When High Accuracy Hides Poor Calibration: Rethinking Confidence Evaluation in Transformer-Based Text Classification with Balanced Brier Score

Guilherme Fonseca[1], Gabriel Prenassi[2], Washington Cunha[3], Leonardo Rocha[2], Marcos André Gonçalves[1]
[1]Federal University of Minas Gerais, [2]Federal University of São João del Rei,
[3]State University of Campinas, Brasil,
{guilhermefonseca, mgoncalv}@dcc.ufmg.br, wcunha@unicamp.br,
prenassigabriel@aluno.ufsj.edu.br, lcrocha@ufsj.edu.br

Abstract

Transformer-based Small (SLMs) and Large Language Models (LLMs) achieve strong effectiveness in text classification (TC), yet deployment requires reliable confidence estimates. Although miscalibration in Transformers has been reported, evidence for TC under fine-tuning remains limited. We evaluate the calibration of fine-tuned SLMs and LLMs against Logistic Regression, a classical, well-calibrated baseline, and find that, despite superior effectiveness, Transformers remain markedly overconfident. Crucially, we show that widely used calibration metrics, such as Expected Calibration Error and Brier Score, become biased in high-effectiveness regimes, where the dominance of correct predictions masks severe miscalibration on errors, sometimes even suggesting better calibration than Logistic Regression, a well-known calibrated method. To address this limitation, we propose the Balanced Brier Score (BBS), which balances the contribution of correct and incorrect predictions within confidence bins. BBS reveals substantially poorer calibration in both SLMs and LLMs, consistent with qualitative evidence from calibration curves. These findings challenge current calibration assessment practices and provide a more reliable alternative for evaluating confidence quality, particularly in high-effectiveness regimes where miscalibration may otherwise be underestimated.

1 Introduction

Transformer-based models achieve remarkable effectiveness in text classification, but their confidence estimates remain systematically unreliable. As a fundamental NLP task, Text Classification (TC) underpins applications such as sentiment analysis (Ribeiro et al., 2016), text categorization (Cunha et al., 2025a), and hate speech detection (Davidson et al., 2017). Small Language Models (SLMs), such as BERT and RoBERTa, and Large Language Models (LLMs), such as Llama and GPT, have substantially advanced TC effectiveness, with fine-tuning consistently emerging as the strongest paradigm over zero-shot and in-context learning (ICL) (Cunha et al., 2025a; Bucher and Martini, 2024). More critically, we show that widely used calibration metrics can contradict qualitative evidence, suggesting good calibration even when models are clearly overconfident.

Beyond high effectiveness, a reliable TC system must also be **well calibrated**, meaning its predicted probabilities should align with the true likelihoods of correctness (Guo et al., 2017; Brier, 1950). Calibration is essential in decision-making scenarios, particularly in high-stakes contexts (Kumar et al., 2019). Even highly accurate models can be unsafe if their confidence estimates are misleading: in healthcare, miscalibration may lead to inappropriate triage decisions, while in credit assessment, it can result in unjust loan denials.

Recent work has increasingly exposed calibration limitations in Transformer-based models. Studies in Question Answering (Tian et al., 2023) and relation extraction (de Oliveira et al., 2025) show that these models are often poorly calibrated, typically displaying strong overconfidence. However, in the context of TC, most prior studies concentrate on zero-shot or ICL scenarios (Li et al., 2025), offering limited insight into calibration behavior when models are fine-tuned.

This leads to a crucial blind spot: current literature rarely examines calibration in extremely high-effectiveness regimes, typically attained after **fine-tuning**. This omission is consequential, as tuning remains the pathway to **maximum** effectiveness, precisely where calibration matters most. Moreover, even without tuning, LLMs already operate in high-effectiveness regimes in TC tasks, making the calibration question practically unavoidable. Importantly, SLMs **require** fine-tuning to reach peak performance and often rival or surpass LLMs in TC (Cunha et al., 2025a;

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 45888–45900

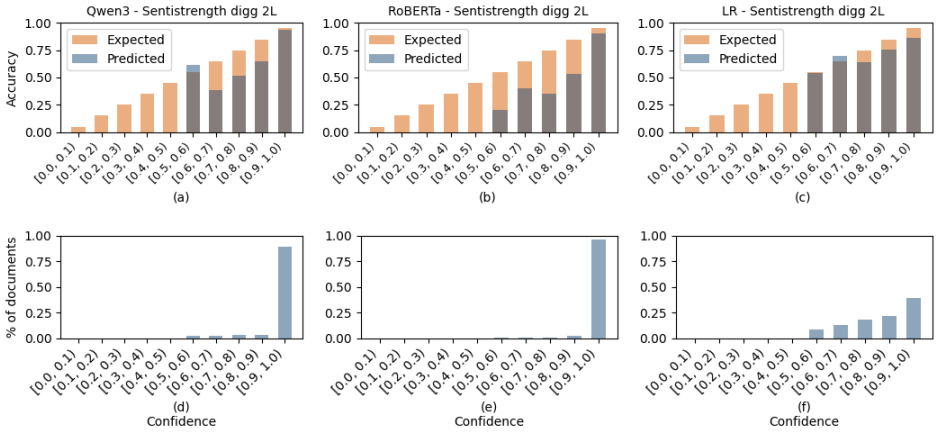

Figure 1: Calibration curves (a, b, c) and confidence histograms (d, e, f) for Qwen3, RoBERTa, and LR on the Sentistrength-Digg2L dataset. In the calibration curves, "Expected" denotes the ideal calibration line, where confidence equals empirical accuracy, and "Predicted" corresponds to the observed empirical accuracy within each confidence bin. A well-calibrated model should have both lines closely aligned. Calibration curves reveal that while LR closely follows the ideal diagonal, both Qwen3 and RoBERTa systematically deviate from it, indicating pervasive overconfidence (RQ1). Confidence histograms show that Transformer-based models concentrate virtually all predictions in the highest-confidence bin (> 0.9), explaining why traditional metrics are dominated by well-calibrated correct predictions and fail to expose miscalibration on errors (RQ2).

Bucher and Martini, 2024), meaning tuned models represent the predominant deployment scenario. Consequently, failing to analyze calibration in such settings leaves a significant gap in our understanding of how these widely deployed models behave when practitioners rely on their predictions.

Accordingly, our first contribution is a comprehensive analysis of the calibration behavior of **fine-tuned** Transformer-based TC models. We begin with **RQ1**: *How do tuned Transformers (SLMs and LLMs) trade off effectiveness and calibration in TC?* To answer this, we evaluate tuned LLMs – Llama3.1 (Dubey et al., 2024), Qwen3 (Yang et al., 2025), Gemma (Team et al., 2024), and Mistral (Jiang et al., 2023) – and compare them against RoBERTa (Liu et al., 2019), a strong SLM baseline (Cunha et al., 2025b), and Logistic Regression (LR), a widely recognized calibrated model (Cunha et al., 2023a). While tuned LLMs consistently achieve superior Macro-F1, our calibration analyses expose a clear and concerning pattern: despite their high effectiveness, **Transformer-based models remain poorly calibrated**.

Figure 1(a–c) makes this explicit. Qwen3 and RoBERTa diverge strongly from ideal calibration across confidence ranges, whereas LR remains closely aligned with expected accuracy, as a truly calibrated model should (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005).

