Absolute Error Is Not Enough: Ranking Model Performance Without Labels to Estimate Disparities at Scale

Veronica Rammouz[1], **Aaron Gonzalez**[1], **Carlos Cruzportillo**[1], **Adrian Tan**[1], **Nicole Beebe**[2], **Ke Yang**[1], **Anthony Rios**[1]
[1]The University of Texas at San Antonio
[2]Illinois Institute of Technology
{Veronica.Rammouz, Anthony.Rios}@utsa.edu

Abstract

Estimating model performance without labeled data is important for deploying NLP systems across diverse domains. Prior work has proposed proxy measures based on distribution shift or predicted correctness, but it remains unclear whether these methods reliably preserve the *ranking* of domains, which is critical for model selection and risk triage. We study this problem from an ordinal perspective using a two-step framework that pairs a base classifier with an error predictor to estimate instance-level failures and disparities and aggregate them into label-free performance estimates. Across multiple sample sizes from the *GeoOLID* and *Amazon Reviews* datasets, large language model–based error predictors substantially outperform drift-based baselines in ρ and τ rank correlations. We further show that lower absolute estimation error ($RMSE$) does not necessarily imply better ranking, highlighting the importance of evaluating ordinal consistency directly.

WARNING: This paper contains offensive and hateful speech in the appendix.

1 Introduction

Evaluating NLP systems across every domain, population, and setting is expensive. In deployment, performance can vary across topics, regions, social groups, and language varieties. A model may perform well on average while still failing for specific cities, dialect communities, or domains. Because collecting labels for every setting is costly, practitioners often rely on proxy signals such as dataset similarity, linguistic drift, or domain shift to estimate generalization (Chang et al., 2023; Ramesh Kashyap et al., 2021; Pudasaini et al., 2025; Elsahar and Gallé, 2019; Sun et al., 2025).

Prior work provides useful starting points. Studies of domain shift show that vocabulary, structure, topic, and linguistic drift can affect model robustness (Chang et al., 2023; Ramesh Kashyap et al., 2021). Other work estimates transfer performance using dataset similarity, including n-gram overlap (Xia et al., 2020), or studies how models such as BERT and GPT adapt across domains (Patankar et al., 2022). Recent work also extends performance prediction to multilingual and low-resource settings (Khiu et al., 2024) and to correctness estimation, where large models have limited knowledge of their own correctness (Xiao et al., 2025).

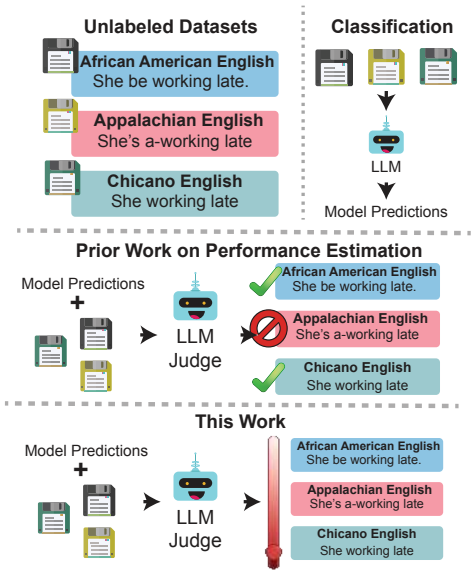


Figure 1: High-level motivation for our study to use ranking as a way for large-scale auditing of performance disparities across multiple factors.

However, most work treats performance estimation as predicting an overall score. Less is known about whether these methods can identify where models work best or worst. This matters for fairness and deployment: practitioners may need to know which cities, dialect communities, or domains should be audited first, where labels are most needed, and where model outputs should be treated with caution. We therefore study performance ranking as a measurement problem for large-scale model auditing. Hence, we ask the following research questions:

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2290–2307
November 6-10, 2026 ©2026 Association for Computational Linguistics

- **RQ1:** How reliably can current performance estimation methods rank datasets and domains by expected model performance?

- **RQ2:** When do performance rankings remain stable or break down across model families, dataset sizes, and domain shifts?

- **RQ3:** How feasible is it to use LLMs as judges to estimate performance at scale across dialects and geographic regions that may be underserved in many datasets?

To answer these questions, we compare auxiliary error models, drift-based heuristics, similarity-based estimates, and LLM-as-judge estimates across multiple classifiers and datasets. We train base classifiers using RoBERTa (Liu, 2019), a linear model, GPT4o-mini (OpenAI, 2023), and LLaMa-3.1-8B-Instruct (Touvron et al., 2023). We then test whether estimated errors recover the relative ordering of model performance across target domains, including settings organized by dialect and geographic region.

Our results show that LLM-as-judge error estimators can often recover reliable performance rankings, especially compared to traditional drift and similarity baselines. On *Amazon Reviews*, LLM judges substantially improve rank correlations and better recover top domains for auditing. On *GeoOLID-Dial*, they produce strong rankings across dialect-mapped targets, suggesting that this approach can help identify language varieties that need further evaluation. However, ranking remains sensitive to dataset structure, domain heterogeneity, and the spread of true performance across domains. We also find that low RMSE does not necessarily imply good ranking, showing that ordinal reliability must be evaluated directly.

Our contributions are: **(1.)** We present a large-scale study of whether label-free performance estimators can rank datasets and domains by expected performance. **(2.)** We introduce a controlled two-step evaluation framework for comparing ranking reliability across model families, datasets, and domain shifts. **(3.)** We show that LLM-as-judge error estimators often provide stronger rankings than traditional drift-based methods, while also identifying cases where ranking remains unstable under realistic distributional variation.

2 Related Work

Instance-Level Complexity. Prior work studies dataset and instance-level complexity to understand task difficulty and its relationship to model performance (Cook et al., 2025; Lorena et al., 2024; Barella et al., 2021). These methods use complexity metrics, such as feature importance or class overlap, or model-based signals, such as cross-validation or training loss, to improve model selection. However, they usually require labels, making them difficult to use in large-scale deployment settings. They also do not directly study fine-grained model failures under distribution shift, though related work has examined error-based estimation and the effects of concept drift on failure prediction (Poenaru-Olaru et al., 2022; Eberlein et al., 2025; Zhou et al., 2023; Kwon et al., 2024; Mallick et al., 2022). We instead study failure-informed, label-free ranking across model variants and shifted domains.

Dataset Drift. A large body of work quantifies dataset or covariate drift using token-frequency divergences, TF-IDF vectors, and embedding-based distances (Ramesh Kashyap et al., 2021; Dredze et al., 2010; Elsahar and Gallé, 2019; Ruder et al., 2017). These measures have been used to predict performance degradation under domain shift (Feldhans et al., 2021), with Jensen-Shannon divergence often used to measure distributional change (Lin, 2002). More recent work decomposes drift into vocabulary, structural, and semantic dimensions, showing that different types of linguistic variation can affect robustness in different ways (Chang et al., 2023). This is especially important for domains with jargon, regional language, or dialectal variation (Yamshchikov et al., 2021; Ramesh Kashyap et al., 2021). In contrast to drift-only estimation, we evaluate whether failure-based estimators can more reliably rank out-of-domain datasets when labels are unavailable.

Model-Based Methods. Model-based approaches use representations, such as contextual embeddings from BERT or RoBERTa, to predict transfer performance across datasets (Feldhans et al., 2021). Fine-tuned embedding distances can capture syntactic and semantic differences missed by simpler frequency metrics and can rank examples by expected performance (Elango et al., 2022). However, these methods are often less reliable for dataset-level performance prediction in out-

of-domain settings (Nigenda et al., 2022). Prior work also notes the need for robust low-data metrics when labels are available for only one domain (Ramesh Kashyap et al., 2021; Patankar et al., 2022). We build on this setting by using error prediction models trained or prompted from source-domain failures to estimate and rank performance across unseen domains.

LLM-as-a-Judge. LLMs are increasingly used as evaluators to replace or augment human annotation in benchmarking and meta-evaluation (Gu et al., 2024). Models such as GPT-4 can approximate human preferences and produce useful rankings across tasks (OpenAI, 2023; Zheng et al., 2023). This has led to both prompted evaluators, such as ALPACAEVAL (Li et al., 2023), and fine-tuned judge models, such as PANDALM, JUDGELM, and PROMETHEUS (Wang et al., 2023; Zhu et al., 2023; Kim et al., 2023). Recent work also studies calibrated and correctness-aware judging through self-evaluation and reasoning-based feedback (Bai et al., 2022; Hao et al., 2023), while meta-evaluation work shows that judge signals may not always align with factual correctness and can lack robustness across domains (Trung et al., 2024; Zhuge et al., 2024). Most related to our work, Generalized Correctness Models show that LLMs do not have privileged self-knowledge, but historical prediction data and calibration can improve cross-model correctness prediction (Xiao et al., 2025). We ask a complementary question: whether correctness estimates preserve ordinal relationships across datasets and models. Rather than focusing only on calibrated probabilities or absolute accuracy, we study ranking robustness under prediction error.

