To Predict or Not to Predict? Towards reliable uncertainty estimation in the presence of noise

Nouran Khallaf, Serge Sharoff
Centre for Translation, Localisation and Interpreting Studies
School of Languages, Cultures and Societies
University of Leeds, UK

Abstract

This study examines the role of uncertainty estimation (UE) methods in multilingual text classification under noisy and non-topical conditions. Using a complex-vs-simple sentence classification task across several languages, we evaluate a range of UE techniques against a range of metrics to assess their contribution to making more robust predictions. Results indicate that while methods relying on softmax outputs remain competitive in high-resource in-domain settings, their reliability declines in low-resource or domain-shift scenarios. In contrast, Monte Carlo dropout approaches demonstrate consistently strong performance across all languages, offering more robust calibration, stable decision thresholds, and greater discriminative power even under adverse conditions. We further demonstrate the positive impact of UE on non-topical classification: abstaining from predicting the 10% most uncertain instances increases the macro F1 score from 0.81 to 0.85 in the Readme task. By integrating UE with trustworthiness metrics, this study provides actionable insights for developing more reliable NLP systems in real-world multilingual environments. See https://github.com/Nouran-Khallaf/To-Predict-or-Not-to-Predict

Keywords: Multilingual NLP; Robustness; Non-topical text classification

1. Introduction

In many real-world NLP applications, it is not enough for a classifier to make predictions; users also need to know when the classifier is more likely to be wrong for an individual instance. This is particularly challenging in tasks involving non-topical classification, noisy data, and multilingual settings. In this paper, we focus on the task of sentence complexity classification, which often experiences performance degradation under cross-lingual transfer and domain shift.

To address these challenges, we investigate the role of Uncertainty Estimation (UE), i.e., the ability to quantify uncertainty of a model when making predictions on instances in a test set (Vazhentsev et al., 2023). UE serves as a critical failure indicator by identifying cases where the model is uncertain and should abstain from making predictions, thereby improving overall robustness. While many UE approaches have recently been proposed, there has been no comprehensive evaluation across multiple languages, noisy contexts, and non-topical classification tasks.

This study makes the following contributions:

1. **Comprehensive evaluation of UE methods:** We benchmark nine popular techniques, including probabilistic, distributional, geometric, and hybrid approaches.

2. **Multilingual and cross-domain investigation:** We evaluate these methods across seven languages (Arabic, Catalan, English, French, Hindi, Russian, and Spanish) and three datasets to assess their robustness to language and domain variation.

3. **Analysis of multiple evaluation metrics for UE assessment:** We compare nine widely used metrics, covering uncertainty discrimination, calibration, and selective prediction perspectives. We further analyse correlations among these measures within and across the perspectives.

4. **Insights into selective prediction:** We show that abstaining from 5–10% of the most uncertain predictions can yield significant improvements in macro F1 score.

The remainder of the paper is organized as follows. Section 2 describes the methodology, including the dataset, and the experimental setup. Section 3 describes the UE methods, and the evaluation metrics. Section 4 presents the results across the multilingual datasets.

2. Methodology

2.1. Datasets

Table 1 lists the datasets used in this study. As our focus is on multilingual detection of segments which are difficult to read, our training dataset is Readme++, a multilingual collection of paragraphs graded by their CEFR level (Naous et al., 2024). We converted it to a binary task: Simple is B1 or below, Complex is C1 or C2.

In addition to testing the multilingual model on its original dataset via cross-validation, we also test it on other datasets to assess its robustness across

Table 1: Statistics of the datasets.

Language	Simple #Ex's	Simple IQR	Complex #Ex's	Complex IQR
ReadMe				
Arabic	766	9–20	533	12–33
English	1,296	9–20	448	13–32
French	810	10–23	367	14–35
Hindi	569	9–20	396	11–29
Russian	702	9–20	328	10–26
Vikidia / Wikipedia				
Catalan	1,236	12–27	1,397	18–38
English	30,553	11–23	26,790	15–32
French	41,127	13–29	41,127	16–36
Italian	32,994	12–26	34,283	13–32
Spanish	4,038	13–29	4,038	18–38
Simplext				
Spanish	167	11–17	167	16–42

languages and domains. One test set comes from Vikidia,[1] a website that maintains Wikipedia-style content aimed at "children and anyone seeking easy-to-read content". We have removed the Vikidia entries marked as stubs (with little content at the moment) and collected the corresponding main Wikipedia entries for the respective languages. This removes topic biases, which often lead to unreasonable accuracy scores in non-topical classification tasks. The set of languages for testing on the Vikidia corpus is determined by the availability of annotators for quality control purposes. As another test set we use Simplext, a manually simplified set of Spanish news articles from the domains of national and international news, society, and culture (Saggion et al., 2015).

Table 1 shows the number of examples for each language and dataset in each category, as well as the interquartile range (IQR) of segment lengths, as this is another potential confounding factor in non-topical text classification.

2.2. Classifiers and Experimental Setup

We trained a multilingual classification model using multilingual Bert (Devlin et al., 2019) with 5-fold cross-validation to estimate the robustness of the classifiers and the respective UE scores. Preliminary experiments with XLM-Roberta showed very similar trends; therefore, we report only mBERT below. Recent studies demonstrate that larger generative LLMs such as GPT or LLaMA do not consistently outperform BERT-sized PLMs in text classification tasks (Edwards and Camacho-Collados, 2024), as was also shown in our own experiments on text complexity (Khallaf et al., 2025). Therefore, we focus on more computationally efficient PLMs in our task.

In terms of hyper-parameters, we fine-tune using the AdamW optimizer with a learning rate of 5×10^{-5}, a cosine decay scheduler (with 10% linear warmup), and early stopping (patience = 5). To promote generalisation and training stability, we apply dropout ($p = 0.3$), mixed-precision training (FP16), and gradient clipping to prevent excessively large gradient updates during unstable training steps, which can occur more frequently with imbalanced or low-resource data.

3. Uncertainty Evaluation

3.1. Uncertainty Estimation Methods

We evaluate UE methods belonging to three groups of approaches: probabilistic, feature geometry, as well as rank-based hybrids of geometric and probabilistic approaches, see Table 2. Formal definitions are listed in the Appendix.

Softmax Response (SR) is a probability-based method that quantifies uncertainty using the maximum predicted class probability from the softmax layer, which is already computed during each inference (Guo et al., 2017).

Sampled Max Probability (SMP) is a probability-based method that extends SR by introducing randomness at inference time using MC Dropout (Gal and Ghahramani, 2016). Instead of relying on a single softmax output, SMP averages the class probabilities over T forward passes with dropout, then computes uncertainty using the maximum of the mean probabilities (Shelmanov et al., 2021).

Entropy (ENT) quantifies uncertainty by measuring the spread of the predicted probability distribution (Ovadia et al., 2019). ENT can be computed directly from the softmax output of a single forward pass, or using averaged probabilities over multiple stochastic passes (ENT-MC). A higher entropy value indicates a prediction which is more uniform over the predicted classes (and thus more uncertain).

Probability Variance (PV) quantifies epistemic uncertainty by checking how much the predicted class probabilities change across T MC–Dropout runs. Unlike ENT-MC or SMP, which rely on the average prediction, PV looks at the spread of those predictions to reflect model instability: bigger spread means the model is more uncertain (Lakshminarayanan et al., 2017).

[1] https://www.vikidia.org/

Table 2: Summary of Uncertainty Estimation methods, grouped by the underlying approaches.

Method	Formulation	Key Characteristics
Class probability-based methods		
SR	Top-class softmax probability.	Directly available, but calibration is questionable.
SMP	MC Dropout-averaged top-class probability.	Captures disagreement across stochastic passes.
ENT	Entropy over softmax probabilities.	Captures prediction ambiguity.
ENT-MC	MC Dropout-averaged entropy over probabilities.	Captures ambiguity across stochastic passes.
PV	MC Dropout-averaged variance in probabilities.	Captures disagreement across stochastic passes.
BALD	Mutual information between model and predictions.	Isolates epistemic uncertainty.
Feature-Based Methods		
MD	Distance of test embeddings.	Captures distance from the training distribution.
ISOF	Isolation-based anomaly detection on embeddings.	Finds OOD inputs using tree partitions.
LOF	Local density deviation on embeddings.	Finds OOD inputs using local density.
Hybrid Methods		
HUQ-MD	Rank-based mix of MD and Mean probability.	Balances model and data uncertainty.

Bayesian Active Learning by Disagreement (BALD) quantifies epistemic uncertainty using information gain in terms of predictive entropies (Houlsby et al., 2011). Variance is approximated via MC-dropout (Gal and Ghahramani, 2016).

Mahalanobis Distance (MD) estimates uncertainty from the model's representations rather than output probabilities. It compares the embeddings of a test sample against the training data. Larger distances mean the test instance looks unlike what has been seen during training (Lee et al., 2018).

Hybrid Uncertainty Quantification (HUQ–MD) combines epistemic and aleatoric uncertainty (Vazhentsev et al., 2023). Epistemic uncertainty comes from Mahalanobis Distance (MD). Aleatoric uncertainty comes from the model's predicted probability for the chosen class: lower confidence (closer to a tie) means higher aleatoric uncertainty. HUQ ranks each test instance by both signals within the dataset and then combines the two ranks. This penalizes both overconfident mistakes and genuinely ambiguous cases.

Local Outlier Factor (LOF) detects uncertainty by measuring how isolated a test example is within its local neighbourhood of training data: outliers are in a less dense region (Breunig et al., 2000).

Isolation Forest (ISOF) identifies examples that look unusual compared to the training data by building a set of binary trees that split the feature space; needing fewer splits to build a tree implies a more anomalous example (Liu et al., 2008).

3.2. Uncertainty evaluation metrics

We organise our evaluation of uncertainty estimators in three groups of metrics (see Appendix B for the formal definitions):

- **Uncertainty discrimination** Ranking ability of uncertainty estimates to distinguish correct from incorrect predictions (Ovadia et al., 2019). High discrimination means the model is better at separating correct vs. incorrect predictions by assigning higher uncertainty to errors.
- **Calibration Metrics** evaluate whether the model's predicted confidence (the inverse of uncertainty) reflects the true likelihood of correctness (Guo et al., 2017; Naeini et al., 2015).
- **Selective Prediction Metrics** measure how effectively uncertainty can improve reliability by rejecting uncertain predictions (Geifman and El-Yaniv, 2017).

