Optimizing the Effectiveness-Efficiency-Interpretability Trade-off in Text Classification via Hard Instance Paraphrasing

Anonymous ACL submission

Abstract

Large Language Models (LLMs) achieve strong text classification performance but incur high training and inference costs. Small Language Models (SLMs) are more efficient yet struggle with ambiguous or low-confidence ("hard") instances. We propose RAP – **R**ewrite-**A**ssess-**P**redict for Hard Instances, a multi-stage framework that combines SLM efficiency with zero-shot LLM adaptability. RAP first detects hard instances using calibrated confidence, then applies prompt-optimized LLM paraphrasing to make class-relevant cues more explicit, with the resulting paraphrases replacing the original texts for subsequent classification. The paraphrased texts are reassessed, and predictions are chosen using calibrated confidence, ensuring performance does not degrade when rewriting is unhelpful. Experiments across multiple datasets show that paraphrases preserve semantic content, mitigate label shifts, and substantially improve SLM effectiveness on hard cases. Qualitative analysis further reveals systematic textual modifications that enhance class salience and interpretability. Overall, RAP provides a more accurate, reliable, interpretable, scalable, and cost-effective alternative to end-to-end LLM classification.

1 Introduction

Text classification (TC) is a core NLP task underpinning applications such as information retrieval, content moderation, and digital archiving (Sebastiani, 2002; Kowsari et al., 2019). Recent advances in Language Models (LMs) have substantially improved classification effectiveness (Devlin et al., 2019; Raffel et al., 2020; Bommasani et al., 2021; Zhang et al., 2025). However, state-of-the-art performance typically relies on fine-tuning large models (Cunha et al., 2025), which increases computational cost and raises concerns about scalability, efficiency, and sustainability (Kaplan et al., 2020).

Although fine-tuned LMs (Small or Large) are highly effective, their training and inference costs hinder widespread deployment. Small Language Models (SLMs), such as RoBERTa (Liu et al., 2019), with fewer than one billion parameters – commonly considered suitable for resource-constrained settings (Liu et al., 2024; Wang et al., 2025a) – offer substantially lower overhead. Yet SLMs often underperform on ambiguous or low-confidence ("hard") documents, where class boundaries are unclear (Desai and Durrett, 2020; Cunha et al., 2025). LLMs typically handle such cases better, but at a significantly higher cost and usually only when fine-tuned; zero-shot or in-context LLMs rarely surpass tuned SLMs in TC (Cunha et al., 2025; Bucher and Martini, 2024). This reveals a persistent effectiveness–efficiency trade-off that remains insufficiently addressed.

Recent work suggests two complementary directions to mitigate this trade-off. First, identifying hard instances–those for which the model assigns low confidence scores–enables targeted allocation of computational resources (Andrade et al., 2024). Second, zero-shot LLMs can generate semantically consistent paraphrases[1] that make class-discriminative cues more explicit, improving downstream classification (Li et al., 2025). These insights motivate hybrid strategies that combine SLM efficiency with the instruction-following flexibility and semantic rewriting capabilities of LLMs. In addition, paraphrasing may enhance interpretability by explicitly surfacing class-relevant information (Shwartz and Dagan, 2018).

Prior work has explored data augmentation for TC via lexical perturbations or LLM-generated rewrites, typically expanding the training set with paraphrased instances (Wei and Zou, 2019; Sarker et al., 2023; Ding et al., 2023; Anaby-Tavor

[1] "paraphrase" and "rewrite" are used interchangeably throughout this text.

et al., 2023). While sometimes effective, such approaches generally apply augmentation uniformly across the dataset, which increases computational cost and requires full model retraining. Moreover, they rarely aim to explicitly highlight class-relevant information, limiting interpretability and obscuring how paraphrasing benefits classification.

By contrast, this paper proposes **RAP** – **R**ewrite-**A**ssess-**P**redict for Hard Instances, a multi-stage classification framework that selectively integrates SLMs and zero-shot LLMs. RAP first relies on SLMs for confident predictions, then applies prompt-optimized LLM paraphrasing to hard instances – defined here as low-confidence texts identified by a calibrated SLM classifier – only, rewriting the original text to make class membership more explicit. Predictions from the SLM on the original and rewritten texts are subsequently reassessed and combined based on calibrated confidence, yielding additional effectiveness and reliability gains. Prompt optimization is performed per dataset using a validation set, enabling effective paraphrasing without fine-tuning the model.

Our proposal's *novelty* lies in integrating text difficulty detection, prompt optimization, and paraphrase generation from the original text into a single pipeline. To the best of our knowledge, this is the first approach that combines these three elements for TC, balancing effectiveness with efficiency and interpretability, paving the way for more sustainable and interpretable NLP solutions.

Our proposal is based on the hypothesis that paraphrasing benefits classification by making class-relevant information more explicit, as long as semantic content, and consequently the underlying label, is preserved. Accordingly, our first research question (**RQ1**) evaluates whether LLM-generated paraphrases remain semantically similar to their source texts. Using established similarity metrics, we consistently observe high semantic alignment, with BERTScore values of around 0.9 across all datasets (Section 5). These high similarity scores indicate that the paraphrases remain meaning-preserving. We further confirmed this hypothesis in a manual sample evaluation (Tab. 4)

While RQ1 shows that paraphrasing preserves meaning, semantic fidelity alone does not guarantee effectiveness gains. Two semantically equivalent texts may still differ in how clearly they expose class-relevant evidence. This motivates **RQ2**, which asks whether paraphrasing hard instances not only preserves meaning but also enhances class salience, enabling SLMs to make correct predictions. To address this, we analyze prompt design and quantify the impact of rewriting on SLM effectiveness, showing that paraphrases that retain semantics while highlighting discriminative cues yield consistent improvements.

When paraphrasing fails to raise confidence, we use a confidence-based decision rule: with the same calibrated model, we score both original and rewritten texts and choose the prediction with higher calibrated confidence. This motivates **RQ3**, assessing whether confidence-guided selection further improves effectiveness. Results show this targeted strategy is competitive while incurring lower overhead than end-to-end LLM-based TC.

Finally, we address interpretability. **RQ4** investigates which textual modifications introduced by the LLM make class-relevant information more explicit in hard instances. We analyze these modifications using an established taxonomy of paraphrase phenomena (Bhagat and Hovy, 2013), complemented by fine-grained linguistic analysis grounded in Systemic-Functional Theory (Halliday and Matthiessen, 2013) and Appraisal Theory (Martin and White, 2003). We provide qualitative evidence showing how paraphrasing clarifies class membership, complementing the quantitative results by explaining *how* and *why* rewriting improves classification in ambiguous cases.

In sum, this paper's contributions are six-fold: (1) a systematic comparison of calibration-based methods for identifying hard instances in TC; (2) a targeted evaluation of zero-shot LLM-based paraphrasing for improving SLM performance on hard instances, with explicit analysis of semantic preservation; (3) a prompt optimization procedure for class-revealing paraphrasing in sentiment and topic classification tasks; (4) a confidence-based combination strategy that leverages both the original text and its corresponding paraphrase to determine the final classification; (5) a qualitative linguistic analysis of LLM-induced textual modifications that increase class salience; and (6) public dataset release with annotations of hard instances, enabling further scientific advances.

2 Related Work

Recent work has explored LLM-based rewriting across multiple NLP tasks. Substantial research has focused on improving human readability through sentence simplification in specialized domains such as medicine (Yang et al., 2025) and governmental

communication (Scalercio et al., 2025). These studies typically assess semantic preservation and fluency using similarity metrics and qualitative analysis, but do not examine how rewriting affects downstream confidence or classification robustness.

Related work uses LLM paraphrasing as data augmentation for classification and QA. Common strategies generate multiple rewrites per instance and inject them into training data (Ding et al., 2023; Sarker et al., 2023; Wang et al., 2025c), including multi-label scenarios with paraphrases or label-conditioned synthetic examples (Zhao et al., 2024). Although effective, these approaches uniformly rewrite entire datasets, increase training cost, require model retraining, and primarily expand coverage rather than explicitly enhancing class evidence at inference time.

More recently, rewriting has been explored as a means to improve interpretability by making implicit attributes (e.g., sentiment or stance) explicit (Li et al., 2025). While showing that LLMs can surface latent cues, these works do not investigate interactions with small-scale classifiers or selective deployment under uncertainty.

Orthogonally, several studies reduce computation by selectively invoking stronger models. For instance, Yue et al. (2024) propose weak–strong model cascades, querying a powerful LLM only when a weaker model is unreliable. Although effective, such approaches mainly target API usage reduction and do not explicitly address inference efficiency, semantic preservation, or interpretability.

Our work differs from prior approaches along four key dimensions. First, instead of uniformly paraphrasing or augmenting data, we selectively rewrite only hard instances identified via calibrated SLM confidence, greatly reducing LLM usage. Second, our goal is not to improve readability or expand the dataset, but to explicitly surface class-relevant cues at inference time so that SLMs can resolve ambiguous inputs. Third, we present a unified multi-stage framework that integrates difficulty detection, prompt optimization, and zero-shot paraphrasing with explicit evaluation of semantic fidelity, performance, and cost. Finally, while most rewriting work focuses only on quantitative gains, we also conduct linguistically grounded qualitative analysis of how LLM-induced changes increase class salience. To our knowledge, this is the first work to jointly combine selective rewriting, efficiency awareness, and interpretable linguistic analysis within a single TC framework.

3 RAP: *Rewrite–Assess–Predict for Hard Instances*

One of our main contributions is RAP—**R**ewrite-**A**ssess-**P**redict for Hard Instances—a unified framework that leverages LLM paraphrasing to enhance Small Language Models (SLMs) while jointly addressing efficiency, effectiveness, and interpretability. RAP operationalizes a central hypothesis: SLMs are efficient and accurate on confident cases but degrade on challenging instances, where correct inference cannot be derived from surface features alone and instead requires deeper semantic or pragmatic interpretation, contextual integration, or domain-specific knowledge, often under limited lexical alignment between input and target labels.

LLM-based rewriting can clarify such instances by making class-relevant cues more explicit, but indiscriminate use is costly. RAP therefore selectively targets hard instances—identified as low-confidence predictions from a calibrated SLM—rewrites them to emphasize discriminative information, and reclassifies them using a max-confidence rule. This improves effectiveness on difficult cases while preserving efficiency and offering linguistic interpretability through explicit textual transformations.

Proposal Framework/Classification Workflow. Figure 1 overviews the RAP framework and its workflow. The process begins by splitting the data into training, validation, and test sets. The training set fine-tunes the SLM (a), which generates predictions (b), whose confidence scores are then calibrated via isotonic regression (c). Instances with calibrated confidence below a threshold L (tuned on validation) are labeled hard (d) and rewritten to make class cues explicit. After optimizing the rewrite prompt (e), the LLM generates paraphrases (f), which are reclassified and recalibrated by the SLM (g). The original and rewritten calibrated confidences (c,g) are then used to select or combine predictions via a confidence-based rule (h), yielding the final classification (i).

Confidence estimation and calibration – Fig.1(c) Identifying hard instances requires reliable confidence estimates from the SLM (i.e., RoBERTa; see Section 4.2 for more details). We therefore rely on predicted class probabilities and explicitly address the need for calibration. A classifier is considered calibrated when its predicted probabilities align

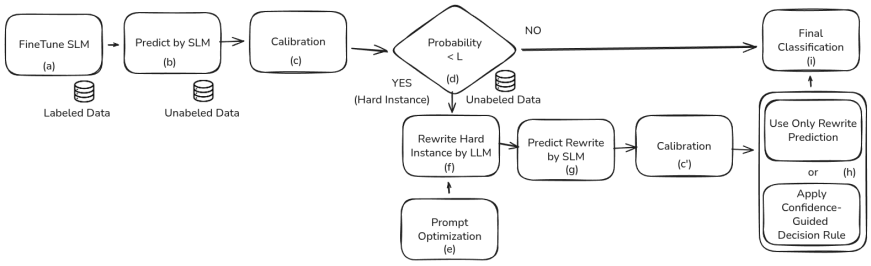

Figure 1: Overview of the proposed RAP Framework.

with empirical correctness frequencies within corresponding probability bins (Guo et al., 2017). To justify using SLM confidence as a decision signal, we show that softmax probabilities provide sufficient discriminative power, consistent with prior findings (Wolfe et al., 2017). We further improve reliability through post-hoc calibration. Table 7 in Appendix B reports Brier Scores (i.e., the closer to 1, the better) for the original softmax outputs and for calibrated probabilities obtained via isotonic regression. As we can see, across **all** datasets, isotonic regression yields the best calibration performance; consequently, we adopt it to obtain the final confidence estimates used throughout the framework.

Hard instance threshold selection. Fig.1(d) A key framework parameter is the confidence threshold (denoted as L), which determines whether an instance is considered hard. After splitting the data into training, validation, and test sets, we generate LLM-based paraphrases for hard instances in the validation set and evaluate a range of candidate values for L, starting at 0.10 and increasing at 0.05 steps. The optimal threshold is selected as the value that maximizes the target effectiveness metric (in this case, MacroF1) on the validation set. This validation-driven selection ensures that rewriting is applied only when it yields measurable gains, directly supporting the framework's efficiency objective. The selected threshold is then fixed and applied to the test set.

Prompt optimization. Fig.1(e) The process begins with an initial prompt designed to rewrite instances in a class-revealing manner, with the goal of improving classification performance through the use of these paraphrases. Prompt optimization proceeds iteratively: at each step, the effectiveness score of the current prompt is computed and compared against the best score observed so far. The prompt achieving the highest validation performance is selected as the final prompt. This dataset-specific optimization step is crucial to ensure that rewriting consistently enhances class salience while preserving semantic similarity. Once the final prompt is fixed, hard instances in the test set are rewritten and forwarded to the classification stage.

Re-write Hard Instance by LLM and predict by SLM. Fig.1(f-g) After identifying the hard instances, we prompt the LLM to paraphrase the original text to make its class membership more explicit (f). The instance is reclassified using the SLM (g), now operating on the rewritten text, thereby producing a new prediction and its corresponding confidence score. As in step (c), the confidence of the paraphrased text is subsequently calibrated (c'), enabling a direct comparison with the confidence obtained from the original text, which is going to be considered in the next step (h).

Choosing the most suitable prediction. Fig.1(h-i) Two complementary variants are instantiated to yield the final classification, both operating on the subset of documents identified as hard by the calibrated SLM. (i) **SLM Reclassification**: hard documents are rewritten using the optimized LLM prompt, and the SLM reclassifies the rewritten texts. This variant isolates the effect of rewriting on SLM performance and directly evaluates whether making class-relevant cues explicit suffices to correct low-confidence predictions. (ii) **Confidence-Guided Decision Rule**: after identifying hard instances and generating their paraphrases, we compute SLM confidence scores for both the original and rewritten texts, calibrating them with the same model to ensure comparability. The final prediction is then selected based on the higher calibrated confidence. This selective fallback mechanism acts as a safety control, preventing performance degradation by automatically reverting to the original text whenever rewriting decreases model confidence.

4 Experimental Methodology and Setup

4.1 Datasets

Our study draws on 14 datasets covering sentiment analysis and topic classification. Sentiment

datasets correspond to binary classification. We include IMDB[2] , PangMovie from Rotten Tomatoes[3] , SemEval17 (Twitter posts), the Stanford Sentiment Treebank (SST and SST-2), and Yelp Reviews. Because large LLM training corpora raise concerns about benchmark contamination, we further curate two post-LLM datasets: RottenT2024 (Jan–Nov 2024) and IMDB2024 (Jan–May 2024), ensuring that evaluation instances were not seen during training. Together, these datasets provide a diverse, contemporary benchmark. For topic classification, we use AGNews (four topical categories), ACM Digital Library and DBLP (academic titles and abstracts), Reddit (10,000 sampled subreddit posts), and Twitter Topic with six topic labels. Table 6 in Appendix A summarizes dataset statistics.

4.2 SLM and LLM Selection

We adopt RoBERTa as the SLM and LLaMA 3.1 as the LLM due to their open-source availability, popularity, and strong empirical performance in TC (Wang et al., 2025b). RoBERTa is extensively validated for sentiment and topic classification, while LLaMA-based models achieve competitive results across multiple benchmarks (Cunha et al., 2025; Bai et al., 2023; Fonseca et al., 2025).

4.3 Prompt Optimization

We perform prompt optimization 1(e) on the validation set. We start from an initial rewriting prompt generated by ChatGPT 4.0 and refine it using optimization strategies inspired by (Chen et al., 2024; França et al., 2025). The prompt optimizer comprises two stages: (i) a meta-prompt that instructs the model to generate alternative prompts that differ from the provided examples and make the target label more explicit; and (ii) a set of candidate prompts, each one associated with an effectiveness score. Table 1 presents an example of the meta-prompt used for the sentiment classification task, where the LLM is expected to generate and complete a new prompt.

Figure 2 illustrates the optimization process, showing how rewriting prompt quality evolves across iterations. The x-axis indicates the prompt iteration (iteration 0 is the initial ChatGPT-generated prompt), and the y-axis reports the Macro-F1 obtained by the SLM when classifying validation hard instances rewritten with each prompt. The highest effectiveness is achieved at iteration 3, with Macro-F1 of 0.67. This prompt is selected as final and used for test-set classification.

[2] https://www.imdb.com/

[3] https://www.rottentomatoes.com/

Table 1: Prompt optimization example.

Your task is to create a single new prompt, distinct from the examples below, which paraphrases texts more directly with respect to their label. The prompts provided below serve as examples along with their corresponding achieved scores. example: <start prompt> Paraphrase the following text more directly and concisely regarding the positive and negative labels: <end prompt> score: 0.43 <start prompt> {Response from LLM}

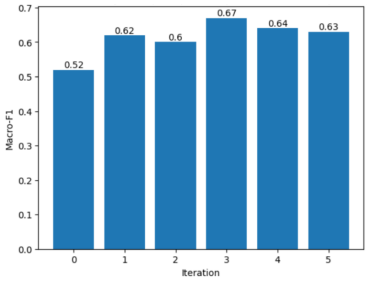

Figure 2: Prompt Optimization of SemEval 2017.

4.4 Experimental Protocol and Tuning

All datasets are partitioned using stratified 5-fold cross-validation. In each fold, the data are split into training, validation, and test sets while preserving the original class distribution, and the roles of these partitions rotate across iterations. The validation set is exclusively used for method-specific parameter tuning. For SLMs, we follow the hyperparameter configuration proposed by (Cunha et al., 2023), fixing the learning rate at 2×10^{-5}, using a batch size of 64 documents, fine-tuning the model for five epochs, and setting the maximum input length to 256 tokens. For the LLM, we adopt a LLaMA 3.1 zero-shot and few-shot setting and retain all default model hyperparameters, except the temperature, which is fixed at 0 to improve reproducibility across runs and facilitate comparison in future work. Under this configuration, the LLM is provided with the original text together with the optimized prompt.

4.5 Metrics and Experimental Protocol

The SLM runs on a 4-core CPU with 32 GB of RAM and an NVIDIA Tesla P100 GPU, while LLM inference is performed via AWS Bedrock. Classification effectiveness is measured using Macro-F1 (Equation 1, Appendix C), appropriate given class imbalance in several datasets. To ensure statistical validity and robustness, we also report 95% confidence intervals computed from five test scores. For pairwise method comparisons, we use the Wilcoxon signed-rank test, which avoids

normality assumptions and is widely recommended for classification evaluation (Demšar, 2006).

We evaluate cost-effectiveness by measuring the end-to-end runtime of each method, from SLM fine-tuning to test-time inference. We first report the computational cost of the RoBERTa SLM baseline, including fine-tuning and inference (Equation 2, Appendix C), which is shared across all variants since the SLM is always used for confidence estimation and hard-instance detection. Equation 3 (Appendix C) captures the cost of threshold estimation followed by direct LLM inference without rewriting. Equation 4 (Appendix C) adds the cost of LLM rewriting plus SLM reclassification of hard instances. Finally, Equation 5 (Appendix C) incorporates the additional cost of calibrating the confidence of the rewritten text and selecting between the prediction of the original text and its rewritten version.

To understand **why** paraphrasing improves effectiveness, we conducted a post-classification qualitative analysis on a randomly selected set of original-paraphrase pairs. All paraphrases were manually verified as semantically equivalent, and the SLM correctly predicted the target label after rewriting. Each pair was analyzed using (i) the paraphrase taxonomy of (Bhagat and Hovy, 2013): lexical substitution, syntactic restructuring, and discourse-level reformulation; and (ii) Systemic-Functional Theory (Halliday and Matthiessen, 2013) and Appraisal Theory (Martin and White, 2003), focusing on evaluative stance (Affect, Judgement, Appreciation), polarity, and explicitness. We paid attention to interpersonal re-realizations, such as ironic texts becoming non-ironic or the removal of stance-marking prefaces. In some cases, the LLM added contextual details that did not change the proposition but made implicit evaluations clearer. Overall, these transformations explain gains: paraphrases make evaluative cues, discourse structure, and topical markers more surface-accessible, reducing high-inference demands and improving classification through increased explicitness.

5 Experimental Results and Analyses

5.1 RQ1:Do paraphrases preserve semantics?

To assess whether LLM-generated paraphrases preserve the underlying semantics of the original texts, enabling their use as surrogate instances for hard cases, we first compute BERTScore (Zhang et al., 2020), which compares contextualized representations rather than surface forms. Table 2 shows consistently high values across datasets, indicating strong semantic fidelity between originals and paraphrases. As a complementary view, we also report BLEU (Papineni et al., 2002). Lower BLEU scores are expected, since it is sensitive to lexical overlap and paraphrases naturally introduce lexical and structural variation. Together, high BERTScore and low BLEU indicate that paraphrases preserve meaning while allowing substantial surface diversity, creating opportunities to make class-relevant cues more explicit.

Table 2: BertScore and BLEU metrics.

Dataset	BERTScore	BLEU
IMDB	0.87	0.04
PangMovie	0.92	0.26
SemEval2017	0.90	0.15
SST	0.92	0.28
SST2	0.88	0.09
Yelp2L	0.89	0.04
IMDB2024	0.88	0.06
RottenT2024	0.89	0.09
ACM	0.90	0.25
AGNews	0.95	0.59
DBLP	0.98	0.89
Reddit	0.92	0.38
Twitter	0.92	0.38

5.2 RQ2: Does paraphrasing hard instances improve SLM effectiveness?

Table 3: Average Macro-F1; **boldface** marks statistically superior results with Wilcoxon tests relative to RoBERTa. Table version with 95% CIs in Appendix G.

Dataset	RoBERTa	LLM Predict	SLM Reclassification	Confidence-Guided Decision Rule	LLM Zero-Shot (Full Test)	LLM Few-Shot (Full Test)	Oracle
IMDB	93.0	**94.4**	**94.6**	**94.6**	**94.3**	92.1	96.0
Pang Movie	88.7	86.4	**90.3**	**90.6**	**91.0**	86.8	97.1
SemEval2017	91.2	91.2	**91.5**	**91.5**	89.7	84.8	92.5
SST	87.3	87.9	**88.2**	**88.2**	**89.4**	83.8	91.6
SST2	94.6	**95.1**	**94.9**	**94.9**	93.0	90.0	96.5
Yelp2L	97.9	97.0	**98.7**	**98.7**	**99.3**	**99.0**	99.3
IMDB2024	97.6	94.0	**98.4**	**98.5**	97.7	93.4	99.5
RottenT2024	93.7	**94.2**	**94.2**	**94.3**	**95.8**	91.0	95.3
ACM	70.7	67.2	70.8	**71.1**	29.5	**71.8**	78.0
AGNews	92.2	91.7	92.2	**92.4**	84.2	92.1	95.5
DBLP	**81.9**	74.4	**81.9**	**81.9**	41.8	58.4	82.8
Reddit	96.0	96.1	**96.7**	**96.8**	91.4	95.7	98.8
Twitter	77.5	77.6	**78.5**	**79.0**	65.6	**79.6**	90.9

Table 3 reports full test set effectiveness. The first column shows the RoBERTa baseline. For all other methods, but LLM Zero-shot/instruction learning, predictions for hard instances are replaced and Macro-F1 is recomputed. **LLM Predict** classifies hard instances directly with the LLM. **SLM Reclassification** paraphrases hard instances and reclassifies them with the SLM. The fourth column presents the confidence-based variant. The Oracle column represents the upper bound if all hard instances were correctly classified; it does not reach 100% because it only concerns the hard subset. Finally, **LLM zero-shot/instruction learning** evaluates the LLM on the full test set

Table 4: Original text, paraphrase, label (0 = neg, 1 = pos; topic), BLEU, BERTScore, and confidence scores for both texts.

ID	Original Text	Paraphrase	Label	BLEU	BERT Score	Original confidence	Paraphrase confidence
1	if sh*t was a movie, it would be this.	This is the worst movie ever.	0	0.02	0.87	0.56	1
2	if the message seems more facile than the earlier films, the images have such a terrible beauty you may not care.	The message may be less profound than in earlier films, but the images are so stunningly beautiful that you may not notice.	1	0.08	0.94	0.56	0.98
3	The movie generates plot points with a degree of randomness usually achieved only by lottery drawing.	The movie's plot is ridiculously random.	0	0.01	0.89	0.63	0.9
4	Don't let the subtitles fool you; the movie only proves that Hollywood no longer has a monopoly on mindless action.	The movie is a mindless action film that proves Hollywood is no longer the only place to find such movies.	0	0.03	0.9	0.63	0.96
5	I just took a few hours break from the internet and now I am missing a lot of stuff I didnt even know jk went live Ahhh {@BTS_twt@} army struggles	I just took a break from social media and now I'm feeling out of the loop on all the latest pop culture news - I had no idea JK went live Ahhh, the struggles of being a part of the BTS ARMY are real!	Pop Culture	0.10	0.88	0.38	1.00
6	The halting problem for turning machines d m wilson	The halting problem for turing machines d m wilson	Computation Theory	0.597	0.973	0.399	0.618
7	What is a solution of an ode? robert m. corless	What is a solution of an ordinary differential equation? robert m. corless	Computing Math	0.616	0.955	0.315	0.896

using the same optimized instruction prompt as RAP, but without examples (zero-shot mode), ensuring fair comparison. LLM Few-Shot (Full Test) classifies the entire test set using **30** in-context examples per prompt, according to (Prenassi et al., 2025). For each test instance, these examples are selected as the most similar training instances to the test input, aiming to improve effectiveness.

For sentiment datasets, paraphrasing consistently helps by making polarity cues more explicit. As a result, SLM Reclassification yields statistically significant gains across **all** datasets, with improvements up to 1.8% on PangMovie. These gains are highly meaningful given RoBERTa's already strong performance (Macro-F1 $>$ 87%) and the fact that improvements occur precisely on the hardest cases. SLM Reclassification also surpasses LLM Predict on several datasets—IMDB 2024, PangMovie, Yelp2L, and SST—achieving up to 4.7% improvement on IMDB 2024.

Topic classification is more challenging due to variation in label granularity and domain vocabulary. In ACM, AGNews, and DBLP, paraphrasing yields performance statistically comparable to RoBERTa. In contrast, for Twitter and Reddit, paraphrasing provides clearer gains, indicating expanded and clarified topical content (Section 5.4).

5.3 RQ3: Does RAP's confidence-based rule further improve model effectiveness?

We analyze Table 3, focusing on the Confidence-Guided decision rule, which combines predictions from the original and paraphrased texts. Both confidence scores are calibrated with the same model to ensure comparability, and the final output is selected based on the higher-calibrated confidence. This strategy is statistically superior to RoBERTa across all datasets except DBLP, whose short titles limit paraphrasing benefits. Compared with SLM Paraphrasing, it generally achieves higher mean effectiveness and narrower confidence intervals, delivering statistically significant gains on ACM and AGNews, where no prior improvements were observed. Indeed, Confidence Rule **consistently outperforms LLM Predict across all sentiment and topic datasets**, where there is a tie, achieving substantial gains of up to 5.8% in ACM. RAP also outperforms or ties LLM Zero-Shot and Few-Shot on **all** sentiment and topic datasets except Yelp2L, with substantial gains over Zero-Shot on topic tasks. LLM Zero-Shot underperforms on topics due to task difficulty and class imbalance, while incurring higher costs. Providing examples improves LLM Few-Shot on topic datasets but degrades effectiveness on most sentiment datasets, as semantic proximity is ineffective for sentiment (Prenassi et al., 2025). Indeed, LLM Few-Shot never statistically outperforms RAP—except on Yelp2L by a negligible margin (0.3%)—while incurring significantly higher costs, as we shall see.

Finally, while datasets such as PangMovie, ACM, and Twitter still show a clear gap to the Oracle—indicating remaining untapped potential—others present a different picture. In SemEval2017, Yelp2L, IMDB 2024, and Reddit, RAP attains performance remarkably close to the Oracle, effectively capturing nearly all the benefit obtainable from focusing on the hardest instances. These results strongly reinforce the robustness and practical relevance of the proposed approach.

5.4 RQ4: How do LLM rewrites clarify class-relevant cues in hard instances?

In Table 4, the analysis of hard instances reveals that the LLM performs targeted textual modifications that make class-relevant information more explicit, enabling the SLM to make the correct class assignment after rewriting. Modifications manually categorized revealed recurrent patterns. These are: (i) item typo correction, as in: "turning",

"turing" (Table 5, line 6); (ii) acronym expansion, as in: "ode", "ordinary differential equation" (line 7); (iii) synonym substitution, as in the following pairs: "sh*t" -> "worst movie ever" (line 1); "facile" -> "less profound" (line 2); "terrible" -> "stunningly" (line 2); "care" -> "notice" (line 2); "seems" -> "may be" (line 2); "has a monopoly" -> "is the only place" (line 4); (iv) changes in clausal mood and logical relation, as in conditional "if" clauses rewritten as single affirmative clauses: "if sh*t was a movie, it would be this", "This is the worst movie ever" (line 1); (v) ironical into non-ironical wording, as in: "a degree of randomness usually achieved only by lottery drawing", "ridiculously random" (line 3); (vi) metaphorical into non-metaphorical wording, as in: "sh*t", "worst movie ever" (line 1); (vii) removal of discourse-level, stance-markers, as in: "don't let subtitles fool you" (line 4); (viii) addition of external knowledge, which clarifies cultural references, as in: "a lot of stuff", "all the latest pop culture news", "Ahhh {@BTS_twt@} army struggles", "Ahhh, the struggles of being a part of the BTS ARMY are real!", which interprets "jk" as Jungkook, a BTS member (line 5).

In particular, these removals do not affect overall sentiment assignment. For instance, in "don't let subtitles fool you," the clause matters to humans because "subtitles" evokes a cultural inference (e.g., foreign films as artistic and high quality versus mainstream Hollywood). However, removing it has little impact on sentiment, since the remaining text still strongly signals negativity by emphasizing that the film is a "mindless action film." It is also worth noting that additions do not introduce contextual hallucinations, as this was explicitly controlled in our prompt. Overall, patterns (i)–(viii) illustrate how LLM paraphrases make class-relevant information more explicit in hard instances.

5.5 Efficiency Analysis

Table 5 reports the total runtime of each method under a fixed machine configuration. RoBERTa corresponds to full training and inference, whose calibrated confidence identifies hard instances. LLM Predict adds the cost of directly classifying these instances with the LLM. SLM Reclassification rewrites them and reclassifies with the SLM, and the Confidence-Guided variant further selects between original and rewritten predictions. As expected, all strategies are more expensive than RoBERTa alone, as they require LLM inference. However, compared to LLM Predict, RAP introduces only a modest overhead while delivering *substantially higher effectiveness by up to 5.8% gains on topic datasets (ACM)* with improved transparency and interpretability. Importantly, the Confidence-Guided variant, which yields the best results, achieves a runtime comparable to SLM Reclassification while adding negligible calibration and lightweight selection steps. RAP is faster than LLM Zero-Shot across all datasets and delivers substantially higher effectiveness on most of them, particularly on topic classification tasks. Finally, LLM Few-Shot is the most expensive option, up to 3x-13x slower than RAP, while never surpassing it in effectiveness. **Overall, RAP delivers the best effectiveness-cost-interpretability trade-off among all methods.**

Table 5: Average total time (in seconds).

Dataset	RoBERTa	LLM Predict	SLM Reclassification	Confidence-Guided Decision Rule	LLM Zero-Shot (Full Test)	LLM Few-Shot (Full Test)
IMDB	2615	2709	3906	3906	14028	16480
Pang Movie	934	1441	1746	1746	6032	7395
SemEval2017	2416	2508	2604	2604	14179	19438
SST	1027	1214	1335	1335	6712	7733
SST2	5817	6272	6677	6678	37461	45453
Yelp2L	510	557	730	730	3013	9578
IMDB2024	681	859	1755	1755	3671	15814
RottenT2024	789	835	915	916	4566	6312
ACM	2665	3660	5010	5010	13829	21901
AGNews	1083	1735	2113	2114	7272	9061
DBLP	4140	4474	4939	4939	22455	27122
Reddit	986	1643	1761	1762	5606	22533
Twitter	651	1196	1521	1521	3818	4905

6 Conclusion

We introduced **RAP**, *Rewrite-Assess-Predict for Hard Instances*, a hybrid framework that combines SLM efficiency with zero-shot LLM adaptability. By identifying low-confidence documents, generating semantically faithful paraphrases that surface class-relevant information, and selecting predictions via calibrated confidence, RAP achieves substantial gains on hard cases while controlling computational cost. Results show that paraphrasing preserves semantics, enhances class salience, and improves SLM effectiveness without sacrificing reliability. Beyond quantitative gains, qualitative analysis demonstrates how rewriting clarifies decision evidence, improving interpretability. RAP offers a scalable, reliable, and interpretable alternative to end-to-end LLM classification, supporting sustainable hybrid NLP pipelines that allocate computation where it matters most. Future work includes extending RAP to multi-label and hierarchical classification, exploring alternative hard-instance criteria, and investigating more explanation-driven paraphrasing strategies.

Limitations

Although RAP significantly reduces cost relative to LLM-based pipelines, it still requires selective LLM inference, meaning its efficiency benefits arise from strategic usage rather than complete elimination of LLM computation. This implies that environments with extremely restrictive compute budgets may still prefer SLM solutions purely. Likewise, RAP relies on calibrated confidence to identify hard instances; while our results show that this strategy is effective and robust, any confidence-based method is ultimately bounded by calibration quality, and further advances in calibration could unlock even stronger gains.

Our evaluation focuses on sentiment and topic classification, which provide a clear and well-controlled setting for investigating the trade-off between efficiency, effectiveness, and interpretability. Extending RAP to more complex scenarios—such as multilingual settings, multi-label or hierarchical classification, or domains with higher subjectivity—remains promising future work rather than a limitation of current evidence. Finally, while our interpretability analysis reveals systematic and linguistically grounded rewriting behaviors, it is naturally shaped by the datasets studied. Broader validation across additional domains and user-centered interpretability assessments represent valuable next steps to strengthen RAP's explanatory potential further.

References

Ateret Anaby-Tavor, Boaz Carmeli, Eyal Goldbraich, Yoav Kantor, George Kour, Tomer Lavee, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2023. Improving text classification with large language model-based data augmentation. *Transactions of the Association for Computational Linguistics (TACL)*.

Claudio Moisés Valiense De Andrade, Washington Cunha, Guilherme Fonseca, Ana Clara Souza Pagano, Luana De Castro Santos, Adriana Silvina Pagano, Leonardo Chaves Dutra Da Rocha, and Marcos André Gonçalves. 2024. Explaining the hardest errors of contextual embedding based classifiers. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 419–434, Miami, FL, USA. Association for Computational Linguistics.

Jun Bai, Xiaofeng Zhang, Chen Li, Hanhua Hong, Xi Xu, Chenghua Lin, and Wenge Rong. 2023. How to determine the most powerful pre-trained language model without brute force fine-tuning? an empirical survey. In *Findings of the EMNLP 2023*.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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

Yuyan Chen, Zhihao Wen, Ge Fan, Zhengyu Chen, Wei Wu, Dayiheng Liu, Zhixu Li, Bang Liu, and Yanghua Xiao. 2024. Mapo: Boosting large language model performance with model-adaptive prompt optimization.

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

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

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302.

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 NAACL-HLT*, pages 4171–4186.

Yudong Ding, Xiaojian Ma, Chen Li, and Wei Wang. 2023. auggpt: Leveraging chatgpt for text data augmentation. In *Proceedings of the 61st Annual Meeting of the ACL*.

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.

Celso França, Gestefane Rabbi, Thiago Salles, Washington Cunha, Leonardo Rocha, and Marcos André Gonçalves. 2025. Optimizing tail-head trade-off for extreme multi-label text classification (xmtc)

with rag-labels and a dynamic two-stage retrieval and fusion pipeline. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 1392–1401, New York, NY, USA. Association for Computing Machinery.

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.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2013. *Halliday's introduction to functional grammar*. Routledge.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kamran Kowsari, Kian Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Xia Li, Junlang Wang, Yongqiang Zheng, Yuan Chen, and Yangjia Zheng. 2025. Paraphrase makes perfect: Leveraging expression paraphrase to improve implicit sentiment learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3631–3647, Abu Dhabi, UAE. Association for Computational Linguistics.

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*.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024. Mobilellm: optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

James R Martin and Peter R White. 2003. *The language of evaluation*, volume 2. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gabriel Prenassi, Guilherme Fonseca, Davi Reis, Washington Cunha, Marcos Gonçalves, and Leonardo Rocha. 2025. Um estudo comparativo de estratégias de seleção de exemplos para in-context learning aplicado à classificação automática de texto com grandes modelos de linguagem. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 921–927, Porto Alegre, RS, Brasil. SBC.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shouvon Sarker, Lijun Qian, and Xishuang Dong. 2023. Medical data augmentation via chatgpt: A case study on medication identification and medication event classification.

Arthur Mariano Rocha De Azevedo Scalercio, Elvis A. De Souza, Maria José Bocorny Finatto, and Aline Paes. 2025. Evaluating LLMs for Portuguese sentence simplification with linguistic insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24452–24477, Vienna, Austria. Association for Computational Linguistics.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, 34(1):1–47.

Vered Shwartz and Ido Dagan. 2018. Paraphrase to explicate: Revealing implicit noun-compound relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1200–1211.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, TzuHao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, Qi He, Yao Ma, Ming Huang, and Suhang Wang. 2025a. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2025b. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87.

Guangzhan Wang, Hongyu Zhang, Beijun Shen, and Xiaodong Gu. 2025c. Transplant then regenerate: A new paradigm for text data augmentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13917–13931, Suzhou, China. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

J. Wolfe, X. Jin, T. Bahr, and N. Holzer. 2017. Application of softmax regression and its validation for spectral-based land cover mapping. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W1:455–459.

Xiongwen Yang, Yi Xiao, Di Liu, Huiyou Shi, Huiyin Deng, Jian Huang, Yun Zhang, Dan Liu, Maoli Liang, Xing Jin, Yongpan Sun, Jing Yao, XiaoJiang Zhou, Wankai Guo, Yang He, Weijuan Tang, and Chuan Xu. 2025. Enhancing physician-patient communication in oncology using gpt-4 through simplified radiology reports: Multicenter quantitative study. *J Med Internet Res*, 27:e63786.

Murong Yue, Jie Zhao, Min Zhang, Liang Du, and Ziyu Yao. 2024. Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Yazhou Zhang, Mengyao Wang, Qiuchi Li, Prayag Tiwari, and Jing Qin. 2025. Pushing the limit of llm capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1524–1528.

Huanhuan Zhao, Haihua Chen, Thomas A. Ruggles, Yunhe Feng, Debjani Singh, and Hong-Jun Yoon. 2024. Improving text classification with large language model-based data augmentation. *Electronics*, 13(13).

A Dataset Statistics

Table 6 summarizes the datasets used in our study. For sentiment analysis, we include eight binary datasets spanning movie reviews, social media posts, and user-generated content, with varying document lengths and degrees of class balance. To mitigate benchmark contamination from large LLM training corpora, we additionally curate two recent post-LLM datasets (IMDB2024 and RottenT2024), ensuring temporal separation from pretraining data.

For topic classification, we rely on five datasets with diverse textual characteristics, ranging from short Twitter posts and news articles to longer academic titles and abstracts. Class distributions vary across datasets, with some exhibiting moderate imbalance, providing a realistic and challenging evaluation scenario.

	Dataset	Domain	Number Docs	Avg Words	Classes	Minor Class	Major Class
Sentiment	IMDB	Movie	24904	234	2	12432	12472
	PangMovie	Movie	10662	21.02	2	5331	5331
	SemEval17	Twitter	27413	19.85	2	7745	19668
	SST	Movie	11841	19.18	2	5905	5936
	SST2	Movie	66973	10.45	2	29643	37330
	Yelp2L	Place	4995	131.8	2	2495	2500
	IMDB2024	Movie	6572	163.02	2	2057	4515
	RottenT2024	Movie	7948	46.13	2	3315	4633
Topic	ACM	Article	24897	63.52	11	63	6562
	AGNews	News	12760	37.52	4	3190	3190
	DBLP	Article	38128	141.43	10	1414	9746
	Reddit	Reddit	10000	96.28	11	339	2346
	Twitter	Twitter	6997	28.68	6	152	2738

Table 6: Datasets Statistics.

B Model's Brier Score with softmax and isotonic

Table 7 reports Brier Scores for the SLM under raw softmax probabilities and after post-hoc calibration with isotonic regression. Across all sentiment and topic datasets, isotonic regression consistently improves probability reliability, yielding higher Brier Scores in every case. These results confirm that calibrated confidence provides a more dependable signal for identifying hard instances and supports our decision to adopt isotonic regression throughout RAP.

Dataset	Softmax	Isotonic
IMDB	0.953	0.957
PangMovie	0.800	0.840
SemEval17	0.887	0.901
SST	0.790	0.820
SST2	0.911	0.920
Yelp2L	0.968	0.977
IMDB2024	0.959	0.963
RottenT2024	0.910	0.916
ACM	0.662	0.680
AGNews	0.875	0.886
DBLP	0.747	0.761
Reddit	0.955	0.956
Twitter	0.828	0.833

Table 7: Model's Brier Score with softmax and isotonic.

C Metrics Equations

Table 8 summarizes the metrics and runtime formulations used in our evaluation. Equation 1 defines the F1 score averaged over all classes (Macro-F1). For cost-effectiveness, we model the total runtime T of each strategy as the sum of its constituent stages. Equation 2 represents the SLM baseline, including fine-tuning and inference. Equation 3 adds threshold estimation and direct LLM inference. Equation 4 incorporates the cost of generating paraphrases for hard instances followed by SLM reclassification, while Equation 5 further accounts

for calibration and confidence-based selection between original and rewritten predictions. Finally, Equation 6 represents the total time required to classify the entire test set using the Zero-shot LLM, while Equation 7 also reports the time to classify the whole test set, including the extra cost of retrieving similar examples to be incorporated into the prompt for the Few-Shot LLM version.

Equations	
Macro-$F_1 = \frac{1}{C}\sum_{i=1}^{C} F_{1,i}$	(1)
$T = Tu_{SLM} + P_{SLM}$	(2)
$T = Tu_{SLM} + P_{SLM} + F_L + P_{LLM}$	(3)
$T = Tu_{SLM} + P_{SLM} + F_L + G_{LLM} + P_{SLM}$	(4)
$T = Tu_{SLM} + P_{SLM} + F_L + G_{LLM} + P_{SLM} + C_{SLM}$	(5)
$T = P_{LLMZeroShot}(FullTest)$	(6)
$T = P_{LLMFewShot}(FullTest)$	(7)

Table 8: Metric equations. Time (T), Tuning (Tu), Find(F), Generate (G), Prediction (P), and C (Calibration).

D Prompt Optimization

This section describes the prompt optimization process. Figure 3 illustrates this procedure: an initial prompt generated by ChatGPT-4.0 is first evaluated to obtain a score. A meta-prompt, which includes both the current prompt and its score, is then used to generate a refined prompt aimed at improving effectiveness. This iterative process is repeated, and the prompt achieving the highest stored score (S) is selected as the final prompt.

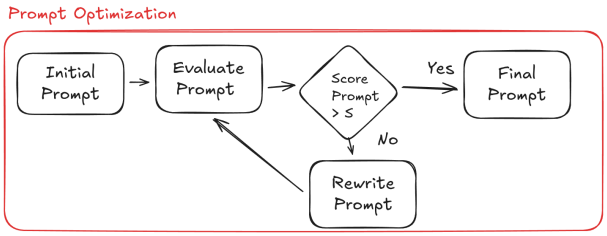

Figure 3: Overview of the prompt optimization workflow.

E Calibration

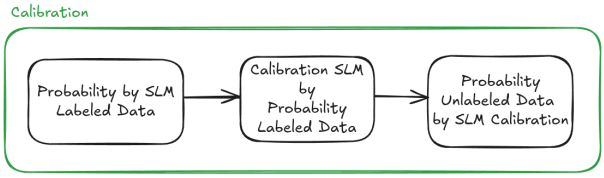

Figure 4: Overview of the calibration workflow.

This section describes the calibration procedure applied to the model. Figure 4 illustrates the overall process. First, we obtain the original class probabilities produced by the SLM via the softmax layer. Using the validation set probabilities, we train an isotonic regression calibrator, which is then applied to the test set probabilities to more reliably identify hard test instances. Table 7 shows that isotonic calibration consistently improves the Brier score across datasets, indicating better probability calibration. In addition to quantitative results, we provide a visual analysis for the SST dataset: Figure 5 shows the softmax outputs, while Figure 6 presents the isotonic-calibrated probabilities. The latter exhibits a stronger alignment between predicted confidence (x-axis) and empirical accuracy, confirming the effectiveness of the calibration process.

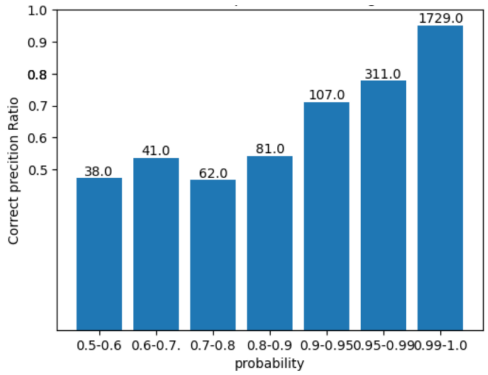

Figure 5: Softmax for SST dataset.

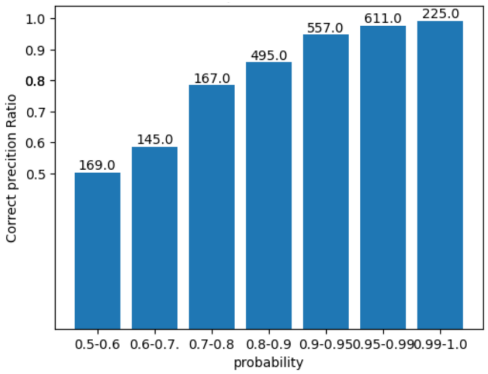

Figure 6: Isotonic for SST dataset.

Table 9: Emission CO_2. Calculation based on the work of Lacoste et al. (2019).

Dataset	**RoBERTa**	**LLM Predict**	**SLM Reclassification**	**Confidence-Guided Decision Rule**	**LLM Zero-Shot (Full Test)**	**LLM Few-Shot (Full Test)**
IMDB	0.1	0.11	0.15	0.15	0.55	0.67
Pang Movie	0.04	0.06	0.07	0.07	0.23	0.29
SemEval2017	0.09	0.1	0.1	0.1	0.55	0.76
SST	0.04	0.05	0.05	0.05	0.26	0.3
SST2	0.23	0.24	0.26	0.26	1.46	1.77
Yelp2L	0.02	0.02	0.03	0.03	0.12	0.37
IMDB2024	0.03	0.03	0.07	0.07	0.14	0.62
RottenT2024	0.03	0.03	0.04	0.04	0.18	0.25
ACM	0.1	0.14	0.19	0.19	0.54	0.85
AGNews	0.04	0.07	0.08	0.08	0.28	0.35
DBLP	0.16	0.17	0.19	0.19	0.87	1.02
Reddit	0.04	0.06	0.07	0.07	0.22	0.91
Twitter	0.03	0.05	0.06	0.06	0.15	0.19

Table 10: Average Macro-F1 (95% CI); **boldface** marks statistically superior results with Wilcoxon tests relative to the RoBERTa model.

Dataset	RoBERTa	LLM Predict	SLM Reclassification	Confidence-Guided Decision Rule	LLM Zero-Shot (Full Test)	LLM In-Context	Oracle
IMDB	93.0 (0.5)	**94.4 (0.8)**	**94.6 (0.8)**	**94.6 (0.7)**	**94.3 (0.6)**	92.1 (0.4)	96.0 (0.8)
Pang Movie	88.7 (0.9)	86.4 (3.2)	**90.3 (0.5)**	**90.6 (0.7)**	**91.0 (0.5)**	86.8 (1.1)	97.1 (0.8)
SemEval2017	91.2 (0.7)	91.2 (0.7)	**91.5 (0.8)**	**91.5 (0.8)**	89.7 (0.8)	84.8 (0.8)	92.5 (1.3)
SST	87.3 (1.0)	87.9 (1.1)	**88.2 (1.2)**	**88.2 (1.1)**	**89.4 (0.4)**	83.8 (0.8)	91.6 (1.1)
SST2	94.6 (0.2)	**95.1 (0.1)**	**94.9 (0.2)**	**94.9 (0.2)**	93.0 (0.5)	90.0 (0.9)	96.5 (0.1)
Yelp2L	97.9 (0.5)	97.0 (1.6)	**98.7 (0.4)**	**98.7 (0.3)**	**99.3 (0.5)**	**99.0 (0.3)**	99.3 (0.2)
IMDB2024	97.6 (1.0)	94.0 (3.8)	**98.4 (1.0)**	**98.5 (0.8)**	97.7 (0.9)	93.4 (0.5)	99.5 (0.5)
RottenT2024	93.7 (1.1)	**94.2 (1.3)**	**94.2 (0.9)**	**94.3 (1.1)**	**95.8 (0.8)**	91.0 (1.2)	95.3 (1.9)
ACM	70.7 (1.5)	67.2 (3.8)	70.8 (1.5)	**71.1 (1.6)**	29.5 (0.5)	**71.8 (2.2)**	78.0 (1.9)
AGNews	92.2 (0.4)	91.7 (0.6)	92.2 (0.4)	**92.4 (0.5)**	84.2 (0.8)	92.1 (0.4)	95.5 (0.9)
DBLP	**81.9 (0.7)**	74.4 (0.6)	**81.9 (0.6)**	**81.9 (0.6)**	41.8 (0.6)	58.4 (0.8)	82.8 (0.8)
Reddit	96.0 (0.5)	96.1 (0.9)	**96.7 (0.6)**	**96.8 (0.7)**	91.4 (0.6)	95.7 (0.7)	98.8 (0.4)
Twitter	77.5 (2.7)	77.6 (1.9)	**78.5 (2.4)**	**79.0 (2.7)**	65.6 (2.4)	**79.6 (3.5)**	90.9 (2.1)

F Carbon Cost (CO_2)

We additionally estimate the CO_2 emissions associated with model inference following the methodology of Lacoste et al. (2019). For this, we use a reference emission rate of approximately 0.14 kg CO_2 per GPU hour for hardware comparable to that used in our experiments[4]. As reported in Table 9, all solutions generate similar CO_2 emissions, with the alternatives that exploit the LLM, either for rewriting or for classification, having a slightly higher emission rate, especially for the topic datasets.

G Complete Result Table

Due to space limitations in the papers‘ main body, we omit the confidence intervals from Table 3. In this section, we report them in detail in Table 10.

[4] https://mlco2.github.io/impact/#co2eq