3 Experimental Design

Problem Statement. We study label-free performance ranking across domains. Intuitively, suppose a classifier is deployed across many cities, product categories, dialect communities, or other domains. Collecting labels for every domain may be expensive or infeasible, but practitioners may still need to know where the model is likely to work well and where it is likely to fail. This is especially important for model auditing, where the goal is not only to estimate average performance, but also to identify which domains should be inspected first.

Formally, let $\mathcal{K}$ denote a set of domains. For each source domain $s \in \mathcal{K}$, we train or prompt a base classifier $f_\theta^{(s)}$ using labeled source data. For each target domain $t \in \mathcal{K} \backslash \{s\}$, we observe only unlabeled inputs $\mathcal{U}^{(t)} = \{x_i^{(t)}\}_{i=1}^{n_t}$. A label-free estimator assigns each target domain an estimated performance score $\hat{p}^{(s,t)}$ without using target-domain labels. We evaluate whether these estimated scores recover the true ordering of domains induced by $p^{(s,t)}$, the target-domain performance computed from gold labels only after estimation.

We therefore frame the task as a benchmark problem rather than a new model proposal. Figure 2 summarizes the benchmark. For each source domain, we train or prompt a base classifier, apply several label-free estimators to held-out target domains, rank the target domains by their estimated scores, and compare this estimated ranking against the true ranking. This lets us test whether existing drift-based and error-prediction estimators can support large-scale model auditing when target-domain labels are unavailable.

Source–Target Evaluation Protocol. Let $\mathcal{K}$ denote a set of domains, such as cities in GeoOLID or product categories in Amazon Reviews. For each source domain $s \in \mathcal{K}$, we assume access to labeled source data $\mathcal{D}^{(s)} = \{(x_i^{(s)}, y_i^{(s)})\}_{i=1}^{n_s}$. The remaining domains $\mathcal{K} \setminus \{s\}$ are treated as target domains. For each target domain t, the estimator observes only the unlabeled inputs $\mathcal{U}^{(t)} = \{x_i^{(t)}\}_{i=1}^{n_t}$. Gold labels for target domains are hidden during estimation and are used only after estimation to compute the true performance ranking. This lets us test whether a method can identify relatively high- and low-performing domains without using target-domain labels.

Base Classifiers. For each source domain s, we train or prompt a base classifier $f_\theta^{(s)}$ on the downstream task. Given an input x_i, the classifier produces a predicted label $\hat{y}_i = f_\theta^{(s)}(x_i)$. For encoder-based models such as ROBERTA-BASE, we fine-tune on labeled source data using standard cross-entropy, $\mathcal{L}_{\text{task}}(\theta) = -\frac{1}{n_s} \sum_i \log P_\theta(y_i^{(s)} \mid x_i^{(s)})$. For instruction-tuned LLMs, we use few-shot prompting rather than parameter updates. The benchmark is model-agnostic: any base classifier that produces task predictions can be evaluated under the same protocol.

Label-Free Performance Estimators. We compare two broad families of label-free performance estimators. The first family consists of distribution-

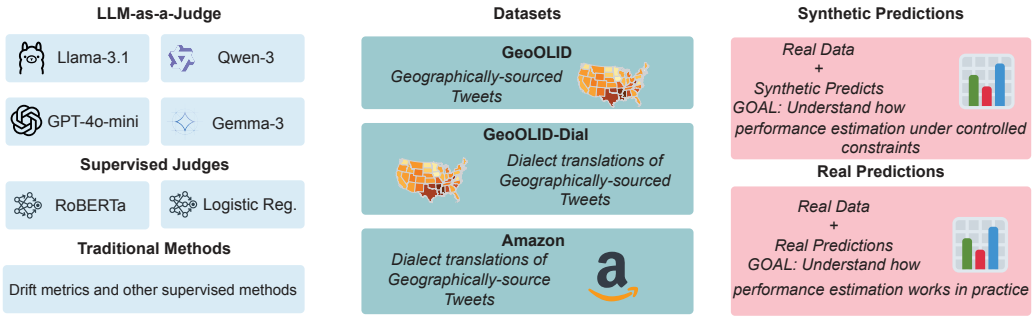


Figure 2: Overview of our experimental setup.

shift estimators. These methods compare the source and target domains using drift or similarity measures, such as token-frequency divergence, semantic distance, or covariate shift. They do not directly model the base classifier's errors. Instead, they assume that larger source–target differences are associated with larger performance degradation.

The second family consists of error-prediction estimators. These methods estimate performance by predicting whether the base classifier's output is likely to be correct. On labeled source data, we can observe the base classifier's errors as $e_i^{(s)} = \mathbb{1}[f_\theta^{(s)}(x_i^{(s)}) \neq y_i^{(s)}]$. An error predictor g_ϕ is then trained or prompted to estimate $P(e_i = 1 \mid x_i, \hat{y}_i)$. This converts the original classification task into a binary correctness-prediction task: given an input and the base classifier's predicted label, predict whether that prediction is correct.

For supervised error predictors, such as linear models or RoBERTa-based models, we train g_ϕ using binary cross-entropy over the observed source-domain error labels $e_i^{(s)}$. For LLM-based error predictors, we do not train parameters. Instead, we provide in-context examples of correct and incorrect base-model predictions and ask the LLM to judge whether a new prediction is likely to be correct. Complete prompts and demonstration-selection details are provided in the Appendix.

Estimating Target-Domain Performance. For each unlabeled target domain $\mathcal{U}^{(t)}$, we first apply the source-trained base classifier $f_\theta^{(s)}$ to obtain predictions $\hat{y}_i^{(t)}$. We then apply a label-free estimator to produce an estimated performance score $\hat{p}^{(s,t)}$ for that target domain.

For error-prediction estimators, we estimate accuracy by subtracting the predicted error rate from one:

$$\hat{p}^{(s,t)} = 1 - \frac{1}{|\mathcal{U}^{(t)}|} \sum_{x_i^{(t)} \in \mathcal{U}^{(t)}} \mathbb{1}[g_\phi(x_i^{(t)}, \hat{y}_i^{(t)}) > 0.5].$$

For LLM-based judges, we aggregate the model's binary correctness judgments over all target-domain examples. For drift-based estimators, we aggregate the corresponding source–target drift or similarity score and use it as a proxy for target-domain performance.

Ranking and Estimation Evaluation. To evaluate how well we estimate performance, we explore multiple metrics and ranking and non-ranking methods. First, for ranking metrics, we want our performance estimates to order the groups from most to least accurate in the same way as the ground-truth accuracy.

Formally, for a base classifier trained on source domain s, the true accuracy on target domain t is $p^{(s,t)} = \frac{1}{|\mathcal{D}^{(t)}|} \sum_i \mathbb{1}[f_\theta^{(s)}(x_i^{(t)}) = y_i^{(t)}]$. Our estimate of the performance of a model trained on domain s and applied to target domain t is $\hat{p}^{(s,t)}$. We rank each target domain $t \in T$ as $\hat{R} = [\hat{p}^{(s,1)}, \hat{p}^{(s,2)}, \ldots, \hat{p}^{(s,T)}]$ and compare this estimated ranking against the true ranking induced by $R = [p^{(s,1)}, p^{(s,2)}, \ldots, p^{(s,T)}]$, where the closer the orders match, the better.

We evaluate estimators from four complementary perspectives. First, we measure overall ordinal agreement using Spearman's ρ (Spearman, 1961) and Kendall's τ (Kendall, 1938). Spearman's ρ measures monotonic agreement between the estimated and true performance rankings, while Kendall's τ measures pairwise concordance: whether pairs of domains are ordered in the same direction by the estimated and true rankings. Statistical significance for Spearman's ρ and Kendall's τ is computed using standard asymptotic correlation tests under the null hypothesis of no ordinal association between estimated and true rankings. We also report Accuracy@3 (Acc@3) and Recall@3 (Rec@3), which measure whether we rank the top-3 performing groups in the top three items of our estimated rank $\hat{R}$.

Second, as stated earlier, prior work estimates how close we can estimate performance compared to the ground truth. In this case, each group's estimated performance $\hat{p}^{(s,t)}$ is directly compared to the ground-truth performance $p^{(s,t)}$. Intuitively, if real accuracy is .80, an estimated accuracy of .81 is better than an estimated accuracy of .75 without comparison to the other group's estimated performance. For this measure, we report RMSE between estimated and true performance scores when the estimator produces accuracy-like values, computed as $\text{RMSE} = \sqrt{\frac{1}{m}\sum_t (\hat{p}^{(s,t)} - p^{(s,t)})^2}$ over the m target domains.

Finally, we evaluate the performance of the error prediction models directly, i.e., how well they predict whether an error occurs. For error-prediction estimators, we report binary F1 for predicting whether base-model outputs are correct or incorrect error labels, e.g., if the ground truth is "offensive" and the base model predicts "not offensive," then the error classifier should predict "Incorrect."

4 Experimental Setup

We present the details of the datasets we evaluate our error modeling and estimation approach to rank performance, as well as the baselines and base models to set up our downstream task.

4.1 Datasets

We use four datasets in our study: *GeoOLID*, *GeoOLID-dial*, *GeoOLID-Synth*, and *Amazon Reviews*. These are described below:

GeoOLID. The *GeoOLID* dataset was introduced by Lwowski et al. (2022) and used to investigate the geographic performance disparities of offensive language classifiers. *GeoOLID* contains over $14,000$ annotated examples of tweets from $n = 15$ geographically and demographically diverse cities across the United States.

GeoOLID-Dial. We introduce *GeoOLID-Dial*, a paired benchmark for measuring dialect-related disparities in offensive language classification. It builds on *GeoOLID* (Lwowski et al., 2022), which contains over $14,000$ annotated tweets from 15 geographically and demographically diverse U.S. cities. We apply `Multi-VALUE` (Ziems et al., 2023) to generate controlled dialectal variants of the original examples through lexical, syntactic, and morphosyntactic transformations while preserving semantic intent. The resulting dataset consists of semantically paired examples that differ mainly in dialectal surface form, allowing us to test whether offensive language classifiers produce different predictions across English dialects. Representative examples are shown in Appendix 12. The sample size in *GeoOLID-Dial* amounts to $n = 52$ dialects.

GeoOLID-Synth. We also introduce further linguistic variation into the *GeoOLID* dataset by mapping existing cities into a list of new cities to reach $n = 100$ sample cities. We randomly sample a candidate city a and conduct example-level transformations into a target city b via LLM inference using LLAMA-3.1-8B-INSTRUCT. Specifically, we ask the LLM to generate the same example as if it were spoken by an individual from target city b.

Amazon Reviews. The second dataset used in our study is the *Amazon Reviews* dataset introduced by (Hou et al., 2026). This dataset comprises over 570 million reviews and 48 million items across 33 product categories. To create a balanced and more manageable subset for analysis, we selected all distinct product categories, assigned each review a positive or negative label based on its numerical rating, and randomly sampled $1,000$ reviews for each label. This dataset is used because it has many categories to test and rank (compared to only 15 from GeoOLID), making it ideal for evaluation as a robustness check. Also, it provides clear, diverse distributional differences compared to GeoOLID-dial (which has more syntactic than semantic differences) and GeoOLID, which is more similar because differences across cities are not always direct.

4.2 Baselines and Models

We explore multiple sets of baselines and models in our experiments. First, we describe our error estimation baselines. These baselines estimate performance in a variety of ways. These methods are not LLM judges. Second, we evaluate on a wide variety of "baseline models." These models predict the main classes in all our datasets, e.g., offensive vs. not offensive for the GeoOLID dataset variations and positive vs. negative for the Amazon review datasets. Finally, we evaluate many "judges," both supervised and modern LLM judges.

Error Estimation Baselines. We evaluate drift-based baselines that estimate performance without labeled data, matching the names used in Tables 3–1 to benchmark our failure-based approach.

	$\rho_{mean} \uparrow$	$\tau_{mean} \uparrow$	$Acc@3 \uparrow$	$Rec@3 \uparrow$	$RMSE \downarrow$	$F1 \uparrow$	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
Estimation Baselines								
LINGUISTIC DATASET DRIFT (Chang et al., 2023)	.169	−.006	.342	.219	.346	—	.232	.226
SEMANTIC DRIFT (cos)	.308	.224	.196	.174	—	—	.206	.197
TOKEN FREQUENCY DIVERGENCE ($JS\text{-}div$)	−.363	.264	.015	.025	—	—	.253	.256
COVARIATE SHIFT ($K\text{-}S$) (Poenaru-Olaru et al., 2025)	.057	.050	.070	.078	—	—	.280	.264
ZERO-SHOT BASELINE MODEL	.004	.007	.068	.098	.129	—	.472	.245
Error Modeling and Estimation								
ROBERTA	.392	.291	—	—	.055	.170	.163	.164
LINEAR MODEL	.024	.008	.015	.059	.134	—	.539	.530
ZERO-SHOT ERROR MODEL	.454	.338	.174	.252	.202	.727	.100	.003
GPT4O-MINI	.848	.711	.822	.662	.029	.785	.004	.004
LLAMA-3.1-8B-INSTRUCT	.720	.571	.719	.568	.098	.734	.011	.010
GEMMA-3-12B-IT	.844	.709	.878	.711	.035	.724	.004	.002
QWEN3-4B-INSTRUCT-2507	.770	.623	.727	.507	.082	.767	.001	.001
LLM AVERAGE	.795	.653	.786	.612	.061	.752	.005	.004

Table 1: Aggregated combinatorial metrics across 4 base models over the *Amazon Reviews* (Hou et al., 2026) dataset, where $n = 33$ real product category datasets with $1,000$ examples per label per domain.

As our first baseline, we employ the LINGUISTIC DATASET DRIFT method (Chang et al., 2023), which leverages fine-grained, learned linguistic features to predict performance. Inspired by this work, we test a more general cosine distance-based SEMANTIC DRIFT *(cos)* as a second baseline to measure the semantic proximity between predicted and reference label texts using SBERT embeddings (Reimers and Gurevych, 2019). For each example i we compute the maximum cosine similarity over candidate labels, $s_i = \max_{y \in \mathcal{Y}} \cos\big(\mathbf{e}_{\hat{y}_i}, \mathbf{e}_y\big)$, and estimate domain-level performance by averaging these scores, $\hat{A}_d = \frac{1}{N_d} \sum_{i=1}^{N_d} s_i$. Also inspired by the previous work, we use the TOKEN FREQUENCY DIVERGENCES as baseline to estimate out-of-domain performance. We also measure the COVARIATE SHIFT using the KOLOGOMOROV-SMIRNOV *(K-S)* statistic (Poenaru-Olaru et al., 2025) over drifting representations, given its wide application in unsupervised drift detection methods (Lu et al., 2018; Poenaru-Olaru et al., 2022; Bayram et al., 2022; Lewis et al., 2022), which are expected to correlate with model performance degradation. Lastly, to verify whether our error model is important, we evaluate a ZERO-SHOT BASELINE using LLAMA-3.1-8B-INSTRUCT to produce label predictions directly; we treat the baseline model's predicted labels as a proxy for correctness and average these performances within each domain to obtain an estimated accuracy.

Baseline Models. Our base classifiers for predicting offensive language and sentiment across both datasets include three models of increasing complexity. (1) A LINEAR model, which uses lexical and confidence-based features (details in the Appendix). (2) A ROBERTA-BASE model, fine-tuned for each task using standard cross-entropy loss (training details in the Appendix). (3) A LLAMA-3.1-8B-INSTRUCT and (4) a GPT4O-MINI model evaluated in a few-shot setting, where task-specific examples are provided through in-context prompting (the full prompt is shown in the Appendix).

Error Estimation Models. We employ four different LLMs and two trained models as error models to be evaluated over our four base classifiers. We evaluate our approach over GPT4O-MINI, LLAMA-3.1-8B-INSTRUCT, GEMMA-3-12B-IT, and QWEN3-4B-INSTRUCT-2507 in a few-shot setting using the baseline errors in the downstream task. The ROBERTA baseline trains a ROBERTA-base error model (Liu, 2019) to predict whether a base classifier's output is correct given the input and predicted label; we average predicted correctness probabilities over each domain at inference. The LINEAR MODEL baseline is a logistic regression error model n-gram features; as with ROBERTA, we average predicted correctness to estimate accuracy. Training and optimization details for the learned error models are in the Appendix.

5 Results

We report results based on two major sets of experiments: Synthetic prediction experiments and real prediction experiments. The synthetic prediction experiments use randomly generated predictions for each group set, with predictions tuned to achieve a target overall accuracy. Intuitively, this allows us to systematically test settings where

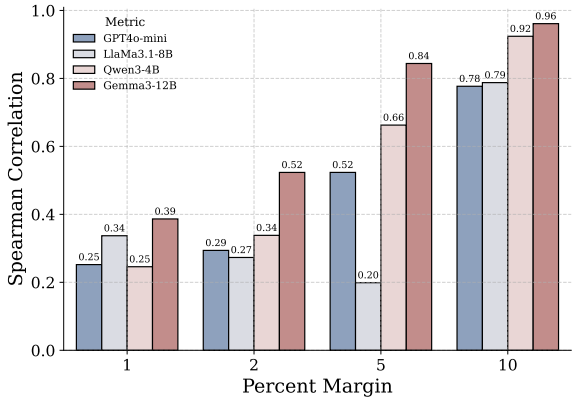


Figure 3: Structured variation study showing the robustness of our estimator across a decreasing margin between accuracy variables over $n = 33$ test categories with the *Software* base category.

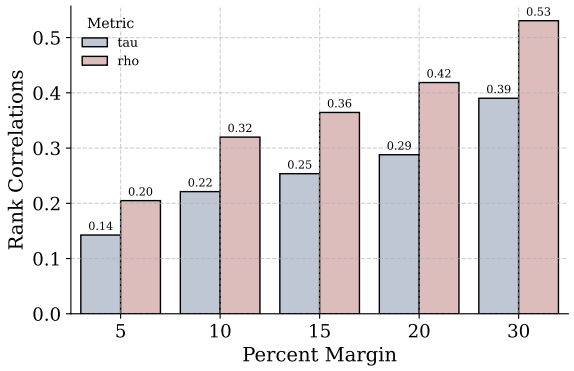


Figure 4: Rank correlations for the synthetically augmented *GeoOLID* (Lwowski et al., 2022) dataset with per-city linguistic variations and structured performance variations over the *Seattle, WA* base city, where $n = (15$ real cities $+ 85$ synthetic cities$)$.

there are very large performance disparities across groups (e.g., accuracy for San Antonio is 90%, while New York is 20%), as well as settings where there are very small differences between each group (e.g., accuracy for San Antonio is 90%, while New York is 89%). The real predictions are not controllable in this way. Here we evaluate real classifiers applied to out-of-domain target datasets.

Synthetic Predictions Results. We first test whether estimators can recover the correct domain ranking (**RQ1**) and how this depends on the size of the accuracy differences between domains (**RQ2**). Instead of using a classifier's predictions, we generate predictions with a chosen accuracy for each domain. We define the *accuracy gap* δ as the difference between the highest and lowest target accuracies. We assign domains evenly spaced accuracies between these endpoints. For example, with three domains, a highest accuracy of 90% and a gap of 10 percentage points ($\delta = 0.10$) give target accuracies of 90%, 85%, and 80%. Reducing the gap to 1 percentage point ($\delta = 0.01$) gives 90%, 89.5%, and 89%. This lets us test both domains with large performance differences and domains whose performance is nearly identical.

To generate predictions for each domain, we calculate the number of incorrect predictions needed for its target accuracy, rounding down to a whole number. We select that many examples uniformly at random and assign them incorrect predictions. Predictions for all remaining examples match their gold labels. For example, a target accuracy of 90% on $1{,}000$ examples requires 100 incorrect predictions and 900 correct predictions. The estimators receive no target-domain gold labels. We rank domains by their estimated performance and compare this ranking with the assigned ordering using Spearman's ρ and Kendall's τ. We repeat this evaluation with different values of δ to measure how ranking reliability changes as performance differences widen.

Small performance differences make rankings less reliable. Figure 3 reports results on *Amazon Reviews* using the *Software* base category, with margins from 1 to 10 percentage points. Rankings generally improve as the margin widens. GEMMA-3-12B-IT performs best throughout, improving from $\rho = .39$ at the smallest margin to .96 at the largest. QWEN3-4B-INSTRUCT improves from .25 to .92, and GPT4O-MINI from .25 to .78. The trend is less consistent for LLAMA-3.1-8B-INSTRUCT, which falls from .34 at 1 percentage point to .20 at 5 before reaching .79 at 10. At the smallest margin, all four judges have weak correlations ($\rho = .25$–$.39$). Thus, even the strongest judges struggle to recover the ordering when true accuracies are closely spaced.

Larger performance differences help, but reliability also varies across datasets. Figure 4 reports results on the synthetically augmented *GeoOLID*, using *Seattle, WA* as the base city across 15 real and 85 synthetic cities. Both rank correlations improve as the margin widens: Spearman's ρ rises from .20 at 5 percentage points to .53 at 30, while Kendall's τ rises from .14 to .39. However, correlations remain lower overall than on *Amazon Reviews*, despite the larger imposed performance gaps. This is consistent with the weaker real-prediction results for *GeoOLID* reported below. Together, these ex-

periments show that wider performance gaps generally improve ranking reliability, but do not ensure equally reliable rankings across datasets.

Real Predictions. We now evaluate rankings using actual predictions from base classifiers on out-of-domain data. Unlike the synthetic experiments, we do not control the accuracy differences between domains. These differences depend on the classifier and the data. We compare estimated and true rankings to assess ranking reliability (**RQ1**), variation across models and domains (**RQ2**), and the potential for dialect and geographic auditing (**RQ3**).

RQ1: How reliably do estimators rank domains? LLM error models produce the strongest rankings on *Amazon Reviews* (Table 1), with results aggregated across four base classifiers. GPT4O-MINI achieves $\rho = .848$ and $\tau = .711$, followed closely by GEMMA-3-12B-IT at $\rho = .844$ and $\tau = .709$. The average Spearman correlation across the four LLM judges is .795. Drift baselines produce much weaker rankings: SEMANTIC DRIFT reaches $\rho = .308$, COVARIATE SHIFT is near zero ($\rho = .057$), and TOKEN FREQUENCY DIVERGENCE is negative ($\rho = -.363$). The reported p-values for both rank correlations are below .05 for all four few-shot LLM judges and above .05 for the drift baselines.

Acc@3 and Rec@3 show a similar advantage for recovering the three highest-performing categories. GEMMA-3-12B-IT achieves the best Acc@3 (.878) and Rec@3 (.711), despite GPT4O-MINI having a slightly higher overall rank correlation. Thus, the best method for recovering the full ordering is not necessarily the best for recovering the top three domains.

Neither absolute estimation error nor instance-level F1 fully captures ranking reliability. ROBERTA has a lower RMSE than LLAMA-3.1-8B-INSTRUCT (.055 versus .098), but a much lower rank correlation ($\rho = .392$ versus .720). Similarly, GEMMA-3-12B-IT has a lower error-prediction F1 than LLAMA-3.1-8B-INSTRUCT (.724 versus .734), yet ranks domains more accurately. The ZERO-SHOT ERROR MODEL also achieves a comparable F1 (.727) but a lower ρ (.454). These comparisons show that domain ranking needs to be evaluated directly rather than inferred from RMSE or F1.

RQ2: How does reliability vary across models and domains? The LLM judges recover strong rankings on *Amazon Reviews* even though true accuracies cluster near .70 (Figure 7). Thus, closely spaced accuracies do not always prevent reliable ranking with real predictions.

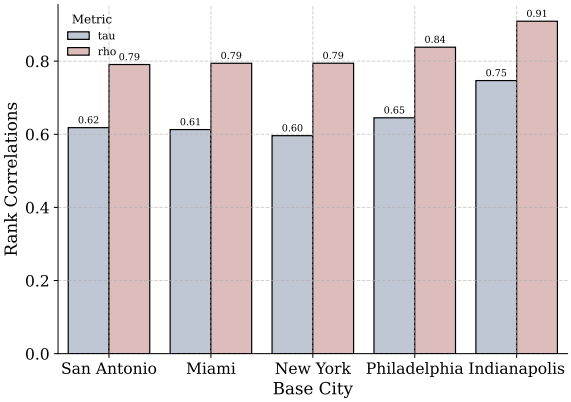


Figure 5: Ranking correlations across the 52 dialect-mapped targets in *GeoOLID-Dial*, generated with `Multi-VALUE` (Ziems et al., 2023). We show the five source cities with the highest Spearman correlations.

However, ranking quality is lower on *GeoOLID-Dial* (Table 16). QWEN and GEMMA achieve the highest Spearman correlations (.594 and .592), with Kendall correlations of .456 and .465, respectively. LLAMA follows at $\rho = .508$ and $\tau = .387$. All three outperform the LINEAR baseline ($\rho = .164$, $\tau = .107$), but their rank correlations are lower than on *Amazon Reviews*. The reported p-values are below .05 for both correlations for GEMMA and QWEN, but not for LLAMA or LINEAR.

The dialect results also show why low RMSE is not enough: LINEAR has the lowest RMSE (.042) but the weakest rankings. GEMMA achieves the highest Acc@3 (.467) and ties with LLAMA for the highest Rec@3 (.378), despite QWEN having a slightly higher ρ.

Together, these results show that ranking reliability varies with both the estimator and the evaluation setting. The comparison between 33 product categories and 52 dialect targets also changes the task and type of domain shift, so it does not isolate the effect of the number of domains.

RQ3: Can LLM judges support dialect and geographic auditing? The results across 52 dialect targets show that LLM judges can recover performance rankings across many language varieties, although reliability varies by source city. Figure 5 shows the five source-city settings with the strongest rankings. Spearman's ρ is approximately .79 for San Antonio, Miami, and New York, ris-

	$\rho_{mean} \uparrow$	ρ_p	$\tau_{mean} \uparrow$	τ_p	$Acc@3 \uparrow$	$Rec@3 \uparrow$	$RMSE \downarrow$
LINEAR	.164	.264	.107	.286	.200	.133	.042
GEMMA	.592	.038	.465	.045	.467	.378	.132
LLAMA	.508	.072	.387	.072	.400	.378	.195
QWEN	.594	.041	.456	.044	.400	.356	.262

Table 2: *GeoOLID-Dial* error model results compared to the Linear baseline.

ing to .84 for Philadelphia and .91 for Indianapolis. Kendall's τ ranges from approximately .60 to .75.

Error Analysis. In *GeoOLID-Dial*, for instance, when using *Chicago* few shots and testing over the African American dialect, we find 127 false negatives, 548 true positives, 183, false positives, and 117 true negatives. For example, the post *"That market gon crash tomorrow morning. Numbers gon continue to rise. Persons gon continue to die"* which was predicted as *not offensive* via the base training city *Phoenix* which is a correct prediction, was misclassified by the error model as *error*. This highlights a major disparity over the African American Vernacular dialect over benign statements. In the *Amazon Reviews*, we find $1,758$ true positives (i.e., positive label is *correct*), 485 false negatives, 219 true negatives, and 538 false positives. For example, *"This says it comes with one of the plastic wipes bins. It did not. I LOVE these wipes. My son has very sensitive skin and we can use them on his face (after meals) without causing redness,"* which is labeled as a negative review, was misclassified as a positive review. Hence, when the error model judge is given the example along with the related prediction as positive, its judgment of its own prediction is *'correct.'* This example is a cascading failure mode where there is an overconfidence in the sentiment and the validity judgment, likely due to *"I LOVE these wipes"*. The impact can be at large, where the estimating model's trustworthiness is questionable under the test-time domain. Other instances include the LLM refusing to give a response, which may impact performance if not accounted for in downstream tasks.

Implications. Our results suggest several implications for large-scale NLP evaluation and auditing. First, LLM-as-a-judge frameworks appear promising for **large-scale disparity analysis** when labeled data is unavailable. Across geographic and dialect-based settings, LLM judges consistently outperform drift and similarity baselines in recovering ordinal performance relationships. Although rankings become unreliable when true performance differences are very small, practitioners often need to identify domains, dialects, or populations with substantially degraded performance. We therefore view LLM-based ranking as a useful triage mechanism for identifying groups that require deeper inspection or additional data collection, while still encouraging real evaluation data for underserved communities.

Second, direct performance prediction and ranking quality should be treated as separate evaluation problems. Several methods achieved low RMSE while producing weak rankings, showing that accurate average estimates do not necessarily preserve which domains perform best or worst. This distinction matters for fairness-sensitive deployment, where relative disparities may matter more than exact accuracy estimates. Moreover, ranking matters when comparing 100s or even thousands of groups.

Third, the results highlight the importance of domain heterogeneity. Ranking was easier when true performance differences were larger, while tightly clustered domains were difficult for all methods. Some benchmarks may therefore underestimate the usefulness of label-free auditing because their domains are too similar to produce a stable ordinal structure. More broadly, large-scale evaluation across dialects, geographic regions, and socially meaningful subpopulations may benefit from combining label-free ranking with targeted human annotation.

6 Conclusion

We studied label-free performance ranking for large-scale NLP auditing. Across domain, geographic, and dialect-based settings, we find that ranking domains by expected performance is difficult for traditional drift and similarity methods. LLM-based error estimators are more reliable, producing better rank correlations and top-k auditing results. However, low RMSE does not always imply good ranking, showing that absolute error and ordinal reliability should be evaluated separately. Overall, our results suggest that label-free ranking is a promising tool for identifying underserved domains and language varieties, but it should be used carefully in fairness-sensitive deployments.

Acknowledgments

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 2145357.

7 Limitations

This study has several limitations. First, our experiments focus on classification tasks. While classification is a common setting for domain-level auditing, the results may not directly extend to generation, question answering, summarization, or multimodal tasks, where correctness is harder to define and may require more complex judging criteria.

Second, the reliability of label-free ranking depends on the quality and stability of the error model. LLM judges may be sensitive to prompt wording, demonstration selection, model family, and repeated sampling. We evaluate several LLM judges and find broadly consistent trends, but future work should more directly study prompt robustness, seed sensitivity, and the effects of varying the number or types of in-context examples.

Third, GeoOLID-Dial uses controlled dialect transformations rather than naturally occurring dialect data. This design allows us to isolate dialectal surface variation while holding meaning mostly fixed, which is useful for controlled evaluation. However, synthetic transformations may not capture the full social, regional, and contextual complexity of real dialect use. Future work should test whether similar ranking patterns hold on naturally occurring dialect corpora and with human validation of dialect quality.

Fourth, some domains have very similar true performance. In these cases, rankings are inherently unstable because small changes in estimated performance can change the ordering. This is not only a limitation of our methods, but a property of the evaluation setting. For auditing, these cases may be less practically important than settings where some domains clearly perform worse than others, but they still show why rank-based evaluation should be interpreted with care.

Finally, our results should be viewed as evidence that label-free ranking can support audit triage, not as a replacement for targeted human annotation or real evaluation data. The goal is to help identify domains, regions, or language varieties that may need closer inspection. Once those areas are identified, human-centered evaluation and data collection remain necessary, especially in fairness-sensitive deployment settings.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Victor H Barella, Luis PF Garcia, Marcilio CP de Souto, Ana C Lorena, and André CPLF de Carvalho. 2021. Assessing the data complexity of imbalanced datasets. *Information Sciences*, 553:83–109.

Firas Bayram, Bestoun S Ahmed, and Andreas Kassler. 2022. From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245:108632.

Tyler A. Chang, Kishaloy Halder, Neha Anna John, Yogarshi Vyas, Yassine Benajiba, Miguel Ballesteros, and Dan Roth. 2023. Characterizing and measuring linguistic dataset drift. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8953–8967, Toronto, Canada. Association for Computational Linguistics.

Ryan A. Cook, John P. Lalor, and Ahmed Abbasi. 2025. No simple answer to data complexity: An examination of instance-level complexity metrics for classification tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2553–2573, Albuquerque, New Mexico. Association for Computational Linguistics.

Mark Dredze, Tim Oates, and Christine Piatko. 2010. We're not in Kansas anymore: Detecting domain changes in streams. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 585–595, Cambridge, MA. Association for Computational Linguistics.

Jonas Eberlein, Daniel Rodriguez, and Rachel Harrison. 2025. The effect of data complexity on classifier performance. *Empirical Software Engineering*, 30(1):16.

Vikram Elango, Tony Chen, and Raghu Ramesha. 2022. Detect NLP data drift using custom Amazon SageMaker Model Monitor. AWS Machine Learning Blog. Accessed: 2022-09-01.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, Hong Kong, China. Association for Computational Linguistics.

Robert Feldhans, Adrian Wilke, Stefan Heindorf, Mohammad Hossein Shaker, Barbara Hammer, Axel-Cyrille Ngonga Ngomo, and Eyke Hüllermeier. 2021. Drift detection in text data with document embeddings. In *Intelligent Data Engineering and Automated Learning*, pages 107–118. Springer International Publishing.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore. Association for Computational Linguistics.

Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3251–3265, San Diego, California, United States. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Eric Khiu, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Armando Parra Flores, Leandro Acros Roman, A Seza Doğruöz, and En-Shiun Annie Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. *arXiv preprint arXiv:2402.02633*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *ArXiv preprint*, abs/2310.08491.

Hyunjin Kwon, Matthew Greenberg, Colin Bruce Josephson, and Joon Lee. 2024. Measuring the prediction difficulty of individual cases in a dataset using machine learning. *Scientific Reports*, 14(1):10474.

Grace A Lewis, Sebastián Echeverría, Lena Pons, and Jeffrey Chrabaszcz. 2022. Augur: A step towards realistic drift detection in production ml systems. In *Proceedings of the 1st Workshop on Software Engineering for Responsible AI*, pages 37–44.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ana C Lorena, Pedro YA Paiva, and Ricardo BC Prudêncio. 2024. Trusting my predictions: on the value of instance-level analysis. *ACM Computing Surveys*, 56(7):1–28.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363.

Brandon Lwowski, Paul Rad, and Anthony Rios. 2022. Measuring geographic performance disparities of offensive language classifiers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6600–6616, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ankur Mallick, Kevin Hsieh, Behnaz Arzani, and Gauri Joshi. 2022. Matchmaker: Data drift mitigation in machine learning for large-scale systems. *Proceedings of Machine Learning and Systems*, 4:77–94.

David Nigenda, Zohar Karnin, Bilal Zafar, Raghu Ramesha, Alan Tan, Michele Donini, and Krishnaram Kenthapadi. 2022. Amazon SageMaker Model Monitor: A system for real-time insights into deployed machine learning models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Shantanu Patankar, Omkar Gokhale, Onkar Litake, Aditya Mandke, and Dipali Kadam. 2022. To train or not to train: Predicting the performance of massively multilingual models. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 8–12.

Lorena Poenaru-Olaru, Luis Cruz, Jan S Rellermeyer, and Arie van Deursen. 2025. Improving the reliability of failure prediction models through concept drift monitoring. In *2025 IEEE/ACM International Workshop on Deep Learning for Testing and Testing for Deep Learning (DeepTest)*, pages 1–8. IEEE.

Lorena Poenaru-Olaru, Luis Cruz, Arie van Deursen, and Jan S Rellermeyer. 2022. Are concept drift detectors reliable alarming systems?-a comparative study. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3364–3373. IEEE.

Shushanta Pudasaini, Luis Miralles, David Lillis, and Marisa Llorens Salvador. 2025. Benchmarking ai text detection: Assessing detectors against new datasets, evasion tactics, and enhanced llms. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 68–77.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, and Roger Zimmermann. 2021. Domain divergences: A survey and empirical analysis. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1830–1849, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2017. Data selection strategies for multi-domain sentiment analysis. *arXiv*, abs/1702.02426.

Charles Spearman. 1961. The proof and measurement of association between two things.

Jun Sun, Xinxin Zhang, Simin Hong, Jian Zhu, and Lingfang Zeng. 2025. Adversarial alignment with anchor dragging drift (a^3d^2): Multimodal domain adaptation with partially shifted modalities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19680–19690, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Luong Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. ReFT: Reasoning with reinforced fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7601–7614, Bangkok, Thailand. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *ArXiv preprint*, abs/2306.05087.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

Hanqi Xiao, Vaidehi Patil, Hyunji Lee, Elias Stengel-Eskin, and Mohit Bansal. 2025. Generalized correctness models: Learning calibrated and model-agnostic correctness predictors from historical patterns. *arXiv preprint arXiv:2509.24988*.

Ivan P. Yamshchikov, Viacheslav Shibaev, Nikolay Khlebnikov, and Alexey Tikhonov. 2021. Style-transfer and paraphrase: Looking for a sensible semantic similarity metric. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14213–14220.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Helen Zhou, Yuwen Chen, and Zachary Lipton. 2023. Evaluating model performance in medical datasets over time. In *Conference on Health, Inference, and Learning*, pages 498–508. PMLR.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *ArXiv preprint*, abs/2310.17631.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, et al. 2024. Agent-as-a-judge: Evaluate agents with agents. *ArXiv preprint*, abs/2410.10934.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-value: A framework for cross-dialectal english nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768.

A Appendix

A.1 RoBERTa Model Hyper-parameters

We set the tokenizer for RoBERTa to ROBERTA-BASE, learning rate of the base model to $1e-5$, learning rate of the error model to $5e-5$, max length to 128, base model epochs to 20, error model epochs to 20, training batch size to 16, evaluation batch size to 32, warmup steps to 500, weight decay to .01, train-test split seed to 42, and internal evaluation size to .1.

A.2 Linear Model

Based on a comparison of prediction accuracies between the COUNT and TF-IDF vectorizers, the selected linear classifier model used TF-IDF vectorization with stop-word elimination and uni-gram

tokenization. The linear error model employed TF-IDF vectorization with stop-word elimination and bi-gram tokenization. Bi-grams were selected based on having the highest Spearman rank coefficients for error prediction of other models, e.g., RoBERTa and LLMs, when compared with uni-gram and tri-gram tokenization.

A.3 LLM Prompts

(a.1) Base Model System Prompt for GeoOLID

SYSTEM
You are a helpful assistant. Your task is to provide sound judgement on the nature of the text that will be provided to you. If you think the text is offensive, please say 'offensive'. If you think the text is not offensive, please say 'not offensive'. Take into consideration human tone. Here are some examples:

User/Assistant Few-shot Examples
Text: {text}
Label: {label}

User
Text: {text}

(a.2) Base Model System Prompt for Amazon

SYSTEM
You are a helpful assistant. Your task is to provide sound judgement on the nature of the text that will be provided to you. Your task is sentiment analysis. If you think the text is positive, please say 'positive'. If you think the text is neutral, please say 'neutral'. If you think it is negative, please say 'negative'. Only say 'positive', 'negative', or 'neutral'. Here are some examples:

User/Assistant Few-shot Examples

Text: {text}
Label: {label}

User
Text: {text}

(b) Error Model System Prompt

SYSTEM
You are a helpful assistant. Your task is to check whether our prediction is an error or not based on sound judgment about the nature of the text. You will check if the predicted label is correct for the given text. If you think the predicted label is correct, please say 'correct'. If you think the predicted label is wrong, please say 'error'. Only say 'error' or 'correct'. Please note that the order of the examples does not matter - the content matters. Here are some examples:

User/Assistant Few-shot Examples
Text: {text} -> Predicted Label: {label}
Error Label: {label}

User
Text: {text} -> Predicted Label: {label}

(c) *GeoOLID-Synth* Prompt

SYSTEM
You are a helpful assistent. Your task is to synthetically generate an adapted version of the text based on the specified original category, to match the new category, based on how they speak there. Only give the related text, do not give any other explanation.

User
Text: {text}
Original Category: {original category}
New Category: {new category}

B Baselines

B.1 Linguistic Dataset Drift

We reproduce the method in (Chang et al., 2023) via their publicized Github repository. This work consists multiple experiments to eventually calculate drift metrics that are used to predict model performance on novel data. We reproduce their method in its entirety. First, after we configure our datasets into their code, we run the JS-distance, cross-entropy, and raw example sequence length experiments, which use the roberta-base Hugging Face model tokenizer. Then, we get the spaCy drift metrics (e.g., POS n-gram divergences, content word frequency divergences) at the dataset level and the example level (i.e., vocabulary and structural drift). Next, we fine-tune and evaluate a RoBERTa-base model over the classification task on all domains. We compute the pre-trained cosine embedding metrics as well as the fine-tuned ones. With these aggregated metrics, we run their logistic regression to predict whether the model will get an example correct. Using their method as a performance estimator, we evaluate it over our baseline models.

B.2 Semantic Drift

Inspired by their work, we implement our version of semantic drift between our training and test sets. Specifically, we get the SBERT embeddings associated with the train and test sets, get the pairwise cosine similarity at the example level, then get the maximum values assigned to the train-test pair, and use that value as an estimator.

B.3 Token Frequency Divergences

We tokenize each train set and test set iteratively to retrieve the token frequency divergence values.

	ρ_{mean}	τ_{mean}	$Acc@3$	$Rec@3$	$RMSE$	$F1$	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
Estimation Baselines								
LINGUISTIC DATASET DRIFT (Chang et al., 2023)	−.022	−.018	.280	.225	.140	—	.480	.495
SEMANTIC DRIFT (*cos*)	.223	.243	.307	.288	—	—	.237	.237
TOKEN FREQUENCY DIVERGENCE ($JS\text{-}div$)	.066	.051	.649	.388	—	—	.584	.554
COVARIATE SHIFT ($K\text{-}S$) (Poenaru-Olaru et al., 2025)	.125	.041	.330	.254	—	—	.488	.545
ZERO-SHOT BASELINE MODEL	.253	.181	.000	.000	.099	—	.002	.007
ZERO-SHOT ERROR MODEL	.608	.469	.899	.516	.332	.454	.122	.070
Error Modeling and Estimation via Supervied Learning								
ROBERTA	.375	.283	.683	.460	.071	—	.281	.275
LINEAR MODEL	.118	.091	.249	.166	.043	—	.372	.397
LLM-as-a-judge Framework								
GPT4O-MINI	.526	.408	.766	.521	.137	.812	.160	.166
LLAMA-3.1-8B-INSTRUCT	.636	.484	.799	.577	.274	.728	.059	.067
GEMMA-3-12B-IT	.635	.491	.833	.488	.213	.487	.081	.090
QWEN3-4B-INSTRUCT-2507	.443	.330	.699	.504	.340	.448	.154	.154
LLM AVERAGE	.560	.428	.774	.522	.241	.618	.113	.119

Table 3: Aggregated combinatorial metrics across 4 base models over the *GeoOLID* (Lwowski et al., 2022) dataset, where $n = 15$ real city datasets with $\sim 1,000$ examples per domain in an imbalanced base class setting.

ERROR MODEL	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4O-MINI	.783	.625	.010	.008
LLAMA-3.1-8B-INSTRUCT	.811	.640	.014	.014
GEMMA-3-12B-IT	.789	.638	.031	.031
QWEN3-4B-INSTRUCT-2507	.711	.542	.000	.000

Table 4: Dataset-aggregated metrics over the GPT4O-MINI baseline model over the *Amazon Reviews* (Hou et al., 2026) dataset, where $n = 33$ real product category datasets with $1,000$ examples per label, per domain in a balanced base class setting.

ERROR MODEL	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4O-MINI	.526	.405	.014	.009
LLAMA-3.1-8B-INSTRUCT	.526	.405	.014	.009
GEMMA-3-12B-IT	.775	.609	.012	.008
QWEN3-4B-INSTRUCT-2507	.647	.492	.000	.000

Table 5: Dataset-aggregated metrics over the LLAMA-3.1-8B-INSTRUCT baseline model over the *Amazon Reviews* (Hou et al., 2026) dataset, where $n = 33$ real product category datasets with $1,000$ examples per label, per domain in a balanced base class setting.

Specifically, for each set of examples, we pool the tokens at the dataset level to get their associated frequency distributions and compute the JS-divergence between the two datasets, and use that value as our estimator.

B.4 Covariate Shift

We retrieve the LLAMA 8B representations for each train/test set and compute Kologomorov-Smirnov test over the distribution shift and use that as our estimator.

ERROR MODEL	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4O-MINI	.757	.590	.005	.008
LLAMA-3.1-8B-INSTRUCT	.766	—	.002	—
GEMMA-3-12B-IT	.806	.649	.001	.002
QWEN3-4B-INSTRUCT-2507	.506	.365	.092	.106

Table 6: Dataset-aggregated metrics over the GPT4O-MINI baseline model over the *GeoOLID* (Hou et al., 2026) dataset, where $n = 15$ real product city datasets with $1,000$ examples per domain in an imbalanced base class setting.

ERROR MODEL	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4O-MINI	.766	.586	.002	.004
LLAMA-3.1-8B-INSTRUCT	.770	.594	.007	.012
GEMMA-3-12B-IT	.775	.588	.002	.005
QWEN3-4B-INSTRUCT-2507	.463	.320	.141	.179

Table 7: Dataset-aggregated metrics over the LLAMA-3.1-8B-INSTRUCT baseline model over the *GeoOLID* (Hou et al., 2026) dataset, where $n = 15$ real product city datasets with $1,000$ examples per domain in an imbalanced base class setting.

C Compute

We use 4 NVIDIA $H100$ GPUs and 4 $A6000$ GPUs spread across 3 servers for our experiments.

D Additional Results

Table 17 reports the average accuracy of the base models across the *Amazon Reviews* and *GeoOLID* datasets. Performance varies by model and dataset, with ROBERTA achieving the highest overall accuracies (.730 on *Amazon Reviews* and .878 on *GeoOLID*). The LINEAR model performs lowest on both datasets (.574 and .790), while the large

Error Model	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4o-mini	.481	—	.120	—
LLaMa-3.1-8B-Instruct	.593	.459	.042	.041
Gemma-3-12B-it	.629	.485	.030	.032
Qwen3-4B-Instruct-2507	-.081	-.066	.738	.752

Table 8: Dataset-aggregated metrics over the Linear baseline model over the *GeoOLID* (Hou et al., 2026) dataset, where $n = 15$ real product city datasets with $1,000$ examples per domain in an imbalanced base class setting.

Error Model	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4o-mini	.206	.158	.492	.518
LLaMa-3.1-8B-Instruct	.415	.299	.186	.211
Gemma-3-12B-it	.333	.245	.293	.322
Qwen3-4B-Instruct-2507	.079	.053	.579	.561

Table 9: Dataset-aggregated metrics over the RoBERTa baseline model over the *GeoOLID* (Hou et al., 2026) dataset, where $n = 15$ real product city datasets with $1,000$ examples per domain in an imbalanced base class setting.

Error Model	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4o-mini	.902	.803	.002	.002
LLaMa-3.1-8B-Instruct	.788	.640	.013	.012
Gemma-3-12B-it	.895	.781	.002	.000
Qwen3-4B-Instruct-2507	.936	.817	.000	.000

Table 10: Dataset-aggregated metrics over the Linear baseline model over the *Amazon Reviews* (Hou et al., 2026) dataset, where $n = 33$ real product category datasets with $1,000$ examples per label, per domain in a balanced base class setting.

Error Model	ρ_{mean}	τ_{mean}	$p\text{-}value_{\rho}$	$p\text{-}value_{\tau}$
GPT4o-mini	.832	.687	.004	.005
LLaMa-3.1-8B-Instruct	.753	.599	.004	.004
Gemma-3-12B-it	.913	.791	.000	.000
Qwen3-4B-Instruct-2507	.768	.622	.004	.003

Table 11: Dataset-aggregated metrics over the RoBERTa baseline model over the *Amazon Reviews* (Hou et al., 2026) dataset, where $n = 33$ real product category datasets with $1,000$ examples per label, per domain in a balanced base class setting.

language models (GPT4o-mini and LLaMa-3.1-8B-Instruct) fall in between, averaging around .75 on *Amazon Reviews* and .79 on *GeoOLID*. These results show that base model performance differs across datasets and model types, with *GeoOLID* generally yielding higher accuracies.

Table 15 presents the average accuracy and $F1$ scores for the error prediction models. The Linear error model achieves the highest accuracies overall (.734 on *Amazon Reviews* and .825 on *GeoOLID*), followed by GPT4o-mini (.674 and .737) and LLaMa-3.1-8B-Instruct (.716 and .644). Gemma and Qwen perform worse across both datasets, with notably low $F1$ scores on *GeoOLID* (.487 and .203). These results indicate that while larger models can learn complex patterns, simpler error models can produce stable and competitive accuracy estimates across datasets.

Figures 3 illustrate the relationship between performance ranking and the margin of true accuracy differences for the Detroit training category in *GeoOLID*. The first figure shows how Spearman rank correlation decreases as the margin between true accuracies narrows, demonstrating that ranking reliability weakens when performance gaps are smaller. The second figure applies the same analysis under an error-informed setting, where performance variation is guided by the model's predicted error distribution. In this case, correlations remain higher across decreasing margins, suggesting that model-informed variation better preserves ranking consistency. Together, the figures highlight how performance differences and error structure influence correlation stability across datasets.

In Table 13, we show the significance of our findings using Fisher's Z transformation. We find significance across models observed via p-values $< .001$ and high correlations with the LLMs.

In Table 14, we show the results of our additional experiments where we aggregate all drift metric baselines using a logistic regression, and observe modest correlations of $\rho = .449$ and $\tau = .339$.

Real Data Results across Domains. Tables 3 and 1 present the Spearman rank correlation between estimated and true model performance across different domains. Each value represents how well the ranking produced by the error-prediction method aligns with the actual ranking of accuracies across cities in *GeoOLID* or product categories in *Amazon Reviews*. The average scores at the rightmost column are computed across all training domains. For comparison, the baseline results are the average of three approaches: semantic drift and a zero-shot baseline. These baselines capture how similarity-based or unsupervised methods perform when no explicit error modeling is used.

Table 3 reports results on *GeoOLID*. The baseline measures yield weak or inconsistent results, with average correlations below .30 across all three baselines. In contrast, the error-prediction approaches show much stronger and more stable correlations. The large language model predictors

Original City	Original Text	Dialect Transformation	Transformed Text
New York	Couldn't agree more! Just wash your damn hands!!	Indian South African Dialect	Couldn't agree more! Just wash y'all's damn hands!!
New York	Fuck not Forrest!	New Zealand Dialect	Are Fucking not Forrest!
New York	Bruhhh more niggas have died from the actual flu then this corona shit im so done with yall	Southeast American Enclave Dialect	Bruhhh more niggas died from that actual flu then this here corona slam died im so did with yall
New York	Don't forget Fake Christian	Southeast American Enclave Dialect	Aren't forgetting Fake Christian
New York	This virus is gonna be so much bigger and worse than we think	Rural African American Vernacular	This here virus gonfor be so much more bigger and more worse for we thinks
New York	Tell that to the man who died in Washington state from coronavirus	Ghanaian Dialect	Are Telling that the man who had died Washington state coronavirus

Table 12: Example Dialect transformations.

Model	Dataset	Mean Fisher's Z	STD Fisher's Z	Z Statistic	P-value
GPT4o-mini	GeoOLID	0.58921	0.07658	28.78837	***
LlaMa-3.1-8B	GeoOLID	0.7351	0.06862	40.08135	***
Gemma-3-12B-it	GeoOLID	0.7542	0.072	39.19211	***
Qwen-3-8B	GeoOLID	0.32757	0.08811	13.91012	***
Linear	GeoOLID	0.12803	0.10935	4.3808	***
RoBERTa	GeoOLID	0.45105	0.1108	15.23199	***

Table 13: Significance tests using Fisher's Z transformation.
Note: We use * if p-value is $<= 0.05$, ** if $<= 0.01$, and *** if less than 0.001

Spearman Rho	Kendall Tau
0.449	0.339

Table 14: Ensembled drift metrics with logistic regression correlations.

Error Model	Dataset	Average Acc.	Average F1
Linear	Amazon	.734	–
Linear	GeoOLID	.825	–
GPT4o-mini	Amazon	.674	.750
GPT4o-mini	GeoOLID	.737	.812
LLaMa-3.1-8B	Amazon	.716	.780
LLaMa-3.1-8B	GeoOLID	.644	.728
Gemma	Amazon	.632	.453
Gemma	GeoOLID	.687	.487
Qwen	Amazon	.372	.457
Qwen	GeoOLID	.644	.203

Table 15: Error models' combinatorial average accuracy and average F1 scores per dataset.

perform best, with LLAMA-3.1-8B-INSTRUCT and GEMMA-3-12B-IT achieving average correlations of .625 and .635, respectively. These models capture consistent ranking differences across cities such as Houston (.707) and Memphis (.562), indicating that ranking reliability improves when true accuracy differences are larger and geographically structured. The smaller ROBERTA and LINEAR models perform less consistently, averaging .419 and .126, respectively, suggesting that they capture less of the structure in regional variation.

Table 1 shows results for the Amazon product review dataset. Here, ranking performance is notably weaker, with most correlations close to zero. The LINEAR model and LLAMA-3.1-8B-INSTRUCT again perform best, with average correlations of .346 and .392, while other models show low or negative values. This pattern aligns with the narrow accuracy distribution seen in Figure 7, where accuracies cluster near .70 with a standard deviation of only .015. When the range of true performance is small, small prediction errors can easily change rank order, leading to unstable results. In contrast, Figure 6 shows a wider spread in *GeoOLID* accuracies ($\mu = 0.81$, $\sigma = 0.028$), producing stronger signals for ranking. The relatively strong performance of the Linear model aligns with the Semantic Drift baseline, as both rely on shallow lexical or

Base City	ρ	τ	$Acc@3$	$Rec@3$	$p\text{-}value_\rho$	$p\text{-}value_\tau$
Indianapolis, IN	.909	.747	1.00	.667	.000	.000
Philadelphia, PA	.838	.645	1.00	.667	.000	.000
Miami, FL	.794	.613	1.00	.667	.000	.000
New York, NY	.794	.596	.000	.667	.000	.000
San Antonio, TX	.791	.618	1.00	1.00	.000	.000
Los Angeles, CA	.729	.539	.000	.000	.000	.000
Chicago, IL	.632	.475	1.00	.333	.000	.000
Baltimore, MD	.566	.405	.000	.333	.000	.000
Phoenix, AZ	.559	.410	1.00	.667	.000	.000
Detroit, MI	.486	.355	.000	.333	.000	.000
Columbus, OH	.365	.261	.000	.333	.008	.007
New Orleans, LA	.300	.235	.000	.000	.031	.016
Houston, TX	.116	.071	.000	.000	.414	.467
El Paso, TX	.073	.057	.000	.000	.605	.563
Memphis, TN	−.304	−.209	.000	.000	.028	.033

Table 16: Ranked evaluation over the *Multi-Value* (Ziems et al., 2023) dialect-mapped *GeoOLID* (Lwowski et al., 2022) dataset with $n = 52$ target dialects.

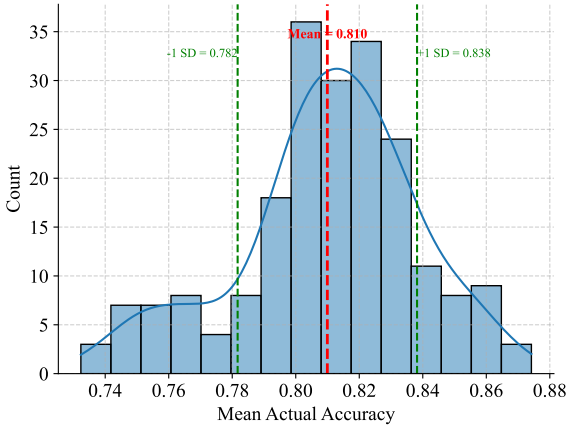


Figure 6: Histogram showing the base model accuracy ranges with their respective statistics on the *GeoOLID-Dial* dataset.

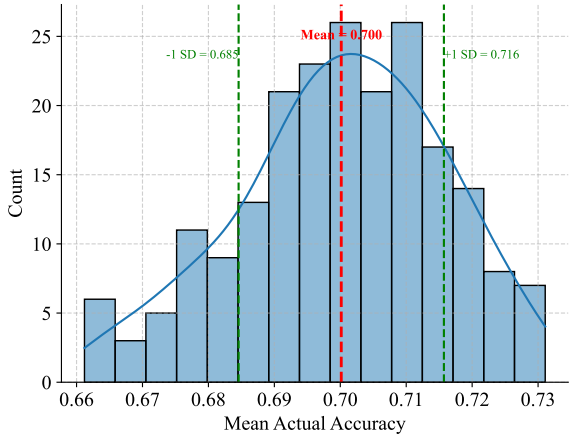


Figure 7: Histogram showing the base model accuracy ranges with their respective statistics on the *Amazon Reviews* (Hou et al., 2026) dataset.

confidence-based cues that generalize well across categories. These features capture broad sentiment polarity and stylistic consistency across product types, allowing both methods to recover smooth, monotonic estimates even when absolute performance differences are minimal.

Table 16 reports ranked evaluation results over the dialect-mapped *GeoOLID-Dial* benchmark with $n = 52$ target dialects. Ranking performance varies substantially depending on the source city used to train the base classifier. *Indianapolis, IN* achieves the strongest overall ranking performance, with $\rho = .909$ and $\tau = .747$, followed by *Philadelphia, PA* ($\rho = .838$, $\tau = .645$). Other cities such as *Miami, FL*, *New York, NY*, and *San Antonio, TX* also produce strong ordinal recovery, all with $\rho \approx .79$. In contrast, cities such as *Houston, TX* ($\rho = .116$), *El Paso, TX* ($\rho = .073$), and especially *Memphis, TN* ($\rho = -.304$) show weak or unstable rankings across dialects. The top-performing cities also achieve strong auditing metrics, with several reaching perfect $Acc@3 = 1.00$, while *San Antonio, TX* achieves both perfect $Acc@3 = 1.00$ and $Rec@3 = 1.00$. Most high-performing cities yield statistically significant correlations ($p < .001$), whereas weaker-performing cities such as Houston and El Paso fail to achieve significance. Over-

all, these results suggest that some source cities produce substantially more transferable ranking signals across dialectal variation than others, likely reflecting differences in linguistic coverage and robustness to dialectal shifts.

Figures 6 and 7 show the empirical distribution of true base-model accuracies across domains for *GeoOLID-Dial* and *Amazon Reviews*. The two datasets exhibit notably different performance spreads, which helps explain the ranking behavior observed in the main results. In *GeoOLID-Dial*, accuracies are distributed more broadly around a mean of $\mu = .810$ with standard deviation $\sigma = .028$, spanning approximately .74 to .88. This wider spread produces clearer ordinal structure between domains, making it easier for error-prediction methods and LLM judges to recover the correct ranking relationships across dialects and cities. This pattern aligns with the stronger rank correlations reported in the main experiments, where several methods achieve $\rho > .75$ and statistically significant ranking consistency across dialect-mapped targets.

In contrast, *Amazon Reviews* exhibits a much narrower distribution centered around $\mu = .700$ with standard deviation $\sigma = .015$, where most domains cluster tightly between roughly .68 and .72. Because the true performance differences between product categories are substantially smaller, even minor estimation errors can change the induced ordering of domains. This helps explain why ranking becomes less stable under small margins and why several methods produce weaker or inconsistent correlations despite relatively competitive RMSE values. Together, these distributions support the central finding of the paper: ranking reliability depends strongly on the underlying heterogeneity and spread of true domain performance, rather than only on absolute estimation accuracy.

Table 11 reports the combinatorial average downstream classification accuracy for each base model across the Amazon Reviews and GeoOLID datasets. On Amazon Reviews, GPT4O-MINI achieves the strongest average accuracy (.751), followed closely by LLAMA-3.1-8B (.745) and ROBERTA (.730), while the LINEAR model performs substantially worse (.574). On GeoOLID, ROBERTA achieves the highest average accuracy (.878), outperforming both GPT4o-MINI (.782) and LLAMA-3.1-8B (.793). These results show that the proposed label-free estimators are evaluated across base classifiers with different architectures and substantially different performance levels, providing a more controlled setting for analyzing ranking reliability.

Base Model	Dataset	Average Acc.
RoBERTa	Amazon	.730
RoBERTa	GeoOLID	.878
Linear	Amazon	.574
Linear	GeoOLID	.790
GPT4o-mini	Amazon	.751
GPT4o-mini	GeoOLID	.782
LLaMa-3.1-8B	Amazon	.745
LLaMa-3.1-8B	GeoOLID	.793

Table 17: Base models' combinatorial average accuracy.

E Dialect Translations

Section 12 presents representative examples from *GeoOLID-Dial* after applying the Multi-VALUE dialect transformations. The examples illustrate how lexical, syntactic, and morphosyntactic variation is introduced while preserving the original semantic intent and offensive-language label. For instance, transformations include features associated with African American English (e.g., habitual "be"), Appalachian English constructions, and Chicano English lexical or grammatical variation. By holding the underlying meaning largely constant while modifying dialectal surface form, *GeoOLID-Dial* provides a controlled setting for evaluating whether classifiers and label-free estimators remain robust across socially and geographically meaningful language varieties. This setup also reduces confounds from topic or label shifts, allowing performance differences to be more directly attributed to dialectal variation.

F Speedup

For a practical test time setting, assuming the incoming data is $1,000$ examples, and a particular model is selected, the Llama error model performs as follows over a server with an NVIDIA $A6000$ GPU, we achieve a running time of 0 minutes and 27.55 seconds. This is much faster and cost-effective than human labelers, and an $A6000$ is not an extremely expensive GPU.

G Use of AI

We used AI only to help write the paper, e.g., to fix grammar and sentences. All ideas are our own.