This exposes a deeper question: *Is this miscalibration intrinsic, or are current calibration metrics failing in high-effectiveness regimes?* We formalize this as **RQ2**: *Are traditional calibration metrics reliable for highly effective TC models?* Surprisingly, although calibration curves clearly reveal overconfidence, Expected Calibration Error (ECE) and Brier Score (BS) often suggest excellent calibration, sometimes even surpassing LR. This contradiction demonstrates that widely used calibration metrics can contradict qualitative evidence, suggesting good calibration even when models are clearly overconfident.

A detailed decomposition explains why. Transformers tend to be reasonably calibrated for correct predictions but severely miscalibrated for errors, which still receive extremely high confidence. Since more than 90% of predictions often lie above 0.9 confidence, incorrect predictions become both rare and dangerously overconfident. As a consequence, global metrics become dominated by correct predictions, masking true calibration weaknesses. When we exclude ultra-high-confidence predictions, calibration scores deteriorate dramatically, confirming that traditional metrics systematically obscure risk.

Grounded in this limitation, we introduce our second main contribution: the **Balanced Brier Score (BBS)** – a calibration metric designed to remain informative in high-effectiveness regimes. BBS balances contributions from correct and incorrect predictions within each confidence bin, preventing dominance bias and producing substantially more faithful calibration assessments under extreme accuracy and confidence concentration.

This motivates **RQ3**: *How does BBS reassess the calibration of Transformer-based TC models?*

Our results demonstrate that BBS consistently reveals substantially poorer calibration than traditional metrics suggest, while strongly aligning with qualitative evidence from curves and prediction distributions. This confirms that BBS provides a more reliable and practically meaningful characterization of calibration in Transformer-based TC.

In summary, we make two key contributions: (1) the first comprehensive calibration study of **fine-tuned** Transformer-based TC models, showing that widely used metrics can significantly misrepresent model behavior in high-effectiveness regimes; and (2) the proposal of the **Balanced Brier Score (BBS)**, a metric designed explicitly for such regimes, offering a more faithful and actionable assessment of calibration for both SLMs and LLMs.

2 Related Work

Research on neural calibration has evolved alongside the shift from traditional classifiers to deep learning. Seminal work by (Guo et al., 2017) showed that, unlike classical models, Deep Neural Networks (DNNs) are often miscalibrated despite strong effectiveness. They linked miscalibration to architectural and training factors (depth, width, weight decay, Batch Normalization) and demonstrated that post-hoc methods such as Temperature Scaling (Platt et al., 1999) can mitigate overconfidence. Although (Guo et al., 2017) includes TC experiments, it predates Transformer architectures (Vaswani et al., 2017), which later reshaped TC (Cunha et al., 2025b).

More recent work examines calibration in pretrained SLMs. Studies such as (Desai and Durrett, 2020) and (Xiao et al., 2022) evaluate BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) across in-domain and out-of-domain NLP tasks. Both report substantially poorer out-of-domain calibration and reaffirm the usefulness of Temperature Scaling, consistent with (Guo et al., 2017). In Section 6.4, we analyze how our proposed metric behaves under Temperature Scaling and find that, although it improves calibration on some datasets, SLMs and LLMs remain poorly calibrated overall.

Parallel efforts investigate LLM calibration, particularly for generation and reasoning. Surveys such as (Shorinwa et al., 2025) and (Liu et al., 2025) summarize challenges and overconfidence risks, while works like (Lyu et al., 2025) and (Ulmer et al., 2024) explore alternative confidence estimation strategies. In TC, (Li et al., 2025) studies low-resource ICL settings, showing systematic miscalibration and proposing self-ensembling.

Our work differs in two key ways. First, unlike prior studies focusing on zero-shot, ICL, or out-of-domain scenarios, we analyze calibration in *fine-tuned* Transformer-based TC models – the high-effectiveness regime in which they are commonly deployed. Second, whereas prior work relies on standard calibration metrics, we show that such metrics can be misleading in these regimes, motivating our **Balanced Brier Score (BBS)**, which remains informative when predictions are highly confident and overwhelmingly correct.

3 Traditional Calibration Metrics

In this section, we review widely adopted traditional calibration metrics. We begin by formally defining TC and model calibration. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the test set, where x_i is an input text and $y_i \in \{1, \ldots, K\}$ its true label. A classifier f maps x_i to a probability vector $\hat{p}_i$, with predicted label $\hat{y}_i = \arg\max_k \hat{p}_i$ and associated confidence $\hat{c}_i = \max_k \hat{p}_i$. A model is **perfectly calibrated** if, for any confidence level $c \in [0, 1]$, the empirical probability of correctness matches c, i.e., $\mathbb{P}(\hat{y} = y \mid \hat{c} = c) = c$.

To quantify model calibration, two metrics are widely used in the literature (Guo et al., 2017; Ulmer et al., 2024): the Brier Score (BS) and the Expected Calibration Error (ECE).

Brier Score (BS) (Brier, 1950): It measures the quality of probabilistic predictions by computing the Mean Squared Error (MSE) between the predicted probability distribution and the true label (encoded as a one-hot vector). It is defined as:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} (\hat{p}_{i,k} - o_{i,k})^2 \qquad (1)$$

where $\hat{p}_{i,k}$ is the predicted probability for class k and $o_{i,k}$ is an indicator equal to 1 if $y_i = k$ and 0 otherwise. BS ranges from $[0, 2]$, with lower values indicating better calibration.

Expected Calibration Error (ECE) (Naeini et al., 2015): It partitions predictions into M bins based on confidence $\hat{c}$. Let B_m be the indices of samples in bin m. ECE is computed as the weighted average of the absolute gap between bin accuracy $\text{acc}(B_m)$ and mean bin confidence $\text{conf}(B_m)$:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \qquad (2)$$

where N is the total number of samples. Following prior work (Li et al., 2025), we use $M = 10$. ECE ranges from $[0, 1]$, and, as in BS, lower values indicate better calibration.

Other works propose variants of ECE, such as the three metrics presented in (Nixon et al., 2019): SCE, which considers all class probabilities rather than only the maximum confidence; ACE, which modifies the bin separation strategy by using equal-mass adaptive bins; and TACE, which adds a minimum confidence threshold for instances. There is also the cw-ECE metric, proposed by (Kull et al., 2019), which, similarly to SCE, computes calibration for each class individually. Although these variants introduce meaningful modifications relative to the original metric, they all share ECE's limitation in high-effectiveness regimes (as detailed in Section 6): the aggregation of predictions without distinguishing between correct and incorrect ones. Consequently, in models with high accuracy and elevated confidence, correct predictions dominate the computation, diluting the contribution of overconfident errors. For this reason, we chose to use only ECE: besides being the most widely adopted metric in the literature (Chidambaram et al., 2024), it is already sufficient to highlight the problem we analyze, whereas using its variants would not substantially change this conclusion.

4 Balanced Brier Score (BBS) Proposal

As we shall see in Section 6.2, in TC scenarios where Transformer-based models (SLMs and LLMs) achieve very high effectiveness, correct predictions overwhelmingly dominate the evaluation space. In this regime, traditional calibration metrics such as ECE and BS become biased: they are influenced by many highly confident correct predictions and therefore suggest that models are better calibrated than they actually are.

More specifically, we observe that correct predictions typically present high confidence and low calibration error (with BS values close to 0). Conversely, incorrect predictions are also made with very high confidence, but incur substantially higher error (often BS > 1). However, because incorrect cases are few relative to the total, their contribution to the global Brier Score and to ECE is heavily diluted. As a result, the final metric values appear artificially low, masking systematic overconfidence on errors – precisely the type of miscalibration most harmful in real applications.

