Journal of the Brazilian Computer Society, 2026, 32:1, doi: 10.5753/jbcs.2026.6153
This work is licensed under a Creative Commons Attribution 4.0 International License.

"Call My Big Sibling (CMBS)" – A Confidence-Based Strategy Leveraging Instance Selection to Combine Small and Large Language Models for Cost-Effective Text Classification

Claudio Moisés Valiense de Andrade [**Federal University of Minas Gerais** | ***claudio.valiense@dcc.ufmg.br***]
Washington Cunha [**State University of Campinas** | ***wcunha@unicamp.br***]
Davi Reis [**Federal University of Sao Joao del Rei** | ***davireisesus@aluno.ufsj.edu.br***]
Celso França [**Federal University of Minas Gerais** | ***celsofranca@dcc.ufmg.br***]
Wasterman Ávila Apolinário [**Federal University of Sao Joao del Rei** | ***wastermanavila@aluno.ufsj.edu.br***]
Luana de Castro Santos [**Federal University of Minas Gerais** | ***lcs2017@ufmg.br***]
Adriana Silvina Pagano [**Federal University of Minas Gerais** | ***apagano@ufmg.br***]
Leonardo Chaves Dutra da Rocha [**Federal University of Sao Joao del Rei** | ***lcrocha@ufsj.edu.br***]
Marcos André Gonçalves [**Federal University of Minas Gerais** | ***mgoncalv@dcc.ufmg.br***]

Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627 - Pampulha, Belo Horizonte, MG, 31270-901, Brazil.

Received: 22 June 2025 • **Accepted:** 02 March 2026 • **Published:** 05 May 2026

Abstract Transformers have achieved state-of-the-art results, with Large Language Models (LLMs) leading many NLP tasks. However, it remains unclear whether LLMs always outperform first-generation Transformers (*aka* Small Language Models, SLMs) across different text classification tasks and scenarios (e.g., movie reviews, topic classification). This study compares four SLMs (BERT, RoBERTa, Qwen, BART) with four open LLMs (LLaMA 3.1, Mistral, Falcon, DeepSeek) across nine sentiment and four topic classification datasets, totaling over 1000 results. Results show that open LLMs only moderately outperform or tie with SLMs when fine-tuned, and at a very high computational cost. To address this trade-off, we propose "Call My Big Sibling" (CMBS), a novel confidence-based framework that integrates *calibrated SLMs and open LLMs using advanced instance selection techniques*. CMBS assigns high-confidence instances to the cheaper SLM, while low-confidence instances are routed to an LLM in zero-shot, in-context, or partially tuned modes, optimizing cost-effectiveness. Experiments show CMBS significantly outperforms SLMs and delivers *LLM-level performance at a fraction of the cost*, offering a cost-sensitive solution for NLP applications.

Keywords: Text Classification, Large Language Model, Cost and Effective

1 Introduction

Automatic Text Classification (ATC) is a central task in Natural Language Processing (NLP), supporting applications such as sentiment analysis, topic categorization, and large-scale content organization. Recent advances in Transformer-based language models have substantially improved ATC performance, establishing strong encoder-based models such as BERT and RoBERTa as competitive and efficient baselines across diverse benchmarks [Devlin *et al*., 2019; Zanotto *et al*., 2021; de Andrade *et al*., 2023; Cunha *et al*., 2025b].

The emergence of Large Language Models (LLMs) has further reshaped the NLP landscape. Built on Transformer architectures but scaled to billions of parameters, LLMs achieve state-of-the-art (SOTA) performance on many generative tasks [Liang *et al*., 2023]. However, their advantages for discriminative tasks such as sentiment and topic classification remain unclear. Prior work reports mixed evidence on whether increased model scale yields statistically or practically meaningful gains over smaller Transformer-based models (a.k.a., Small Language Models (SLMs)), which are often better suited to resource-constrained settings [Cunha *et al*., 2023b; Zhang *et al*., 2024; Cunha *et al*., 2025b].

From an architectural perspective, LLMs encompass encoder-based, decoder-based, and hybrid encoder–decoder models. While these architectures differ in training objectives, all can be applied to ATC: encoder-based models typically rely on classification heads over learned representations, whereas decoder-based models cast classification as label generation via prompting. Beyond architecture, LLMs also differ in their degree of task adaptation, ranging from zero-shot and in-context learning to partially and fully fine-tuned approaches [Hu *et al*., 2021; Lepagnol *et al*., 2024; Han *et al*., 2024].

These developments raise a fundamental and still unresolved question: *Do open LLMs actually outperform strong SLM baselines in ATC, and if so, at what cost?* This motivates our first research question: RQ1: "Are (open) LLMs more effective than SLMs in sentiment and topic classification?" To investigate this, we evaluate four SLMs (BERT, RoBERTa, Qwen[1], BART) and four open LLMs

[1]We use the smallest Qwen version (0.5B parameters), which we consider an SLM due to its effectiveness and fast inference time.

(LLaMA 3.1 8B, Mistral 7B, Falcon 7B, DeepSeek 8B) across 9 sentiment and 4 topic datasets, including post-LLM releases to mitigate data contamination [Liang *et al*., 2023].

Effectiveness alone, however, is insufficient for model selection in practice. Fully fine-tuning LLMs incurs substantially higher computational and environmental costs, raising concerns about their feasibility for routine ATC deployment. This leads to our second research question: RQ2: "How does the computational cost of using open LLMs for ATC compare to SLMs' cost?" We analyze cost-effectiveness across zero-shot, in-context, partially tuned, and fully tuned strategies, considering training and inference time as well as carbon emissions. Our results show that fully fine-tuned LLMs are orders of magnitude more expensive than SLMs and often yield only moderate gains.

Current ATC approaches face a trade-off between efficient SLMs, which perform well on most instances but struggle with ambiguous cases, and LLMs, which better handle difficult instances at significantly higher computational cost. Existing methods typically rely on one model family in isolation, either overusing LLMs or limiting performance with SLMs. This reveals a methodological gap in leveraging their complementary strengths. These findings motivate our final research question: RQ3: "Is it possible to combine SLMs and (open) LLMs to achieve a better effectiveness–cost trade-off than using either alone?" To address this, we propose *Call My Big Sibling* (CMBS), a confidence-based strategy that leverages calibrated SLM predictions to selectively invoke LLMs only on low-confidence instances. High-certainty cases are handled efficiently by fully tuned SLMs, while challenging instances are deferred to zero-shot, in-context, or partially tuned LLMs. For the latter, CMBS integrates state-of-the-art instance selection techniques [Cunha *et al*., 2025a] to further reduce tuning costs while preserving effectiveness.

Our contributions are summarized as follows:

- A comprehensive empirical comparison of SLMs and open LLMs for ATC, focusing on effectiveness–cost trade-offs.
- The proposal of *Call My Big Sibling* (CMBS), a novel confidence-based framework that combines calibrated SLMs with selectively invoked LLMs.
- An extensive evaluation over 13 datasets spanning sentiment and topic classification, covering multiple architectures and adaptation strategies.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents CMBS. Section 4 describes the experimental setup, Section 5 discusses results, and Section 6 concludes the paper. Section 7 outlines limitations.

2 Related Work

This section reviews prior work addressing the computational efficiency, reproducibility, and trade-offs involved in using Large Language Models (LLMs) and Small Language Models (SLMs) for text classification tasks. We discuss studies focusing on the financial and environmental costs of large-scale models, the use of hybrid strategies combining LLMs and SLMs, and the importance of transparency in ensuring reproducible and responsible experimentation.

Several studies have emphasized the growing financial and environmental burden associated with training and deploying LLMs. Strubell *et al*. [2019] demonstrated how the escalating demand for specialized hardware substantially increases both monetary and energy costs, leading to greater CO_2 emissions and limiting accessibility to advanced models. These concerns have reinforced the need for cost-aware strategies that maintain effectiveness while reducing computational overhead.

A related line of research explores the combination of SLMs and LLMs to balance efficiency and accuracy. For instance, Xu *et al*. [2024] proposed a hybrid system in which both model types are employed to enhance effectiveness. However, their approach uses LLMs to classify the entire test set—without assessing computational cost or efficiency. In contrast, our method strategically forwards only low-confidence instances from the SLM to the LLM, significantly reducing inference costs while preserving accuracy. Moreover, their experiments rely on a single sentiment dataset and employ a closed API-based LLM, which limits transparency and control over computational factors.

Liang *et al*. [2023] conducted an extensive benchmarking study involving multiple LLMs across various tasks, datasets, and metrics. Similar to their work, we analyze the effectiveness–cost trade-off of LLMs. However, whereas their focus lies on breadth—covering multiple domains with limited depth—our approach provides a deeper investigation into sentiment and topic classification using multiple, diverse datasets. Additionally, unlike their evaluation, we include SLM baselines such as RoBERTa, a strong reference model for text classification [Bai *et al*., 2023a; Cunha *et al*., 2021, 2020; França *et al*., 2024; Belém *et al*., 2024]. Importantly, they do not propose a concrete mechanism for optimizing this trade-off, whereas our work explicitly addresses it.

Similarly, Yue *et al*. [2024] proposed a two-stage pipeline involving GPT-3.5 (as a weak model) and GPT-4 (as a strong model), where the stronger model is invoked when the weaker one's responses lack consistency. While this approach can improve accuracy, it primarily measures financial cost (API calls) and overlooks computational efficiency. In contrast, we base our decision policy on SLM confidence scores, invoking the LLM only when necessary and jointly considering time and financial costs.

