	0.191	0.427	0.150	0.448	0.000	0.000
Human-B.0	0.236	0.516	0.150	0.359	0.139	0.452	0.000	0.000
Human-P.0	0.328	0.581	0.207	0.478	0.197	0.491	0.000	0.000
Huoshan_Translate	0.385	0.660	0.384	0.644	0.105	0.398	0.000	0.000
OPPO	0.425	0.691	0.388	0.649	0.160	0.470	0.000	0.000
Online-A	0.441	0.703	0.391	0.691	0.112	0.375	0.000	0.000
Online-B	0.369	0.646	0.399	0.641	0.116	0.365	0.000	0.000
Tencent_Translation	0.435	0.690	0.398	0.642	0.160	0.461	0.000	0.000
Tohoku-AIP-NTT	0.377	0.656	0.360	0.636	0.130	0.405	0.000	0.000
eTranslation	0.449	0.707	0.436	0.673	0.130	0.441	0.000	0.000
Pooled	0.369	0.657	0.287	0.557	0.167	0.497	0.000	0.000

floor, whereas $\sqrt{\text{Var}}$ with Hurdle achieves $R^2 = 0.24$. For MT systems with fewer zeros, both encodings perform similarly.

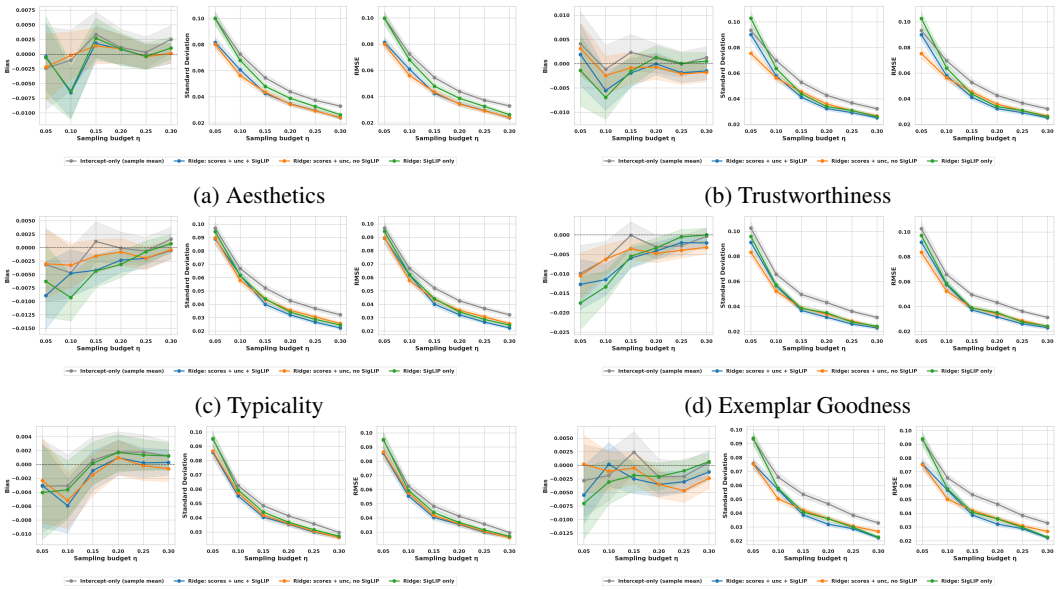
Per-system ablation. Table 6 reports the Hurdle model with $\sqrt{\text{Var}}$ encoding across all ten translation systems and the pooled corpus. The full model (LLM + uncertainty + embeddings) consistently dominates the LLM-only and embedding-only ablations. LLM scores alone explain much of the variance for MT systems (R^2 0.36–0.44) but substantially less for human translations (R^2 0.15–0.21), where inter-rater noise is higher. Embeddings alone provide a complementary signal (R^2 0.10–0.20) that is especially valuable for the human-translation systems where LLM scores are weakest. Combining both sources yields consistent gains across all systems.

A.3 WebDesign evaluation

For the Universities subset, we ran the same judge ablation protocol across six outcomes (Aesthetics, Trustworthiness, Typicality, Exemplar Goodness, Family Resemblance, Usability), including a constant intercept-only baseline. Table 7 reports average 5-fold OOF metrics across outcomes. The full hybrid model (VLM scores + uncertainty + SigLIP intercept) achieves the best overall performance, improving over both the VLM-only and SigLIP-only ablations. As uncalibrated baselines, using raw VLM scores directly yields highly negative R^2 (due to a scale mismatch: human ratings are z-scored while VLM scores are on their original integer scale), though Spearman correlation remains moderate (0.56), confirming that outcome-model calibration is essential in this domain.

Table 7: WebDesign (Universities) judge-model ablation results. Values are averaged across six outcomes. Lower RMSE is better; higher R^2 and Spearman are better. Uncalibrated rows use raw VLM scores directly (no outcome-model fitting); note that human ratings are z-scored while VLM scores are on their original integer scale, so scale mismatch explains the negative R^2.

Configuration	Type	Avg. R^2	Avg. RMSE	Avg. Spearman
VLM scores + uncertainty + SigLIP	Calibrated	0.555	0.462	0.727
VLM scores + uncertainty (no SigLIP)	Calibrated	0.390	0.544	0.591
SigLIP-only	Calibrated	0.497	0.492	0.690
Intercept-only constant	Calibrated	−0.001	0.699	−0.047
Raw avg. of all judges	Uncalibrated	−9.719	2.258	0.561
Raw single-judge avg.	Uncalibrated	−10.151	2.291	0.340



(a) Aesthetics (b) Trustworthiness

(c) Typicality (d) Exemplar Goodness

(e) Usability (f) Average (across all outcomes)

Figure 8: WebDesign (Universities) AEE estimator quality: bias, SD, and RMSE vs. sampling budget η, per perceptual outcome and averaged. Lower is better. The full hybrid model achieves the lowest estimator variance for most outcomes.

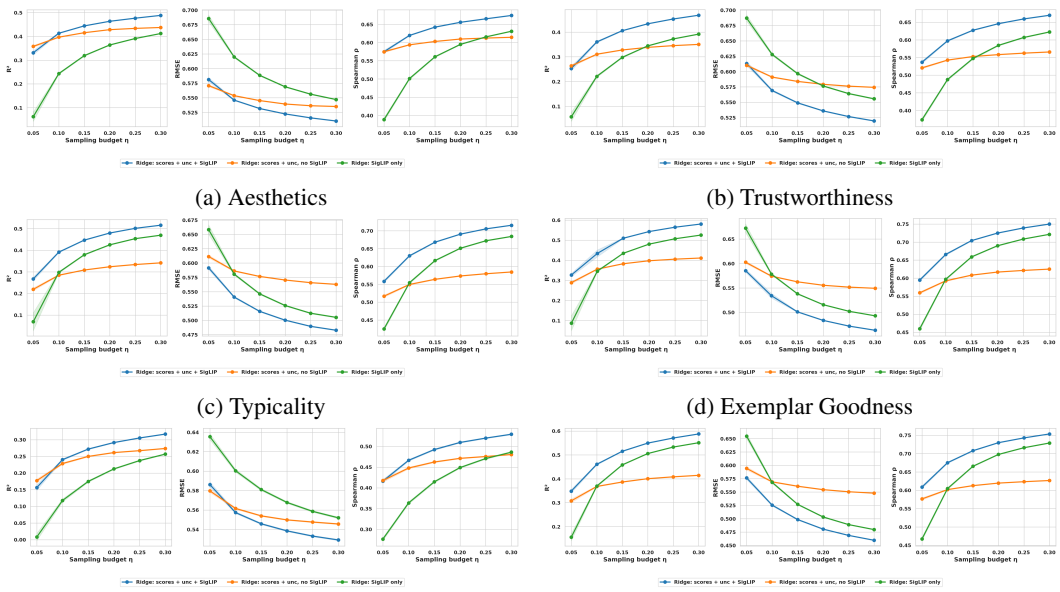
Simulation on summary metrics. We evaluate each configuration as an AEE outcome model under uniform sampling budgets $\eta \in \{0.05, \ldots, 0.30\}$ across all six perceptual outcomes. In each trial a fraction η of websites is sampled uniformly; the AEE estimator combines population-level predictions from the judge model with an in-sample bias correction. Figures 8 and 9 report estimator and outcome-model quality per outcome and on average.

Estimator quality (Figure 8). All models are approximately unbiased across budgets. The full hybrid model (VLM scores + uncertainty + SigLIP intercept) achieves the lowest standard deviation and RMSE for most outcomes, confirming that the variance-reduction gains from the judge model carry through to the AEE estimator. Perceptual outcomes vary considerably in difficulty: Aesthetics and Usability are better predicted than Exemplar Goodness and Typicality, and this ordering is reflected in estimator efficiency. The AVG panel shows that the full model consistently outperforms both the VLM-only and SigLIP-only ablations when averaged across outcomes.

Outcome model quality (Figure 9). Figure 9 traces the out-of-fold predictive performance of each outcome model as a function of η. Even at the smallest budget (η=0.05, roughly 30–40 websites), the full model achieves nontrivial R^2 and Spearman correlation for several outcomes, demonstrating that VLM judge scores and SigLIP embeddings together provide a strong prior that makes efficient use of limited human labels. Outcome-model quality improves consistently with budget, and the full hybrid model leads on R^2 and Spearman for most outcomes, matching the pattern observed in the static ablation (Table 7).

Number of judges. Table 8 presented the marginal utility of different panel sizes for the WebDesign dataset. Each row averages over all sampled judge panels of the given size, using the full hybrid feature set (scores + uncertainty + embeddings), 5-fold cross-validated.

Cheaper-panel allocation on WebDesign. Table 9 shows the comparison between GPT-4.1 and the intentionally cheaper three-judge panel across different labeling budgets. The three-judge panel attains lower RMSE in 21 of the 24 cells.



(a) Aesthetics (b) Trustworthiness

(c) Typicality (d) Exemplar Goodness

(e) Usability (f) Average (across all outcomes)

Figure 9: WebDesign (Universities) outcome model quality vs. sampling budget η: out-of-fold R^2, RMSE, and Spearman ρ per perceptual outcome and averaged. Higher R^2/Spearman and lower RMSE are better. The full model leads on most outcomes.

Table 8: Marginal utility of panel size for the WebDesign dataset. The reported metrics are averaged over its six perceptual outcomes.

Dataset	Judges	R^2	RMSE	Spearman
WebDesign (mean)	1	0.527	0.477	0.708
WebDesign (mean)	2	0.542	0.469	0.717
WebDesign (mean)	3	0.546	0.467	0.720
WebDesign (mean)	4	0.553	0.463	0.724
WebDesign (mean)	5	0.559	0.460	0.729
WebDesign (mean)	6	0.562	0.458	0.731

Table 9: Cheaper-panel allocation on WebDesign: GPT-4.1 ($3.17) vs. an intentionally cheaper three-judge panel (Gemini 2.5 Flash-Lite, GPT-4.1 mini, GPT-4.1 nano; $1.52, i.e. 48% of the GPT-4.1 API cost). Entries are RMSE against the population mean over 100 paired trials, for each of the six perceptual outcomes and four labeling budgets; "G" is GPT-4.1, "3J" the three-judge panel, and **bold** marks the lower-RMSE arm. Coverage stays within 0.91–1.00 for both arms throughout. The three-judge panel attains lower RMSE in 21 of the 24 cells.

	$\eta = 5\%$		$\eta = 10\%$		$\eta = 20\%$		$\eta = 30\%$	
Outcome	G	3J	G	3J	G	3J	G	3J
AE	0.0806	**0.0798**	**0.0615**	0.0633	0.0357	**0.0347**	0.0272	**0.0268**
TRU	0.0952	**0.0879**	0.0595	**0.0557**	0.0394	**0.0352**	0.0242	**0.0228**
TYP	0.0897	**0.0792**	0.0488	**0.0464**	**0.0298**	0.0301	0.0259	**0.0251**
EXMPL	0.0926	**0.0866**	0.0545	**0.0480**	**0.0324**	0.0345	0.0255	**0.0253**
AVG	0.0977	**0.0860**	0.0532	**0.0531**	0.0330	**0.0304**	0.0259	**0.0231**
US	0.0822	**0.0800**	0.0495	**0.0491**	0.0380	**0.0379**	0.0279	**0.0262**

Table 10: Robustness of BACON's AEE population-mean estimator to noisy calibration labels. RMSE is against the clean-population mean; Cov. is empirical 95%-CI coverage; η is the labeling budget (pool fraction). All numbers average over 100 trials. Bias is negligible throughout ($|\text{bias}| \leq 0.02$ on the native score scale). Higher coverage (near 0.95) and lower RMSE are better.

	$\eta = 0.05$		$\eta = 0.10$		$\eta = 0.20$	
Setting	RMSE	Cov.	RMSE	Cov.	RMSE	Cov.
PERSUADE — ordinal ±1 flip noise						
flip 5%	0.049	0.98	0.030	0.98	0.023	0.91
flip 15%	0.058	0.94	0.033	0.96	0.025	0.94
flip 30%	0.067	0.93	0.039	0.94	0.030	0.91
WebDesign — additive Gaussian noise (std)						
std 0.10	0.076	0.95	0.053	0.94	0.032	0.98
std 0.25	0.080	0.95	0.063	0.96	0.037	0.95
std 0.50	0.094	0.97	0.074	0.97	0.050	0.94
MQM — real multi-rater labels (raters/item, budget-matched)						
1 rater	0.113	0.96	0.074	0.98	0.054	0.94
2 raters	0.129	0.95	0.100	0.94	0.064	0.96
3 raters	0.150	0.93	0.093	0.97	0.073	0.94

A.4 Inter-annotator disagreement and noisy labels

A central assumption behind budgeted human calibration is that the small set of human labels used to anchor the outcome model is reliable. In practice human labels are themselves noisy: annotators disagree, and label errors are pervasive even in curated benchmarks (Northcutt et al., 2021). Our WMT 2020 MQM data make this concrete: every segment is triple-rated, and the three raters disagree on 69–88% of segments depending on the system, with a mean per-segment standard deviation of the MQM error score ranging from 0.81 (Human-B) to 1.93 (Online-A). We therefore ask whether BACON's population-level estimates remain valid when the calibration labels are corrupted.

We run a Monte Carlo audit of the augmented estimating-equation (AEE) mean estimator. For each labeling budget $\eta \in \{0.05, 0.10, 0.20\}$ (fraction of the pool that receives human labels), we draw a random sample, replace the sampled labels with *noisy* observations, fit the cross-fitted outcome model and residual correction on those noisy labels only, and compare the resulting estimate against the *clean*-population mean. We use three noise processes matched to each modality: (i) PERSUADE (ordinal essay scores 1–6): each sampled label moves ± 1 with total probability equal to the flip rate; (ii) WebDesign (continuous AVG perceptual score): additive Gaussian noise of increasing standard deviation, clipped to range; and (iii) MQM: instead of synthetic noise, we sample $k \in \{1, 2, 3\}$ *real* human raters per item and average them, holding the *total* human-rating budget fixed (so k=3 labels one third as many distinct items as k=1). Each cell aggregates 100 trials; coverage is the empirical rate at which the nominal 95% confidence interval contains the clean mean.

Table 10 supports three conclusions. **(1) Noisy calibration labels do not bias the estimate.** Across all three modalities the bias stays negligible ($|\text{bias}| \leq 0.02$ on the native score scale) and the empirical coverage remains close to the nominal 0.95 (0.91–0.98), even under 30% ordinal flips, Gaussian noise with std up to 0.5, and the genuine multi-rater disagreement of MQM. The residual-correction step keeps the estimator anchored to the clean mean, so calibration labels need not be error-free for the population-level guarantees to hold. **(2) Noise inflates variance gracefully.** RMSE grows monotonically with the noise level (flip rate, Gaussian std, or raters-per-item) but only modestly, and it is absorbed by budget: increasing η from 0.05 to 0.20 roughly halves RMSE at every noise level. Label noise thus costs efficiency, not validity. **(3) For population estimation, breadth beats redundancy.** Under a *fixed* total human-rating budget, single-labeling more distinct items (1 rater/item) clearly dominates repeatedly labeling fewer items (RMSE 0.054 vs. 0.073 at $\eta = 0.20$; the 2- and 3-rater arms are uniformly worse). Repeated labeling reduces per-item measurement noise but shrinks the effective sample of items seen by the outcome model, which hurts the population estimate. Redundant annotation is therefore most useful when *item-level* scores are the deliverable rather than the population summary.

Table 11: Outcome-model transfer under leave-one-group-out distribution shift. $\widehat{g}$ is trained on all-but-one group and scored on the held-out group. $\bar{r}$ is the signed residual bias (human − model) on the held-out group—the bias a surrogate-only estimator would report as truth, and the term BACON's residual correction cancels using target-drawn human labels. Scales are task-native (PERSUADE 1–6, WebDesign ≈ 1–5, MQM error points), so compare R^2/ρ across tasks and $\bar{r}$/RMSE within a task.

Held-out group	n_{test}	R^2	RMSE	ρ	$\bar{r}$
PERSUADE — shift across student grade levels					
Grade 10	4654	0.25	0.95	0.65	+0.27
Grade 11	1863	0.57	0.63	0.74	+0.11
WebDesign — shift across website domains (AVG outcome)					
University sites	1056	0.49	0.51	0.72	−0.07
Commercial-bank sites	1031	−0.13	0.65	0.60	+0.43
MQM — shift across translation systems (leave-one-system-out)					
MT engines (mean, 7)	9926	0.42	1.90	0.73	+0.43
Human references (mean, 3)	4249	−0.86	1.75	0.52	−1.12

A.5 Generalization under distribution shift

Budgeted calibration assumes the outcome model trained on the labeled items transfers to the unlabeled pool. This is safe when the two are exchangeable, but weakens when the pool is drawn from a different population than the calibration items—a new cohort, a new content domain, or a new system. We stress-test this transfer directly with leave-one-group-out audits. For each task we train the cross-fitted outcome model $\widehat{g}$ on all-but-one group and evaluate its predictions on the held-out group, whose human labels are used only for scoring. We study three shifts of increasing severity: PERSUADE across student *grade levels* (Grade 10 ↔Grade 11), WebDesign across website *domains* (University sites ↔ commercial-bank sites, on the AVG perceptual outcome), and WMT 2020 MQM across the ten translation *systems* (leave-one-system-out, summarized as MT engines vs. human reference translations). Besides predictive fit (R^2, Spearman ρ, RMSE) we report the signed residual bias $\bar{r} = \frac{1}{n}\sum_i \big(y_i - \widehat{g}(x_i)\big)$ on the held-out group. This is precisely the systematic error a surrogate-only "impute-then-average" estimator would inherit, and precisely the quantity BACON's residual-correction term removes once it observes human labels drawn from the target.

Table 11 supports three conclusions. **(1) Surrogate accuracy degrades under shift, sometimes catastrophically.** In-distribution the panel-conditioned outcome model is strong, but held out to an unseen group its R^2 falls and can go *negative*—to -0.13 on commercial-bank sites and -0.86 on held-out human reference translations—meaning the transferred surrogate predicts worse than the target-group mean. Rank correlation is more robust than absolute fit (ρ stays 0.52–0.74), so an autorater may still *order* items acceptably even where its *level* is badly off. **(2) Shift induces systematic bias, not just added noise.** The residual bias $\bar{r}$ departs from zero by up to $+0.43$ points (commercial-bank aesthetics) and -1.12 points (MQM human references, roughly 60% of a within-group standard deviation). A prediction-only pipeline would report these offsets as if they were the truth; they do not average out with more unlabeled data. This is the failure mode budgeted human calibration is designed for: because BACON's AEE estimator subtracts the mean human−model residual measured on labels sampled from the *target*, it removes exactly $\bar{r}$ regardless of how poorly $\widehat{g}$ transfers, converting a bias problem into a (small) variance cost. **(3) Severity is asymmetric and interpretable.** The two directions of every shift differ: transfer onto the cleaner, more homogeneous group (Grade 11, University sites) is far easier than the reverse, and the MQM calibrator generalizes across MT engines ($R^2 = 0.42$) but systematically *under*-rates human reference translations, which are quality outliers absent from the MT-dominated training pool ($\bar{r} = -1.12$). The sign and size of $\bar{r}$ are thus an auditable readout of *how much* the human anchor is doing under a given shift, and directly motivate the residual monitoring and periodic recalibration on fresh human labels that we recommend for deployment (Limitations, above) rather than an unqualified trust in the surrogate.

Table 12: Sensitivity of the embedding-only and full-hybrid five-fold held-out fits to the number of retained unsupervised embedding principal components. WebDesign entries average its six outcomes. The full hybrid (scores + uncertainty + embedding PCs) exceeds embeddings alone at every dimension on every task.

Dataset	PCs	Embed-only R^2	Full-hybrid R^2	Full-hybrid RMSE	Full-hybrid ρ
PERSUADE	32	0.614	0.654	0.645	0.802
	64	0.673	0.699	0.602	0.832
	128	0.703	0.717	0.584	0.842
MQM	32	0.078	0.351	1.877	0.559
	64	0.170	0.393	1.815	0.650
	128	0.219	0.414	1.785	0.677
WebDesign (mean)	32	0.494	0.559	0.460	0.730
	64	0.498	0.562	0.458	0.731
	128	0.492	0.549	0.465	0.721

Table 13: Held-out R^2 for five feature families, isolating the incremental value of uncertainty summaries. WebDesign entries average its six outcomes. Uncertainty helps score-only models on all three tasks and the full MQM model, but adds little once embeddings are present on PERSUADE and WebDesign.

Feature family	PERSUADE R^2	MQM R^2	WebDesign mean R^2
Embeddings only	0.673	0.170	0.498
Scores only	0.387	0.318	0.394
Scores + uncertainty	0.396	0.329	0.413
Scores + embeddings	0.700	0.387	0.561
Scores + uncertainty + embeddings	0.699	0.393	0.562

A.6 PCA SENSITIVITY

The full hybrid stacks the judge scores and uncertainties on top of an unsupervised low-dimensional projection of the item embeddings, so its behavior could in principle be sensitive to the number of retained principal components. We therefore repeated the five-fold held-out embedding-only and full-hybrid analyses with 32, 64, and 128 unsupervised embedding PCs on all three tasks (Table 12; WebDesign values average its six outcomes).

The full hybrid exceeds embeddings alone at every PCA dimension on every dataset, so the benefit of adding judge scores and uncertainties is not an artifact of a particular projection size. The exact optimum, however, is not universal: PERSUADE and MQM keep improving through 128 PCs, whereas WebDesign is effectively flat from 32 to 64 and declines at 128. We therefore treat 64 PCs as a compact common operating point rather than a uniquely optimal choice, and recommend that a new deployment validate the dimension on its own held-out calibration data.

A.7 EVIDENCE FOR UNCERTAINTY FEATURES

BACON can augment each judge's point score with cross-judge and token-level uncertainty summaries. To isolate what these add, we compare five feature families under the same five-fold held-out protocol, building up from embeddings alone and from scores alone (Table 13; WebDesign averages its six outcomes).

Cross-judge and token-level uncertainty improves the score-only model on every dataset (PERSUADE $0.387 \rightarrow 0.396$, MQM $0.318 \rightarrow 0.329$, WebDesign $0.394 \rightarrow 0.413$) and also lifts the full MQM model ($0.387 \rightarrow 0.393$). Once embeddings are present, however, its incremental effect is negligible on PERSUADE ($0.700 \rightarrow 0.699$) and WebDesign ($0.561 \rightarrow 0.562$), where the embedding representation already captures most of what the uncertainty channel encodes. We read this as useful negative evidence: uncertainty summaries should be treated as an optional auxiliary feature validated per task, not a mandatory component of the estimator.

Table 14: Quantile estimation on WebDesign (median, 100 Monte Carlo draws per cell). Residual-CDF AEE versus the sample-quantile baseline, reported as median MSE at the 10% and 20% label budgets; "Cov" is residual-CDF 95%-CI coverage. Lower MSE per cell in **bold**.

	$\eta = 10\%$			$\eta = 20\%$		
Outcome	Sample MSE	Resid-CDF MSE	Cov	Sample MSE	Resid-CDF MSE	Cov
AE	0.00821	**0.00703**	0.95	0.00380	**0.00350**	0.98
TRU	0.00879	**0.00749**	0.98	0.00421	**0.00360**	0.97
TYP	0.00905	**0.00564**	1.00	0.00344	**0.00220**	1.00
EXMPL	0.01071	**0.00681**	0.94	0.00576	**0.00293**	0.97
AVG	0.00979	**0.00575**	0.97	0.00648	**0.00444**	0.97
US	0.00606	**0.00533**	0.97	0.00310	**0.00296**	0.97

Table 15: Quantile estimation on MQM across Q25/Q50/Q75 (100 Monte Carlo draws per cell). Residual-CDF AEE versus the sample-quantile baseline; "Cov" is residual-CDF coverage. Lower MSE per row in **bold**.

Quantile	Budget	Sample MSE	Resid-CDF MSE	Cov
Q25	10%	0.000257	**0.000033**	1.00
Q50	10%	0.014414	**0.013978**	0.95
Q75	10%	**0.018336**	0.020056	1.00
Q25	20%	0.000011	**0.000000**	1.00
Q50	20%	0.010378	**0.009544**	0.96
Q75	20%	0.008633	**0.006956**	0.97

A.8 QUANTILE AEE: CROSS-FITTED RESIDUAL-CDF MODEL

Beyond the population mean, BACON's augmented estimating-equation (AEE) machinery extends to quantiles of the outcome. For each labeled calibration draw we fit the full ridge conditional-mean model, obtain cross-fitted residuals on the labeled items, and estimate the conditional CDF as $\widehat{F}(t \mid X) = \widehat{G}(t - \widehat{\mu}(X))$, where $\widehat{G}$ is the empirical CDF of the cross-fitted residuals. We then apply the AEE correction at each threshold, monotonize the resulting CDF, and invert it to recover quantiles. This uses a single distributional outcome-model fit per draw, avoids a Gaussian residual assumption, and retains the observed outcome support for inversion. The sample-quantile baseline and the residual-CDF estimator are evaluated on the same uniform calibration draws.

We take WebDesign as the primary quantile evidence: 100 Monte Carlo draws per cell, all six outcomes, median only (Table 14). It is intentionally separated from PERSUADE, whose coarse six-point scores make this budget range a ceiling-effect check rather than a clean test.

Residual-CDF AEE reduces median MSE in all 12 outcome-by-budget cells, by between 4% and 49%, while retaining 94–100% coverage. The largest gains occur at the tighter 10% budget, where the direct sample-quantile estimator has the highest variance. This supports the quantile formulation in a non-degenerate multi-outcome setting; it does not imply that every quantile, outcome model, or domain will benefit equally.

MQM provides a broader check across the lower, median, and upper quartiles (100 draws per cell; Table 15). Residual-CDF AEE improves MSE in five of six cells, including both median budgets; the sole exception is Q75 at the 10% budget. Coverage is 95–100%.

We also evaluated a 16-bin monotone cumulative-link (ordered-logit) CDF model. It was not competitive overall and showed weak MQM and WebDesign tail coverage in some cells, so it is not used for the claims above.

B MODEL-JUDGE SCORING PROTOCOL

For each item, we query a panel of language or vision-language judges through a unified API endpoint. Prompts enforce strict scalar outputs to reduce parse ambiguity and improve comparability across models. We use fixed templates:

Essay scoring (LLMs).

```
You are an essay grader.  You will be given a student essay
and its writing prompt.
Rate the essay's holistic quality on a scale of 1 to 6:
1 = Very poor, 2 = Poor, 3 = Below average, 4 = Above average,
5 = Good, 6 = Excellent.
You MUST respond with ONLY a single digit (1, 2, 3, 4, 5, or
6).
No explanation, no reasoning, no other text - just the number.

Essay prompt:  {prompt_name}

Essay:
{full_text}
```

Each judge receives this system and user text; we take the first substring matching a digit in $\{1, \ldots, 6\}$ as the score (same rule as in the scoring script).

Web-design scoring (VLMs).

```
You are a strict visual evaluator.  Return only the requested
six lines.

Rate this screenshot from a {domain_label} on six dimensions.
Return ONLY six integer scores in [-3, -2, -1, 0, 1, 2, 3],
one per line, exactly:
AE: <score>
TRU: <score>
TYP: <score>
EXMPL: <score>
AVG: <score>
US: <score>

Definitions:
- AE (Aesthetics):  visually appealing overall design.
- TRU (Trustworthiness):  appears credible and trustworthy.
- TYP (Typicality):  {prompt_typ}
- EXMPL (Exemplar Goodness):  {prompt_exmpl}
- AVG (Family Resemblance):  {prompt_avg}
- US (Usability):  easy to navigate and understand.
```

The placeholders {domain_label}, {prompt_typ}, {prompt_exmpl}, and {prompt_avg} are filled from a domain table (university/college, commercial-bank, online fashion-shopping, online homeware-shopping) matching our dataset splits. Each judge receives this user text together with one screenshot (system role as above); responses are parsed as six lines of the form `AE: ...`, `TRU: ...`, etc.

Machine translation scoring (LLMs).

```
You are a professional MQM annotator following the official
WMT MQM human evaluation guidelines at the segment level.
Identify up to five target-side errors, each with a category
(Accuracy, Fluency, Terminology, Style, Locale conventions,
or Other, including subcategories such as Mistranslation
or Grammar; also Source error or Non-translation when
```

```
applicable), a severity (Major, Minor, or Neutral), and the
exact span from the source or hypothesis.
If the translation is too garbled or unrelated to the source,
return exactly one error:  Non-translation with severity
Major.
Source errors are listed separately and do not count toward
the five target-error limit.
Do not double-count the same underlying issue; when multiple
issues overlap one span, keep only the most severe; when
severities tie, prefer Accuracy > Fluency > Terminology >
Style > Locale conventions > Other.
Be conservative and return JSON only (no prose outside the
object).

[Source]
{source}

[Hypothesis]
{hypothesis}

Return JSON only:  a single object with an errors array; each
entry has string fields category, severity, and span.
```

Each judge receives the source segment and the system hypothesis (machine translation) and returns structured segment-level MQM annotations in that JSON format.

For every API response, we log token-level statistics (including per-token log probabilities and top alternatives when available) and derive uncertainty summaries such as mean token entropy and perplexity. Failed requests are retried with bounded concurrency and backoff; unresolvable failures are marked `NA` and excluded from successful-completion counts during resume runs.

C Choice of Outcome Models

Table 16: Example outcome models in BACON. All models use embeddings e_i, judge scores ℓ_i, and uncertainty features u_i as inputs.

Score pattern	Example model	Key idea
Real-valued	Ridge regression	Linear predictor $\widehat{y}_i = \alpha_0 + \alpha_e^\top e_i + \alpha_\ell^\top \ell_i + \alpha_u^\top u_i$; ℓ_2 penalty controls judge collinearity.
Non-negative, zero-inflated	Hurdle model (logistic gate + conditional ridge)	Gate $\Pr(z_i{=}1)$ predicts whether $y_i{>}0$; conditional ridge predicts $\widehat{y}_i^+$ on positive examples; final prediction $\widehat{y}_i = \Pr(z_i{=}1) \cdot \widehat{y}_i^+$.
Ordinal	Proportional-odds (ordered logit)	Linear predictor $\eta_i = \alpha_0 + \alpha_e^\top e_i + \alpha_\ell^\top \ell_i + \alpha_u^\top u_i$ with $K{-}1$ ordered cutpoints $\theta_1 < \cdots < \theta_{K-1}$; $\Pr(Y_i{\le}k) = \sigma(\theta_k - \eta_i)$.
Categorical	Multinomial regression	$\Pr(Y_i{=}k) = \exp(\eta_i^{(k)}) / \sum_{j=1}^{K} \exp(\eta_i^{(j)})$, where $\eta_i^{(k)} = \alpha_0^{(k)} + \alpha_e^{(k)\top} e_i + \alpha_\ell^{(k)\top} \ell_i + \alpha_u^{(k)\top} u_i$.

D Summary metrics estimation

D.1 An estimating equation framework

Let Y be the observed human scores (with `NA` if no human label is collected), S be the indicator variable indicating whether a human label is collected, X be some measurements for the context, such as the embedding features, the judge scores and uncertainty summary metrics, etc. Below we provide an estimating equation framework (Hardin & Hilbe, 2002).

Suppose we are interested in a target parameter $\theta^\star = \theta(\mathbb{P}_{Y,S,X})$, which is the solution to the following estimating equation: $\mathbb{E}\{m(Y, S, X; \theta^\star, \gamma^\star)\} = 0$, where $m(Y, S, X; \theta, \gamma)$ is the estimating equation, and $\gamma^\star$ is a nuisance parameter. For example, for the mean and τ-th quantile, we have

$$m_{\text{mean}}(Y, S, X; \theta, \gamma) = \frac{SY}{\pi(X)} - \theta, \quad \text{and} \quad m_{\text{quantile}}(Y, S, X; \theta, \gamma) = \frac{S\mathbf{1}\{Y \le \theta\}}{\pi(X)} - \tau,$$

respectively. Here $\pi(X)$ is the labeling propensity and serves as a given nuisance parameter $\gamma^\star$. In practice, we can apply many classical sampling schemes to collect the human labels, which corresponds to different choices of $\pi(X)$. For example, we can draw items uniformly at random, or perform stratified sampling based on some context features X. Alternatively, we may learn the labeling propensity from some pilot data using a propensity model. See Appendix E for more details.

It is usually desirable to incorporate a regression adjustment term to the estimating equation to improve the efficiency of the estimator. For the above examples, we can augment the estimating equation as:

$$m_{\text{mean}}(Y, S, X; \theta, \gamma) = \frac{S(Y - \mu(X))}{\pi(X)} + \mu(X) - \theta,$$
$$m_{\text{quantile}}(Y, S, X; \theta, \gamma) = \frac{S(\mathbf{1}\{Y \le \theta\} - \phi(X; \theta))}{\pi(X)} + \phi(X; \theta) - \tau,$$

respectively. Here $\mu(X)$ is a regression adjustment for the human scores and $\phi(X; \theta)$ is a regression adjustment for the quantile. We can then estimate the target parameter $\theta^\star$ by solving the following estimating equation: $\mathbb{E}_N(m(Y, S, X; \widehat{\theta}, \widehat{\gamma})) = N^{-1}\sum_{i=1}^N m(Y_i, S_i, X_i; \widehat{\theta}, \widehat{\gamma}) = 0$, which we call as the augmented estimating-equation (AEE) estimator. Under some regularity conditions (Appendix D.2), the estimator $\widehat{\theta}$ is consistent and asymptotically normal: $\sqrt{N}(\widehat{\theta} - \theta^\star) \rightsquigarrow \mathcal{N}(0, \Sigma_\theta)$. The asymptotic covariance takes the sandwich form $\Sigma_\theta = A^{-1}BA^{-T}$, where A and B are given by

$$A = \mathbb{E}\{\nabla_\theta m(Y, S, X; \theta^\star, \gamma^\star)\}, \quad \text{and} \quad B = \mathbb{E}\left\{m(Y, S, X; \theta^\star, \gamma^\star)\, m(Y, S, X; \theta^\star, \gamma^\star)^\top\right\}.$$

Accordingly, a plug-in sandwich estimator of $\text{Var}\{\widehat{\theta}\}$ is $\widehat{v} = \frac{1}{n}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-T}$, with $\widehat{A}$ and $\widehat{B}$ given by

$$\widehat{A} = \mathbb{E}_N \nabla_\theta m(Y, S, X; \widehat{\theta}, \widehat{\gamma}), \quad \text{and} \quad \widehat{B} = \mathbb{E}_N m(Y, S, X; \widehat{\theta}, \widehat{\gamma})\, m(Y, S, X; \widehat{\theta}, \widehat{\gamma})^\top.$$

Then we can construct a confidence interval for $\theta^\star$ as $\widehat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{v}}$ with confidence level $1 - \alpha$. When the target is the population mean, the AEE estimator is equivalent to the inverse-propensity weighted estimator upon regression adjustment in survey sampling (Särndal et al., 1992; Cassel et al., 1976) or predictive powered inference (Angelopoulos et al., 2023a). For a rigorous technical presentation, please refer to Appendix D.2.

D.2 Properties of the estimating equation estimator

In this section, we will present the theoretical properties of the estimating equation estimator.

Theorem 1 (Consistency and asymptotic normality of the AEE estimator)**.** *Let $\theta^\star \in \Theta \subset \mathbb{R}^p$ be the unique solution to*

$$\Psi(\theta, \gamma^\star) := \mathbb{E}\{m(Y, S, X; \theta, \gamma^\star)\} = 0.$$

Assume:

1. ***Identification.*** *$\theta^\star$ is the unique zero of $\Psi(\theta, \gamma^\star)$ in a neighborhood of the truth.*
2. ***Positivity and MAR.*** *There exists $\epsilon > 0$ such that $\pi^\star(X) \ge \epsilon$ almost surely, and $S \perp Y \mid X$.*
3. ***Smoothness.*** *The map $\theta \mapsto m(Y, S, X; \theta, \gamma)$ is continuously differentiable near $\theta^\star$, and*
$$A := \mathbb{E}\{\nabla_\theta m(Y, S, X; \theta^\star, \gamma^\star)\}$$
exists and is nonsingular.
4. ***Nuisance estimation.*** *$\|\widehat{\gamma} - \gamma^\star\| = o_p(1)$ and*
$$\mathbb{E}_N[m(Y, S, X; \theta^\star, \widehat{\gamma})] - \mathbb{E}_N[m(Y, S, X; \theta^\star, \gamma^\star)] = o_p(N^{-1/2}).$$
This holds, for example, under Neyman orthogonality together with standard consistency rates for the nuisance estimators.

5. ***Empirical process control.*** *A uniform law of large numbers and a central limit theorem hold for the relevant class of score functions in a neighborhood of* $(\theta^\star, \gamma^\star)$.

If $\widehat{\theta}$ *satisfies*

$$\mathbb{E}_N\Big[m(Y, S, X; \widehat{\theta}, \widehat{\gamma})\Big] = o_p(N^{-1/2}),$$

then $\widehat{\theta} \xrightarrow{p} \theta^\star$. *Moreover,*

$$\sqrt{N}(\widehat{\theta} - \theta^\star) = -A^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} m(Y_i, S_i, X_i; \theta^\star, \gamma^\star) + o_p(1),$$

and hence

$$\sqrt{N}(\widehat{\theta} - \theta^\star) \rightsquigarrow \mathcal{N}(0, \Sigma_\theta), \qquad \Sigma_\theta = A^{-1}BA^{-T},$$

where

$$B = \mathbb{E}\left\{m(Y, S, X; \theta^\star, \gamma^\star)m(Y, S, X; \theta^\star, \gamma^\star)^\top\right\}.$$

Corollary 1 (Consistency of the sandwich variance estimator)**.** *If, in addition,* $\widehat{A} \xrightarrow{p} A$ *and* $\widehat{B} \xrightarrow{p} B$, *then*

$$\widehat{\mathrm{Var}\left\{\widehat{\theta}\right\}} = \frac{1}{n}\widehat{A}^{-1}\widehat{B}\widehat{A}^{-T}$$

is a consistent estimator of the asymptotic covariance of $\widehat{\theta}$.

Remark 1. *This theorem applies directly to the augmented estimating equations for the population mean and quantiles introduced above. In those cases, the nuisance functions are the labeling propensity* $\pi(X)$ *together with the regression adjustments* $\mu(X)$ *or* $\phi(X; \theta)$. *Neyman orthogonality is especially useful because it makes the estimator insensitive, to first order, to regularization bias from flexible nuisance fits.*

D.3 Proof of Theorem 1

Proof. Consistency follows from standard Z-estimation arguments. By the uniform law of large numbers and nuisance consistency,

$$\sup_{\theta \in \Theta} \|\mathbb{E}_N[m(Y, S, X; \theta, \widehat{\gamma})] - \Psi(\theta, \gamma^\star)\| = o_p(1).$$

Since $\Psi(\theta, \gamma^\star)$ has a unique zero at $\theta^\star$, any approximate empirical root must converge in probability to $\theta^\star$.

For asymptotic normality, expand the empirical moment around $(\theta^\star, \gamma^\star)$:

$$0 = \mathbb{E}_N[m(Y, S, X; \widehat{\theta}, \widehat{\gamma})] = \mathbb{E}_N[m(Y, S, X; \theta^\star, \gamma^\star)] + A(\widehat{\theta} - \theta^\star) + r_N,$$

where $r_n = o_p(n^{-1/2})$ by differentiability, stochastic equicontinuity, and the assumed negligible first-order effect of nuisance estimation. Rearranging gives

$$\sqrt{N}(\widehat{\theta} - \theta^\star) = -A^{-1}\sqrt{N}\,\mathbb{E}_N[m(Y, S, X; \theta^\star, \gamma^\star)] + o_p(1).$$

The multivariate central limit theorem yields

$$\sqrt{N}\,\mathbb{E}_N[m(Y, S, X; \theta^\star, \gamma^\star)] \rightsquigarrow \mathcal{N}(0, B),$$

and Slutsky's theorem implies the stated Gaussian limit with covariance $A^{-1}BA^{-T}$. □

E Extension: constrained Weighting Optimization for Adaptive Sampling

E.1 Framework

Suppose we can learn a good outcome model from the one round of labeling, we can then solve a constrained weighting objective to optimize future sampling schema under a fixed budget when

the goal is to estimate the population summary metrics. We use population means as an example. Formally, we want to solve the following problem:

$$\min_{\sigma} \sum_i \frac{r_i^2}{e(x_i)} \quad \text{subject to} \quad \frac{1}{N}\sum_i e(x_i) \le \eta,$$

where $e(x_i) = \pi + (1-\pi)\sigma(b + x_i^\top \beta)$ is the inclusion probability for item i, and b and β are the parameters of the outcome model. We use an augmented Lagrangian approach with a hinge-squared penalty for constraint violations. This formulation encourages larger weights on high-residual items while enforcing a global sampling-rate cap.

We propose the following finite-population labeling workflow:

1. Draw an initial random sample (stage 1), collect human labels, and fit an outcome model f.
2. Fit a propensity model σ from features (primarily embedding PCs) to approximate hardness-informed inclusion scores.
3. Sample additional items adaptively (stage 2) using probabilities derived from σ under an overall budget η.
4. Estimate the population mean using the AEE estimator that combines inverse-propensity correction with outcome-model augmentation.

This objective prioritizes high-residual items while enforcing a global sampling-rate cap. This is relevant to active statistical inference literature: (Li et al., 2025; Hamilton et al., 2025).

E.2 Variance characterization of the AEE estimator

First, let's think about the variance of the AEE estimator. For simplicity, let's assume that the outcomes are i.i.d. draws from a super population (y, x) from some distribution P. Then, we can write the AEE estimator as

$$\widehat{\mu} = \frac{1}{N}\sum_i \frac{S_i(y_i - f(x_i))}{e(x_i)} + f(x_i).$$

Due to independence, the variance of $\widehat{\mu}$ is given by

$$\frac{1}{N}\mathrm{Var}\left\{\frac{S(y-f(x))}{e(x)} + f(x)\right\}.$$

Now we compute the variance term. First, notice the following decomposition:

$$\begin{aligned}
&\frac{S(y-f(x))}{e(x)} + f(x) \\
=&\underbrace{\frac{S(y-f^\star(x))}{e(x)}}_{\text{Term I}} + \underbrace{f^\star(x)}_{\text{Term II}} + \underbrace{\frac{(S-e(x))(f^\star(x)-f(x))}{e(x)}}_{\text{Term III}},
\end{aligned}$$

Note that Term I, II and III are mutually orthogonal, i.e.

$$\begin{aligned}
\mathrm{Cov}\left\{\text{Term I}, \text{Term II}\right\} =& 0 \\
\mathrm{Cov}\left\{\text{Term I}, \text{Term III}\right\} =& 0 \\
\mathrm{Cov}\left\{\text{Term II}, \text{Term III}\right\} =& 0.
\end{aligned}$$

Therefore, the variance of $\widehat{\mu}$ is given by

$$\frac{1}{N}\mathrm{Var}\left\{\text{Term I}\right\} + \frac{1}{N}\mathrm{Var}\left\{\text{Term II}\right\} + \frac{1}{N}\mathrm{Var}\left\{\text{Term III}\right\}.$$

Now we compute the variance of each term.

$$\mathrm{Var}\left\{\text{Term I}\right\} = \mathbb{E}\left\{\frac{\sigma^{*2}(x)}{e(x)}\right\}.$$

For Term II, we have

$$\mathrm{Var}\left\{\text{Term II}\right\} = \mathbb{E}\left\{(f^\star(x) - \mu)^2\right\}.$$

For Term III, we have

$$\mathrm{Var}\left\{\text{Term III}\right\} = \mathbb{E}\left\{\frac{(1-e(x))(f^\star(x) - f(x))^2}{e(x)}\right\}.$$

Therefore we summarize two forms of the variance of the AEE estimator:

- **Form 1:**

$$\mathrm{Var}\left\{\widehat{\mu}\right\} = \frac{1}{N}\mathbb{E}\left\{\frac{\sigma^{*2}(x)}{e(x)} + (f^\star(x) - \mu)^2 + \frac{(1-e(x))(f^\star(x) - f(x))^2}{e(x)}\right\}.$$

- **Form 2:**

$$\mathrm{Var}\left\{\widehat{\mu}\right\} = \frac{1}{N}\mathbb{E}\left\{\frac{(y - f(x))^2}{e(x)} + (f^\star(x) - \mu)^2 - (f^\star(x) - f(x))^2\right\}.$$

What are we actually optimizing when we optimize the propensity score? From Form 1, we can see that given a fixed propensity model, we can benefit from fitting a better outcome model to reduce the variance of the AEE estimator.

From Form 2, we can see that given a fixed outcome model, we can benefit from fitting a better propensity model to reduce the variance of the AEE estimator. More importantly, the best choice of propensity model should be proportional to the square root of the conditional quadratic error of the outcome model. To show this, let's try to optimize the propensity model based on Form 2. Define the conditional squared error as:

$$\sigma_f^2(x) = \mathbb{E}\left\{(y - f(x))^2 \mid x\right\}.$$

When $f = f^\star$, we have $\sigma_f^2(x) = \sigma^{*2}(x)$. By Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left\{\sigma_f(x)\right\}^2 \leq \mathbb{E}\left\{\frac{\sigma_f^2(x)}{e(x)}\right\} \cdot \mathbb{E}\left\{e(x)\right\} \leq \eta\mathbb{E}\left\{\frac{\sigma_f^2(x)}{e(x)}\right\}.$$

Therefore, we have the following lower bound:

$$\mathbb{E}\left\{\frac{\sigma_f^2(x)}{e(x)}\right\} \geq \frac{\mathbb{E}\left\{\sigma_f(x)\right\}^2}{\eta}.$$

The equality holds when

$$e(x) \propto \sigma_f(x).$$

That is, the best choice of propensity model should be proportional to the square root of the conditional quadratic error of the outcome model. This is consistent with the intuition that we should sample more items that are harder to predict.

E.3 Synthetic experiments on summary metrics

To isolate the benefit of adaptive sampling under controlled heteroskedasticity, we construct a synthetic labeling environment that inherits the covariate structure of the real PERSUADE corpus but replaces human scores with outcomes drawn from a known data-generating process (DGP).

E.3.1 Data-Generating Process

Let $e_i \in \mathbb{R}^d$ denote the OpenAI text embedding for essay i, and let $\bar{\ell}_i$ be the average LLM-judge score across the model panel. We compute principal components from the standardised embedding matrix and define the conditional mean

$$\mu_i \;=\; \sum_{j=1}^{5} \mathrm{PC}_{ij} \;+\; \bar{\ell}_i,$$

so that the outcome depends on both intrinsic essay features (through the first five PCs) and the aggregate model-judge signal. Conditional variance is heteroskedastic in the first principal component:

$$\sigma_i^2 \;=\; 0.1\,\exp\bigl(\lambda\,\mathrm{PC}_{i1}\bigr),$$

with $\lambda > 0$ governing the severity of heteroskedasticity. For each item we draw

$$Y_i \;\sim\; \mathcal{N}\bigl(\mu_i,\ \sigma_i^2\bigr).$$

This DGP makes high-PC1 items intrinsically noisier while low-PC1 items remain relatively clean, creating a directional heteroskedasticity that the propensity model can exploit by selectively oversampling the noisy tail.

E.3.2 Simulation Protocol

We sweep $\lambda \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$ to trace the effect of increasing heteroskedasticity. For each value of λ, the synthetic outcomes are generated once and held fixed across Monte Carlo trials. Each trial then executes one of three estimation strategies under a common total labeling budget η:

1. **Baseline 1 (Direct Sample Mean).** Draw $\lfloor \eta N \rfloor$ items uniformly at random, collect their labels, and report the sample mean.
2. **Baseline 2 (One-Stage estimation).** Draw $\lfloor \eta N \rfloor$ items uniformly, fit an outcome model $\widehat{f}$ by ridge regression on embedding PCs, and form an AEE estimator with constant propensity $e(x_i) = \eta$.
3. **Two-Stage Adaptive estimation.** Draw a stage-1 pilot sample at rate $\pi < \eta$. Fit an outcome model $\widehat{f}$ and compute squared residuals $r_i^2 = (Y_i - \widehat{f}(x_i))^2$ on the pilot. Fit a propensity model $\sigma(x_i)$ on embedding PCs to predict hardness, then draw stage-2 items with inclusion probability $e(x_i) = \pi + (1-\pi)\sigma(x_i)$ subject to a mean-budget constraint $\frac{1}{N}\sum_i e(x_i) \le \eta$. Combine both stages via an AEE estimator with item-specific propensities $e(x_i)$.

The propensity model solves the constrained weighting problem described above using an augmented Lagrangian optimizer, where $e(x_i) = \pi + (1-\pi)\,\sigma(b + x_i^\top \beta)$ and (b, β) are chosen to minimize $\sum_i r_i^2 / e(x_i)$ subject to $\frac{1}{N}\sum_i e(x_i) \le \eta$. We allow separate PCA dimensionalities for the outcome model and the propensity model.

Performance is assessed by the root mean squared error (RMSE) of each estimator relative to the true population mean $\mu = N^{-1}\sum_i \mu_i$, computed over repeated Monte Carlo trials. We additionally report bias and standard deviation to diagnose whether improvements stem from variance reduction or bias correction.

E.3.3 Synthetic Simulation Results

Figure 10 reports the bias, variance, and MSE of all three estimators as λ increases. The direct sample mean (Baseline 1) maintains near-zero bias but its variance stays flat and high across all λ values, as uniform sampling cannot adapt to the heteroskedastic noise structure. The one-stage AEE estimator (Baseline 2) substantially reduces variance relative to the direct mean at low λ by leveraging the outcome model, but its variance grows steadily with λ as residual heteroskedasticity overwhelms the constant-propensity correction. The two-stage adaptive AEE estimator achieves the lowest MSE across all λ values. Its advantage widens as heteroskedasticity increases: at $\lambda = 3.0$ the two-stage MSE is roughly half that of the one-stage AEE estimator and over an order of magnitude below the direct mean. All three estimators remain approximately unbiased throughout the sweep, confirming that the MSE gains are driven entirely by variance reduction through adaptive allocation of the labeling budget toward the noisy tail of the PC1 distribution.

E.3.4 Sensitivity to Pilot Rate and Budget

We next fix $\lambda = 1.5$ and study how estimation quality depends on the pilot rate π and the total budget η.

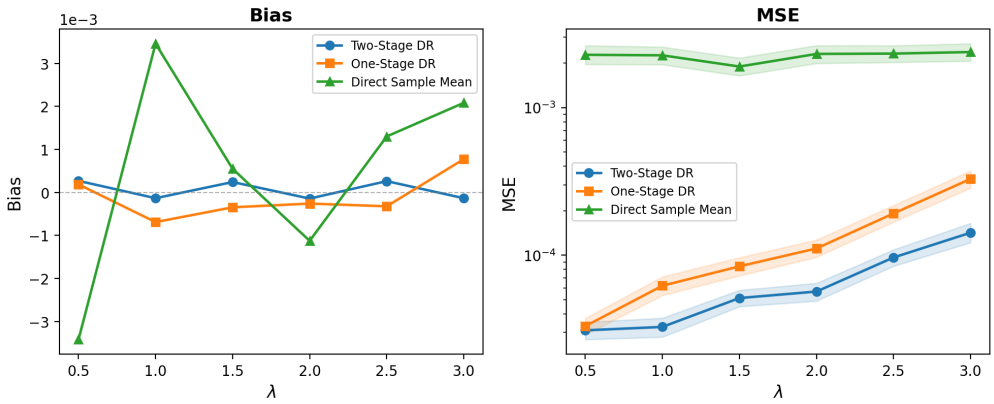


Figure 10: Synthetic simulation: bias and MSE of three population-mean estimators as a function of the heteroskedasticity parameter λ. Budget $\eta = 0.40$, pilot rate $\pi = 0.15$, 400 Monte Carlo trials per λ value.

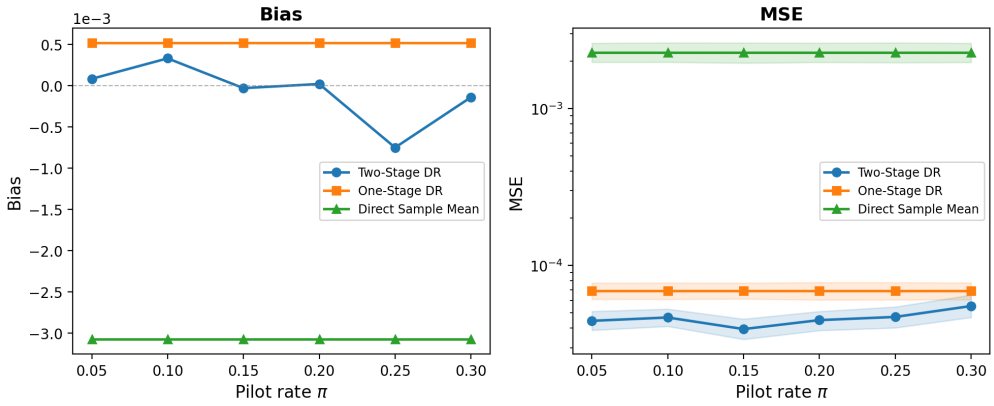


Figure 11: Sensitivity to pilot rate π with fixed budget $\eta = 0.40$ and $\lambda = 1.5$. Left: bias; right: MSE (log scale). 400 trials per setting.

Varying π (fixed $\eta = 0.40$). Figure 11 sweeps $\pi \in \{0.05, 0.10, \ldots, 0.30\}$. Because the baselines use the full budget η in a single uniform draw, their MSE is invariant to π. The two-stage AEE estimator is relatively stable across this range, with a slight optimum near $\pi = 0.15$; very small pilots ($\pi = 0.05$) leave the outcome model under-trained, while large pilots ($\pi \geq 0.25$) consume too much budget before the adaptive stage can act. All estimators remain approximately unbiased.

Varying η (fixed $\pi = 0.15$). Figure 12 sweeps $\eta \in \{0.20, 0.25, \ldots, 0.45\}$. As expected, MSE decreases monotonically with budget for all three methods. The two-stage AEE estimator maintains a consistent advantage over both baselines at every budget level: even at the tightest budget ($\eta = 0.20$) it achieves lower MSE than the direct mean at $\eta = 0.45$. The relative improvement of two-stage AEE estimator over one-stage AEE estimator is largest at small η, where the adaptive allocation of the scarce remaining budget after the pilot is most valuable.

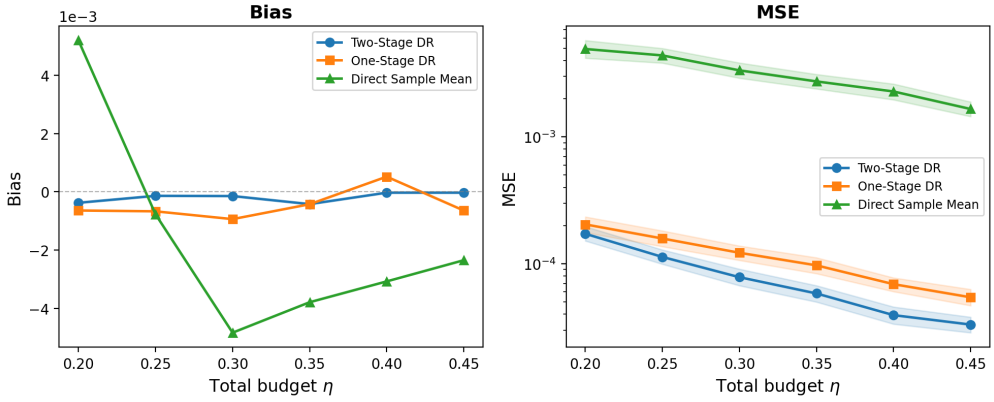


Figure 12: Sensitivity to total budget η with fixed pilot rate $\pi = 0.15$ and $\lambda = 1.5$. Left: bias; right: MSE (log scale). 400 trials per setting.