To address this limitation, we propose the **Balanced Brier Score (BBS)**, a calibration metric designed to remain reliable in high-effectiveness, high-confidence regimes. The key idea is to ensure that the contribution of incorrect predictions is not overwhelmed by the abundance of correct ones. Instead of averaging over all samples, BBS computes the Brier Score on a *balanced* evaluation set where correct and incorrect predictions contribute symmetrically within each confidence region.

The **BBS** computation proceeds as follows in six steps: **(1) Binning:** First, given a set of N instances $\mathcal{D} = \{(y_i, p_i)\}_{i=1}^{N}$, where $y_i \in \{1, \ldots, K\}$ is the ground truth label and p_i is the probability vector assigned by a classifier f, the predictions are partitioned into M confidence *bins*. This grouping is based on the prediction confidence $\hat{c}_i = \max p_i$, following a process similar to the one used in ECE[1]. **(2) Splitting:** For each *bin* B_m, split the samples into two groups: correct predictions (C_m, where $\hat{y}_i = y_i$) and incorrect predictions (I_m, where $\hat{y}_i \neq y_i$). **(3) Balancing Size:** For each bin, we define the balancing size k_m as the minimum between the number of correct predictions and the number of incorrect predictions: $k_m = \min(|C_m|, |I_m|)$. **(4) Bin Balancing:** Apply random undersampling to the majority group in B_m, such that both groups contain exactly k_m instances. **(5) Aggregation:** Let B'_m be the resulting balanced subset for *bin* m, and let $\mathcal{D}' = \bigcup_{m=1}^{M} B'_m$ be the union of all balanced *bins*. **(6) Calculation:** Compute the standard Brier Score over $\mathcal{D}'$, resulting in the Balanced Brier Score (BBS). Due to this final step, BBS ranges from [0, 2], with lower values indicating better calibration.

Algorithm 1: BBS Metric

Input: true label and predictions $\mathcal{D} = \{(y_i, p_i)\}_{i=1}^{N}$, Number of bins M
Output: BBS value

$\hat{c} \leftarrow \{\max p_i\}$;
$B \leftarrow$ Divide $\mathcal{D}$ into M bins based on confidence $\hat{c}$;
$\mathcal{D}_{balanced} \leftarrow \emptyset$;
for *each bin* $B_m \in B$ **do**
    // Correct predictions set
    $C_m \leftarrow \{i \in B_m \mid \hat{y}_i = y_i\}$;
    // Incorrect predictions set
    $I_m \leftarrow \{i \in B_m \mid \hat{y}_i \neq y_i\}$;
    $k_m \leftarrow \min(|C_m|, |I_m|)$;
    // Random sampling from majority set
    $B'_m \leftarrow$ randomSampler(C_m, I_m, k_m);
    $\mathcal{D}_{balanced} \leftarrow \mathcal{D}_{balanced} \cup B'_m$;
end
$BBS \leftarrow BrierScore(\mathcal{D}_{balanced})$;
return BBS;

[1]Following the ECE metric, we set the $M = 10$.

By construction, **BBS** prevents the dominance of correct predictions from suppressing the contribution of high-confidence errors. It penalizes overconfident mistakes with the same magnitude that it rewards confident correct predictions, thereby producing a calibration estimate that remains meaningful even when model effectiveness is extremely high, as we shall see in our experiments. In other words, while traditional metrics converge toward optimistic values in such regimes, BBS continues to reveal miscalibration that is consequential in practice. Algorithm 1 summarizes the calculation procedure.

5 Experimental Setup

5.1 Datasets

We evaluate our approach on 6 widely used text datasets covering different domains, sizes, and label configurations. Three datasets are binary, while the others are multi-class. Table 1 summarizes dataset statistics, including the number of instances and number of classes for each collection.

dataset	Size	# Classes
sentistrength **digg** 2L	782	2
sentistrength **myspace** 2L	834	2
sentistrength **bbc** 2L	752	2
ohsumed	18,302	23
trec	5,952	6
twitter	6,997	6

Table 1: Dataset statistics used in the experiments, including number of instances and classes.

5.2 Classification Models and Fine-Tuning

dataset	llama3.1	gemma	Mistral	qwen3	RoBERTa	LR
digg	84.8(4.2)•	86.7(3.7)•	81.3(10.4)•	86.8(3.6)▲	85.8(5.1)•	63.2(7.3)
myspace	89.5(3.7)•	90.4(3.8)▲	89.5(2.8)•	88.9(4.7)•	81.1(6.2)	59.0(4.3)
bbc	74.1(5.7)	75.6(6.4)•	74.5(6.4)•	80.3(4.3)▲	77.9(6.3)•	48.6(2.6)
ohsumed	83.2(0.8)•	83.2(0.7)•	83.1(0.7)•	84.0(1.0)▲	77.5(1.2)	71.4(1.2)
trec	95.7(1.1)•	96.1(0.8)▲	95.7(1.2)•	95.8(0.9)•	95.4(0.8)•	68.0(1.9)
twitter	78.9(1.8)•	79.6(1.7)▲	78.4(1.8)•	79.2(1.4)•	77.8(2.3)•	60.8(1.6)

Table 2: MacroF1. Numbers in () represent 95% CIs. "▲" indicates the best dataset result, "•" indicates a statistical tie, and unmarked, a statistical loss compared to the best result.

As the SLM representative, we adopt **RoBERTa** (Liu et al., 2019), a strong and well-established baseline for TC (Cunha et al., 2025a; Zanotto et al., 2021; Fonseca et al., 2025). For LLMs, we evaluate four open-source models from distinct families: **Llama3.1** (Llama-3.1-8B) (Dubey et al., 2024), **Gemma** (gemma-7B) (Team et al., 2024), **Mistral** (Mistral-7B-v0.1) (Jiang et al., 2023), and **Qwen3** (Qwen3-8B) (Yang et al., 2025), all of which have been widely adopted in recent NLP research (Reis et al., 2024; Fonseca et al., 2025; Li et al., 2025). To provide a classical calibrated reference, we also include **Logistic Regression (LR)** (Wright, 1995), a model consistently shown to exhibit strong calibration properties in TC (Cunha et al., 2023a).

All LMs are **fine-tuned** on each dataset to adapt pre-trained representations to task- and domain-specific distributions. Fine-tuning learns a task-specific fully connected classification head on top of the `[CLS]` representation, enabling supervised TC prediction. For LR, we use TF-IDF features as input. Before constructing the TF-IDF matrix, we remove stopwords and retain features that appear in at least two documents. We then apply L2 normalization to the resulting TF-IDF vectors (Cunha et al., 2021).

5.3 Metrics and Experimental Protocol

We evaluate systems from two complementary perspectives: (1) **classification effectiveness**, and (2) **model calibration**. Effectiveness is measured using **Macro-F1**, which is appropriate in the presence of class imbalance. Calibration is assessed using calibration curves (DeGroot and Fienberg, 1983; Niculescu-Mizil and Caruana, 2005), Expected Calibration Error (**ECE**) (Naeini et al., 2015), Brier Score (**BS**) (Brier, 1950), and our proposed **Balanced Brier Score (BBS)**. Experiments were conducted on AWS using a **G6e.2xlarge** instance. Datasets were evaluated under **10-fold cross-validation**, and statistical comparisons were performed using a paired **t-test with Bonferroni correction** (Cunha et al., 2023b) to ensure robust significance assessment.

6 Experimental Results

6.1 RQ1: Effectiveness and Calibration

To investigate **RQ1**, we jointly evaluate effectiveness (Macro-F1) and calibration using calibration curves. Table 2 reports Macro-F1 scores. As indicated by "▲" (best), "•" (statistically tied), and unmarked (inferior) cells, results reveal a clear trend: Transformer-based models consistently outperform traditional classifiers, confirming prior evidence (Fonseca et al., 2025; Cunha et al., 2025b). Logistic Regression (LR) never statistically matches SLMs or LLMs, probably due to the use of the TF-IDF purely frequentist representation. All Transformer-based models exceed 74% Macro-F1, reaching 96.1% (Gemma-7B on TREC), reinforcing their strong effectiveness. Gemma and Qwen3 are the most successful overall, each leading across three datasets, whereas

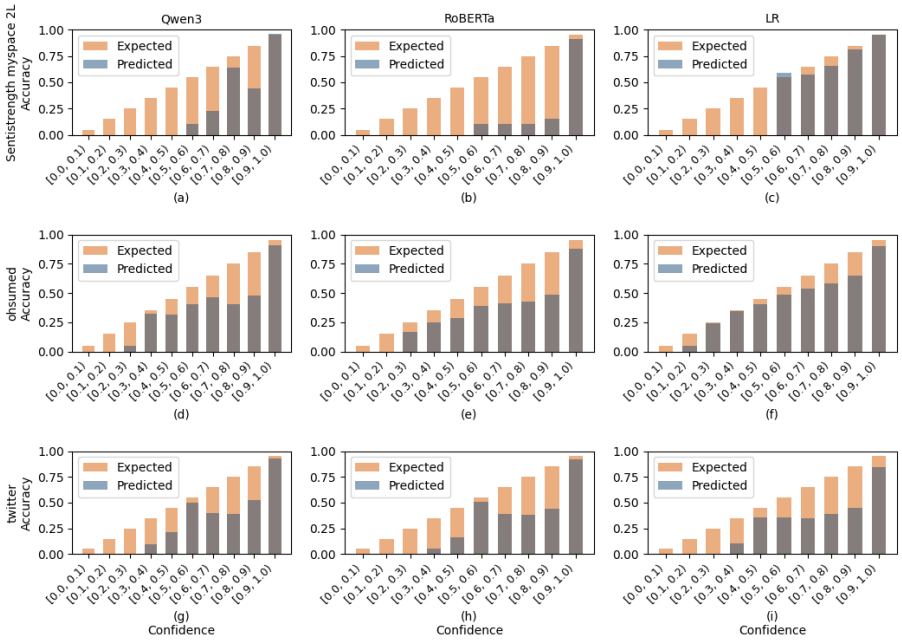

Figure 2: Calibration curves for Qwen3, RoBERTa, and LR on the SentistrengthMyspace2L, Ohsumed, and Twitter datasets. Consistent with Figure 1, LR maintains close alignment with the diagonal across datasets, deviating from the ideal pattern primarily on the Twitter dataset, while Transformer-based models systematically exhibit overconfidence.

dataset	Llama3.1	Gemma	Mistral	Qwen3	RoBERTa	LR
			BS			
Sentistrength digg 2L	0.19(0.05)•	0.18(0.05)•	0.21(0.07)	0.16(0.05)▲	0.21(0.07)	0.35(0.06)
Sentistrength myspace 2L	0.09(0.03)•	0.09(0.03)▲	0.10(0.02)•	0.10(0.03)•	0.17(0.05)	0.22(0.02)
Sentistrength bbc 2L	0.16(0.03)•	0.17(0.05)•	0.17(0.04)	0.13(0.03)▲	0.18(0.06)•	0.23(0.02)
ohsumed	0.22(0.01)▲	0.23(0.01)•	0.24(0.01)	0.23(0.01)•	0.31(0.01)	0.36(0.01)
trec	0.07(0.01)•	0.06(0.01)▲	0.07(0.01)•	0.06(0.01)•	0.07(0.02)•	0.49(0.01)
twitter	0.17(0.01)•	0.17(0.01)▲	0.17(0.01)•	0.17(0.01)•	0.20(0.02)	0.37(0.02)
			ECE			
Sentistrength digg 2L	0.09(0.02)•	0.09(0.02)•	0.10(0.03)	0.08(0.02)▲	0.10(0.03)	0.11(0.04)•
Sentistrength myspace 2L	0.05(0.01)•	0.04(0.02)▲	0.05(0.01)•	0.05(0.02)•	0.09(0.03)	0.06(0.02)•
Sentistrength bbc 2L	0.08(0.01)•	0.09(0.02)•	0.09(0.02)	0.07(0.01)▲	0.09(0.03)•	0.08(0.03)•
ohsumed	0.09(0.00)▲	0.10(0.01)•	0.11(0.01)	0.10(0.01)	0.13(0.00)	0.10(0.02)•
trec	0.03(0.01)•	0.03(0.01)▲	0.03(0.00)•	0.03(0.00)•	0.04(0.01)•	0.12(0.01)
twitter	0.07(0.01)▲	0.07(0.01)•	0.08(0.01)	0.08(0.01)	0.09(0.01)	0.16(0.01)
			BBS			
Sentistrength digg 2L	0.67(0.15)▲	0.84(0.07)•	0.82(0.04)•	0.73(0.18)•	0.89(0.10)	0.68(0.04)•
Sentistrength myspace 2L	0.71(0.14)•	0.79(0.12)	0.82(0.09)	0.79(0.12)	0.95(0.06)	0.62(0.13)▲
Sentistrength bbc 2L	0.70(0.09)•	0.84(0.05)	0.88(0.08)	0.77(0.08)•	0.92(0.06)	0.66(0.14)▲
ohsumed	0.87(0.01)	0.88(0.01)	0.91(0.00)	0.91(0.01)	0.91(0.01)	0.79(0.02)▲
trec	0.88(0.02)	0.90(0.03)	0.89(0.03)	0.89(0.04)	0.94(0.02)	0.68(0.01)▲
twitter	0.87(0.01)▲	0.90(0.01)	0.91(0.01)	0.92(0.01)	0.92(0.02)	0.89(0.01)

Table 3: BS, ECE, and BBS results for all models. While traditional metrics (BS and ECE) consistently suggest that Transformer-based models are better calibrated than LR, BBS reverses this ranking in most datasets, revealing that LR is in fact the best-calibrated model. This contradiction exposes the dominance bias of traditional metrics in high-effectiveness regimes (RQ2) and demonstrates that BBS provides a more faithful calibration assessment (RQ3). Numbers in parentheses represent 95% confidence intervals. "▲" indicates the best dataset result, "•" indicates a statistical tie, and unmarked, a statistical loss compared to the best result.

Mistral remains statistically competitive across all scenarios. Even RoBERTa and Llama, which underperform statistically in only two and one datasets, still substantially outperform traditional baselines. In short, **Transformers establish a remarkably high standard of effectiveness**.

Calibration results, however, tell a different story. Figures 1 and 2 present representative calibration curves for Qwen, RoBERTa, and LR across Sentistrength Digg 2L (Figure 1) and Sentistrength Myspace 2L, Ohsumed, and Twitter (Figure 2); full results appear in Appendix A. Ideally, empirical accuracy ("Predicted") should match expected accuracy ("Expected") across confidence bins. LR closely approximates this behavior in Sentistrength Digg 2L and Myspace 2L, deviating mainly in Twitter, reinforcing prior findings that LR is generally very well calibrated (Cunha et al., 2023a).

In stark contrast, **SLMs and LLMs systematically deviate from ideal calibration**. Except

dataset	qwen3-8B			roberta			LR		
	BS	BS Correct	BS Error	BS	BS Correct	BS Error	BS	BS Correct	BS Error
Sentistrength digg 2L	0.16(0.05)	0.01(0.00)	1.49(0.19)	0.21(0.07)	0.00(0.00)	1.87(0.07)	0.35(0.06)	0.09(0.01)	1.22(0.13)
Sentistrength myspace 2L	0.10(0.03)	0.01(0.00)	1.67(0.10)	0.17(0.05)	0.00(0.00)	1.94(0.04)	0.22(0.02)	0.04(0.01)	1.25(0.09)
Sentistrength bbc 2L	0.13(0.03)	0.01(0.00)	1.57(0.13)	0.18(0.06)	0.00(0.00)	1.90(0.10)	0.23(0.02)	0.03(0.01)	1.50(0.09)
ohsumed	0.23(0.01)	0.01(0.00)	1.73(0.02)	0.31(0.01)	0.01(0.00)	1.69(0.03)	0.36(0.01)	0.05(0.01)	1.37(0.05)
trec	0.06(0.01)	0.00(0.00)	1.71(0.11)	0.07(0.02)	0.00(0.00)	1.83(0.05)	0.49(0.01)	0.26(0.01)	0.95(0.02)
twitter	0.17(0.01)	0.00(0.00)	1.73(0.05)	0.20(0.02)	0.00(0.00)	1.74(0.03)	0.37(0.02)	0.01(0.00)	1.63(0.04)

Table 4: BS, BS Correct, and BS Error results. BS Correct refers to the BS calculated only on correctly predicted instances; BS Error refers to the BS calculated only on misclassified instances. Numbers in parentheses are 95% confidence intervals. Transformer-based models show near-perfect calibration on correct predictions but overconfidence on errors, whereas LR exhibits a smaller and more balanced gap between the two.

for the highest-confidence bin $[0.9, 1.0)$, observed accuracy is consistently lower than predicted confidence, revealing pervasive overconfidence. This is especially evident for RoBERTa in Sentistrength Myspace 2L (Figure 2(b)), where predictions collapse almost entirely into the highest-confidence bin while still making errors. Moreover, calibration gaps tend to be larger for SLMs than for LLMs, indicating even weaker reliability despite strong effectiveness.

In sum, answering **RQ1**, Transformer-based models deliver excellent effectiveness but poor calibration, creating a clear trade-off: despite achieving SOTA Macro-F1, their probability estimates are unreliable compared to traditional models such as LR. **This misalignment between confidence and correctness exposes a critical reliability limitation of modern Transformer-based TC systems**.

6.2 RQ2: Traditional Calibration Metrics

While calibration curves clearly reveal overconfidence, we now examine whether standard metrics capture this behavior. Table 3 reports ECE, BS, and BBS for all classifiers (BBS analyzed in Section 6.3). When examining Brier Score (recall that for all calibration metrics, the lower the value, the more calibrated), results initially appear consistent with earlier findings: **LLMs seem more calibrated than SLMs**, as Llama3.1, Gemma, Mistral, and Qwen3 consistently achieve lower BS values than RoBERTa. However, a more surprising and concerning pattern emerges. Contrary to the calibration curves in Section 6.1, **BS suggests that LR is actually worse calibrated than all Transformer-based models**, despite visually appearing the most reliable one. This contradiction becomes evident in datasets such as Sentistrength Myspace 2L (Figure 2(a–c)), where LR closely follows the ideal curve while Transformers clearly do not – yet BS still favors them. A similar mismatch is observed with ECE: Transformer models achieve excellent ECE scores, with LLMs obtaining the best calibration results across all datasets, outperforming LR or remaining statistically equivalent to it in every dataset. Despite matching the best results in only 2 datasets, SLMs achieve extremely low ECE values, with the worst recorded value being 0.13 (ohsumed dataset), which can still be considered very good calibration. Taken together, ECE and BS convey a misleading message that **highly effective Transformers are well calibrated**, in direct contradiction with calibration curve evidence, strongly suggesting that **traditional metrics systematically mask severe overconfidence in high-effectiveness regimes**.

To explain this discrepancy, we decompose BS into contributions from correct (BS Correct) and incorrect (BS Error) predictions. Results in Table 4[2] reveal a consistent pattern: Transformer-based models show **near-perfect calibration on correct predictions** (BS between 0.0 and 0.1), but calibration collapses on errors, where BS often exceeds 1.49 and reaches 1.94[3], indicating **extreme overconfidence when wrong**. In other words, they are confidently right, but also confidently wrong. LR behaves more evenly: although still poor on errors, its BS never reaches such extremes, and the gap between BS Correct and BS Error is notably smaller.

Confidence histograms in Figures 1 (d–f) and 3[4] reinforce this interpretation. Transformer probabilities concentrate almost entirely in the highest-confidence bin (> 0.9). Thus, BS and ECE appear excellent because most predictions are both confident and correct. However, when errors occur, they do so with very high confidence, a behavior that reveals severe overconfidence and poor calibration.

Table 5 reports ECE and BS after removing predictions with confidence exceeding 0.9. In this scenario, Transformer models exhibit substantially worse calibration than suggested by traditional

[2]For space reasons, we report Qwen3, RoBERTa, and LR. Full results appear in Appendix B.

[3]BS ranges in [0,2] (Brier, 1950).

[4]Full histograms in Appendix A

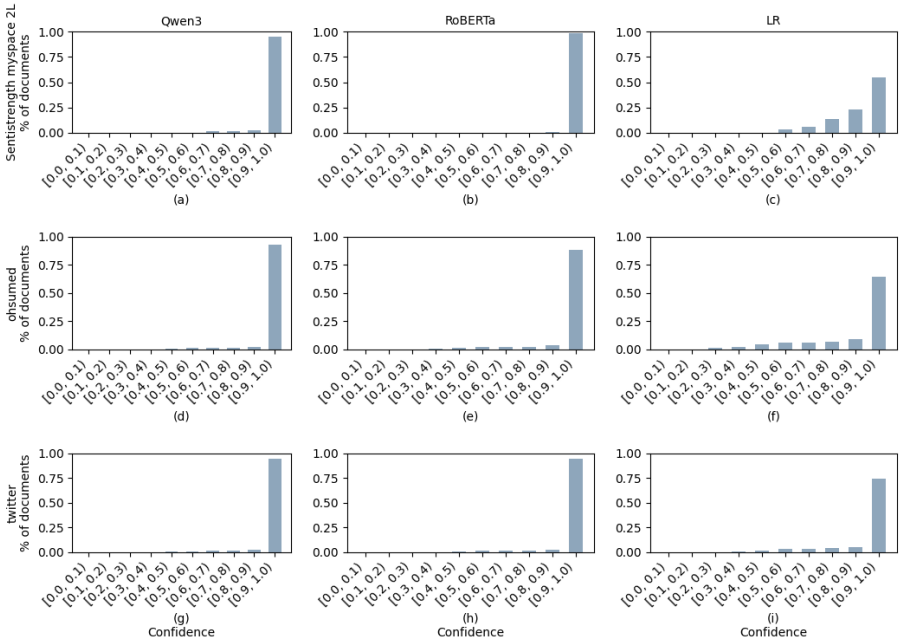

Figure 3: Confidence histograms for Qwen3, RoBERTa, and LR on the SentistrengthMyspace2L, Ohsumed, and Twitter datasets. Transformer-based models concentrate more than 90% of their predictions in the highest confidence interval, while LR distributes predictions more uniformly across the entire confidence spectrum.

metrics. In the case of BS, values that previously ranged around 0.10 and at most 0.31 now reach a minimum of 0.44. ECE displays similar behavior, with values that did not exceed 0.13 rising to above 0.20. LR, in turn, also deteriorates, but with a noticeably smaller variation: the average ECE increases from 0.10 to 0.15, remaining more closely aligned with the calibration curves presented earlier. In summary, answering **RQ2**, the evidence indicates traditional metrics are dominated by well-calibrated correct predictions and fail to expose catastrophic miscalibration on errors. Our results suggest that current calibration practices may systematically underestimate miscalibration, particularly in modern NLP systems.

6.3 RQ3: BBS

To answer **RQ3**, we revisit the BBS results presented in Table 3. Upon applying the BBS, we observe a contrast with respect to traditional metrics. While BS and ECE suggest that Transformers are better calibrated than Logistic Regression (LR), the BBS reveals the opposite reality. As demonstrated in Table 3, LR achieves the best BBS values on four out of six datasets (myspace, bbc, ohsumed, and trec), consistently outperforming Transformer-based models in terms of calibration. The only exceptions occur on the Sentistrength digg 2L dataset, where Llama3.1 obtains the best result but is statistically tied with LR, and on the Twitter dataset, where the calibration curves (Figure 2 (i)) had already indicated that LR did not exhibit good calibration. Focusing on the LLM results, we observe that Llama3.1 emerges as the most calibrated Transformer model, achieving the best result on 2 datasets (digg and Twitter) and being statistically comparable to the best result on 2 others (myspace and bbc).

Conversely, BBS reinforces the observation that **SLMs are less calibrated than LLMs**, with RoBERTa yielding the worst results in four of the six datasets and statistically tying for the worst in the remaining two. This finding may have important practical consequences, as recent work shows that, in TC, fine-tuned SLMs often offer a better cost–effectiveness trade-off than LLMs in zero-shot, in-context, or fine-tuned settings (Cunha et al., 2025b; Bucher and Martini, 2024). If LLMs consistently provide better-calibrated confidence than SLMs, model choice should weigh not only effectiveness and inference cost, but also uncertainty quality. In risk-sensitive contexts, even modest calibration gains can materially improve thresholding, abstention, and human-in-the-loop decisions. In contrast, an overconfident SLM can be operationally risky despite strong Macro-F1, since confidence-driven policies may amplify its errors. Overall, our findings motivate viewing SLM–LLM selection as a three-way trade-off among effectiveness, efficiency, and calibrated confidence.

dataset	Llama3.1	Gemma	Mistral	Qwen3	RoBERTa	LR
			BS<90			
Sentistrength digg 2L	0.48(0.13)•	0.61(0.19)	0.51(0.10)•	0.51(0.11)•	0.58(0.36)•	0.42(0.06)▲
Sentistrength myspace 2L	0.55(0.29)•	-	0.42(0.21)•	0.52(0.23)•	-	0.39(0.06)▲
Sentistrength bbc 2L	0.45(0.09)•	0.43(0.19)•	0.65(0.25)	0.56(0.17)	-	0.35(0.07)▲
ohsumed	0.70(0.03)	0.74(0.04)	0.76(0.03)	0.78(0.04)	0.82(0.04)	0.66(0.03)▲
trec	0.58(0.08)•	0.57(0.10)•	0.59(0.15)•	0.63(0.12)•	0.70(0.14)	0.52(0.01)▲
twitter	0.70(0.07)▲	0.70(0.07)•	0.78(0.07)•	0.74(0.11)•	0.83(0.08)•	0.86(0.04)
			ECE<90			
Sentistrength digg 2L	0.23(0.08)	0.36(0.06)	0.29(0.10)	0.32(0.08)	0.42(0.17)	0.13(0.06)▲
Sentistrength myspace 2L	0.34(0.15)	-	0.34(0.11)	0.33(0.16)	-	0.12(0.05)▲
Sentistrength bbc 2L	0.22(0.05)	0.30(0.12)	0.40(0.15)	0.39(0.09)	-	0.12(0.06)▲
ohsumed	0.21(0.03)	0.24(0.03)	0.27(0.02)	0.28(0.03)	0.27(0.03)	0.13(0.03)▲
trec	0.24(0.05)	0.23(0.06)	0.31(0.07)	0.30(0.09)	0.36(0.07)	0.13(0.01)▲
twitter	0.25(0.06)•	0.23(0.05)▲	0.30(0.04)	0.30(0.07)•	0.33(0.07)	0.32(0.04)

Table 5: BS and ECE results when excluding instances with confidence exceeding 90%. Once high-confidence correct predictions are removed, Transformer-based models exhibit worse calibration than reported by metrics computed on the full data, while LR aligns more closely with the evidence from the calibration curve. Numbers in parentheses represent 95% CIs. Cells with '-' indicate that, for one or more folds, the classifier produced no predictions with confidence levels below 90%. "▲" indicates the best dataset result, "•" indicates a statistical tie, and unmarked, a statistical loss compared to the best result.

In summary, by balancing correct and incorrect predictions in its computation, BBS prevents high effectiveness from masking severe miscalibration on errors, as can occur with traditional metrics. As a result, BBS characterizes calibration more faithfully in the high-effectiveness regime typical of SLMs and LLMs.

6.4 BBS Responsiveness to Recalibration

dataset	llama3.1	gemma	Mistral	qwen3	RoBERTa	LR
digg	0.47(0.2)	0.47(0.2)	0.43(0.2)	0.43(0.2)	0.63(0.2)	0.59(0.0)
myspace	0.48(0.2)	0.56(0.2)	0.59(0.1)	0.42(0.2)	0.77(0.1)	0.60(0.1)
bbc	0.47(0.1)	0.57(0.2)	0.58(0.1)	0.66(0.1)	0.73(0.2)	0.57(0.2)
ohsumed	0.75(0.0)	0.75(0.0)	0.75(0.0)	0.75(0.0)	0.80(0.0)	0.73(0.0)
trec	0.77(0.0)	0.75(0.1)	0.76(0.0)	0.77(0.1)	0.88(0.0)	0.67(0.0)
twitter	0.72(0.0)	0.73(0.0)	0.73(0.0)	0.74(0.0)	0.80(0.0)	0.70(0.0)

Table 6: Models' BBS when applying Temperature Scaling calibration. Calibration systematically reduces BBS values, confirming that the metric is responsive to genuine improvements in reliability. Numbers in parentheses represent 95% CIs.

We further assess BBS robustness by examining its sensitivity to mitigation strategies. Specifically, we apply post-hoc recalibration via Temperature Scaling (Guo et al., 2017), chosen for its consistent calibration improvements in Transformer-based models (Guo et al., 2017), as discussed in Section 2. Results show that recalibration systematically reduces BBS values, with an average decrease of 17p.p.. In the best-case scenario, this reduction reaches 39 p.p., observed for Mistral on the *digg* dataset, where BBS drops from 0.82 to 0.43. This confirms that the metric is responsive to genuine improvements in reliability. However, even after correction, BBS indicates that high-effectiveness Transformers often fail to match the calibration stability of LR. This reinforces the utility of BBS: unlike traditional metrics that can be dominated by accuracy, BBS distinguishes truly well-calibrated models from those whose confidence estimates remain unreliable despite strong performance.