When a quality metric requires *confidence* rather than uncertainty, we derive it as the opposite to normalised uncertainty ($c_i = 1 - u_i$).

3.2.1. Uncertainty discrimination

Receiver Operating Characteristic- Area Under the Curve (ROC-AUC) measures the discriminative ability of an uncertainty score, i.e. whether the incorrect predictions have higher uncertainty. We treat prediction correctness as the positive class ($y_i = 1$ if the prediction is correct, 0 otherwise). Higher ROC-AUC indicates that correct predictions receive higher confidence.

Area Under Precision-Recall Curve (AU-PRC) is similar to ROC-AUC, but it uses the Precision-Recall curve, it also treats incorrect predictions as the positive class. This metric is claimed to be particularly informative for higher-accuracy models (Davis and Goadrich, 2006).

Table 3: Summary of UE quality metrics grouped by the evaluation perspective.

Metric	Range	Interpretation
Uncertainty Discrimination		
ROC-AUC	[0, 1]	Indicates whether incorrect predictions have higher uncertainty. Higher is better.
AU-PRC	[0, 1]	Identifies incorrect predictions via uncertainty. Higher is better.
Calibration Metrics		
C-Slope	$[0, \infty]$	Regression of accuracy on confidence. <1 = underconfident; >1 = overconfident.
CITL	$[-1, 1]$	Measures the difference between confidence and accuracy. The best value is 0.
ECE	[0, 1]	Expected calibration error. Lower is better.
Selective Prediction Metrics		
RC-AUC	[0, 1]	Area under the risk-coverage curve. Lower is better.
N.RC-AUC	[0, 1]	Normalised Area under the risk-coverage curve. Higher is better.
E-AUoptRC	[0, 1]	Area under the RC curve up to the full-set performance threshold. Lower is better.
Trust Index	[0, 1]	Performance of macro-F1 on the most confident c^* fraction of predictions. Higher is better.

Table 4: Macro F1 with standard deviations across the folds. The source dataset for training the classifier is Readme. The domain shift tested on Vikidia/Wikipedia and Simplext (Spanish only)

Readme

Language	Class	P	± Std	R	± Std	F1	± Std
Arabic	complex	0.77	0.050	0.86	0.058	0.81	0.035
	simple	0.90	0.023	0.81	0.067	0.85	0.034
English	complex	0.80	0.056	0.79	0.080	0.79	0.036
	simple	0.93	0.022	0.93	0.029	0.93	0.013
French	complex	0.79	0.047	0.85	0.084	0.81	0.037
	simple	0.93	0.032	0.89	0.047	0.91	0.021
Hindi	complex	0.79	0.048	0.74	0.128	0.76	0.075
	simple	0.83	0.069	0.86	0.044	0.84	0.033
Russian	complex	0.75	0.083	0.78	0.113	0.75	0.057
	simple	0.90	0.037	0.87	0.067	0.88	0.028
Macro averaged		0.84	0.013	0.84	0.016	**0.83**	0.014

Vikidia/Wikipedia

Language	Class	P	± Std	R	± Std	F1	± Std
Catalan	complex	0.72	0.044	0.49	0.137	0.57	0.101
	simple	0.62	0.044	0.80	0.075	0.69	0.017
English	complex	0.79	0.020	0.62	0.026	0.69	0.021
	simple	0.68	0.016	0.83	0.017	0.75	0.015
French	complex	0.69	0.006	0.60	0.008	0.64	0.006
	simple	0.64	0.005	0.73	0.008	0.68	0.005
Italian	complex	0.66	0.017	0.64	0.102	0.65	0.058
	simple	0.66	0.049	0.68	0.048	0.66	0.012
Spanish	complex	0.68	0.017	0.72	0.007	0.70	0.011
	simple	0.71	0.011	0.66	0.024	0.68	0.017
Macro averaged		0.69	0.035	0.68	0.036	**0.68**	0.040

Simplext–Spanish

Language	Class	P	± Std	R	± Std	F1	± Std
Spanish	complex	0.79	0.152	0.54	0.174	0.62	0.112
Simplext	simple	0.65	0.064	0.83	0.150	0.72	0.067
Macro averaged		0.72	0.108	0.69	0.162	**0.67**	0.090

3.2.2. Calibration metrics

Calibration Slope (C-slope) evaluates how well the confidence score aligns with prediction accuracy. It is computed by fitting a linear regression model between the **confidence** and the correctness indicator.

Calibration-in-the-large (CITL) captures the average bias in the confidence score. *CITL* $= 0$ is ideal; positive values indicate the model is overconfident, and negative values indicate underconfident (Naeini et al., 2015; Mukhoti et al., 2020).

Expected Calibration Error (ECE) measures how well predicted confidences match observed accuracy (Ao et al., 2023). Predictions are grouped into confidence bins; within each bin, we compute the absolute gap between mean confidence and empirical accuracy, then take a (sample-weighted) average over bins. Lower ECE indicates better calibration; it is important to note that the values depend on the chosen binning strategy. We compute ECE with 15 equal-width confidence bins over $[0, 1]$ and take a sample-weighted average of the absolute gap between mean confidence and empirical accuracy within each bin.

3.2.3. Selective prediction

Risk-Coverage Area Under Curve (RC-AUC): RC-AUC assesses how prediction risk accumulates as we increase the fraction of retained (non-rejected) predictions. We rank samples by decreasing confidence c_i and define risk using the error indicator $r_i = \mathbb{1}[\hat{y}_i \neq y_i]$. Lower RC-AUC indicates that the most confident predictions are less risky.

Normalised RC-AUC: Vazhentsev et al. (2025) claim that the RC-AUC absolute values are not normalised, so they are dataset- and model-

dependent, which makes them difficult to interpret in isolation across the models and the datasets. They suggest rescaling RC-AUC between a random ranking (worst reasonable baseline) and an oracle ranking by true risk (best case).

Expected Area-Under-Optimal-Risk-Coverage (E-AUoptRC) measures the area under the risk-coverage curve up to the accuracy of the model, corresponding to the optimal coverage point beyond which further rejections do not improve performance (Geifman and El-Yaniv, 2017). In our setting, c^* is defined by the model's macro-F_1 (rather than accuracy).

Trust Index (TI). TI evaluates performance when keeping only the most confident predictions (Ao et al., 2023). We adapt the original formulation with accuracy to macro-F_1 to handle class imbalance. Let n be the number of test samples and c^* the coverage at which the model's macro-F_1 on the full set is reached. This involves ranking the predictions by confidence (descending), keeping the top $k = \lfloor c^* \cdot n \rfloor$, and measuring macro-$F_1$ on this subset. We report two settings: the optimal selective threshold (often $c^* \approx 0.50$) and $c = 0.95$ (abstaining from prediction on the least-confident 5%).

4. Results and Error Analysis

4.1. Evaluation of the classifier

Table 4 reports the quality of the baseline classifier for each language via cross-validation with the respective standard deviation values across the folds. Overall, the macro F1 score is acceptable. However, the classifiers have a greater problem in detection of more complex examples, this is true across all languages, irrespectively of the amount of data for pre-training the mBert model. In terms of the domain and language shift, there is a consistent drop, in both cases when the classifier is applied to a new domain with the languages present in the training set (English and French) and when it is applied to a new language (Catalan, Italian and Spanish). Now we want to understand how different UE scores impact classification and how this is influenced by the domain and language shift.

4.2. Uncertainty Estimation scores

The heatmap in Figure 1 aggregates uncertainty-evaluation results across languages by converting each metric to a benefit score on a shared "higher-is-better" axis before computing z-scores. For metrics where larger values are better (e.g., ROC-AUC, AU-PRC, TI), we use the raw values; for metrics where smaller values are better (e.g., ECE, RC-AUC, E-AUoptRC), we invert the sign so that higher always indicates better performance; and for metrics with an ideal target value (e.g., CITL with target 0), we score systems by closeness to that target (e.g., $-|\mathrm{CITL}|$). The transformed scores are then standardised into z-scores to enable direct comparison across methods and languages. Colour intensity reflects relative performance; horizontal whiskers indicate variability across folds and languages. Per-language values appear in Appendix in the full version of the paper.

Our first observation is the inconsistency of UE scores across the quality metrics. One of the surprises is that while the softmax classifiers have been claimed to be poorly calibrated, as they seem to output high probabilities for incorrect predictions (Guo et al., 2017), our experiment shows that SR is a strong measure according to nearly all of the quality metrics across the tested languages, achieving the best CITL scores with relatively low variance across the folds and languages (shorter whiskers). SMP (an MC-dropout version of SR) only shows clear advantage over SR in TI@95 scores.

Dropout-based approaches (PV, BALD) show strong calibration according to C-Slope and ECE, particularly in English, where C-slope reaches 0.75 and 0.72 and ECE ≈ 0.08, but again with higher variance across the folds.

Older outlier detection methods (ISOF, LOF) often achieve the highest discrimination and selective prediction scores in many settings, in particular the best scores for ROC-AUC, RC-AUC and N.RC-AUC. For instance, in the per-language results (Appendix Table 9), *ISOF* attains ROC-AUC ≈ 0.73 (AR), 0.74 (RU), and 0.67 (HI), yet its aggregate stability is weaker. This instability highlights their dataset-dependent behaviour.

Overall, *MD* is the most reliable and consistent choice across languages and evaluation lenses. The hybrid score *HUQ-MD* which combines MD and SR uncertainty is consistently good on discrimination and selection quality measures, but its calibration is much weaker. *ISOF* is often the best or near-best for discrimination/selection but is unstable across languages. Finally, the simple *SR* and *ENT* baselines remain surprisingly effective and robust across languages, offering competitive performance at no additional computational cost.