The issue of model transparency and reproducibility has also gained attention. Proprietary, closed-source LLMs such as GPT models operate as black boxes, providing no access to architecture, training data, or parameters. This opacity hampers reproducibility, cost estimation, and fair comparison across experiments, while also raising privacy concerns due to API-based data transmission [Spirling, 2023]. Consequently, several authors advocate prioritizing open-source, locally executed LLMs for scientific evaluation, ensuring greater transparency and experimental control. Our study adheres to this principle by relying exclusively on open models whose costs can be directly measured and compared.

Finally, the emergence of DeepSeek [DeepSeek-AI *et al*., 2025], which achieves competitive performance while substantially reducing computational demands, underscores the

community's growing attention to sustainable modeling. We include DeepSeek among our evaluated LLMs, reflecting its relevance to the broader efficiency–effectiveness debate.

In summary, while prior research has explored hybrid use of LLMs and SLMs or large-scale benchmarking, none has systematically optimized the effectiveness–cost trade-off through confidence-based instance selection and transparent experimentation with open, locally executed models. Our work uniquely combines these dimensions to propose a reproducible, cost-efficient framework for text classification under resource constraints.

3 The CBMS Solution

One of the main contributions of our work is the proposal of a novel strategy to combine simpler, more efficient, but perhaps less effective SLMs with potentially more effective but costly LLMs, aiming to promote effectiveness while minimizing computational costs. Our solution, "Call-My-Big-Sibling" (CMBS), metaphorically evokes the image of a small (but smart) child who, in a challenging situation, seeks help from a bigger sibling. CBMS pursues the best trade-off between effectiveness and costs with a confidence-based pipeline of Language Models.

CMBS seamlessly integrates SLMs and (open) LLMs by leveraging instance selection and calibrated confidences. Figure 1 presents the workflow of our framework: the dataset is first split using k-fold cross-validation (Fig. 1(a)), enabling the training of *fully-tuned SLMs models*[2] (Fig. 1(b)), which are already highly effective in some classification tasks (and faster to tune compared to LLMs); the tuned SLM then generates predictions on the test split (Fig. 1(c)), and test documents whose confidence scores fall below a predefined threshold (a tunable parameter) are routed to an open LLM for classification—these are referred to as hard instances (Fig. 1(d))—whereas predictions with confidence above the threshold are directly accepted from the SLM; to classify hard instances, we consider five LLM configurations: Zero-shot, In-Context, Partially-Tuned-IS, biO-IS and Aggressive, and Fully-Tuned.

As mentioned above, hard instances are classified by one of the LLM variants. In the Zero-shot LLM setting (Fig. 1(e)), the document is classified using a prompt without any training examples, whereas in the In-Context LLM setting (Fig. 1(f)), the prompt includes the most similar training example to the instance being evaluated. In particular, in the case of partially-tuned LLMs, we have two options. Partially-Tuned-IS biO-IS (Fig. 1(g)) leverages the state-of-the-art just-released instance selection method biO-IS [Cunha *et al*., 2025a] to reduce the training set size, and thus the computational cost, while trying to maintain effectiveness. In biO-IS, the "optimal" training set reduction is variable and determined by its own algorithm [Cunha *et al*., 2025a]. Partially-tuned-IS Aggressive (Fig. 1(h)) always applies a randomly stratified reduction of 50% in the training set despite potential effectiveness losses in the LLM. We use this percentage, drawing on recent work in Instance Selection [Cunha *et al*., 2023a] that determined this is the maximum reduction rate that can still assure good efficiency while producing minimal effectiveness losses. In the final stage, documents that are not hard instances receive their final predictions directly from the SLM, while hard instances are classified using one of the four CMBS LLM-based variants (Fig. 1(i)).

For CBMS to properly work, we have to trust the probability outputs, or, in other words, the probabilities need to be calibrated[3]. Wolfe *et al*., 2017 argue that RoBERTa's softmax function provides calibrated probabilities as it is a generalization of logistic regression. To demonstrate this, Table 1 presents the Brier [BRIER, 1950] score, a scoring rule used to measure model calibration and the accuracy of probabilistic predictions. Brier [BRIER, 1950] defines $BS = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}(P(Y = y_c|x_i) - o_{ci})^2$, which we computed on the datasets used in our experiments with two transformers: BERT and RoBERTa. This score is calculated based on the model probabilities and actual labels. The score ranges from 0 to 1, with values closer to 1 indicating a better alignment between probabilistic predictions and actual outcomes.

Results in Table 1 confirm that RoBERTa exhibits a high degree of calibration (Brier score > 0.8), achieving levels comparable to well-calibrated classifiers such as Logistic Regression and Random Forests, which report similar scores on benchmark datasets [Cunha *et al*., 2025a]. Following the experimental protocol in [Cunha *et al*., 2025a], we employed the PangMovie, SST2, Yelp Review, DBLP, ACM, Twitter, and WebKB datasets, which collectively encompass a wide range of domains, including movie and business reviews, research abstracts, social media posts, and web page classification. Complementing Table 1, Figure 2 depicts the calibration curves for BERT (red) and RoBERTa (blue), with the dashed diagonal line representing perfect calibration. Across all datasets, RoBERTa's curve remains consistently closer to the ideal diagonal, corroborating the superior calibration performance indicated in Table 1.

Table 1. Brier score for BERT and RoBERTa.

Dataset	BERT	RoBERTa
Finance	0.784	0.989
IMDB	0.854	0.873
PangMovie	0.793	0.804
SemEval17	0.879	0.887
Sst	0.783	0.792
Sst2	0.919	0.912
Yelp2L	0.968	0.968
IMDB2024	0.944	0.959
RottenT2024	0.898	0.911
ACM	0.691	0.662
DBLP	0.765	0.747
Twitter	0.810	0.828
Webkb	0.779	0.80

We identify a subset of documents for which the SLM exhibits low classification confidence (i.e., predicted probability < L) and forward them to an LLM for final prediction. To manage computational costs, the LLM operates under one of three regimes: zero-shot, in-context, or partially-tuned

[2]Tuned with the full training data.

[3]A calibrated classifier has a strong correlation between class prediction probabilities and frequency of correctly predicted instances belonging to each probability range.

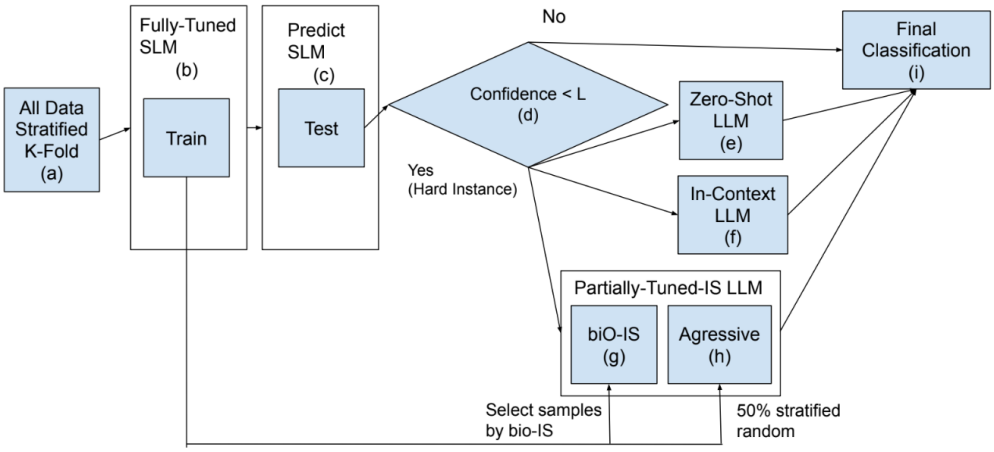


Figure 1. CBMS FlowChart.

inference. The final prediction set is constructed through the following procedure: (1) compute the confidence score produced by the SLM and compare it against the threshold L; and (2) determine whether the instance should be classified by the SLM or escalated to the LLM, employing the selected inference strategy (zero-shot, in-context, or partially-tuned).

In the proposed framework, the selection of the confidence threshold L is crucial in determining which documents are forwarded to the LLM for reclassification. To illustrate this relationship, Figure 3 depicts the variation in model effectiveness as a function of prediction confidence on the SST2 dataset. The Y-axis represents RoBERTa's effectiveness, while the X-axis corresponds to its confidence scores. As shown, higher confidence levels are consistently associated with greater predictive accuracy. This observation underscores the importance of appropriately setting the threshold L, as forwarding only low-confidence instances to the LLM is more cost-effective—given that high-confidence predictions are already handled accurately by the more efficient SLM.

3.1 Zero-shot, In-context, Partially or Fully-tuned Strategies for ATC

Within the CBMS framework, the application of SLMs or LLMs to ATC can be carried out through five strategies: *zero-shot*, *in-context*, *partially-tuned–IS biO-IS*, *partially-tuned–IS Aggressive*, and *fully-tuned*. In the *zero-shot* strategy, the model predicts text classes without access to labeled examples or fine-tuning. The *in-context* approach, in turn, leverages a prompt containing the nearest neighbors of the evaluated instance to provide contextual information for prediction without updating model parameters. This nearest-neighbor mechanism has been widely employed in the construction of Retrieval-Augmented Generation (RAG) systems [Xu *et al*., 2025; Wu *et al*., 2025].

The *partially-tuned (IS)* strategies fine-tune the model using only labeled data subset, simulating a data-scarce setting. As mentioned, we explore two variants: (i) *biO-IS*, which reduces the training set through bi-objective instance selection at dataset-dependent reduction rates; and (ii) *Aggressive*, in which 50% of the training partition is used, selected through stratified random sampling. Additional experiments with varying training set sizes are found in Appendix A, further supporting the choice of the 50% reduction rate.

The *fully-tuned* strategy employs all available labeled data in the training partition to maximize task-specific adaptation. This process involves updating all model weights using the complete training set, with a softmax classification layer whose output dimensionality matches the number of target classes [Hu *et al*., 2021]. Although this approach generally yields the highest effectiveness, it incurs substantially greater computational cost. In this work, the fully-tuned configuration is adopted as a *baseline* for comparison with the proposed CBMS strategies.

For SLMs, we employ only the fully-tuned configuration, as this step is essential for achieving high effectiveness [de Andrade *et al*., 2023]. In this setting, fine-tuning consists of optimizing the SLM's text representation (CLS token) along with a fully connected layer responsible for class prediction, using all available training samples.

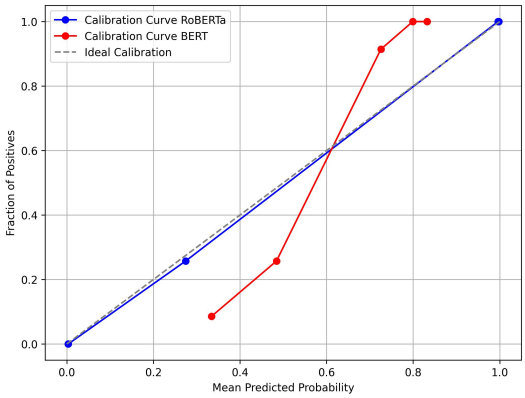
4 Experimental Methodology and Setup

In this section, we describe the experimental methodology adopted in this study. Section 4.1 presents the datasets used in our experiments. Section 4.2 describes the prompts submitted to the LLMs. Section 4.3 details the parameters employed for both SLMs and LLMs. Finally, Section 4.4 outlines the evaluation metrics and the experimental protocol.

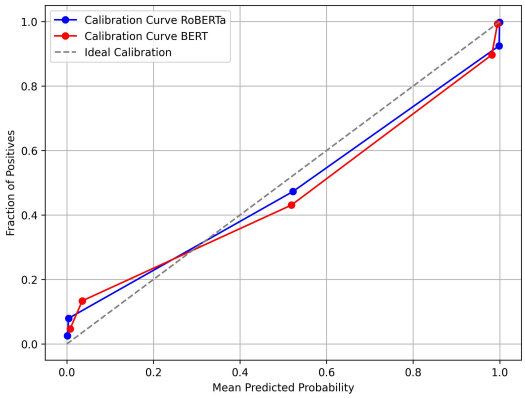
4.1 Datasets

Our study draws on thirteen datasets developed for topic and sentiment classification. Our choice was strategic because we aimed to conduct an in-depth analysis of this task. The datasets include Finance [Malo *et al*., 2014] focusing on economic news, IMDB [Maas *et al*., 2011][4] compiling movie reviews as well as PangMovie [Pang and Lee, 2005] including Rotten Tomatoes[5] data, SemEval17 [Rosenthal *et al*., 2019] containing Twitter texts used in a significant text classification challenge, and the Stanford Sentiment Treebank (SST) [Socher *et al*., 2013] and SST2 [Socher *et al*.,

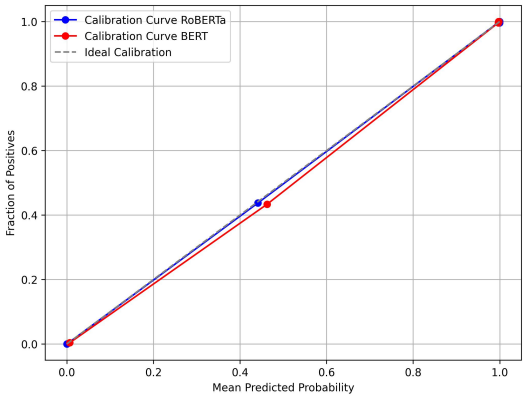
[4] https://www.imdb.com/
[5] https://www.rottentomatoes.com/



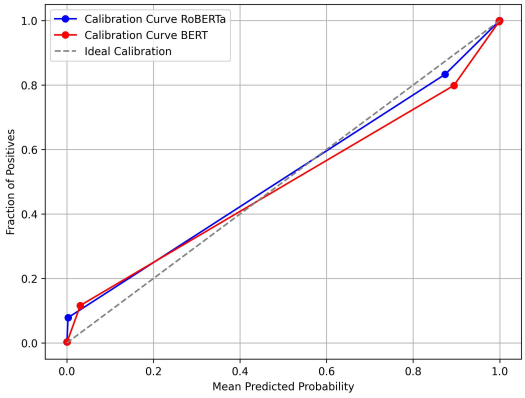
(a) Finance



(b) Pang Movie



(c) IMDB2024



(d) RottenT2024

Figure 2. Calibration curve for BERT and RoBERTa.

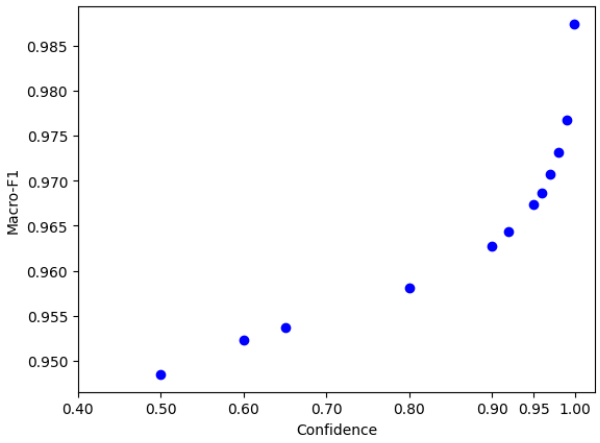


Figure 3. RoBERTa's Macro-F1 vs Confidence for SST2.

2013], where sentiment classification relies on a *treebank*, a corpus with sentiment labels and labeled parse trees. Yelp Review is a subset of Yelp data widely used in sentiment classification studies [Canuto *et al*., 2016; Viegas *et al*., 2023; Mendes *et al*., 2020]. IMDB2024 and RottenT2024 were collected to avoid data contamination by LLM. For topic classification, we have ACM Digital Library [Cunha *et al*., 2021], DBLP [Tang *et al*., 2008], Twitter Topic [Antypas *et al*., 2022] and WebKB [Craven *et al*., 1998].

In Table 2, we can observe an ample diversity in many aspects of these datasets: domain, number of documents (|D|), density (the average number of words per document), etc.

4.2 Prompt Template

We evaluated four open LLMs (Falcon 7B, Mistral 7B, DeepSeek 8B, and LLaMA 3.1 8B) using the prompt template from Liang *et al*. [2023], who found that the most effective format includes: (i) task description, (ii) examples with expected responses, and (iii) the text to be evaluated. We adapted this for sentiment (Table 3) and topic classification (Table 4). The prompt provides instructions, class examples, and the text to evaluate, with the LLM generating the class as the "next word." The In-Context LLM template (Table 5) modifies this by replacing a generic example with the closest training document, determined via cosine similarity using RoBERTa embeddings.

The following examples illustrate the prompt structures adopted in our experiments for the Zero-Shot and In-Context configurations of the evaluated LLMs. Owing to its simplicity and widespread adoption, the prompt design follows the approach proposed by Liang *et al*. [2023]. Prior studies [Wang *et al*., 2025; Sorensen *et al*., 2022] have shown that the most effective prompts typically integrate clear task instructions, representative examples, and the input text to be analyzed. Although more elaborate prompt engineering could potentially enhance LLM performance, our design choice prioritized simplicity and comparability with previous work. This decision aligns with our primary objective of evaluating the proposed method rather than optimizing prompt formulation. Future research will examine the influence of more sophisticated prompt designs on model performance, with the aim of further refining and generalizing our findings.

Table 3 provides the prompt used for sentiment classification, while Table 4 presents the prompt for topic classification.

Table 2. Datasets Statistics.

	Dataset	Domain	|D|	Avg Words	Classes	Minor Class	Major Class
Sentiment	Finance	Finance	873	24.88	2	303	570
	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
	Rot-tenT2024	Movie	7948	46.13	2	3315	4633
Topic	Acm	Article	24897	63.52	11	63	6562
	Dblp	Article	38128	141.43	10	1414	9746
	Twitter	Twitter	6997	28.68	6	152	2738
	Webkb	Pages	8199	208.81	7	137	3705

Both prompts include the tag [Evaluate Text], which represents the (test) text to be classified, and the tag [Response from LLM], which contains the model's output. If the model's output does not match any of the given alternatives (due to hallucination), we predict the majority class from the training set.

We employ a straightforward heuristic that considers only the token with the highest predicted probability. While this approach offers simplicity, we recognize its limitations and intend to explore alternative methodologies that incorporate a broader spectrum of high-probability tokens, thereby transcending the constraints of a singular, most likely prediction.

The third example of prompt, shown in Table 5, is tailored for in-context learning. For the evaluated test document, "I spent a day at a 5-star hotel, which was amazing." the most similar example from the training set included in the prompt was "5-star hotels have many food options.". A vector representation is generated for each evaluated example using the fully-tuned RoBERTa as an encoder. By comparing the vector of the evaluated (test) document with the vectors of the training set documents, we identify the most similar document based on the cosine similarity between the vectors and use it as a training example in the prompt.

This similarity-based retrieval step effectively transforms the approach into a form of Retrieval-Augmented Generation (RAG), in which relevant information from the training set is retrieved and integrated into the prompt without altering the model's internal parameters. The process relies on a nearest-neighbor search algorithm, consistent with prior studies that employ this technique to construct RAG-based systems [Xu *et al*., 2025; Wu *et al*., 2025]. Accordingly, we adopt the RAG terminology to characterize this mechanism throughout our work, emphasizing its role in enhancing contextualization without explicit fine-tuning.

Table 3. Prompt template for sentiment classification.

Classify the sentiment in the text exclusively as positive or negative:
Input: I love you.
Reference:
A. Positive
B. Negative
Answer: A
Input: The product is bad.
Reference:
A. Positive
B. Negative
Answer: B
Input: {**Evaluate Text**}
Reference:
A. Positive
B. Negative
Answer: {**Response from LLM**}

Table 4. Prompt template for topic classification.

Classify the topic of the text exclusively with one of the references:
Input: Messi scored a goal against France.
Reference:
A. Pop culture
B. Sports or gaming
C. Daily life
D. Science or technology
E. Business or entrepreneurs
F. Arts or culture
Answer: B
Input: {**Evaluate Text**}
Reference:
A. Pop culture
B. Sports or gaming
C. Daily life
D. Science or technology
E. Business or entrepreneurs
F. Arts or culture
Answer: {**Response from LLM**}

4.3 Method-Specific Parameter Tuning

All data is divided using stratified 5-fold cross-validation, a widely accepted technique in model evaluation. This method enhances the robustness and reliability of the model by splitting the dataset into five parts: three for training, validation, and testing. In each of the five iterations, the roles of the partitions alternate between training, validation, and testing,

Table 5. Prompt template for sentiment classification for In-Context Llama and CMBS In-Context.

Classify the sentiment in the text exclusively as positive or negative: Input: 5-star hotels have many food options. Reference: A. Positive B. Negative Answer: A Input: I spent a day at a 5-star hotel, which was amazing. Reference: A. Positive B. Negative Answer: {**Response from LLM**}

ensuring that the class distribution is preserved in the test partition. The validation set is crucial for parameter tuning, as detailed below.

For SLMs, we adopted Cunha *et al*. [2023b]'s hyperparameterization, fixing the learning rate in 2×10^{-5}, the batch size with 64 documents, adjusting the model for five epochs and set the maximum size of each document to 256 tokens. We adopted the following parameters for the LLM models: all LLMs use 4-bit quantization, with QLoRA and PEFT enabling fine-tuning on reasonably equipped machines. For LLaMA, we used 1024 maximum tokens, a learning rate of 2×10^{-4}, and a temperature equal to 0.6. All other parameters were set at their default values. For fully-tuning processes, which are more costly due to the model's weight adjustment (backpropagation), we had to reduce the maximum number of tokens to 256. We performed training for three epochs.

We introduce a confidence threshold parameter: if the SLM's confidence is below this threshold, predictions are forwarded to the LLM, the "Big Sibling". This ensures that the complex LLM takes over only when the SLM is not confident. Using the validation set, we vary this parameter to optimize Macro-F1 without increasing cost. Table 15 shows the selected threshold for sample datasets, the percentage of forwarded instances, and LLM/SLM effectiveness. For example, in SST2, documents with confidence below 0.9 go to the LLM; otherwise, the SLM classifies them. Higher thresholds mean more documents are sent to the LLM. Notably, the optimal threshold (around 0.9) is consistent across most datasets, and the LLM outperforms the SLM in these difficult cases, supporting CBMS gains.

4.4 Metrics and Experimental Protocol

We evaluated SLMs and (open) LLMs regarding the effectiveness/cost tradeoff. All models were assessed on an identical hardware configuration and executed similarly across all experiments, allowing for a fair comparison between the solutions: a 4-core processor, 32GB of system memory, and an Nvidia Tesla P100 GPU. Classification effectiveness is assessed using Macro-F1 (Equation 1) due to imbalance in several datasets. To ensure statistical validity of the results and demonstrate model generality, models were evaluated using the test set from a 5-fold stratified cross-validation methodology and a t-test with 95% confidence with Bonferroni correction to account for multiple comparisons.

To assess the cost–effectiveness trade-off, we evaluate the total model-building time for each method, encompassing both the learning phase and the subsequent test set predictions. Regarding computational time, we first report the total duration required to obtain the results for `RoBERTa` (SLM), which includes fine-tuning and test-set inference, as expressed in Equation 2. For the zero-shot configuration, the time associated with `LLaMA` (Equation 3) corresponds to the direct inference of the entire test set without any task-specific adaptation. The in-context `LLaMA` time (Equation 4) reflects the process of identifying the most semantically similar document to the evaluated instance using a k-nearest neighbors (k-NN) retrieval and generating predictions from the composite prompt containing both documents. For the partially-tuned `LLaMA`, the time measurement includes fine-tuning on a subset of labeled data—under either the *biO-IS* or *Aggressive* configuration—followed by test-set inference, as formalized in Equation 5. Finally, the fully-tuned `LLaMA` time corresponds to model training over the complete training set and subsequent prediction of the test instances.

For the `CMBS` variants, `CMBS Zero-Shot` involves tuning the SLM, determining the optimal threshold L to identify the subset of hard instances, predicting the majority of the test set with the SLM, and delegating a smaller portion to the LLM for zero-shot inference, as formalized in Equation 6. The `CMBS In-Context` configuration additionally accounts for the retrieval of the k nearest neighbors for each hard instance (Equation 7), which are incorporated into the prompt to provide contextual grounding for the LLM's predictions. Finally, the `CMBS Partially-Tuned` variant includes the fine-tuning of the LLM using either the *biO-IS* instance-selected subset or a 50% stratified random sample of the training data in the *Aggressive* configuration, as defined in Equation 8.

Table 6. Metric equations. Due to space constraints, we abbreviated the words Time (T), Tuning (Tu) and Prediction (P).

Equations
Macro-$F_1 = \frac{1}{C}\sum_{i=1}^{C} F_{1,i}$(1)
$T = Tu_{SLM} + P_{SLM}$(2)
$T = P_{LLM}$(3)
$T = KNN + P_{LLM}$(4)
$T = Tu_{LLM} + P_{LLM}$(5)
$T = Tu_{SLM} + Find_L + P_{SLM} + P_{LLM}$(6)
$T = Tu_{SLM} + Find_L + P_{SLM} + KNN + P_{LLM}$(7)
$T = Tu_{SLM} + Find_L + P_{SLM} + Tu_{LLM} + P_{LLM}$(8)

5 Results and Discussion

We now present and discuss the experimental results. Section 5.1 evaluates four SLMs, demonstrating that *RoBERTa* achieves the highest overall effectiveness among the models considered. Section 5.2 subsequently examines four LLMs, with *LLaMA* exhibiting the most stable and consistent performance across datasets. In Section 5.3, we address **RQ1** by comparing the effectiveness of SLMs and multiple LLM-based strategies. Section 5.4, we address **RQ2** by analyzes computational costs associated with both SLM and LLM approaches. Section 5.5, we address **RQ3** by introducing our proposed `CMBS` framework and discusses its results in terms

Table 7. Average Macro-F1 and 95% confidence interval for SLMs . Best results (including statistical ties) marked in **bold**.

Dataset	BERT	BART	Qwen 0.5B	RoBERTa
Finance	94.1±3.8	**97.0±1.7**	67.8±14.2	**98.1±1.9**
Imdb	91.7±0.4	**92.8±0.4**	81.6±1.3	**93.0±0.5**
PangMovie	87.5±0.7	**88.4±1.0**	72.5±1.8	**88.7±0.9**
SemEval17	90.3±0.3	**91.0±0.4**	79.9±0.7	**91.2±0.7**
Sst	86.1±0.4	**87.7±1.1**	69.5±8.7	**87.3±1.0**
Sst2	**94.8±0.1**	94.2±0.3	87.9±0.7	**94.6±0.2**
Yelp2L	96.8±0.4	**97.7±0.2**	89.4±3.7	**97.9±0.5**
IMDB2024	96.6±0.5	**97.5±0.6**	90.2±2.2	**97.6±1.0**
RottenT2024	92.5±1.0	**93.5±0.5**	84.6±0.8	**93.7±1.1**
ACM	**69.8±1.8**	68.0±2.8	46.4±13.7	**70.7±1.5**
DBLP	**82.1±0.9**	**81.9±0.6**	68.5±8.4	**81.9±0.7**
Twitter	**76.6±4.4**	**76.9±3.3**	39.1±5.7	**77.5±2.7**
Webkb	**80.8±3.8**	**81.7±3.5**	45.9±13.6	**82.3±2.6**

Table 8. Average Macro-F1 for sentiment and topic classification tasks with the LLMs in Zero-shot version - Falcon 7B, Mistral 7B, and Llama 3.1 8B. Best results (including statistical ties) in **bold**.

Dataset	Falcon	Mistral	Llama 3.1	DeepSeek
Finance	46.7±4.8	**94.3±1.9**	**95.4±1.2**	**95.6±2.1**
Imdb	68.4±0.7	68.4±0.7	**93.0±0.3**	84.2±0.5
PangMovie	43.6±0.5	82.3±0.9	**88.8±0.9**	82.1±0.5
SemEval17	54.4±0.6	81±0.9	**89.7±0.6**	87.2±0.9
Sst	47.0±1.2	82±0.8	**87.9±0.7**	82.3±0.8
Sst2	38.6±0.1	86.2±0.5	**91.4±0.4**	84.4±0.6
Yelp2L	79.9±1.3	96.2±0.9	**98.6±0.3**	96.4±0.6
IMDB2024	78.4±0.8	94.9±0.9	**96.5±1.0**	94.6±0.8
RottenT2024	65.8±1.3	93.8±1.2	**95.3**±1.0	92.1±1.0
ACM	2.6±0.2	18.2±0.9	**35.6±1.1**	14.2±0.4
DBLP	3.1±0.2	50.2±0.6	**53.7±0.8**	39.2±0.5
Twitter	13.0±0.3	**62.2±2.1**	**63.5±1.7**	**63.6±1.2**
Webkb	3.8±0.3	**42.1±0.6**	37.0±2.1	24.1±3.5

of both effectiveness and efficiency. Section 5.6 investigates the influence of the confidence threshold on predictive performance and on the proportion of instances delegated to the LLM. Finally, section 5.7 extends the evaluation of CMBS to a distinct NLP task beyond sentiment and topic classification.

5.1 Evaluating SLMs

We evaluate four fully-tuned SLMs: BART, BERT, Qwen 0.5B and RoBERTa. Table 7 presents the results regarding Macro-F1, with the best outcomes highlighted in **bold**. As observed, RoBERTa is consistently the best performer, alone or tied with another SLM, across **all** datasets, with no exception, confirming findings reported in the literature [Cunha *et al*., 2025b; Bai *et al*., 2023b; Fonseca *et al*., 2025].

Due to its effectiveness, we selected RoBERTa for evaluation in the subsequent stages, including comparisons with LLMs and their training variants, as well as for inclusion as the SLM component in our proposed combination approach.

5.2 Evaluating LLMs

We evaluate four Large Language Models (LLMs) in the zero-shot setting (Falcon 7B, Mistral 7B, LLaMA 3.1-8B, and DeepSeek 8B) on sentiment and topic classification tasks. Table 8 reports their Macro-F1 scores, with the best results highlighted in bold. Among them, LLaMA 3.1-8B consistently achieves the highest or tied performance across all datasets except WebKB, demonstrating superior robustness and generalization. Considering both its strong effectiveness and the high computational demands of full LLM fine-tuning, we select LLaMA 3.1-8B as the representative LLM for all subsequent analyses. Accordingly, the following sections focus on LLaMA-based configurations and its role in handling hard instances within our proposed framework.

5.3 RQ1: SLMs vs. LLMs - Effectiveness

To address RQ1, we compare RoBERTa with five LLaMA 3.1 variants—zero-shot, in-context, partially-tuned (biO-IS and Aggressive), and fully-tuned—using the complete training set (Table 9). Zero-shot LLaMA performs comparably or worse than RoBERTa on sentiment datasets and substantially worse on topic classification. Similarly, the in-context variant fails to surpass RoBERTa in sentiment analysis. These results are expected, as RoBERTa is fully fine-tuned on labeled data, whereas zero-shot and in-context LLaMA operate without explicit supervision. The performance gap widens for topic classification due to its higher class cardinality and intrinsic complexity.

Among the LLaMA variants, only the partially-tuned (biO-IS and Aggressive) and fully-tuned models outperform

Table 9. Average Macro-F1 and 95% confidence interval for SLMs and versions Llama 3.1 8B. Best results (including statistical ties) are marked in **bold**.

Dataset	RoBERTa	Zero-Shot LLaMA	In-Context LLaMA	Partially-Tuned-IS biO-IS LLaMA	Partially-Tuned-IS Agressive LLaMA	Fully-Tuned LLaMA
Finance	**98.1±1.9**	95.4±1.2	**98.6±1.8**	88.6±1.3	**98.6±0.1**	**98.7±1.6**
Imdb	93.0±0.5	93.0±0.3	78.9±1.2	**95.7±0.3**	**95.8±0.2**	**95.9±0.4**
PangMovie	88.7±0.9	88.8±0.9	89.9±0.7	**93.5±0.4**	**93.1±0.4**	**93.7±0.5**
SemEval17	91.2±0.7	89.7±0.6	90.1±0.7	**92.7±0.6**	**92.7±0.6**	**93.5±0.3**
Sst	87.3±1.0	87.9±0.7	88.5±1.0	**90.7±0.9**	**90.9±0.8**	**91.1±1.0**
Sst2	94.6±0.2	91.4±0.4	93.5±0.4	**95.8±0.3**	**95.7±0.2**	**96.0±0.1**
Yelp2L	**97.9±0.5**	**98.6±0.3**	92.1±1.0	**98.7±0.3**	**98.5±0.6**	**98.5±0.5**
IMDB2024	97.6±1.0	96.5±1.0	93.9±1.0	**98.4±0.7**	**98.6±0.7**	**98.7±0.7**
RottenT2024	93.7±1.1	95.2±1.4	95.3±1.0	**96.3±0.4**	**96.6±0.7**	**96.7±0.4**
ACM	70.7±1.5	35.6±1.1	50.5±1.6	**74.6±2.7**	72.4±1.6	**76.6±2.1**
DBLP	81.9±0.7	53.7±0.8	53.2±1.0	**86.7±1.0**	85.9±0.8	**87.8±0.7**
Twitter	**77.5±2.7**	67.4±2.7	72.9±1.6	70.4±2.3	73.5±3.1	**77.7±2.5**
Webkb	**82.3±2.6**	41.9±1.5	64.0±1.8	83.4±1.0	82.4±2.1	**86.0±1.3**

Table 10. Average Total Time for RoBERTa and versions of Llama3.1-8B. Best results are marked in **bold**.

Dataset	RoBERTa	Zero-Shot LLaMA	In-Context LLaMA	Partially-Tuned-IS biO-IS LLaMA	Partially-Tuned-IS Agressive LLaMA	Fully-Tuned LLaMA
Finance	**79**	103	123	514	484	896
Imdb	**2615**	6295	11548	33554	25176	39257
PangMovie	**934**	1200	1490	7169	5892	10921
SemEval17	**2416**	3160	4251	19541	15154	28087
Sst	**1027**	1230	1562	7913	6544	11791
Sst2	**5817**	7800	10936	47813	37435	65428
Yelp2L	**510**	1161	1736	3743	2407	5116
IMDB2024	**681**	1623	2538	9015	5822	12304
RottenT2024	**789**	983	1708	5766	4393	8130
ACM	**2665**	3163	7896	18539	16877	28207
DBLP	**4140**	8113	17311	47501	27564	139250
Twitter	**651**	892	1478	7105	6664	11406
Webkb	**910**	2877	3274	12547	10150	26021

RoBERTa in most datasets, with the fully-tuned version showing a clearer advantage, particularly in topic classification. Nonetheless, in certain cases—such as Finance and Yelp2L (sentiment) and Twitter (topic)—RoBERTa and fully-tuned LLaMA exhibit statistically similar performance, largely due to RoBERTa's wider confidence intervals indicating higher variability.

The instance selection (IS) versions of the LLM are also very competitive with the fully-tuned version in most datasets, but at a much cheaper cost (between 30%-50%). When comparing both IS alternatives, we can see that biO-IS preserves effectiveness in more datasets, which is consistent with Cunha *et al*. [2025a], probably due to its lower training set reduction rate (around 40% on average when compared to Aggressive (always 50%). These results further motivate us to combine SLMs and LLMs with our proposed CMBS pipeline for the sake of optimizing the effectiveness-cost trade-off. This trade-off is the core of our subsequent analyses.

5.4 RQ2: SLMs vs. LLMs - Computational Cost

Table 10 presents the total time (in seconds) required to obtain final predictions for each solution. The Table shows that RoBERTa's time is the shortest, followed by LLM Zero-Shot, which is around 76% more expensive than the SLM, on average. LLM In-context, in turn, is 176% slower than RoBERTa and 56% costlier than LLM Zero-Shot.

In the Partially-Tuned-IS setting, computational cost rises sharply due to gradient-based weight updates during LLM fine-tuning. The Aggressive variant is consistently 9–35% faster than biO-IS, reflecting its greater reduction of the training set and consequently shorter optimization time. As expected, the fully-tuned LLM is by far the most expensive configuration—approximately 1700% costlier than RoBERTa. Despite yielding an average effectiveness gain of only 3.3% across datasets (reaching 8.3% in ACM), such improvements may not justify the substantial computational overhead. In practical scenarios, these costs can render full fine-tuning infeasible, motivating our proposed

Table 11. Average Macro-F1 and 95% confidence intervals for RoBERTa, versions of CMBS and Fully-Tuned LLaMA. Best results (including statistical ties) are marked in **bold**.

Dataset	RoBERTa	CMBS Zero-Shot	CMBS In-Context	CMBS Partially-Tuned-IS biO-IS	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLaMA
Finance	**98.1±1.9**	**98±2.1**	**98.2±1.7**	**97.7±1.6**	**98.3±1.3**	**98.7±1.6**
Imdb	93±0.5	94±0.6	92.5±0.6	**95.7±0.3**	**95.8±0.2**	**95.9±0.4**
PangMovie	88.7±0.9	90.2±0.9	89.9±0.8	**93.5±0.4**	**93.1±0.3**	**93.7±0.5**
SemEval17	91.2±0.7	92.2±0.6	92±0.5	**92.9±0.5**	**92.9±0.6**	**93.5±0.3**
Sst	87.3±1	89±0.6	88.5±1.2	**90.8±0.9**	**90.9±0.9**	**91.1±1**
Sst2	94.6±0.2	95.1±0.2	94.8±0.3	**95.8±0.2**	**95.7±0.2**	**96±0.1**
Yelp2L	97.9±0.5	**98.5±0.2**	**98.1±0.2**	**98.8±0.2**	**98.6±0.5**	**98.5±0.5**
IMDB2024	97.6±1	**98.2±0.9**	97.3±1.2	**98.5±0.6**	**98.7±0.8**	**98.7±0.7**
Rot-tenT2024	93.7±1.1	95.6±1	95.7±0.7	**96±0.7**	**96.3±0.7**	**96.7±0.4**
ACM	70.7±1.5	70.5±1.2	70.6±1.2	**74.7±2.7**	**73.3±2.4**	**76.6±2.1**
DBLP	81.9±0.7	81.9±0.6	82±1.6	**86.7±0.9**	86±0.8	**87.8±0.7**
Twitter	**77.5±2.7**	**79.4±2.7**	**78.7±2.5**	**77.7±2.2**	**78.2±1.8**	**77.7±2.5**
Webkb	**82.3±2.6**	82.1±2.3	**82.2±2.7**	**83.6±1.1**	**83.8±2.5**	**86±1.3**

Table 12. Average Total Time for RoBERTa and versions of CMBS and Fully-Tuned LLaMA. Best results are marked in **bold**.

Dataset	RoBERTa	CMBS Zero-Shot	CMBS In-Context	CMBS Partially-Tuned-IS biO-IS	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLM
Finance	**79**	84	89	542	515	896
Imdb	**2615**	2930	3245	32813	25273	39257
PangMovie	**934**	994	1054	7386	6237	10921
SemEval17	**2416**	2574	2732	20004	16055	28087
Sst	**1027**	1089	1150	8149	6917	11791
Sst2	**5817**	6207	6597	48848	39508	65428
Yelp2L	**510**	568	626	3879	2676	5116
IMDB2024	**681**	762	844	8795	5921	12304
Rot-tenT2024	**789**	838	887	5978	4743	8130
ACM	**2665**	2823	2981	19349	17854	28207
DBLP	**4140**	4546	4951	46891	28948	139250
Twitter	**651**	695	740	7045	6648	11406
Webkb	**910**	1054	1198	12202	10045	26021

approach, which aims to retain LLM-level effectiveness while significantly reducing computational demands.

5.5 RQ3: Evaluating CBMS

Focusing now on our proposals, we assess the four CBMS implementations: CMBS Zero-Shot, In-Context, Partially-Tuned-IS (bio-IS) and Partially-Tuned-IS (Aggressive). Starting with sentiment classification, Table 11 presents results for RoBERTa, each CMBS version, and Fully-Tuned LLaMA. CMBS Zero-Shot outperforms RoBERTa in 8 out of 9 sentiment datasets, tied only in the Finance dataset. These gains come with a small increase in computational cost over SLMs of only 8%. Moreover, in 4 of the 9 datasets, CMBS Zero-Shot ties with fully-tuned LLM, with minimal losses in others (on average, just 2% less effective). These excellent effectiveness results come at 10% of the fully-tuned cost, as demonstrated in Table 12, which presents total time results for all alternatives. Moreover, both CMBS partially-tuned-IS versions tie with fully-tuned LLama in *all* sentiment datasets at 30%-50% of the fully-tuned cost.

For topic classification, with more categories (up to 11) and uneven distributions, CBMS Zero-Shot and CBMS In-Context struggle with effectiveness. Significant gains over SLMs occur only with CBMS Partially-Tuned-IS versions. Among the four topic datasets, both CBMS Partially-Tuned-IS (Aggressive) and (biO-IS) outperform RoBERTa in two datasets, tying with the other two. Both also surpass partially-tuned LLaMA in all cases, with up to 6.4% gains in Twitter. Compared to fully-tuned LLaMA, CMBS Partially-Tuned-IS (bio-IS) achieves statistical equivalence in all datasets and (Aggressive) in three, having just a small deficit of around 2% in the fourth (DBLP). (Bio-IS) cuts computational costs by 40% on average while Aggressive achieves 50% of cost reduction.

Between the two partially-tuned-IS CBMS variants, the Aggressive version is preferable. While both achieve comparable effectiveness in sentiment and topic classification, Aggressive offers a superior effectiveness–cost trade-off. Its main efficiency gain results from a 50% random reduction

Table 13. Emission CO_2. Calculation based on the work of Lacoste *et al*. [2019].

Dataset	RoBERTa	Zero-Shot LLaMA	In-Context LLaMA	Partially-Tuned-IS biO-IS LLaMA	Partially-Tuned-IS Agressive LLaMA	CMBS Zero-Shot	CMBS In-Context	CMBS Partially-Tuned-IS biO-IS	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLaMA
Finance	**0.02**	0.02	0.02	0.1	0.09	0.02	0.02	0.11	0.1	0.17
Imdb	**0.51**	1.22	2.25	6.52	4.9	0.57	0.63	6.38	4.91	7.63
PangMovie	**0.18**	0.23	0.29	1.39	1.15	0.19	0.21	1.44	1.21	2.12
SemEval17	**0.47**	0.61	0.83	3.8	2.95	0.5	0.53	3.89	3.12	5.46
Sst	**0.2**	0.24	0.3	1.54	1.27	0.21	0.22	1.58	1.34	2.29
Sst2	**1.13**	1.52	2.13	9.3	7.28	1.21	1.28	9.5	7.68	12.72
Yelp2L	**0.1**	0.23	0.34	0.73	0.47	0.11	0.12	0.75	0.52	0.99
IMDB2024	**0.13**	0.32	0.49	1.75	1.13	0.15	0.16	1.71	1.15	2.39
RottenT2024	**0.15**	0.19	0.33	1.12	0.85	0.16	0.17	1.16	0.92	1.58
ACM	**0.52**	0.62	1.54	3.6	3.28	0.55	0.58	3.76	3.47	5.48
DBLP	**0.81**	1.58	3.37	9.24	5.36	0.88	0.96	9.12	5.63	27.08
Twitter	**0.13**	0.17	0.29	1.38	1.3	0.14	0.14	1.37	1.29	2.22
Webkb	**0.18**	0.56	0.64	2.44	1.97	0.2	0.23	2.37	1.95	5.06

Table 14. Finance Cost in dollars ($) for RoBERTa, Zero-Shot LLaMA, In-Context LLaMA, Partially-Tuned LLaMA, CMBS Zero-Shot, CMBS In-Context, CMBS Partially-Tuned, and Fully-Tuned LLaMA.

Dataset	RoBERTa	Zero-Shot LLaMA	In-Context LLaMA	Partially-Tuned-IS biO-IS LLaMA	Partially-Tuned-IS Agressive LLaMA	CMBS Zero-Shot	CMBS In-Context	CMBS Partially-Tuned-IS biO-IS	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLaMA
Finance	**0.08**	0.11	0.13	0.54	0.51	0.09	0.09	0.57	0.54	0.94
Imdb	**2.73**	6.57	12.06	35.05	26.29	3.06	3.39	34.27	26.4	41
PangMovie	**0.98**	1.25	1.56	7.49	6.15	1.04	1.1	7.71	6.51	11.41
SemEval17	**2.52**	3.3	4.44	20.41	15.83	2.69	2.85	20.89	16.77	29.33
Sst	**1.07**	1.28	1.63	8.26	6.84	1.14	1.2	8.51	7.22	12.31
Sst2	**6.08**	8.15	11.42	49.94	39.1	6.48	6.89	51.02	41.26	68.34
Yelp2L	**0.53**	1.21	1.81	3.91	2.51	0.59	0.65	4.05	2.8	5.34
IMDB2024	**0.71**	1.7	2.65	9.42	6.08	0.8	0.88	9.19	6.18	12.85
RottenT2024	**0.82**	1.03	1.78	6.02	4.59	0.88	0.93	6.24	4.95	8.49
ACM	**2.78**	3.3	8.25	19.36	17.63	2.95	3.11	20.21	18.65	29.46
DBLP	**4.32**	8.47	18.08	49.61	28.79	4.75	5.17	48.97	30.23	145.44
Twitter	**0.68**	0.93	1.54	7.42	6.96	0.73	0.77	7.36	6.94	11.91
Webkb	**0.95**	3.01	3.42	13.1	10.6	1.1	1.25	12.74	10.49	27.18

of the training set, substantially lowering fine-tuning costs. Remarkably, within the CBMS framework, its effectiveness matches that of the more sophisticated biO-IS approach. It is important to note, however, that the partially-tuned Aggressive LLaMA in CBMS is applied only to the most challenging instances. Outside the CBMS framework, biO-IS remains more effective (Table 9), but within CBMS, the Aggressive variant provides the optimal balance between performance and computational efficiency.

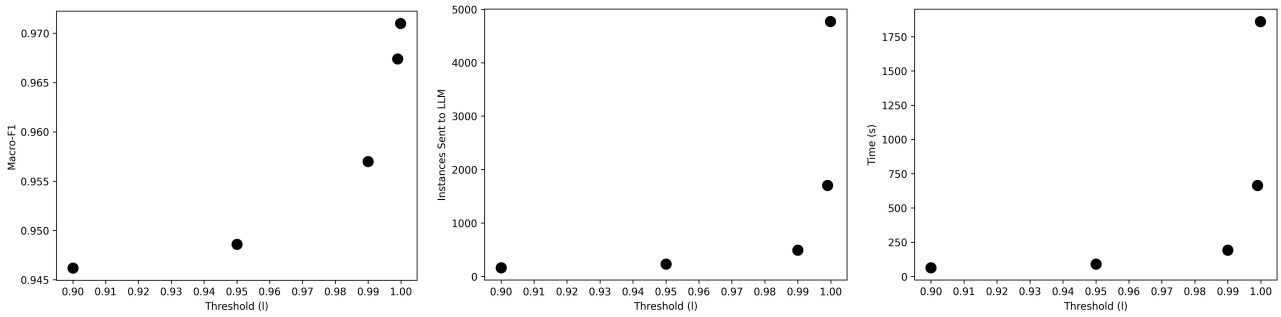
Summarizing, for sentiment, the best effectiveness tradeoff is achieved by CBMS Zero-Shot. If effectiveness is mandatory, the choice is CBMS Partially-Tuned-IS (Aggressive), which ties with LLaMA Fine-tuned at half the cost. For topics, the choice is also CBMS Partially-Tuned-IS (Aggressive), which ties with LLaMA fine-tuned in 3 out of four datasets, losing minimally (by 2%) in the fourth, being twice more efficient.

We also estimated the CO_2 emissions associated with model inference following the methodology proposed by Lacoste *et al*. [2019]. As shown in Table 13, LLMs produce emissions several orders of magnitude higher than SLMs, reflecting their substantially greater computational demands. Using a reference emission rate of approximately 0.14 kg CO_2 per GPU hour for hardware comparable to that used in our experiments[6], we observe a direct correlation between model size, inference time, and environmental cost.

Finally, financial analyses confirm the same effectiveness–sustainability trade-off observed in CO_2 emissions. Based on the pricing methodology of Griggs *et al*. [2024], Table 14 reports the estimated dollar cost of running each evaluated method. Costs were computed using the hourly rate of a cloud configuration comparable to our experimental setup[7], priced at approximately 0.752 per GPU hour.

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

[7] https://aws.amazon.com/ec2/instance-types/g4/



(a) Effectiveness **(b)** Instances Sent to LLM **(c)** Efficiency

Figure 4. Effectiveness, Size of the Test Set Sent to LLM and Efficiency for IMDB dataset.

5.6 Confidence Threshold Sensitivity

We further investigate the effect of the confidence threshold on the performance of the CMBS framework. As shown in Figure 4, the Partially-Tuned-IS (Aggressive) configuration achieves an excellent effectiveness–cost balance, surpassing RoBERTa and matching the Fully-Tuned LLaMA at approximately half the computational cost. Figures 4a–4c summarize the trends in effectiveness, number of instances forwarded to the LLM, and associated costs. Despite differences in metrics, the patterns are consistent: increasing the confidence threshold generally enhances effectiveness but at a higher computational expense.

We also assess the impact of parameter L, which controls the proportion of test instances forwarded to the LLM. Larger L values correspond to lower confidence thresholds, thereby increasing the number of instances processed by the LLM. Table 15 reports results for four representative datasets, including the fraction of forwarded instances and the corresponding SLM and LLM effectiveness. Notably, the optimal cost–effectiveness point occurs around $L = 0.9$ across datasets. Moreover, the LLM consistently outperforms the SLM on these hard-to-classify instances, confirming the advantage of the CBMS Partially-Tuned-IS (Aggressive) configuration in selectively leveraging LLM capabilities.

Table 15. Evaluation Threshold L. Best results are marked in **bold**.

Dataset	Percentage of Instances	SLM Macro-F1	LLM Macro-F1	Threshold (L)
Sst	27.0	0.65	**0.76**	0.9
Sst2	25.3	0.82	**0.85**	0.9
IMDB2024	7.8	0.72	**0.87**	0.9
Webkb	13.9	0.56	**0.67**	0.9
Twitter	13.9	0.51	**0.53**	0.9

5.7 CBMS Applied to Other NLP Tasks- CoLA GLUE

To illustrate the generalizability of our approach beyond sentiment and topic classification, we applied the best-performing configuration—CBMS Partially-Tuned (Aggressive)—to the Corpus of Linguistic Acceptability (CoLA) task from the GLUE benchmark. This task assesses a model's ability to determine whether a sentence is grammatically acceptable (e.g., "The cat sat on the mat") or not (e.g., "On the mat sat cat the"). As shown in Table 16, our method surpasses RoBERTa and achieves statistically equivalent performance to the Fully-Tuned LLaMA, while requiring only about half the computational cost.

Table 16. Average Macro-F1 and 95% confidence interval RoBERTa, CMBS Partially-Tuned-IS Agressive and Fully-Tuned LLaMA. Best results (including statistical ties) marked in **bold**.

Dataset	RoBERTa	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLM
GlueCola	80.4±1.8	**83.4±2.1**	**84.5±0.4**

6 Conclusion

We introduced Call-My-Big-Sibling (CMBS), an Adaptive Text Classification (ATC) framework designed to optimize the trade-off between effectiveness and computational cost by combining efficient, calibrated Small Language Models (SLMs) with more powerful yet expensive Large Language Models (LLMs). Our results reveal that, addressing RQ1 (Are open LLMs more effective than SLMs in sentiment and topic classification?), open LLMs surpass SLMs only when fully fine-tuned, whereas smaller models such as RoBERTa remain highly competitive—particularly for sentiment and topic classification—at a fraction of the cost. Regarding RQ2 (How does the computational cost of using open LLMs for ATC compare to that of SLMs?), we found that fine-tuning LLMs incurs computational expenses several orders of magnitude higher than SLMs, often making their limited performance gains impractical. To address RQ3 (Can combining SLMs and open LLMs yield a superior effectiveness–cost trade-off compared to using either model alone?), CMBS leverages calibrated confidence to selectively invoke LLMs only for uncertain cases, employing zero-shot inference or instance-selected fine-tuning to further minimize cost.

Empirical evaluation across 13 sentiment and topic classification datasets demonstrated that CMBS achieves a robust effectiveness–efficiency balance. The CMBS Zero-Shot variant outperformed SLMs in 8 out of 9 sentiment datasets with negligible additional cost, while the CMBS Partially-Tuned-IS (Aggressive) configuration matched the effectiveness of fully fine-tuned LLMs at roughly half the computational expense. In topic classification, CMBS exceeded the performance of

partially-tuned LLaMA and achieved near-equivalent accuracy to the fully-tuned model with twice the efficiency. These results highlight CMBS as a practical, cost-effective strategy for real-world NLP deployment.

Future directions include extending CMBS to broader NLP tasks—such as hate speech detection, irony recognition, name disambiguation [Santana *et al*., 2017; de Carvalho *et al*., 2011], summarization and question answering —enhancing instance-selection mechanisms to further reduce training overhead, and exploring alternative LLM architectures, calibration methods, and dynamic decision strategies to enhance both scalability and sustainability.

7 Limitations

Despite its contributions, this study presents some limitations. Our experiments focused on two classification tasks – sentiment and topic analysis – across thirteen datasets, plus one GLUE task (CoLA) for generalization. While this ensured a broad and controlled evaluation, Automatic Text Classification (ATC) remains challenging in domains such as hate-speech detection, misinformation identification, and imbalanced multi-class scenarios, as illustrated by the ACM dataset, where the best Macro-F1 reached 76.6.

Only a few prior works – mostly referenced in our related studies – have systematically examined the trade-off between effectiveness and computational cost in NLP, particularly through hybrid use of Small and Large Language Models (SLMs and LLMs). This remains a complex issue, as improvements in accuracy often entail significant computational overhead. Recent advances, such as DeepSeek [DeepSeek-AI *et al*., 2025], which approach state-of-the-art performance at lower cost, reinforce the relevance of this problem.

Our evaluation included four SLMs, four LLMs (four LLaMA variants), and three CMBS versions – over 1,000 runs across 13 datasets. For reproducibility, we restricted experiments to open-source, mid-scale models (Falcon 7B, Mistral 7B, DeepSeek 8B, and LLaMA 3.1 8B)[8] , excluding closed or extremely large models due to transparency and cost constraints. Future work will extend this analysis to newer, more efficient LLMs and explore richer probability aggregation strategies beyond the current single-token heuristic to enhance calibration and robustness.

Declarations

Authors' Contributions

Claudio Andrade, Washington Cunha, Celso França, Davi Reis, Wasterman Apolinário, Luana Santos: Conceptualization, Writing (review & editing), Methodology, Validation. *Claudio Andrade, Washington Cunha, Celso França, Davi Reis, Wasterman Apolinário, Luana Santos*: Writing (review), Methodology, Validation. *Adriana Pagano, Leonardo Rocha and Marcos Gonçalves*: Conceptualization, Writing (review & editing), Validation, Project Management, and Supervision.

[8] `https://huggingface.co/models`

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We would like to thank the Federal University of Minas Gerais for its support to this project.

Funding

This work was partially supported by grants from CAPES, CNPq, FAPEMIG, FAPESP, Google, Unimed-BH, AWS, Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação (INCT-TILDIAR) (process No. 408490/2024-1) and CIIA-Saude (process No. PPE-00030-21) grants.

Availability of data and materials

The datasets and codes generated and/or analysed during the current study are available in `https://github.com/claudiovaliense/cmbs`.

References

Antypas, D., Ushio, A., Camacho-Collados, J., Neves, L., Silva, V., and Barbieri, F. (2022). Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. Available at: `https://aclanthology.org/2022.coling-1.299/`.

Bai, J., Zhang, X., Li, C., Hong, H., Xu, X., Lin, C., and Rong, W. (2023a). How to determine the most powerful pre-trained language model without brute force fine-tuning? an empirical survey. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5369–5382, Singapore. Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-emnlp.357.

Bai, J., Zhang, X., Li, C., Hong, H., Xu, X., Lin, C., and Rong, W. (2023b). How to determine the most powerful pre-trained language model without brute force fine-tuning? an empirical survey. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the EMNLP 2023*. DOI: 10.18653/v1/2023.findings-emnlp.357.

Belém, F., Cunha, W., França, C., Andrade, C., Rocha, L., and Gonçalves, M. A. (2024). A novel two-step fine-tuning pipeline for cold-start active learning in text classification tasks. *arXiv preprint arXiv:2407.17284*. DOI: 10.48550/arXiv.2407.17284.

BRIER, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3. DOI: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Canuto, S. D., Gonçalves, M. A., and Benevenuto, F. (2016). Exploiting new sentiment-based meta-level features for

effective sentiment analysis. In Bennett, P. N., Josifovski, V., Neville, J., and Radlinski, F., editors, *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, San Francisco, CA, USA, February 22-25, 2016*, pages 53–62. ACM. DOI: 10.1145/2835776.2835821.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery, S. (1998). Learning to extract symbolic knowledge from the world wide web. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, page 509–516. DOI: 10.5555/295240.295725.

Cunha, W., Canuto, S., Viegas, F., Salles, T., Gomes, C., Mangaravite, V., Resende, E., Rosa, T., Gonçalves, M. A., and Rocha, L. (2020). Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling. *Information Processing & Management*, 57(4):102263. DOI: 10.1016/j.ipm.2020.102263.

Cunha, W., França, C., Fonseca, G., Rocha, L., and Gonçalves, M. A. (2023a). An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–674. Available at: `https://openreview.net/forum?id=W5dD8KmbRj`.

Cunha, W., Mangaravite, V., Gomes, C., Canuto, S., Resende, E., Nascimento, C., Viegas, F., França, C., Martins, W. S., Almeida, J., Rosa, T., Rocha, L., and Goncalves, M. A. (2021). On the cost-effectiveness of neural and non-neural approaches and representations for text classification: A comprehensive comparative study. *Information Processing Management*, 58(3):102481. DOI: 10.1016/j.ipm.2020.102481.

Cunha, W., Moreo, A., Esuli, A., Sebastiani, F., Rocha, L., and Gonçalves, M. A. (2025a). A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2). DOI: 10.1145/3705000.

Cunha, W., Rocha, L., and Gonçalves, M. A. (2025b). A thorough benchmark of automatic text classification: From traditional approaches to large language models. DOI: 10.48550/arXiv.2504.01930.

Cunha, W., Viegas, F., França, C., Rosa, T., Rocha, L., and Gonçalves, M. A. (2023b). A comparative survey of instance selection methods applied to non-neural and transformer-based text classification. *ACM CSUR*. DOI: 10.1145/3582000.

de Andrade, C. M., Belém, F. M., Cunha, W., França, C., Viegas, F., Rocha, L., and Gonçalves, M. A. (2023). On the class separability of contextual embeddings representations – or "the classifier does not matter when the (text) representation is so good!". *Information Processing & Management*, 60(4):103336. DOI: 10.1016/j.ipm.2023.103336.

de Carvalho, A. P., Ferreira, A. A., Laender, A. H. F., and Gonçalves, M. A. (2011). Incremental unsupervised name disambiguation in cleaned digital libraries. *J. Inf. Data Manag.*, 2(3):289–304. DOI: 10.5753/jidm.2011.1410.

DeepSeek-AI *et al*. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature 645*, page 633–638. DOI: 10.1038/s41586-025-09422-z.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. DOI: 10.18653/v1/N19-1423.

Fonseca, G., Cunha, W., Prenassi, G., Gonçalves, M. A., and Da Rocha, L. C. D. (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. DOI: 10.18653/v1/2025.acl-long.458.

França, C., Lima, R. C., Andrade, C., Cunha, W., de Melo, P. O. V., Ribeiro-Neto, B., Rocha, L., Santos, R. L., Pagano, A. S., and Gonçalves, M. A. (2024). On representation learning-based methods for effective, efficient, and scalable code retrieval. *Neurocomputing*, 600:128172. DOI: 10.1016/j.neucom.2024.128172.

Griggs, T., Liu, X., Yu, J., Kim, D., Chiang, W., Cheung, A., and Stoica, I. (2024). M\'elange: Cost efficient large language model serving by exploiting GPU heterogeneity. *CoRR*, abs/2404.14527. DOI: 10.48550/ARXIV.2404.14527.

Han, Z., Gao, C., Liu, J., Zhang, J., and Zhang, S. Q. (2024). Parameter-efficient fine-tuning for large models: A comprehensive survey. DOI: 10.48550/arXiv.2403.14608.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. DOI: 10.48550/arXiv.2106.09685.

Lacoste, A., Luccioni, A., Schmidt, V., and Dandres, T. (2019). Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*. DOI: 10.48550/arXiv.1910.09700.

Lepagnol, P., Gerald, T., Ghannay, S., Servan, C., and Rosset, S. (2024). Small language models are good too: An empirical study of zero-shot classification. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14923–14936, Torino, Italia. ELRA and ICCL. Available at: `https://aclanthology.org/2024.lrec-main.1299/`.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Featured Certification, Expert Certification. Available at: https://openreview.net/forum?id=iO4LZibEqW.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R., editors, *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, pages 142–150. The Association for Computer Linguistics. Available at: https://aclanthology.org/P11-1015/.

Malo, P., Sinha, A., Korhonen, P. J., Wallenius, J., and Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4):782–796. DOI: 10.1002/ASI.23062.

Mendes, L. F., Gonçalves, M., Cunha, W., Rocha, L., Couto-Rosa, T., and Martins, W. (2020). "keep it simple, lazy" – metalazy: A new metastrategy for lazy text classification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1125–1134, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3340531.3412180.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Knight, K., Ng, H. T., and Oflazer, K., editors, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics. DOI: 10.3115/1219840.1219855.

Rosenthal, S., Farra, N., and Nakov, P. (2019). Semeval-2017 task 4: Sentiment analysis in twitter. *CoRR*, abs/1912.00741. DOI: 10.48550/arXiv.1912.00741.

Santana, A. F., Gonçalves, M. A., Laender, A. H. F., and Ferreira, A. A. (2017). Incremental author name disambiguation by exploiting domain-specific heuristics. *J. Assoc. Inf. Sci. Technol.*, 68(4):931–945. DOI: 10.1002/ASI.23726.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL. Available at: https://aclanthology.org/D13-1170/.

Sorensen, T., Robinson, J., Rytting, C., Shaw, A., Rogers, K., Delorey, A., Khalil, M., Fulda, N., and Wingate, D. (2022). An information-theoretic approach to prompt engineering without ground truth labels. In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.60.

Spirling, A. (2023). Why open-source generative ai models are an ethical way forward for science. *Nature*, 616(7957):413–413. DOI: 10.1038/d41586-023-01295-4.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*. DOI: 10.48550/arXiv.1906.02243.

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., and Su, Z. (2008). Arnetminer: Extraction and mining of academic social networks. KDD '08, page 990–998, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/1401890.1402008.

Viegas, F., Canuto, S., Cunha, W., França, C., Valiense, C., Rocha, L., and Gonçalves, M. A. (2023). Clusent – combining semantic expansion and de-noising for dataset-oriented sentiment analysis of short texts. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, WebMedia '23, page 110–118, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3617023.3617039.

Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G., and Guo, C. (2025). GPT-NER: Named entity recognition via large language models. In Chiruzzo, L., Ritter, A., and Wang, L., editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics. DOI: 10.18653/v1/2025.findings-naacl.239.

Wolfe, J., Jin, X., Bahr, T., and Holzer, N. (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. DOI: 10.5194/isprs-archives-XLII-1-W1-455-2017.

Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., Huang, L., Liu, X., Kuo, T.-W., Guan, N., and Xue, C. J. (2025). Retrieval-augmented generation for natural language processing: A survey. DOI: 10.48550/arXiv.2407.13193.

Xu, C., Xu, Y., Wang, S., Liu, Y., Zhu, C., and McAuley, J. (2024). Small models are valuable plug-ins for large language models. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 283–294, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-acl.18.

Xu, S., Yan, Z., Dai, C., and Wu, F. (2025). Mega-rag: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of llms in public health. *Frontiers in Public Health*, Volume 13 - 2025. DOI: 10.3389/fpubh.2025.1635381.

Yue, M., Zhao, J., Zhang, M., Du, L., and Yao, