

Cost-Aware Selective Text Classification with Calibrated Lightweight Models

Saif Alhusseny¹⁾, Omar Alniemi^{2)*}, Alaa Saadoon³⁾, Hanaa F. Mahmood⁴⁾

¹⁾Department of Physics, College of Education for Pure Science, University of Mosul, Mosul, Iraq

^{2,3,4)} Department of Computer Science, College of Education for Pure Science, University of Mosul, Mosul, Iraq

¹⁾saifalhusseny@uomosul.edu.iq, ²⁾omaralniemi@uomosul.edu.iq, ³⁾ alaa.mohsin.saadoon@uomosul.edu.iq,

⁴⁾dr.hanah@uomosul.edu.iq

Submitted : Jul 19, 2026 | Accepted : Aug 24, 2026 | Published : Oct 5, 2026

Abstract: Reliable text classification requires calibrated confidence, abstention for uncertain cases, and explicit computational-cost reporting. This study evaluates cost-aware selective prediction for calibrated lightweight text classifiers on commodity CPU runtimes. The goal is to find a light-weight classifier that is both deployment friendly and achieves a balance between predictive accuracy, calibration, selective reliability, and computational efficiency. The pipeline trains Multinomial Naive Bayes, Logistic Regression, and LightGBM on TF-IDF features, applies post-hoc probability calibration, and accepts only predictions above a validation-selected confidence threshold. Calibrated Logistic Regression achieves the best practical accuracy-calibration-selective-risk-latency-review-budget trade-off in SMS spam, IMDb sentiment and AGNews topic classification experiments. Calibrated Logistic Regression achieves 0.9857 accuracy and 0.0070 ECE on SMS, 0.8718 accuracy and 0.0162 ECE on IMDb, and 0.8998 accuracy on AGNews. On IMDb, Logistic Regression sacrifices about one accuracy point relative to LightGBM but reduces p95 CPU latency from 2546.06 ms to 3.26 ms. Abstaining on the lowest confidence predictions moves accepted risk down from 0.0143 to 0.00457 (SMS) and 0.1282 to 0.1006 (IMDb). Thresholds are chosen on validation/calibration data, then frozen for a single test run. More generally, the results establish calibrated lightweight classifiers as strong and transparent baselines for cost-effective low-latency text classification in real-world deployments.

Keywords: Confidence Threshold; Logistic Regression; Probability Calibration; Selective Prediction; Text Classification.

INTRODUCTION

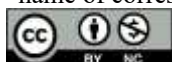
Text classification has continued to be an integral part of modern information systems for spam filtering, content moderation, customer-support routing, news categorization and sentiment monitoring. Such environments restrict model selection based on the accuracy, compute, latency, auditability and cost of deployment. Recent research on resource efficient NLP and model compression has shown that in order to convey practical language systems, one should not consider resource-aware evaluation as an option (Zheng et al., 2025; Dantas et al., 2024).

Because users have the option to check in on their operations (or not), operational decision-making also requires a confidence signal that users can read. So if the classifier only outputs a hard label, then policies such as automatic acceptance, human review or escalation to a slower model cannot be supported either (Steyvers et al., 2025). Note that model competence is often misinterpreted by the user, and occasionally calibration metrics are sensitive to estimator design decisions, binning (histogram discretization), or sample size (Silva Filho et al., 2023).

In this setup, selective prediction provides a clean framework to perform prediction in which the model simply predicts for confident cases and abstains from uncertain ones (Campos et al., 2024). We have recently seen increased study of conformal prediction and abstention in NLP, because uncertainty-aware outputs can mitigate unsafe commitments, hallucination risk, and overconfident automation. However, while many of these studies focus on large models or formal uncertainty sets, production text pipelines often still need simple, low-latency classifiers (Swaminathan et al., 2024).

Thus, a combined deployment-oriented evaluation should incorporate predictive quality, calibration, risk-coverage behavior, validation-based threshold selection and tail latency (Perez et al., 2025). Uncertainty-quantification benchmarks and learning-to-defer work demonstrate that reliability and escalation cost should be

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

evaluated in tandem, but such reporting is still uncommon for lightweight classical text classifiers (Alves et al., 2025).

Thus, this work seeks to design and test a cost-aware selective prediction approach for lightweight text classification that considers predictive performance, calibration quality, selective reliability, computational cost, and CPU latency under realistic deployment conditions.

The main contributions of this study are as follows: A comprehensive, resource-aware evaluation framework for low-complexity text classification that evaluates predictive performance, calibration, risk-coverage trade-offs, latency, validation-selected operating points and robustness, A calibrated TF-IDF selective-prediction pipeline using classical probabilistic classifiers, confidence-threshold abstention, and explicit expected operational cost, Reproducible results on SMS, IMDb and AGNews datasets Comparing the predictive utility, calibration, selective reliability, and CPU cost of Logistic Regression vs Naive Bayes vs LightGBM, Small evidence package that contains the conformal-set comparisons, bootstrap uncertainty for a few metrics, and mild robustness analysis against text noise.

LITERATURE REVIEW

Abstention has a theoretical basis in selective prediction and reject-option classification: we treat it as a way for the model to avoid making predictions if doing so would incur a cost. Classic reject-option studies formalize it as a risk-control problem, and recent surveys further show that rejection is not an error-handling trick, since designing for a reject option affects coverage/workload and impacts end-user trust (Karlsson & Hössjer, 2024; Hendrickx et al., 2024). In a way distinct from theory-oriented analyses, the present work both applies this risk-coverage logic to realistic text pipelines and reports latency & review-budget implications alongside predictive quality.

In contrast to these studies, calibration research asks a complementary question: whether the confidence of a model can be interpreted as the probability that it is right. Several recent statistical comparisons of calibration techniques concluded that although post-hoc methods can provide significant improvements in such probability estimates, they also cautioned against assessing calibration based on a single summary statistic instead of through proper scoring rules. This gives rise for the report not only of accuracy but also ECE, NLL, Brier score and selective risk in an integrated manner (Ojeda et al., 2023; Dimitriadis et al., 2024).

NLP calibration studies show that confidence estimation is task-dependent (Stengel-Eskin & Van Durme, 2023). Notably, studies in semantic parsing and answer-selection also suggest that confidence could inform the final prediction (both detection) as well as whether a particular pseudo-label or generated output or sequence of actions downstream could be trusted (Deng et al., 2023).

Unlike those task-specific approaches, this manuscript is about a benchmark for deployment in broadly used text-classification scenarios, where there is no intention to create a new calibrator but rather evaluate if uncontroversial calibrated baselines still prove robust under exploratory analysis.

Conformal prediction provides stronger distribution-free guarantees under exchangeability and has influenced works in NLP evaluation and set-valued prediction. The motivation for conformalized reject-option classifiers is particularly useful as they link prediction sets with abstention and offer a way to estimate performance at different levels of rejection (Fontana et al., 2023; Johansson & Sönströd, 2026).

This study utilizes confidence-threshold abstention as its primary lightweight mechanism but factors prior contextual work on a split conformal prediction-set comparison to contextualize this proposed pipeline against this more formal domain of uncertainty.

It also follows that cost-aware deferral and robustness under shift are closely related. Learning-to-defer frameworks learn it when to forward a case to experts in high-load settings, and methods that learn reject imply that instead of predicting a label, refusing to predict is more reliable when the test distribution is different than the training (Kwon & Kim, 2026; Tosaki et al., 2025). However, the study focuses more on the practical aspect. Based on validation data, appropriate thresholds are established prior to the testing phase, and the cost-risk balance is then recorded for each dataset.

Recent studies in remote sensing and environmental science emphasize the importance of calibration, particularly when tasks exceed the capabilities of traditional natural language processing applications (Dussert et al., 2024). The focus on calibration increases when automated models are used for decision-making in environmental monitoring and management. Therefore, the idea that integrating calibration probabilities with uncertainty thresholds helps in interpreting automated decisions has been adopted (Zhang et al., 2025). This study considers reliability in classification without neglecting calibration, as it, along with non-reliability, cost, and reproducibility, are all influential factors in performance evaluation.

METHOD

Problem Definition and Selective Prediction

Given a labeled dataset of texts $\{(x_i, y_i)\}_{i=1}^N$, a probabilistic classifier outputs $p(y=k|x)$ over K classes. The predicted label is $\hat{y}(x)=\arg \max_{k \in \{1, \dots, K\}} p(y=k|x)$, and the confidence score is the maximum predicted class probability:

$$c(x) = \max_{k \in \{1, \dots, K\}} p(y = k|x) \quad (1)$$

For binary tasks, $K=2$, for AGNews, $K=4$. A selective classifier accepts a prediction only when $c(x)$ meets or exceeds a threshold $\tau \in [0.5, 1]$. Otherwise, the instance is abstained and routed to a review or fallback process:

$$accept(x, \tau) = \mathbb{1}\{c(x) \geq \tau\} \quad (2)$$

Coverage is the accepted fraction of the workload, and risk is the error rate on the accepted subset:

$$coverage(\tau) = \frac{1}{N} \sum_{i=1}^N accept(x_i, \tau) \quad (3)$$

$$risk(\tau) = \frac{\sum_{i=1}^N \mathbb{1}\{\hat{y}(x_i) \neq y_i\} accept(x_i, \tau)}{\sum_{i=1}^N accept(x_i, \tau)} \quad (4)$$

This definition is relevant to deployment because a system only commits to predictions it accepts. By sweeping τ from high to low, we obtain a risk-coverage curve and the area under this curve (AURC) is a threshold-free measure of selective reliability, where lower AURC values are better since it means fewer errors among low-confidence cases.

Calibration, Models, and Cost-Aware Thresholding

Calibration is evaluated with expected calibration error (ECE) in 15 equal-width bins on $[0, 1]$. Where B_m represents the set of samples in bin m , while $acc(B_m)$ is the empirical accuracy and $conf(B_m)$ be the mean confidence for bin m . ECE is computed as:

$$ECE = \sum_{m=1}^{15} \left(\frac{|B_m|}{N} \right) |acc(B_m) - conf(B_m)| \quad (5)$$

For the representation of texts, we use TF-IDF over unigrams and bigrams. Three different types of probabilistic classifiers are used: Multinomial Naive Bayes (MNB), Logistic Regression (LR) with L2 regularization and LightGBM (LGBM). MNB is a fast generative baseline, LR provides a strong linear discriminative baseline and LGBM serves as a nonlinear ensemble benchmark which can capture interactions but be expensive on sparse high-dimensional inputs.

Model probabilities are calibrated post-hoc for each classifier after applying sigmoid calibration (Platt scaling) with CalibratedClassifierCV using internal three-fold cross-validation. We also ablated isotonic calibration and uncalibrated scores. Calibrator as a post-processing component that is consistent with the deployment practice where calibration can be updated without redesigning the base classifier.

We explicitly detail how an operating-point selection is made by using an expected operational cost as: $EOC(\tau) = C_{err} risk(\tau) + C_{rev} [1 - coverage(\tau)] + C_{lat} p_{95_latency}$. These terms give weights to accepted errors (C_{err}), abstention / manual review (C_{rev}), and tail latency (C_{lat}). Selection of threshold τ only on validation/calibration data & then fixed for one held-out test evaluation. A zero-leakage protocol which removes any opportunity for test-set tuning.

Latency and Reproducibility Protocol

Operational latency is based on wall clock single-instance end-to-end inference timing CPU, which includes the time it takes to perform TF-IDF transformation and probability scoring (If calibration is enabled, the calibrated mapping is included). A warm-up of 50 inferences is used, followed by 300 randomly sampled test instances with seed=42 and batch size 1. We report p50 and p95 latency because tail latency is often more operationally important than the mean.

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

For profiling, a component-wise p95 breakdown over 100 sampled test instances separates TF-IDF transform, model scoring, and calibration mapping. All presented accuracy, ranking, calibration, risk-coverage and latency metrics are obtained using fixed splits and seeds. Key package versions were pinned in the run log, including python=3.11.2, numpy=1.24.0, pandas=2.2.3, scikit_learn=1.4.2, and lightgbm=4.6.0.

Datasets and Experimental Setup

Three public benchmarks are used. SMS Spam consists of 5,572 messages with spam/ham annotations trimmed to two invalid lines, spam accounts for 13.4%, justifying AUROC and AUPRC reporting. IMDb Sentiment employs a fixed 15,000-review sample that is computationally inexpensive for reproducible experimentation on CPU. AGNews is a four-class news-topic benchmark as it constitutes a stratified CPU-friendly subset. Evaluation uses held-out test splits for SMS (4,457 train and 1,115 test instances), IMDb (10,000 train and 5,000 test instances), and AGNews (20,000 train and 4,000 test instances). The top-class probability based ECE is used, for AGNews multi-class calibration, with the use of predicted-class correctness to compute the selective reliability.

Table 1 Datasets and evaluation setup

Dataset	Task type	Train	Test	Primary metrics	Notes
SMS Spam	Binary, imbalanced	4,457	1,115	Acc, F1, AUROC, AUPRC, ECE, AURC	Spam share 13.4%; 5,572 records returned by the core loader
IMDb Sentiment	Binary	10,000	5,000	Acc, F1, AUROC, AUPRC, ECE, AURC	15,000 deterministic sample
AGNews Topic	Multiclass (K=4)	20,000	4,000	Acc, macro-F1, ECE_topprob, AURC	Stratified subset, seed=42

RESULT

Predictive, Calibration and Selective Quality

In Table 2 we look at comparisons of classical baselines on SMS and IMDb, where the three models were trained all under the same pipeline. Logistic Regression is most balanced model on SMS with the best accuracy, F1, AUROC, AUPRC, ECE and AURC. While LightGBM gets the best raw accuracy (type I), its calibration error and operational cost (overhead) are substantially worse, making Logistic Regression more appealing for low-latency calibrated deployment.

Table 2 Classical baseline results on SMS and IMDb

Task	Model	Acc	F1	AUROC	AUPRC	ECE	AURC
SMS	NB	0.9740	0.8922	0.9721	0.9448	0.0140	0.00538
SMS	LogReg	0.9857	0.9452	0.9890	0.9725	0.0070	0.00191
SMS	LightGBM	0.9776	0.9097	0.9747	0.9483	0.0094	0.00471
IMDb	NB	0.8712	0.8697	0.9457	0.9435	0.0188	0.03575
IMDb	LogReg	0.8718	0.8705	0.9470	0.9462	0.0162	0.0416
IMDb	LightGBM	0.8818	0.8812	0.9537	0.9517	0.0408	0.02960

Table 3: The calibrated Logistic Regression pipeline (on the three datasets including AGNews). It retains the central cross-task evidence : the same lightweight family maintains high predictive quality, low calibration error and defined risk reduction through abstention in both binary and multiclass settings.

Table 3 Calibrated Logistic Regression summary across all datasets

Task	Acc	F1 / Macro-F1	ECE	AURC	Risk@0.90	Risk@0.95	N_train	N_test
SMS	0.9857	0.9452	0.0070	0.00191	0.00457	0.00658	4,457	1,115
IMDb	0.8718	0.8705	0.0162	0.0416	0.1006	0.1182	10,000	5,000
AGNews	0.8998	0.8986	0.0493	0.0314	0.0635	0.0812	20,000	4,000

Calibration Ablation and Uncertainty

While for ranking based selective metrics, these improvements are less straightforward due to the fact that they do not necessarily improve monotonically with finer probabilistic reliability information, calibration leads to significant gains in this regard. However this is a critical distinction: calibration is a remapping of probabilities whereas AURC relies heavily on the presorted ordering of confidence values. Consequently, we report calibration results directly rather than assuming an improvement across all downstream metrics.

*name of corresponding author



This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Table 4. Logistic Regression calibration ablation on SMS and IMDB

Task	Setting	Acc	ECE	AURC	Risk@0.90	Risk@0.95
SMS	uncalibrated	0.97130	0.07741	0.00362	0.0103	0.0183
SMS	sigmoid	0.98565	0.00700	0.00191	0.00457	0.00658
SMS	isotonic	0.98565	0.00995	0.00289	0.00443	0.00684
IMDb	uncalibrated	0.88180	0.16932	0.0370	0.0924	0.1094
IMDb	sigmoid	0.87180	0.01623	0.0416	0.1006	0.1182
IMDb	isotonic	0.87180	0.01613	0.0411	0.1009	0.1192

Bootstrap intervals support the stability of the main selective findings. For sigmoid calibration, SMS AURC is 0.00191 [0.00049, 0.00415], Risk@0.90 is 0.00457 [0.00100, 0.00896], and Risk@0.95 is 0.00658 [0.00283, 0.0113]. For IMDb, the corresponding values are 0.0416 [0.0365, 0.0473], 0.1006 [0.0911, 0.1093], and 0.1182 [0.1086, 0.1278]. For AGNews, sigmoid AURC is 0.0314 [0.0259, 0.0377], Risk@0.90 is 0.0635 [0.0553, 0.0717], and Risk@0.95 is 0.0812 [0.0726, 0.0900]. Proper scoring rules show the same reliability trend. On SMS, sigmoid calibration reduces NLL from 0.1372 to 0.0561 and Brier score from 0.0307 to 0.0139. On IMDb, it reduces NLL from 0.4141 to 0.3186 and Brier score from 0.1247 to 0.0973.

Operational Cost

Table 5 reports CPU cost. On both SMS and IMDb, p95 latency with Logistic Regression is in the milliseconds while it becomes orders of magnitude more expensive per sample on sparse inputs with LightGBM. The calibrated Logistic Regression pipeline profile for components shows that calibration mapping can overwhelm tail latency, and especially so for AGNews, thus motivating a deliberate implementation of calibration in latency-critical settings.

Table 5 Operational cost and latency profile

Task	Model / profile	p50 ms	p95 ms	Train s	Cal s	Notes
SMS	NB	2.95	5.08	0.007	0.089	end-to-end calibrated pipeline
SMS	LogReg	2.66	3.90	2.202	1.608	best cost-quality balance
SMS	LightGBM	58.06	187.35	4.056	5.904	high tail latency
IMDb	NB	8.28	9.90	0.042	0.228	end-to-end calibrated pipeline
IMDb	LogReg	1.93	3.26	2.095	6.384	low-latency baseline
IMDb	LightGBM	1880.17	2546.06	253.143	497.647	extreme sparse-input cost
AGNews	LogReg profile	--	11.777	--	--	component p95: TF-IDF 0.733, scoring 2.689, calibration 11.044 ms

This is where the first major operational difference comes in being that LightGBM "beating" Logistic Regressions raw accuracy on IMDb by about a full percentage point at a p95 latency of ~2,546 ms versus ~3.26 ms with several permutations along the way - This trade off would be difficult to justify for real time systems unless that higher-capacity model was only employed for those deferred cases.

Risk–Coverage Behavior, Conformal Sets and Operating Points

In this paper, we argue that risk-coverage curves measure the amount of error avoided by abstaining on the least-confident cases. For SMS risk, full coverage is 0.0143 and at 90% coverage it is reduced to 0.00457 which represents a 68.1% reduction. IMDb risk falls from 0.1282 to 0.1006 (21.5% reduction). A second experiment on AGNews reveals a promising accepted-risk measure at the 90% coverage level (0.0635), which suggests that top-class confidence is still informative for multiclass selective reliability.

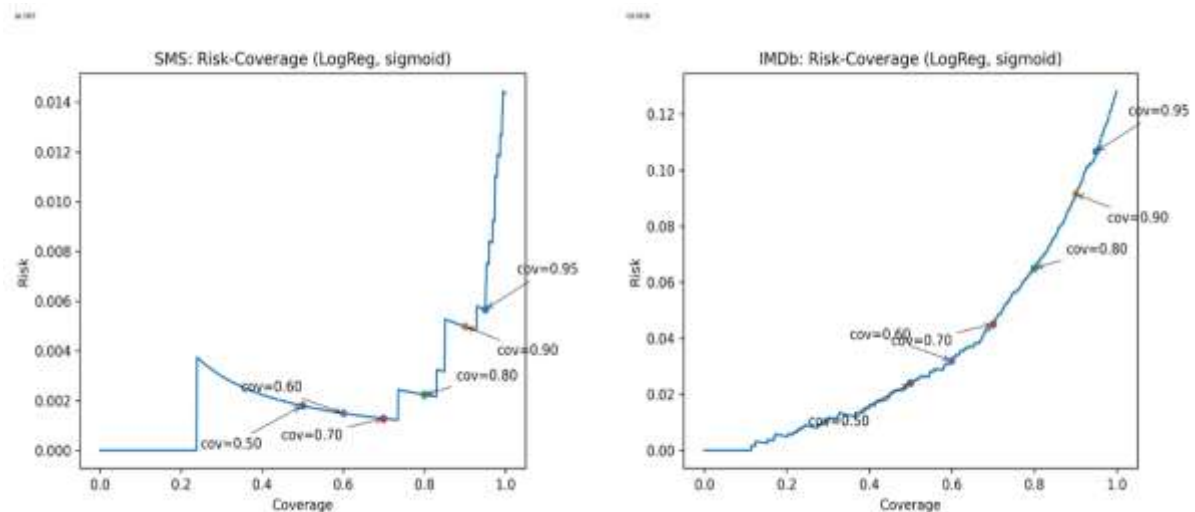


Fig. 1 Risk-coverage curves for calibrated Logistic Regression on SMS and IMDb

Table 6 Risk at practical operating coverages for Logistic Regression

Coverage	SMS Risk	IMDb Risk	SMS reduction vs 100%	IMDb reduction vs 100%
0.25	0.0036	0.0088	74.8%	93.1%
0.50	0.0018	0.0240	87.4%	81.3%
0.75	0.0024	0.0560	83.2%	56.3%
0.90	0.00457	0.1006	68.1%	21.5%
0.95	0.00658	0.1182	54.0%	7.8%
1.00	0.0143	0.1282	0.0%	0.0%

A useful baseline for comparison is set up by using the split conformal prediction-set with $\alpha=0.10$. Singleton-rate/risk pairs for SMS, IMDb and AGNews are 0.904/0.005, 0.956/0.117 and 0.953/0.084 correspondingly for the conformal baseline. However, they differ in terms of uncertainty object output, since conformal sets can output a set of labels rather than 1. Thus these findings are largely consistent with threshold abstention (both methods avoid committed errors by opting-out on uncertain instances).

Table 7 Validation-selected operating points under representative cost scenarios

Task	Calibration	Scenario	C err	C rev	tau val	Cov test	Risk test	EOC test
SMS	sigmoid	error-costly	10	1	0.911	0.948	0.004	0.088
SMS	sigmoid	equal-cost	1	1	0.576	0.996	0.013	0.016
IMDb	sigmoid	error-costly	10	1	0.8972	0.6082	0.0381	0.6238
IMDb	sigmoid	equal-cost	1	1	0.5025	0.9988	0.1262	0.1272
AGNews	sigmoid	error-costly	10	1	0.786	0.749	0.034	0.503
AGNews	sigmoid	equal-cost	1	1	0.501	0.999	0.099	0.100

These operating points will be intuitively seen as validation selection chooses high thresholds and low risk, whilst review being costly also demands for large coverage. Notice that τ must be defined before evaluating the test set.

Cost-quality Frontier and Robustness Analysis

The result is a cost limit, determined by combining response time and high reliability simultaneously. Logistic regression represents the optimal point, combining low response time with risk-hedging behavior. The framework is a two-stage pipeline, where an easy cases were served through fast (but excellently calibrated) classifier, and hard instances to a slow model or human reviewer. Standard Robustness-lite include just a single random deletion and one adjacent-character swap on all of the different instances but with the same test example, corruption method and seed. Only a handful of metrics changed only slightly: $\Delta\text{AURC} \leq 0.0010$, $\Delta\text{Risk}@0.90 \leq 0.0042$, and $\Delta\text{Risk}@0.95 \leq 0.0029$.

*name of corresponding author



This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

DISCUSSIONS

The results show that text classifiers perform strongly across all criteria in terms of prediction accuracy, response speed, calibration, selectivity, and operating cost. The TF-IDF based logistic regression model gave the best balance reliability and reproducibility on all the datasets employed (SMS, IMDb and AGNews). The results are not meant to suggest that traditional models are better than neural models, but that they are still an appropriate choice if interpretability is a key consideration without compromising on speed and accuracy. The results show that accuracy is not the only measure for evaluating the model and other performance and efficiency measures like response time, computational resource consumption, should be taken into account. The LightGBM model performed better with an accuracy score of 0.8818 on IMDb dataset, while the Logistic Regression model's accuracy score was 0.8718. This boost in accuracy came at the expense of a substantial increase in response time: CPU p95 latency went from 3.26 ms to 2546.06 ms. This shows that greater accuracy does not necessarily imply greater performance in practice, particularly when execution time or computational resources grow in proportion to the accuracy. This is particularly important in applications that require rapid or immediate decisions. Thus, a stepwise or sequential design seems to be a better option, in which higher confidence scenarios are covered by fast and well-calibrated models, whereas more expensive models are applied when uncertainties are present.

So the quality of the model needs to be judged from a wider perspective than just accuracy, it needs to be judged on calibration and selective prediction as well. Sigmoid calibration reduces the ECE value on the SMS platform from 0.07741 to 0.00700 and on the IMDb platform from 0.16932 to 0.01623, thus encouraging the possibility of increased reliability. Conversely, abstaining from evaluation leads to a further reduction in acceptable risk; at 90% coverage, the risk decreases from 0.0143 to 0.00457 on the SMS platform and from 0.1282 to 0.1006 on the IMDb platform. All these values demonstrate that it is possible to improve the ease of automation decision-making based on confidence calibration scores, although the benefits vary depending on the task. There are a number of limitations to this study. These include deliberately limiting the robustness analysis, using top probability calibration for AGNews, evaluation on a reduced subset of IMDb, and experiments on public English-language benchmarks. Stronger distribution shifts, multilingual and domain-specific datasets, neural encoders, learned deferral, ensembles, and conformal selective-prediction methods should be assessed in future research.

In general, the results are in favour of deployment-oriented model comparison: the optimal model is not necessarily the one with the highest accuracy, but rather the one that satisfies the intended reliability, latency, review-budget, and reproducibility constraints.

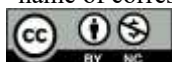
CONCLUSION

The aim of this research is to suggest a cost-aware framework of text classification based on a simple idea: Not all intelligent systems need to be complex in order to be efficient. Thus, the emphasis has been on lightweight models and several criteria apart from accuracy, such as reliability, response time, calibration quality and decision thresholds, have been taken into account. Results indicated that the best overall balance was attained by the calibrated Logistic Regression model that used the TF-IDF features across the three different datasets: SMS, IMDb, and AGNews. The LightGBM model, in turn, achieved superior accuracy on IMDb, at the expense of high time consumption. This highlights that there is not necessarily a direct correlation between accuracy and performance, particularly if the model is slower or demands more resources.

The study also underlined the benefit of making selective predictions in which the model is not compelled to come to a decision every time. If there is high confidence, the prediction is taken as correct, else if there is low confidence, the decision is not made and the situation is moved to a safe procedure like human review or a more complex model. Specific thresholds can be developed using the validation data to help clarify and interpret this process. It demonstrates that these models, properly calibrated and validated to the needs of the actual application, can prove to be a powerful and efficient alternative in the applications where speed, transparency and computational cost are important.

REFERENCES

- Alves, J. V., Leitão, D., Jesus, S., Sampaio, M. O. P., Liébana, J., Saleiro, P., Figueiredo, M. A. T., & Bizarro, P. (2025). A benchmarking framework and dataset for learning to defer in human-AI decision-making. *Scientific Data*, 12, Article 506. <https://doi.org/10.1038/s41597-025-04664-y>
- Campos, M., Farinhas, A., Zerva, C., Figueiredo, M. A. T., & Martins, A. F. T. (2024). Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12, 1497–1516. https://doi.org/10.1162/tacl_a_00715
- Dantas, P. V., da Silva Jr., W. S., Cordeiro, L. C., & Carvalho, C. B. (2024). A comprehensive review of model compression techniques in machine learning. *Applied Intelligence*, 54(22), 11804–11844. <https://doi.org/10.1007/s10489-024-05747-w>



- Deng, W., Pei, J., Ren, Z., Chen, Z., & Ren, P. (2023). Intent-calibrated self-training for answer selection in open-domain dialogues. *Transactions of the Association for Computational Linguistics*, 11, 1232–1249. https://doi.org/10.1162/tacl_a_00599
- Dimitriadis, T., Gneiting, T., Jordan, A. I., & Vogel, P. (2024). Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting*, 40(3), 1101–1122. <https://doi.org/10.1016/j.ijforecast.2023.09.007>
- Dussert, G., Chamaillé-Jammes, S., Dray, S., & Miele, V. (2024). Being confident in confidence scores: Calibration in deep learning models for camera trap image sequences. *Remote Sensing in Ecology and Conservation*, 11(1), 88–99. <https://doi.org/10.1002/rse2.412>
- Fontana, M., Zeni, G., & Vantini, S. (2023). Conformal prediction: A unified review of theory and new challenges. *Bernoulli*, 29(1), 1–23. <https://doi.org/10.3150/21-BEJ1447>
- Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., & Davis, J. (2024). Machine learning with a reject option: A survey. *Machine Learning*, 113(5), 3073–3110. <https://doi.org/10.1007/s10994-024-06534-x>
- Johansson, U., & Sönströd, C. (2026). Conformalized classifiers with reject option. *Machine Learning with Applications*, 23, Article 100838. <https://doi.org/10.1016/j.mlwa.2026.100838>
- Karlsson, M., & Hössjer, O. (2024). Classification under partial reject options. *Journal of Classification*, 41(1), 2–37. <https://doi.org/10.1007/s00357-023-09455-x>
- Kashani Motlagh, N., Davis, J., Anderson, T., & Gwinnup, J. (2025). Naturally constrained reject option classification. *Machine Vision and Applications*, 36, Article 9. <https://doi.org/10.1007/s00138-024-01620-5>
- Kwon, H., & Kim, D.-J. (2026). Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift. *Scientific Reports*, 16, Article 10016. <https://doi.org/10.1038/s41598-026-40637-w>
- Narejo, K. R., Zan, H., Mostafa, S. M., Karim, F. K., Mehmood, F., & Yaseen, A. (2026). Enhancing multimodal sentiment analysis reliability: SentiGuard+ with Dirichlet evidence and selective prediction. *Journal of King Saud University Computer and Information Sciences*, 38, Article 89. <https://doi.org/10.1007/s44443-025-00447-y>
- Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., & Ziegler, A. (2023). Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*, 42(29), 5451–5478. <https://doi.org/10.1002/sim.9921>
- Perez, D. M., Kaiser, B. E., & Boureima, I. (2025). Uncertainty quantification by large language models. *Machine Learning with Applications*, 22, Article 100773. <https://doi.org/10.1016/j.mlwa.2025.100773>
- Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., & Flach, P. (2023). Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9), 3211–3260. <https://doi.org/10.1007/s10994-023-06336-7>
- Stengel-Eskin, E., & Van Durme, B. (2023). Calibrated interpretation: Confidence estimation in semantic parsing. *Transactions of the Association for Computational Linguistics*, 11, 1213–1231. https://doi.org/10.1162/tacl_a_00598
- Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231. <https://doi.org/10.1038/s42256-024-00976-7>
- Swaminathan, A., Lopez, I., Wang, W., Srivastava, U., Tran, E., Bhargava-Shah, A., Wu, J. Y., et al. (2024). Selective prediction for extracting unstructured clinical data. *Journal of the American Medical Informatics Association*, 31(1), 188–197. <https://doi.org/10.1093/jamia/ocad182>
- Tosaki, T., Uchino, E., Kojima, R., Mineharu, Y., Okamoto, Y., Arita, M., Miyai, N., et al. (2025). Out-of-distribution reject option method for dataset shift problem in early disease onset prediction. *Scientific Reports*, 15, Article 19240. <https://doi.org/10.1038/s41598-025-01811-8>
- Zhang, M., Fernández-Torres, M.-Á., Cohrs, K.-H., & Camps-Valls, G. (2025). Calibration and uncertainty quantification for deep learning-based drought detection. *International Journal of Applied Earth Observation and Geoinformation*, 140, Article 104563. <https://doi.org/10.1016/j.jag.2025.104563>
- Zheng, Y., Chen, Y., Qian, B., Shi, X., Shu, Y., & Chen, J. (2025). A review on edge large language models: Design, execution, and applications. *ACM Computing Surveys*, 57(8), Article 209. <https://doi.org/10.1145/3719664>