4.3. Efficiency estimates

We report inference-time overhead per uncertainty method. Experiments were executed on an NVIDIA L40S GPU with 48 GB VRAM. In our implementation, the fastest methods are the feature-based OOD detectors LOF (≈1.34s per fold) and ISOF (≈1.35s). Class probability-based scores (SR and ENT) are slightly slower (≈1.60s per fold) and show higher variability across the folds (much longer for longer examples). However, they come for

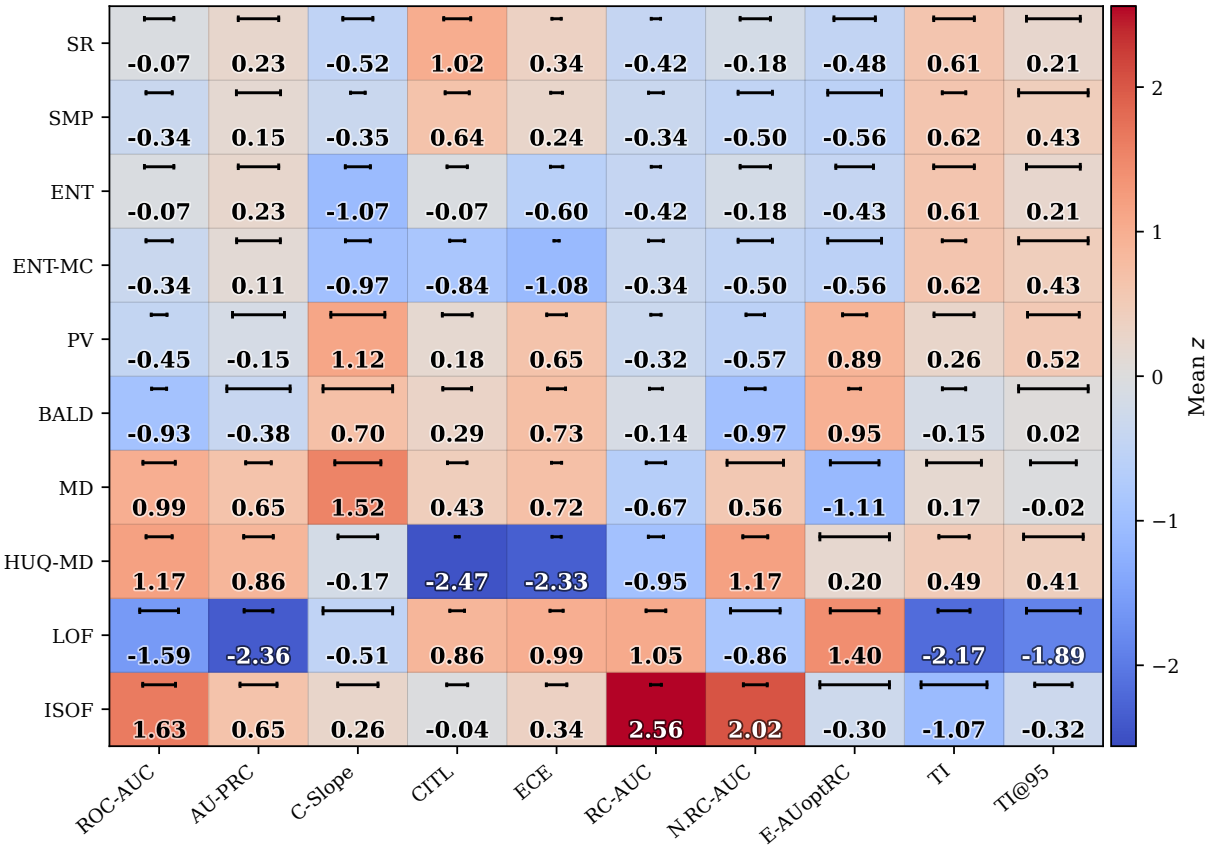


Figure 1: Cross-language average z-scores after applying a direction-aware benefit transform across languages. Cell colors indicate relative performance (better → warmer), and horizontal whiskers show the standard deviation of z across languages.

Table 5: Kendall's τ within metric perspectives.

Metric Pair	AR	EN	FR	HI	RU
Uncertainty Discrimination					
AU-PRC vs ROC-AUC	**0.52**	**0.61**	**0.65**	**0.53**	**0.58**
Calibration Metrics					
CITL vs ECE	**-0.39**	**-0.61**	**-0.55**	**-0.27**	**-0.32**
C-Slope vs CITL	**0.31**	**0.42**	**0.33**	-0.03	0.11
C-Slope vs ECE	**-0.37**	**-0.46**	**-0.45**	**-0.28**	**-0.29**
Selective Prediction Metrics					
E-AUoptRC vs RC-AUC	-0.08	**-0.35**	**-0.34**	**-0.40**	**-0.30**
E-AUoptRC vs NRC-AUC	**-0.33**	-0.22	**-0.46**	**-0.44**	**-0.54**
E-AUoptRC vs TI	-0.07	-0.21	-0.16	**-0.38**	-0.21
NRC-AUC vs TI	**0.39**	0.03	0.13	**0.40**	0.04
RC-AUC vs TI	**0.43**	0.23	**0.31**	**0.55**	**0.34**
RC-AUC vs NRC-AUC	**0.41**	**0.47**	**0.38**	**0.45**	**0.29**
E-AUoptRC vs TI@95	-0.24	-0.20	-0.16	**-0.31**	-0.19
NRC-AUC vs TI@95	**0.33**	-0.09	0.08	**0.32**	0.02
RC-AUC vs TI@95	**0.42**	0.11	**0.27**	**0.47**	**0.28**
TI vs TI@95	**0.60**	**0.64**	**0.67**	**0.76**	**0.64**

free when the prediction task is needed anyway. Mahalanobis-based scoring (MD and HUQ-MD) is moderately more expensive (≈5.78s per fold), with HUQ-MD closely matching MD because it adds only a lightweight rank-combination step. Finally, MC-dropout probability-based methods (SMP, PV, BALD, ENT_MC; T=20) have the highest runtime overhead (≈21.12s per fold), dominated by repeated stochastic forward passes; their near-identical runtimes reflect that they share the same MC-dropout prediction block and differ mainly in inexpensive post-processing.

4.4. Comparison of quality metrics

Tables 5 and 6 compare how the different quality evaluation metrics relate to each other by computing Kendall's τ correlations on the concatenated folds for each language. Bold if $p < 0.01$, underline if $p < 0.05$. The use of Kendall's τ rank-based correlation is justified by the need to compare non-linear patterns in the quality metrics.

Within each perspective (Table 5), the discrimination (ROC-AUC and AU-PRC) and calibration (C-Slope, CITL and ECE) metrics measure very similar properties of the UE scores within each other. Correlation of AU-PRC with ROC-AUC shows that detection of errors (AU-PRC) correlates with ranking of positives and negatives errors (ROC-AUC). ROC-AUC is slightly more stable across the folds and languages (as shown in Figure 1).

Table 6: Kendall's τ across metric perspectives.

Metric Pair	AR	EN	FR	HI	RU
Discrimination vs Calibration					
AU-PRC vs C-Slope	**0.45**	**0.47**	**0.31**	**0.51**	**0.57**
AU-PRC vs CITL	-0.02	0.05	-0.16	-0.22	-0.09
AU-PRC vs ECE	-0.03	-0.07	0.03	0.00	-0.03
C-Slope vs ROC-AUC	**0.38**	**0.40**	**0.31**	**0.59**	**0.46**
CITL vs ROC-AUC	-0.14	-0.01	-0.14	-0.24	-0.13
ECE vs ROC-AUC	-0.15	-0.02	0.00	-0.17	-0.08
Discrimination vs Selection					
AU-PRC vs E-AUoptRC	-0.16	-0.23	**-0.32**	-0.23	-0.10
AU-PRC vs NRC-AUC	**0.30**	**0.50**	**0.47**	**0.41**	0.22
AU-PRC vs RC-AUC	0.18	**0.31**	**0.40**	0.12	0.15
AU-PRC vs TI	**0.34**	0.15	**0.53**	0.16	0.14
AU-PRC vs TI@95	0.21	0.01	**0.36**	-0.00	-0.08
E-AUoptRC vs ROC-AUC	**-0.37**	-0.23	**-0.45**	**-0.42**	**-0.39**
NRC-AUC vs ROC-AUC	**0.72**	**0.75**	**0.76**	**0.80**	**0.57**
RC-AUC vs ROC-AUC	**0.34**	**0.55**	**0.41**	**0.37**	**0.32**
ROC-AUC vs TI	**0.44**	0.16	**0.30**	**0.41**	0.23
ROC-AUC vs TI@95	**0.34**	0.01	0.18	**0.28**	0.07
Calibration vs Selection					
C-Slope vs E-AUoptRC	**-0.32**	**-0.33**	-0.24	**-0.37**	-0.17
C-Slope vs NRC-AUC	**0.26**	**0.30**	0.19	**0.50**	0.16
C-Slope vs RC-AUC	0.19	**0.35**	0.20	0.25	0.12
C-Slope vs TI	**0.30**	0.20	0.25	**0.31**	0.23
C-Slope vs TI@95	**0.30**	0.10	0.18	0.17	0.02
CITL vs E-AUoptRC	-0.10	-0.13	0.05	-0.01	0.00
CITL vs NRC-AUC	-0.19	-0.05	-0.17	**-0.29**	-0.16
CITL vs RC-AUC	-0.06	0.06	-0.14	-0.12	-0.03
CITL vs TI	-0.13	-0.06	-0.19	-0.12	-0.02
CITL vs TI@95	-0.02	-0.12	-0.19	-0.12	-0.13
E-AUoptRC vs ECE	**0.39**	0.18	0.10	**0.32**	0.23
ECE vs NRC-AUC	-0.19	0.01	0.03	-0.16	-0.12
ECE vs RC-AUC	-0.07	0.01	0.06	-0.11	0.08
ECE vs TI	-0.11	-0.03	-0.01	-0.15	-0.10
ECE vs TI@95	-0.24	-0.04	-0.04	-0.09	-0.07

The similarity of calibration measures differs considerably across the languages. The datasets for Hindi and Russian have been developed for the same purpose in the same team. Also the performance of the respective classifiers is consistent for these languages (see Table 4). Nevertheless, CITL differs greatly in their case, which indicates limited robustness of this UE quality metric. ECE is the most stable measure in this group.

The selective prediction evaluation measures show the greatest variation. Only closely related measures (RC-AUC vs N.RC-AUC and TI vs TI@95) show significant correlation across all languages, while there is little agreement in others, for example, N.RC-AUC correlates well with TI in AR/HI but it is close to zero in EN and RU.