7 Conclusions and Future Work

In this paper, we presented a comprehensive calibration study of Transformer-based text classification models, showing that, despite strong effectiveness, both SLMs and LLMs remain markedly overconfident. We also showed that widely used metrics such as ECE and Brier Score fail to capture this behavior in high-effectiveness regimes, often contradicting calibration curves and misleadingly suggesting good calibration. To address this limitation, we proposed the Balanced Brier Score (BBS), which balances contributions from correct and incorrect predictions and remains informative even under strong confidence concentration. BBS aligns with qualitative evidence and provides a more faithful calibration assessment for modern Transformer-based TC models. Future work includes extending BBS to broader NLP tasks and integrating it into confidence-aware decision pipelines. We also plan to investigate further the effectiveness–efficiency–calibration trade-off between SLMs and LLMs, and to evaluate new calibration methods using our proposed metric. This highlights the need to rethink how confidence is evaluated in modern NLP systems.

Acknowledgements

This work was supported by CNPq, CAPES, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILDIAR, grant no. 408490/2024-1), FAPEMIG, AWS, Google, NVIDIA, CI-IASaúde, and FAPESP.

Limitations

Despite these contributions, our study has limitations. First, due to resource constraints, our experimental evaluation focused on LLMs in the 7–8B parameter range (Llama-3.1-8B, Gemma-7B, Mistral-7B, and Qwen3-8B). Although larger models are also expected to achieve high effectiveness, it would be valuable to analyze how calibration varies with model size (i.e., number of parameters).

Second, the proposed Balanced Brier Score and our analyses are restricted to Text Classification. While TC is a fundamental NLP task, we did not evaluate the applicability or behavior of BBS in open-ended generation, complex reasoning, question answering, or other settings. Extending BBS to tasks where "correct" and "incorrect" are less binary than in classification requires further investigation, which we leave for future work.

References

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Martin Juan José Bucher and Marco Martini. 2024. Fine-tuned 'small' llms (still) significantly outperform zero-shot generative ai models in text classification. *arXiv*.

Muthu Chidambaram, Holden Lee, Colin McSwiggen, and Semon Rezchikov. 2024. How flawed is ece? an analysis via logit smoothing. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023a. An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *the 46th ACM SIGIR*.

Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Felipe Viegas, Celso França, Martins, Jussara M Almeida, et al. 2021. On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *IP&M*.

Washington Cunha, Leonardo Rocha, and Marcos André Gonçalves. 2025a. A thorough benchmark of automatic text classification: From traditional approaches to large language models. *arXiv preprint arXiv:2504.01930*.

Washington Cunha, Leonardo Rocha, and Marcos André Gonçalves. 2025b. A thorough benchmark of automatic text classification: From traditional approaches to large language models. *arXiv preprint arXiv:2504.01930*.

Washington Cunha, Felipe Viegas, Celso França, Thierson Rosa, Leonardo Rocha, and Marcos André Gonçalves. 2023b. A comparative survey of instance selection methods applied to nonneural and transformer-based text classification. *ACM CSUR*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Rodrigo de Oliveira, Matthew Garber, James M Gwinnutt, Emaan Rashidi, Jwu-Hsuan Hwang, William Gilmour, Jay Nanavati, Khaldoun Zine El Abidine, and Christina DeFilippo Mack. 2025. A study of calibration as a measurement of trustworthiness of large language models in biomedical natural language processing. *JAMIA open*, 8(4):ooaf058.

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guilherme Fonseca, Washington Cunha, Gabriel Prenassi, Marcos André Gonçalves, and Leonardo Chaves Dutra Da Rocha. 2025. Instance-selection-inspired undersampling strategies for bias reduction in small and large language models for binary text classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9323–9340.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in neural information processing systems*, 32.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2025. Large language models are miscalibrated in-context learners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11575–11596.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2025. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19260–19268.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632.

Jeremy Nixon, Michael W Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *CVPR workshops*, volume 2.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Zilma Silveira Nogueira Reis, Adriana Silvina Pagano, Isaias Jose Ramos de Oliveira, Cristiane dos Santos Dias, et al. 2024. Evaluating large language model–supported instructions for medication use: First steps toward a comprehensive model. *Mayo Clinic Proceedings: Digital Health*, 2(4):632–644.

Filipe N Ribeiro, Matheus Araújo, Pollyanna Gonçalves, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Sentibench-a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ DS*.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*.

Raymond E Wright. 1995. Logistic regression.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. *arXiv preprint arXiv:2210.04714*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Bruna Stella Zanotto, Ana Paula Beck da Silva Etges, Renata Ruschel, Washington Luiz, et al. 2021. Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. *JMIR Med. Infor.*

dataset	llama3.1-8B			gemma-7b			Mistral-7B-v0.1		
	BS	BS Acerto	BS Erro	BS	BS Acerto	BS Erro	BS	BS Acerto	BS Erro
Sentistrength digg 2L	0.19(0.05)	0.02(0.00)	1.49(0.12)	0.18(0.05)	0.01(0.00)	1.65(0.12)	0.21(0.07)	0.01(0.01)	1.58(0.14)
Sentistrength myspace 2L	0.09(0.03)	0.01(0.00)	1.65(0.15)	0.09(0.03)	0.00(0.00)	1.74(0.18)	0.10(0.02)	0.00(0.00)	1.81(0.12)
Sentistrength bbc 2L	0.16(0.03)	0.02(0.01)	1.47(0.15)	0.17(0.05)	0.01(0.00)	1.73(0.13)	0.17(0.04)	0.01(0.00)	1.81(0.05)
ohsumed	0.22(0.01)	0.01(0.00)	1.62(0.03)	0.23(0.01)	0.01(0.00)	1.65(0.03)	0.24(0.01)	0.01(0.00)	1.72(0.01)
trec	0.07(0.01)	0.00(0.00)	1.67(0.06)	0.06(0.01)	0.00(0.00)	1.74(0.05)	0.07(0.01)	0.00(0.00)	1.78(0.04)
twitter	0.17(0.01)	0.01(0.00)	1.63(0.02)	0.17(0.01)	0.01(0.00)	1.70(0.04)	0.17(0.01)	0.01(0.00)	1.70(0.03)

Table 7: BS, BS Correct, and BS Error model results. BS Correct refers to the BS calculated only on correctly predicted instances, while BS Error refers to the BS calculated only on misclassified instances. Numbers in parentheses represent 95% CIs. Results mirror those reported in Table 4: all LLMs exhibit near-perfect calibration on correct predictions but severe overconfidence on errors, confirming that the dominance bias observed in traditional metrics is consistent across all evaluated Transformer-based models.

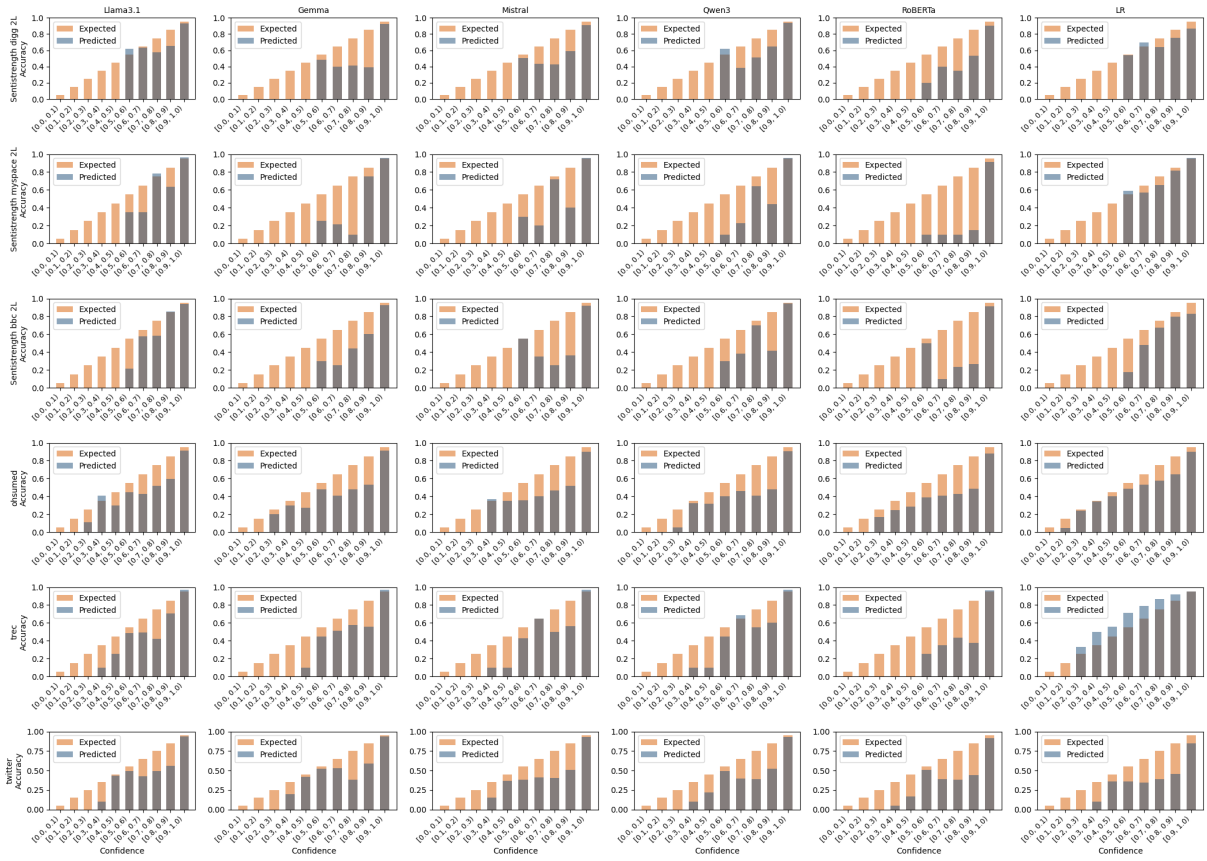

Figure 4: Calibration curves for all models across all datasets. "Expected" denotes the ideal calibration line, where confidence equals empirical accuracy, and "Predicted" corresponds to the observed empirical accuracy within each confidence interval, a well-calibrated model should have both lines closely aligned.

A Calibration curves and Confidence histograms

Figures 4 and 5 present, respectively, the calibration curves and confidence histograms for all models across all datasets. These results are omitted from the main text due to space constraints. The same patterns reported in Sections 6.1 and 6.2 are observed here: Transformer models display calibration curves consistent with poor calibration, and their histograms show that over 80% of predictions are made with confidence above 0.9.

B BS decomposition

Table 7 reports the BS, BS correct, and BS error results for the Llama 3.1, Gemma, and Mistral classifiers, which are omitted from the main text due to space constraints. The conclusions drawn in Section 6.2 for Table 4 also apply here. Specifically, LLMs exhibit near-perfect calibration on correct predictions (BS correct) but very poor calibration on errors (BS error), indicating systematic overconfidence: the models consistently assign high confidence even to incorrectly classified instances.

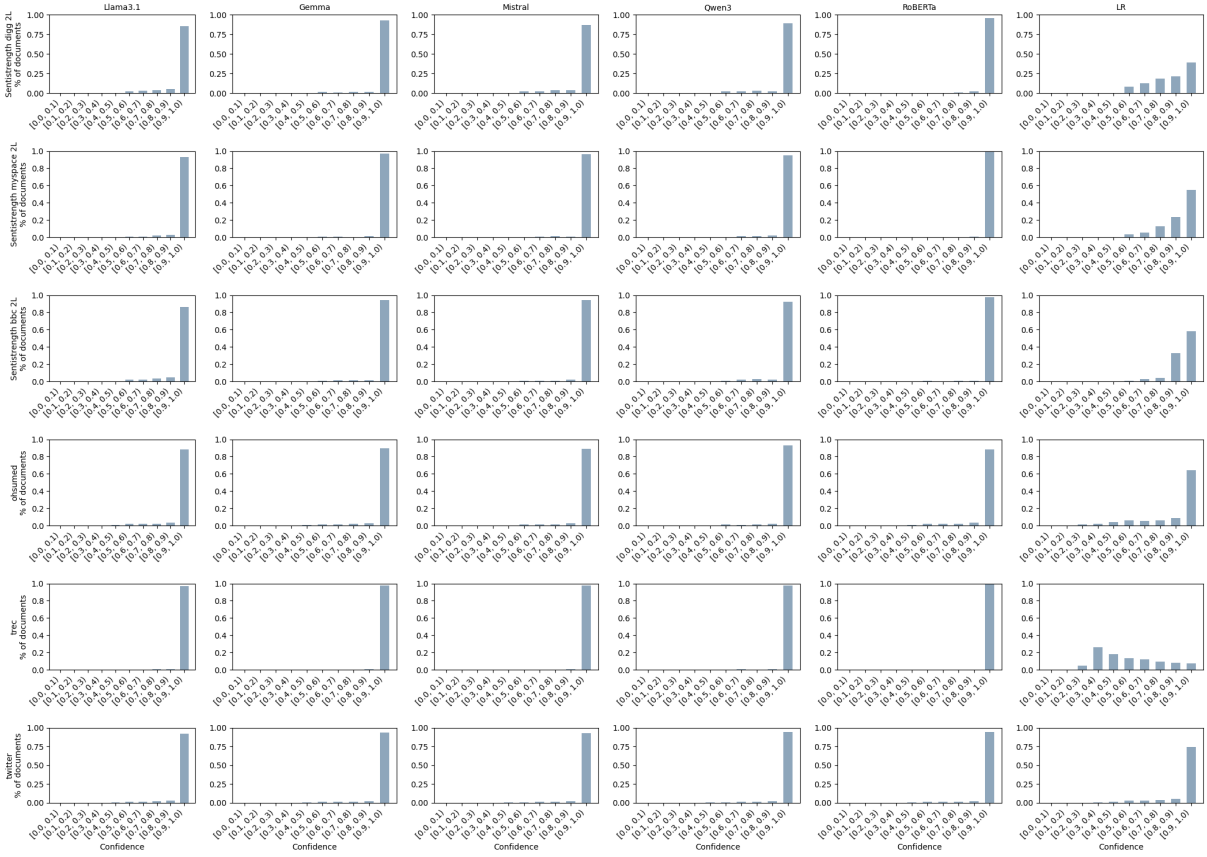

Figure 5: Confidence histograms for all models across all datasets. Each bar represents the proportion of predictions falling within a given confidence bin.