Across the perspectives (Table 6), there is little agreement with some exceptions. For the Discrimination vs Selection perspectives, N.RC-AUC and ROC-AUC correlate strongly, suggesting that normalised rejection coverage is quite close to ROC-AUC in identifying well-separated predictions. ROC-AUC correlates with other Selection measures such as E-AUoptRC slightly better than AU-PRC, which is a reason to adopt it as the primary evaluation measure.

Calibration tends to disagree with either discrimination or selection if all languages are considered, except for C-Slope, which does demonstrate statistically significant correlations with ROC-AUC and AU-PRC (Discrimination) and TI (Selection).

The Selective Prediction metrics are inherently different between each other and across the languages.

4.5. Selection at low rejection rates

While Selective Prediction metrics vary, they are inherently interesting from the practical viewpoint because of their direct impact on robustness. Here we want to investigate two questions:

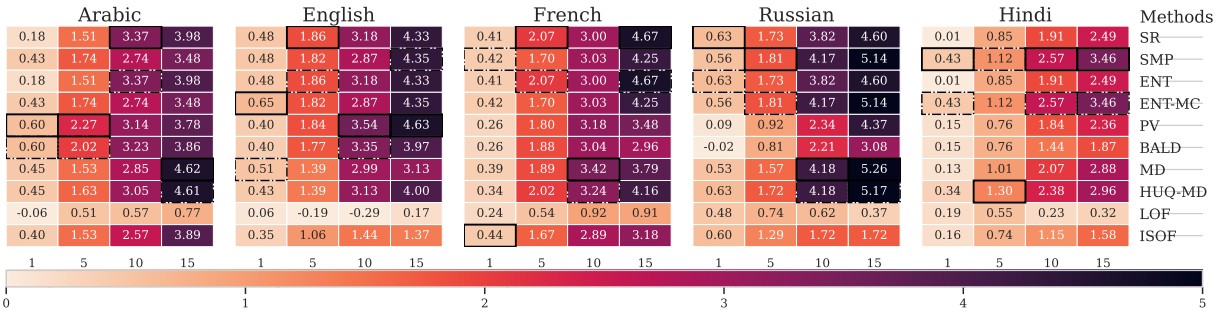
1. the ability of the UE scores to improve prediction quality by abstaining from the decision;
2. the ability of metrics to capture this improved prediction quality.

Metrics for evaluating selective predictions often place emphasis on the best rejection range. In practice, however, it is more realistic to consider low rejection thresholds, since abstaining from predicting a large fraction of the test set is impractical.

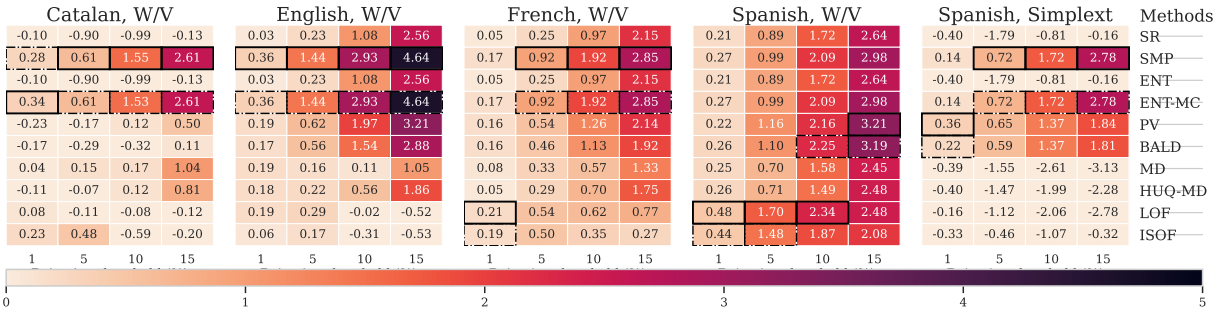
We test across the five core languages in Readme as in-domain, and add two languages (Catalan, Spanish) out-of-domain on Vikidia/Wikipedia and Simplext. This setup probes robustness under both language and domain shift, with in-domain vs. OOD results visualised in Fig. 2 (panels 2a and 2b). The ΔF1 gain is the difference between predicting on the full test set and after abstaining for the rejected test items.

In the in-domain setting, abstention at low rejection thresholds (1–15%) can meaningfully improve robustness, especially for the higher resource languages (the gains are a bit lower for Hindi), as using UE scores yields consistent ΔF1 gains, with no method as a clear winner. The cheap SR and ENT offer good performance, especially for English and French. However, robustness under distribution shift remains a major challenge: in the OOD setting, improvements shrink and become less stable. Especially distance-based methods (LOF, ISOF, MD) show high variability, sometimes with negative improvements, while MC-Dropout based approaches (SMP and ENT-MC) remain comparatively strong.

Error analysis of the most uncertain incorrect predictions (10% threshold) shows that the UE methods detect atypical examples, for example, narrative multi-clause or punctuation-dense structures, atypical lexical patterns, as well as shifts in



(a) In-domain (cross-validation on the Readme dataset)



(b) Out-of-domain (training on Readme; testing on Vikidia/Wikipedia and Simplext)

Figure 2: Improvement measured via ΔF1 over the baseline from Table 4 across rejection thresholds (Columns for 1, 5, 10, 15%) per language per UE score. A solid black box marks the best UE score per column and a dash–dot box marks the second best. The UE scores are labelled on the right.

domains (health/politics/tech) or genres (technical texts), which were missing from the training set.

5. Related Work

Uncertainty estimation (UE) has become a crucial component in assessing the reliability of modern neural networks, particularly in safety-critical or noisy natural language processing (NLP) tasks. Early approaches derived uncertainty directly from model outputs, for example using the maximum *softmax response (SR)* as a confidence proxy (Hendrycks and Gimpel, 2017). However, such probabilistic measures are often poorly calibrated and overconfident under dataset shift (Guo et al., 2017; Liang et al., 2018).

To better capture epistemic uncertainty, stochastic approximation techniques such as Monte Carlo Dropout (MC-Dropout) (Gal and Ghahramani, 2016) and Deep Ensembles (Lakshminarayanan et al., 2017) have been proposed. These approaches generate predictive distributions by repeated stochastic forward passes or model ensembles. Recent studies show that MC-Dropout-derived scores, such as *Sampled Max Probability (SMP)* and entropy-based variants (*ENT-MC*), achieve more reliable uncertainty discrimination and calibration than SR, especially under domain or language shift (Ovadia et al., 2019).

Complementary approaches evaluate uncertainty via feature-space geometry. Mahalanobis Distance (MD) (Lee et al., 2018) measures the distance of a representation to class centroids, while unsupervised detectors such as Local Outlier Factor (LOF) (Breunig et al., 2000) and Isolation Forest (ISOF) (Liu et al., 2008) identify samples lying far from the training distribution. These methods often improve error detection but may exhibit instability across datasets due to sensitivity to embedding space variability.

More recent methods combine complementary uncertainty signals. The *Hybrid Uncertainty Quantification* (HUQ-MD) approach (Vazhentsev et al., 2023) integrates probabilistic and geometric cues, balancing aleatoric and epistemic factors. Hybrid models have been shown to maintain robust calibration and selective prediction performance across multiple NLP benchmarks.

Prior work has primarily focused on English and in-domain evaluations, leaving multilingual and noisy settings underexplored (Lang et al., 2023). Moreover, UE metrics, such as calibration error (ECE), ROC-AUC, and risk–coverage AUC, often correlate weakly across datasets, complicating fair comparisons(Ovadia et al., 2019). Our study addresses these gaps by systematically evaluating nine UE methods across multiple languages and noise conditions.

6. Conclusions and future work

Our results show that:

- Classifier performance degrades under domain and language shifts, and the same holds for the quality of UE scoring.
- Stronger UE performance on in-domain data does not necessarily translate into better prediction under dataset shift.
- SR and ENT are accurate and computationally efficient UE measures, which are often close to the best-performing methods, which makes them a safer choice in higher-resource in-domain settings (rejecting 10% of the lowest SR scores boosts the F1 score from 0.81 to 0.85). However, they become far less robust under domain and language shifts.
- MC-Dropout-based UE methods consistently exhibit better performance under both domain and language variation, particularly in lower-resource languages where softmax confidence begins to break down.
- Traditional outlier detection methods (such as ISOF) as well as BALD, and PV can perform well as UE scorers, often rivalling more recent approaches, but their high variance and sensitivity to data shifts mean they cannot always be trusted in practice, while getting trustable predictions is precisely the reason for using UE scores. Also they require access to the training set to estimate the distance from it, which can be limiting in practice.

Ultimately, our findings return us to the central question of this paper: **to predict or not to predict** is less about finding a single best UE score and more about balancing robustness and computational efficiency.

As future work, we plan to develop a meta-uncertainty framework capable of adaptively selecting or combining uncertainty metrics based on data context. By learning from fold-level and input-level features, such a meta-model could choose the most suitable metric or blend multiple scores to improve uncertainty estimation across domains and conditions. Future research should not only report discrimination and calibration scores, but also directly test how uncertainty estimates translate into selective prediction under both in-domain and OOD conditions. Only by aligning these perspectives can we design UE methods that are not merely theoretically sound but also practically reliable for real-world decision-making. More research is also needed to determine how our findings generalise to multi-class predictions across a wider range of text classification tasks.

In a related study (Khallaf and Sharoff, 2026), we have also investigated the impact of dataset noise in a similar text classification task. This shows that, as expected, noise in the training set reduces performance, while automatic detection of noise (for example, using Gaussian Mixture Models) can improve it. In the next step, it is important to investigate how noise relates to uncertainty on a scale beyond manual error analysis conducted in our current study above.

7. Acknowledgements

This document is part of a project that has received funding from the European Union's Horizon Europe research and innovation program under Grant Agreement No. 101132431 (iDEM Project). The University of Leeds was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant Agreement No. 10103529). The views and opinions expressed in this document are solely those of the author(s) and do not necessarily reflect the views of the European Union. Neither the European Union nor the granting authority can be held responsible for them. We are grateful to Alex Panchenko, Artem Shelmanov and Artem Vazhentsev for their comments on the earlier drafts of the paper.

8. Bibliographical References

Shuang Ao, Stefan Rueger, and Advaith Siddharthan. 2023. Empirical optimal risk to quantify model trustworthiness for failure detection. *arXiv preprint arXiv:2308.03179*.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 29(2):93–104.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification:

Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 1050–1059.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Nouran Khallaf, Stefan Bott, Carlo Eugeni, John O'Flaherty, Serge Sharoff, and Horacio Saggion. 2025. Democracy made easy: Simplifying complex topics to enable democratic participation. In *Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI & EL/PL)*, pages 108–124, Geneva, Switzerland. European Association for Machine Translation.

Nouran Khallaf and Serge Sharoff. 2026. How much noise can BERT handle? insights from multilingual sentence difficulty detection. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Hao Lang, Yinhe Zheng, Yixuan Li, Jian Sun, Fei Huang, and Yongbin Li. 2023. A survey on out-of-distribution detection in nlp. *arXiv preprint arXiv:2305.03236*.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *arXiv preprint arXiv:1807.03888*.

Shiyu Liang, Yixuan Li, and Raghav Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. 2020. Calibrating deep neural networks using focal loss.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907.

Tarek Naous, Michael J Ryan, Anton Lavrouk, Mohit Chandra, and Wei Xu. 2024. ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12230–12266, Miami, Florida, USA. Association for Computational Linguistics.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Andrey Podolskiy, Dmitry Lipin, Artem Bout, Ekaterina Artemova, and Irina Piontkovskaya. 2021. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Trans. Access. Comput.*, 6(4).

Artem Shelmanov, Evgenii Tsymbalov, Dmitri Puzyrev, Kirill Fedyanin, Alexander Panchenko, and Maxim Panov. 2021. How certain is your Transformer? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1833–1840, Online. Association for Computational Linguistics.

Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. Hybrid uncertainty quantification for selective text classification in ambiguous tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

Artem Vazhentsev, Ivan Sviridov, Alvard Barseghyan, Gleb Kuzmin, Alexander Panchenko, Aleksandr Nesterov, Artem Shelmanov, and Maxim Panov. 2025. Uncertainty-aware abstention in medical diagnosis based on medical texts.

A. Uncertainty Measures

Code	Type	Formula	Description
Class Probability–based Methods			
SR	Aleatoric	$u_{\text{SR}} = 1 - \max_{c \in C} p^c$ where C is the number of classes, p_t^c is the predicted probability for class c during the t-th forward pass.	Uses the maximum softmax probability as a proxy for confidence. Higher maximum $\Rightarrow$ lower uncertainty (Guo et al., 2017).
SMP	Epistemic	$u_{\text{SMP}} = 1 - \max_{c \in C} \left(\frac{1}{T} \sum_{t=1}^{T} p_t^c \right)$ where p_t^c denotes the predicted probability for class c during the t-th forward pass.	Averages top-class probability over T MC-dropout passes; captures parameter uncertainty (Gal and Ghahramani, 2016).
ENT	Aleatoric	$u_{\text{ENT}} = -\sum_{c=1}^{C} p^c \log p^c$	Ambiguity of the class probability distribution (Ovadia et al., 2019).
ENT-MC	Total (Predictive)	$u_{\text{ENT-MC}} = -\sum_{c=1}^{C} \overline{p^c} \log \overline{p^c}$	Entropy of mean probabilities across MC passes; captures aleatoric+epistemic (Ovadia et al., 2019).
PV	Epistemic	$u_{\text{PV}} = \frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{T} \sum_{t=1}^{T} (p_t^c - \overline{p^c})^2 \right)$ where C is the number of classes, p_t^c is the predicted probability for class c during the t-th forward pass.	Variability in predicted probabilities across MC passes (Lakshminarayanan et al., 2017).
BALD	Epistemic	$u_{\text{BALD}} = u_{\text{ENT-MC}} + \frac{1}{T} \sum_{t=1}^{T} \sum_{c=1}^{C} p_t^c \log p_t^c$ the first term corresponds to the entropy of the mean prediction (i.e., total predictive uncertainty), while the second term represents the expected entropy over all passes (i.e., data or aleatoric uncertainty)	Mutual information between predictions and weights; separates total vs. aleatoric parts (Gal and Ghahramani, 2016).
Feature–based Methods			
MD	Epistemic	$\min_{c \in C} (h_i - \mu_c)^\top \Sigma^{-1} (h_i - \mu_c)$ where h_i is the hidden representation of the i-th test instance, μ_c is the mean embedding (centroid) for class c, and Σ is the shared covariance matrix estimated from training data.	Distance in latent space to class centroids; effective for OOD (Podolskiy et al., 2021).
LOF	Epistemic / OOD	$\frac{1}{\lvert N_{\min}(x) \rvert} \sum_{o \in N_{\min}(x)} \frac{lrd(o)}{lrd(x)}$ comparing their local reachability density (lrd) to nearby training examples N_{Min}	Local density deviation vs. neighbors (Breunig et al., 2000).
ISOF	Epistemic / OOD	$-2^{-\frac{1}{N} \sum_{i=1}^{N} l_i(x)}$ where l_i is the length of paths of trees in the set, see (Liu et al., 2008) for the method of their construction. To match our "higher = more uncertain" convention, we invert the LOF density output and normalise the result, so larger values mean more outlier-like (and thus more uncertain).	Isolation depth in random trees; shorter paths $\Rightarrow$ more anomalous (Liu et al., 2008).
Hybrid Methods			
HUQ-MD	Combined	$U_T(x) = (1-\alpha) R(U_E(x), \mathcal{D}) + \alpha R(U_A(x), \mathcal{D})$ where h_i is the hidden representation of the i-th test instance, μ_c is the mean embedding (centroid) for class c, and Σ is the shared covariance matrix estimated from training data. For consistency with other measures, we convert the model's scores so that higher values indicate greater uncertainty, and we normalise them to a common scale.	Rank-based mix of epistemic and aleatoric; $\alpha \in [0, 1]$ tunes the trade-off.

Table 7: Summary of uncertainty estimation methods split by formula and description. Methods are grouped by family and labeled by the uncertainty type—**aleatoric**, **epistemic**, and **total/predictive**.

Notes. Aleatoric: data noise/ambiguity (e.g., overlapping classes); captured by single-pass entropy (ENT). **Epistemic:** model uncertainty due to limited data; estimated via variability across stochastic passes (PV, BALD). **Total (Predictive):** overall uncertainty after marginalizing model parameters (aleatoric+epistemic); entropy of averaged predictive distribution (ENT-MC).

B. Uncertainty evaluation metrics

Code	Formula	Description
Discrimination / Ranking		
ROC-AUC	$\text{ROC-AUC} = \text{roc\_auc\_score}(\{y_i\}, \{c_i\})$ y_i=1 if correct else 0	Ranks confidence of correct vs. incorrect predictions across thresholds.
AU-PRC	$\text{AU-PRC} = \int_0^1 \text{Precision}(r)\, d\text{Recall}(r)$	Area under precision–recall curve; preferable to ROC when positives (or correctness) are imbalanced.
Calibration		
Cal. Slope	$y_i = \alpha + \beta\, c_i + \varepsilon_i$ β by OLS	Ideal β=1; β<1 over-confident, β>1 under-confident.
CITL	$\text{CITL} = \frac{1}{n}\sum_{i=1}^{n} c_i - \frac{1}{n}\sum_{i=1}^{n} y_i$	Calibration-in-the-large; 0 indicates perfect average calibration (Naeini et al., 2015; Mukhoti et al., 2020).
ECE	$\text{ECE} = \sum_{m=1}^{M} \frac{\lvert B_m \rvert}{n} \lvert \text{acc}(m) - \text{conf}(m) \rvert$	Expected Calibration Error with M bins; compares accuracy vs. mean confidence per bin (Ao et al., 2023).
Selective Prediction / Risk–Coverage		
RC-AUC	$\text{RC-AUC} = \sum_{k=1}^{n} \frac{1}{k} \sum_{i=1}^{k} r_{(i)}$ $r_{(i)}$: risks ordered by decreasing confidence	Area under the risk–coverage curve; lower cumulative risk at higher coverage is better.
NRC-AUC	$\text{NRC-AUC} = \frac{\text{RC-AUC}_{\text{model}} - \text{RC-AUC}_{\text{random}}}{\text{RC-AUC}_{\text{oracle}} - \text{RC-AUC}_{\text{random}}}$	Normalizes RC-AUC to $[0, 1]$ between random and oracle sorting; 1 is oracle-level (Vazhentsev et al., 2025).
E-AUoptRC	$\text{E-AUoptRC} = \int_0^{c^*} r(c)\, dc$ c^*: coverage defined by full-set macro-F_1	Expected area under risk–coverage up to the optimal coverage c^* (Geifman and El-Yaniv, 2017).
Trust / Thresholded Performance		
TI	$\text{TI} = \text{F}_1(\{(\hat{y}_i, y_i)\}_{i=1}^{k})$ $k = \lfloor c^* n \rfloor$	F1 on the k most confident predictions (coverage c^*); measures trustworthiness of the most confident subset (Ao et al., 2023).

Table 8: Summary of uncertainty evaluation metrics split by *Formula* and *Description*. Metrics are grouped into discrimination/ranking, calibration, selective prediction (risk–coverage), and trust-related measures.

Note: Notation: $y_i \in \{0, 1\}$ (correctness), $c_i \in [0, 1]$ (confidence), $r(c)$ risk at coverage c, $r_{(i)}$ risks sorted by decreasing confidence, c^* coverage at overall accuracy, $\hat{y}_i$ predicted label, n dataset size, $k = \lfloor c^* n \rfloor$.

Methods Summary

Arabic																
Metric	SR		SMP		ENT		ENT-MC		PV		BALD		MD		HUQ-MD		LOF		ISOF	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
ROC-AUC	0.64	0.055	0.63	0.077	0.64	0.055	0.63	0.077	0.64	0.069	0.62	0.056	0.67	0.075	0.67	0.064	0.58	0.046	**0.73**	0.067
AU-PRC	0.30	0.042	0.32	0.035	0.30	0.042	0.32	0.035	0.36	0.069	0.36	0.059	0.34	0.047	0.34	0.032	0.22	0.046	**0.38**	0.114
C-Slope	0.37	0.212	0.35	0.204	0.29	0.188	0.27	0.188	0.74	0.282	**0.77**	0.279	0.56	0.134	0.34	0.100	0.27	0.182	0.50	0.218
CITL	**-0.00**	0.055	-0.04	0.056	-0.09	0.091	-0.16	0.069	0.12	0.036	0.11	0.034	0.09	0.058	-0.33	0.036	0.05	0.025	-0.15	0.064
ECE	0.15	0.082	0.16	0.074	0.20	0.126	0.24	0.105	0.14	0.043	0.13	0.038	0.14	0.025	0.34	0.029	**0.11**	0.020	0.17	0.054
RC-AUC	0.88	0.030	0.87	0.038	0.88	0.030	0.87	0.038	0.87	0.050	0.86	0.052	0.87	0.053	0.88	0.046	0.80	0.038	**0.72**	0.063
N.RC-AUC	0.35	0.114	0.30	0.152	0.35	0.114	0.30	0.152	0.28	0.203	0.23	0.196	0.33	0.206	0.39	0.180	0.24	0.141	**0.51**	0.113
E-AUoptRC	0.23	0.067	0.23	0.081	0.22	0.067	0.23	0.081	0.17	0.112	0.16	0.104	0.22	0.056	0.18	0.063	**0.11**	0.020	0.17	0.084
TI	0.87	0.039	0.87	0.039	0.87	0.039	0.87	0.039	0.87	0.041	0.87	0.039	0.88	0.040	**0.88**	0.036	0.83	0.035	0.87	0.055
TI@95	0.84	0.036	0.85	0.034	0.84	0.036	0.85	0.034	**0.85**	0.038	0.85	0.040	0.84	0.035	0.84	0.034	0.83	0.037	0.84	0.038
English																				
ROC-AUC	0.70	0.109	0.65	0.115	0.70	0.109	0.65	0.115	0.63	0.082	0.61	0.085	0.72	0.051	**0.73**	0.047	0.55	0.037	0.73	0.112
AU-PRC	**0.31**	0.060	0.29	0.055	0.31	0.060	0.28	0.055	0.28	0.045	0.27	0.054	0.29	0.065	0.30	0.032	0.13	0.020	0.29	0.129
C-Slope	0.44	0.302	0.40	0.283	0.33	0.258	0.32	0.259	**0.75**	0.212	0.72	0.250	0.68	0.326	0.31	0.044	0.14	0.150	0.36	0.228
CITL	-0.08	0.090	-0.11	0.069	-0.18	0.159	-0.25	0.113	0.07	0.019	0.06	0.028	0.05	0.033	-0.39	0.011	**-0.05**	0.099	-0.15	0.091
ECE	0.15	0.096	0.16	0.096	0.24	0.155	0.28	0.137	0.08	0.021	**0.08**	0.017	0.09	0.035	0.39	0.015	0.10	0.054	0.16	0.089
RC-AUC	0.92	0.034	0.91	0.034	0.92	0.034	0.91	0.034	0.91	0.025	0.91	0.022	0.93	0.024	0.94	0.023	0.88	0.016	**0.81**	0.072
N.RC-AUC	0.31	0.281	0.22	0.313	0.31	0.281	0.22	0.313	0.20	0.308	0.18	0.338	0.41	0.231	0.47	0.198	0.13	0.138	**0.51**	0.198
E-AUoptRC	0.14	0.062	0.14	0.071	0.14	0.063	0.14	0.071	0.10	0.058	**0.10**	0.046	0.23	0.094	0.19	0.091	0.14	0.053	0.25	0.095
TI	0.90	0.019	**0.90**	0.012	0.90	0.019	0.90	0.012	0.90	0.007	0.89	0.012	0.88	0.034	0.89	0.024	0.86	0.019	0.87	0.036
TI@95	0.88	0.022	0.87	0.022	0.88	0.022	0.87	0.022	0.88	0.018	**0.88**	0.016	0.87	0.020	0.87	0.020	0.85	0.021	0.87	0.009
French																				
ROC-AUC	0.64	0.141	0.62	0.123	0.64	0.141	0.62	0.123	0.63	0.049	0.61	0.066	0.75	0.051	0.76	0.057	0.59	0.069	**0.78**	0.097
AU-PRC	0.31	0.074	0.26	0.050	0.31	0.073	0.26	0.050	0.28	0.037	0.27	0.069	0.33	0.053	**0.33**	0.037	0.20	0.055	0.33	0.106
C-Slope	0.34	0.243	0.36	0.145	0.24	0.268	0.27	0.174	0.59	0.193	0.58	0.189	**0.81**	0.298	0.38	0.088	0.35	0.209	0.43	0.180
CITL	-0.05	0.116	-0.09	0.100	-0.15	0.189	-0.22	0.137	0.07	0.029	0.06	0.032	0.07	0.037	-0.38	0.022	**0.01**	0.080	-0.13	0.105
ECE	0.16	0.075	0.17	0.077	0.25	0.134	0.28	0.119	0.10	0.035	**0.09**	0.036	0.10	0.032	0.38	0.021	0.10	0.023	0.16	0.069
RC-AUC	0.91	0.049	0.90	0.047	0.91	0.049	0.90	0.047	0.90	0.028	0.90	0.026	0.94	0.023	0.95	0.021	0.83	0.048	**0.76**	0.051
N.RC-AUC	0.26	0.365	0.20	0.383	0.26	0.365	0.20	0.383	0.19	0.222	0.19	0.212	0.60	0.161	0.63	0.159	0.25	0.152	**0.67**	0.150
E-AUoptRC	0.16	0.105	0.17	0.102	0.16	0.107	0.17	0.102	0.12	0.031	0.11	0.025	0.15	0.135	**0.08**	0.037	0.09	0.019	0.15	0.116
TI	**0.92**	0.030	0.91	0.031	0.92	0.030	0.91	0.031	0.90	0.030	0.90	0.038	0.90	0.048	0.91	0.032	0.88	0.020	0.90	0.048
TI@95	**0.89**	0.036	0.88	0.032	0.89	0.036	0.88	0.032	0.89	0.031	0.89	0.031	0.89	0.039	0.89	0.038	0.87	0.029	0.88	0.036
Russian																				
ROC-AUC	0.65	0.115	0.64	0.141	0.65	0.115	0.64	0.141	0.64	0.049	0.62	0.051	0.75	0.058	**0.75**	0.062	0.62	0.042	0.74	0.101
AU-PRC	0.37	0.148	0.36	0.170	0.37	0.148	0.36	0.170	0.29	0.080	0.27	0.084	0.39	0.123	**0.40**	0.136	0.25	0.021	0.36	0.140
C-Slope	0.40	0.357	0.42	0.382	0.35	0.340	0.34	0.332	0.42	0.305	0.38	0.274	**0.52**	0.395	0.45	0.137	0.48	0.152	0.48	0.194
CITL	**-0.02**	0.060	-0.05	0.046	-0.12	0.116	-0.19	0.065	0.11	0.030	0.10	0.033	0.09	0.060	-0.33	0.030	0.03	0.039	-0.11	0.044
ECE	0.15	0.083	0.16	0.086	0.22	0.129	0.26	0.115	0.14	0.028	0.14	0.029	0.13	0.047	0.35	0.022	**0.09**	0.014	0.12	0.041
RC-AUC	0.87	0.051	0.87	0.050	0.87	0.051	0.87	0.050	0.88	0.027	0.87	0.031	0.90	0.048	0.91	0.040	0.78	0.037	**0.73**	0.086
N.RC-AUC	0.20	0.393	0.16	0.356	0.20	0.393	0.16	0.356	0.25	0.117	0.21	0.161	0.46	0.221	0.51	0.178	0.33	0.179	**0.61**	0.180
E-AUoptRC	0.18	0.086	0.19	0.102	0.18	0.086	0.19	0.102	0.13	0.068	0.13	0.065	0.18	0.108	0.15	0.099	**0.12**	0.027	0.13	0.113
TI	0.87	0.047	**0.87**	0.042	0.87	0.047	0.87	0.042	0.87	0.022	0.85	0.032	0.86	0.045	0.86	0.040	0.83	0.045	0.83	0.081
TI@95	0.83	0.037	**0.83**	0.039	0.83	0.037	0.83	0.039	0.83	0.032	0.82	0.030	0.83	0.036	0.83	0.038	0.82	0.037	0.83	0.038
Hindi																				
ROC-AUC	0.61	0.115	0.62	0.118	0.61	0.115	0.62	0.118	0.60	0.099	0.57	0.117	0.63	0.015	0.65	0.022	0.54	0.063	**0.67**	0.082
AU-PRC	0.30	0.096	0.33	0.108	0.30	0.096	0.33	0.105	0.29	0.074	0.28	0.079	0.32	0.100	**0.34**	0.123	0.23	0.041	0.31	0.067
C-Slope	0.19	0.373	0.24	0.342	0.17	0.341	0.22	0.324	0.39	0.444	0.29	0.561	**0.41**	0.400	0.30	0.082	0.24	0.255	0.31	0.120
CITL	**-0.01**	0.038	-0.04	0.062	-0.11	0.093	-0.18	0.081	0.14	0.051	0.13	0.046	0.10	0.082	-0.31	0.045	0.07	0.040	-0.09	0.117
ECE	0.16	0.068	0.16	0.080	0.23	0.107	0.26	0.106	0.17	0.044	0.16	0.045	0.15	0.043	0.34	0.044	**0.12**	0.041	0.16	0.073
RC-AUC	0.84	0.059	0.85	0.064	0.84	0.059	0.85	0.064	0.84	0.075	0.82	0.086	0.84	0.045	0.86	0.034	0.79	0.038	**0.73**	0.021
N.RC-AUC	0.23	0.288	0.26	0.369	0.23	0.288	0.26	0.368	0.22	0.255	0.14	0.256	0.20	0.088	0.31	0.049	0.12	0.121	**0.44**	0.213
E-AUoptRC	0.21	0.094	0.20	0.100	0.21	0.094	0.20	0.100	0.16	0.099	0.17	0.102	0.28	0.090	0.23	0.075	**0.14**	0.039	0.23	0.078
TI	0.83	0.058	**0.84**	0.057	0.83	0.058	0.84	0.057	0.83	0.074	0.83	0.083	0.84	0.044	0.84	0.041	0.81	0.065	0.81	0.099
TI@95	0.81	0.057	0.82	0.051	0.81	0.057	0.82	0.051	0.81	0.063	0.81	0.063	0.81	0.049	**0.82**	0.046	0.81	0.056	0.81	0.064

Table 9: Summary of the mean (μ) and standard deviation (σ) of each metric across folds for all methods grouped by the evaluation perspective. Bold indicates the best method; underlined methods are not significantly different from the best at $p = 0.05$ (paired tests).

C. $\triangle$F1 across rejection thresholds

Method	1%				5%				10%				15%			
	MacroF1		Percent		MacroF1		Percent		MacroF1		Percent		MacroF1		Percent	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Arabic																
SR	0.18	0.41	33.3	29.8	1.51	0.49	44.5	9.1	**3.37**	0.44	**46.2**	4.4	3.98	0.97	38.4	4.0
SMP	0.43	0.19	53.3	16.3	1.74	0.36	47.7	6.9	2.74	0.71	40.1	2.9	3.48	1.49	35.4	6.8
ENT	0.18	0.40	33.3	29.8	1.51	0.49	44.5	9.1	3.37	0.44	46.2	4.4	3.98	0.97	38.4	4.0
ENT-MC	0.43	0.19	53.3	16.3	1.74	0.36	47.7	6.9	2.74	0.71	40.1	2.9	3.48	1.49	35.4	6.8
PV	**0.60**	0.37	**66.7**	29.8	**2.27**	0.47	**58.2**	7.3	3.14	1.40	44.7	11.4	3.78	1.54	37.8	7.1
BALD	0.60	0.37	66.7	29.8	2.02	0.64	53.6	9.2	3.23	1.12	45.4	9.6	3.86	1.66	38.3	7.9
MD	0.45	0.18	53.3	16.3	1.53	0.19	44.7	3.9	2.85	1.08	41.7	10.3	**4.62**	1.71	**42.4**	9.4
HUQ-MD	0.45	0.20	53.3	16.3	1.63	0.40	46.0	7.1	3.05	0.84	43.2	8.3	4.61	1.02	41.9	5.5
LOF	-0.06	0.34	13.3	26.7	0.51	0.57	27.1	9.7	0.57	1.02	23.6	8.4	0.77	1.40	22.3	8.8
ISOF	0.40	0.37	53.3	26.7	1.53	1.21	48.0	18.0	2.57	2.56	42.5	17.5	3.89	3.80	41.3	15.9
English																
SR	0.48	0.20	55.0	10.0	**1.86**	0.45	**43.8**	5.0	3.18	0.31	37.4	6.6	4.33	0.39	32.5	5.5
SMP	0.48	0.50	50.0	27.4	1.82	0.69	40.5	8.7	2.87	0.51	34.0	8.6	4.35	1.07	31.7	9.5
ENT	0.48	0.20	55.0	10.0	1.86	0.45	43.8	5.0	3.18	0.31	37.4	6.6	4.33	0.39	32.5	5.5
ENT-MC	**0.65**	0.39	**60.0**	25.5	1.82	0.69	40.5	8.7	2.87	0.51	34.0	8.6	4.35	1.07	31.7	9.5
PV	0.40	0.50	45.0	29.2	1.84	0.67	41.7	9.7	**3.54**	0.81	**38.9**	9.0	**4.63**	1.07	**33.0**	8.5
BALD	0.40	0.50	45.0	29.2	1.77	0.53	41.5	7.4	3.35	1.06	37.6	10.4	3.97	1.64	29.9	9.8
MD	0.51	0.27	50.0	22.4	1.39	0.64	38.8	10.5	2.99	0.76	37.8	8.4	3.13	1.41	32.0	5.0
HUQ-MD	0.43	0.24	45.0	18.7	1.39	0.60	37.7	8.6	3.13	0.64	37.4	8.2	4.00	0.57	32.5	4.2
LOF	0.06	0.30	25.0	15.8	-0.19	0.50	15.8	7.1	-0.29	0.51	14.2	2.7	0.17	0.51	15.1	4.2
ISOF	0.35	0.45	45.0	29.2	1.06	1.22	37.1	17.0	1.44	2.66	30.4	16.1	1.37	4.08	27.3	13.8
French																
SR	0.41	0.48	**46.7**	26.7	**2.07**	0.91	**46.2**	11.9	3.00	1.15	35.3	8.2	**4.67**	2.00	**33.8**	9.9
SMP	0.42	0.37	40.0	24.9	1.70	0.49	40.9	4.2	3.03	0.74	36.6	2.2	4.25	1.32	32.4	6.4
ENT	0.41	0.48	46.7	26.7	2.07	0.91	46.2	11.9	3.00	1.15	35.3	8.2	4.67	2.00	33.8	9.9
ENT-MC	0.42	0.37	40.0	24.9	1.70	0.49	40.9	4.2	3.03	0.74	36.6	2.2	4.25	1.32	32.4	6.4
PV	0.26	0.36	33.3	21.1	1.80	0.55	42.6	7.7	3.18	0.83	38.3	4.0	3.48	0.81	29.0	4.0
BALD	0.26	0.36	33.3	21.1	1.88	0.59	44.1	8.9	3.04	1.24	36.6	9.1	2.96	1.11	26.2	4.9
MD	0.39	0.58	46.7	34.0	1.89	1.19	42.9	17.4	**3.42**	1.38	**40.1**	6.7	3.79	1.77	33.5	4.0
HUQ-MD	0.34	0.51	40.0	24.9	2.02	1.13	44.6	16.0	3.24	1.75	37.8	11.3	4.16	1.37	33.6	4.6
LOF	0.24	0.05	33.3	0.0	0.54	0.94	25.8	15.3	0.92	1.39	21.4	11.7	0.91	1.35	18.4	8.6
ISOF	**0.44**	0.43	46.7	26.7	1.67	1.34	42.3	18.1	2.89	2.40	37.7	13.4	3.18	2.94	32.0	8.2
Russian																
SR	**0.63**	0.53	**63.3**	37.1	1.73	0.75	**49.5**	13.2	3.82	2.19	49.3	16.9	4.60	2.26	40.9	12.7
SMP	0.56	0.48	56.7	38.9	**1.81**	1.33	49.1	21.6	4.17	2.14	50.3	17.6	5.14	2.59	42.8	14.0
ENT	0.63	0.53	63.3	37.1	1.73	0.75	49.5	13.2	3.82	2.19	49.3	16.9	4.60	2.26	40.9	12.7
ENT-MC	0.56	0.48	56.7	38.9	1.81	1.33	49.1	21.6	4.17	2.14	50.3	17.6	5.14	2.59	42.8	14.0
PV	0.09	0.17	23.3	20.0	0.92	0.83	36.2	14.5	2.34	1.72	41.4	14.0	4.37	1.89	42.5	11.2
BALD	-0.02	0.25	13.3	16.3	0.81	0.91	34.5	14.4	2.21	1.88	40.5	14.7	3.08	2.71	36.1	14.1
MD	0.53	0.43	56.7	32.7	1.57	1.26	45.5	21.8	**4.18**	1.77	**51.5**	14.3	**5.26**	1.68	**44.9**	9.5
HUQ-MD	0.63	0.53	63.3	37.1	1.72	1.27	47.3	19.8	4.18	1.45	50.5	12.4	5.17	1.50	43.6	9.9
LOF	0.48	0.43	56.7	22.6	0.74	0.58	32.4	8.4	0.62	1.38	26.7	7.3	0.37	1.48	22.6	4.5
ISOF	0.60	0.53	63.3	37.1	1.29	1.25	46.4	19.3	1.72	2.56	38.5	16.3	1.72	4.50	35.8	14.1
Hindi																
SR	0.01	0.54	30.0	40.0	0.85	1.02	38.4	15.8	1.91	1.24	37.8	10.7	2.49	1.01	34.7	7.3
SMP	**0.43**	0.34	**53.3**	32.3	1.12	1.01	40.0	20.8	**2.57**	1.19	**42.7**	12.2	**3.46**	1.33	**39.2**	9.2
ENT	0.01	0.54	30.0	40.0	0.85	1.02	38.4	15.8	1.91	1.24	37.8	10.7	2.49	1.01	34.7	7.3
ENT-MC	0.43	0.34	53.3	32.3	1.12	1.01	40.0	20.8	2.57	1.19	42.7	12.2	3.46	1.33	39.2	9.2
PV	0.15	0.52	40.0	37.4	0.76	0.99	36.5	11.3	1.84	1.57	37.9	7.0	2.36	2.44	34.3	7.5
BALD	0.15	0.52	40.0	37.4	0.76	0.99	36.5	11.3	1.44	1.85	34.0	10.9	1.87	3.07	30.4	11.6
MD	0.13	0.42	36.7	37.1	1.01	0.78	40.0	16.5	2.07	1.06	39.6	10.5	2.88	1.24	37.3	9.3
HUQ-MD	0.34	0.58	50.0	44.7	**1.30**	1.16	**43.7**	21.7	2.38	1.41	40.5	14.5	2.96	1.42	37.3	10.2
LOF	0.19	0.39	33.3	27.9	0.55	0.61	29.7	7.8	0.23	1.01	21.3	5.2	0.32	1.59	21.9	4.3
ISOF	0.16	0.25	36.7	19.4	0.74	1.18	38.5	14.0	1.15	2.17	33.9	12.5	1.58	3.11	33.0	8.9

Table 10: Change in Macro ($\triangle$F1, in %) after rejecting the most uncertain samples at each threshold, and the percentage of rejected incorrect predictions. Values are mean (μ) and std (σ) over 5 folds.

Method	1%				5%				10%				15%			
	Macro		Percent		Macro		Percent		Macro		Percent		Macro		Percent	
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Catalan, Wikipedia/Vikidia																
SR	-0.10	0.45	40.0	25.5	-0.90	0.75	32.2	6.5	-0.99	1.74	38.3	6.9	-0.13	1.98	44.2	2.7
SMP	0.28	0.34	65.0	25.5	**0.61**	1.01	**52.2**	13.9	**1.55**	0.95	**53.7**	5.5	**2.61**	0.75	**54.2**	4.1
ENT	-0.10	0.45	40.0	25.5	-0.90	0.75	32.2	6.5	-0.99	1.74	38.3	6.9	-0.13	1.98	44.2	2.7
ENT-MC	**0.34**	0.39	**70.0**	29.2	0.61	1.01	52.2	13.9	1.53	0.94	53.7	5.5	2.61	0.75	54.2	4.1
PV	-0.23	0.24	25.0	15.8	-0.17	1.07	40.0	8.2	0.12	1.38	42.3	6.6	0.50	2.46	44.6	6.6
BALD	-0.17	0.23	30.0	10.0	-0.29	1.00	35.6	10.3	-0.32	1.81	36.6	12.8	0.11	3.11	40.4	13.0
MD	0.04	0.42	45.0	33.2	0.15	0.93	44.4	14.9	0.17	0.95	41.7	9.7	1.04	1.24	45.4	7.3
HUQ-MD	-0.11	0.31	35.0	25.5	-0.07	0.75	41.1	13.4	0.12	0.78	41.7	6.9	0.81	1.30	45.0	6.8
LOF	0.08	0.12	45.0	18.7	-0.11	0.57	37.8	11.3	-0.08	0.39	38.3	8.2	-0.12	0.53	37.7	8.3
ISOF	0.23	0.28	60.0	20.0	0.48	1.21	52.2	15.2	-0.59	2.62	41.1	12.7	-0.20	4.01	43.5	13.4
English, Wikipedia/Vikidia																
SR	0.03	0.13	35.0	11.7	0.23	0.64	36.4	10.2	1.08	0.88	40.6	7.4	2.56	0.97	43.6	5.7
SMP	**0.36**	0.24	**61.7**	22.4	**1.44**	0.36	**54.4**	6.4	**2.93**	0.54	**53.6**	5.5	**4.64**	0.94	**53.3**	6.2
ENT	0.03	0.13	35.0	11.7	0.23	0.64	36.4	10.2	1.08	0.88	40.6	7.4	2.56	0.97	43.6	5.7
ENT-MC	0.36	0.24	61.7	22.4	1.44	0.35	54.4	6.4	2.93	0.54	53.6	5.5	4.64	0.94	53.3	6.2
PV	0.19	0.30	48.3	28.3	0.62	0.61	41.8	11.5	1.97	0.98	47.2	8.7	3.21	0.93	47.5	5.2
BALD	0.17	0.29	46.7	27.7	0.56	0.60	41.1	11.4	1.54	0.99	44.2	8.7	2.88	1.12	46.1	6.1
MD	0.19	0.26	48.3	22.9	0.16	0.65	35.0	10.6	0.11	0.75	33.2	5.9	1.05	1.05	37.2	5.4
HUQ-MD	0.18	0.23	48.3	20.3	0.22	0.54	36.1	8.9	0.56	0.80	36.5	6.7	1.86	1.12	40.3	6.3
LOF	0.19	0.15	48.3	13.8	0.29	0.26	35.5	5.5	-0.02	0.50	30.2	4.4	-0.52	0.88	27.6	4.6
ISOF	0.06	0.19	38.3	16.7	0.17	0.38	35.8	6.5	-0.31	0.71	31.0	4.9	-0.53	0.68	31.4	3.0
French, Wikipedia/Vikidia																
SR	0.05	0.05	41.6	4.3	0.25	0.15	41.4	2.8	0.97	0.21	44.2	1.8	2.15	0.26	46.4	1.7
SMP	0.17	0.04	49.4	4.4	**0.92**	0.15	**50.9**	3.0	**1.92**	0.18	**50.5**	1.8	**2.85**	0.29	**49.3**	1.9
ENT	0.05	0.04	41.2	4.2	0.25	0.15	41.4	2.8	0.97	0.21	44.2	1.8	2.15	0.26	46.4	1.7
ENT-MC	0.17	0.04	49.4	4.4	0.92	0.15	50.9	3.0	1.92	0.18	50.5	1.8	2.85	0.29	49.3	1.9
PV	0.16	0.05	50.5	5.3	0.54	0.14	45.6	2.6	1.26	0.25	46.6	2.2	2.14	0.30	47.1	1.8
BALD	0.16	0.05	50.6	4.9	0.46	0.16	44.4	3.0	1.13	0.22	45.9	2.0	1.92	0.35	46.3	2.0
MD	0.08	0.05	43.7	4.8	0.33	0.14	43.0	2.8	0.57	0.20	41.7	1.8	1.33	0.31	43.0	1.7
HUQ-MD	0.05	0.06	41.6	5.2	0.29	0.18	42.2	3.4	0.70	0.21	42.4	1.8	1.75	0.28	44.6	1.7
LOF	**0.21**	0.04	**56.0**	4.0	0.54	0.12	45.1	2.1	0.62	0.19	40.4	1.7	0.77	0.24	39.2	1.2
ISOF	0.19	0.06	53.9	5.5	0.50	0.11	46.0	2.2	0.35	0.15	40.6	1.1	0.27	0.24	39.3	1.3
Spanish, Wikipedia/Vikidia																
SR	0.21	0.10	49.4	10.3	0.89	0.27	46.9	5.7	1.72	0.55	45.7	5.1	2.64	0.65	45.4	3.6
SMP	0.27	0.15	56.5	13.2	0.99	0.40	49.6	7.4	2.09	0.51	49.8	4.9	2.98	0.56	48.1	2.9
ENT	0.21	0.10	49.4	10.3	0.89	0.27	46.9	5.7	1.72	0.55	45.7	5.1	2.64	0.65	45.4	3.6
ENT-MC	0.27	0.15	56.5	13.2	0.99	0.40	49.6	7.4	2.09	0.51	49.8	4.9	2.98	0.56	48.1	2.9
PV	0.22	0.10	50.6	9.6	1.16	0.28	52.1	5.8	2.16	0.36	49.6	3.8	**3.21**	0.50	**48.5**	3.1
BALD	0.26	0.10	54.1	9.4	1.10	0.29	50.9	5.9	2.25	0.45	50.4	4.8	3.19	0.56	48.3	3.2
MD	0.25	0.14	52.9	11.8	0.70	0.24	43.2	5.4	1.58	0.16	44.4	1.8	2.45	0.30	44.2	2.1
HUQ-MD	0.26	0.07	54.1	5.8	0.71	0.18	43.5	4.1	1.49	0.37	43.6	3.2	2.48	0.46	44.4	2.6
LOF	**0.48**	0.13	**75.3**	12.0	**1.70**	0.25	**62.5**	5.3	**2.34**	0.20	**51.4**	2.3	2.48	0.15	44.4	1.6
ISOF	0.44	0.11	71.8	10.8	1.48	0.22	58.0	4.5	1.87	0.38	47.0	4.2	2.08	0.48	42.4	3.4
Spanish, Simplext																
SR	-0.40	0.50	15.0	20.0	-1.79	1.38	15.3	10.3	-0.81	2.62	36.5	9.9	-0.16	6.21	43.9	12.8
SMP	0.14	0.57	50.0	35.4	**0.72**	2.47	**52.9**	30.0	**1.72**	4.06	**52.9**	22.7	**2.78**	6.04	**52.9**	19.8
ENT	-0.40	0.50	15.0	20.0	-1.79	1.38	15.3	10.3	-0.81	2.62	36.5	9.9	-0.16	6.21	43.9	12.8
ENT-MC	0.14	0.57	50.0	35.4	0.72	2.47	52.9	30.0	1.72	4.06	52.9	22.7	2.78	6.04	52.9	19.8
PV	**0.36**	0.55	**65.0**	30.0	0.65	1.84	51.8	12.0	1.37	3.10	48.8	14.0	1.84	4.17	46.3	12.0
BALD	0.22	0.50	55.0	29.2	0.59	1.75	50.6	10.9	1.37	2.94	48.2	13.9	1.81	4.08	44.7	11.8
MD	-0.39	0.49	15.0	20.0	-1.55	1.82	18.8	19.5	-2.61	5.20	27.6	15.8	-3.13	8.01	34.1	11.7
HUQ-MD	-0.40	0.50	15.0	20.0	-1.47	1.93	20.0	21.6	-1.99	4.15	30.6	12.0	-2.28	8.26	37.6	13.5
LOF	-0.16	0.41	20.0	29.2	-1.12	0.90	12.9	12.6	-2.06	1.57	14.1	9.0	-2.78	1.97	15.3	5.7
ISOF	-0.33	0.60	20.0	29.2	-0.46	2.36	34.1	34.6	-1.07	4.25	34.1	31.2	-0.32	5.07	40.0	25.6

Table 11: Change in Macro (ΔF_1, in %) and the percentage of rejected incorrect predictions for the OOD datasets.

The corresponding tables (Tables 10–11) provide the full details, reporting mean values and standard deviations across folds. As before, we highlight the best mean and underline results that are not significantly

different from the best ($p = 0.05$). For example, with an average in-domain accuracy of 81.8% across languages, a random uncertainty estimator would detect 19.2% incorrect predictions. By contrast, rejecting only 1% of the most uncertain predictions with the best methods captures 58.0% of errors on average, peaking at 66.7% for Arabic. This translates into a modest $+0.60$ improvement in F1. At a 10% rejection rate, the overall gain grows to $+3.4$ F1.

The SR method performs reasonably in-domain at very low rejection rates (1–5%), but fails to scale to higher thresholds and degrades sharply in OOD scenarios, where its MC-Dropout equivalents prove more reliable. In our experiments this is especially clear for under-resourced languages (Hindi in-domain, Catalan out-of-domain), where SR becomes unstable and SMP/ENT-MC consistently emerge as the most reliable uncertainty estimation methods. Our results highlight an important contrast: good uncertainty discrimination, as measured by standard metrics, does not always translate into reliable performance gains when using abstention. In particular, methods like LOF and ISOF often score highly in aggregate evaluation, yet their selective prediction behaviour is unstable across languages and domains. Conversely, simpler baselines such as SR and ENT, while less impressive in raw metrics, prove more dependable when applied to real selective prediction. This discrepancy underscores the need to go beyond traditional UE metrics when evaluating practical usefulness.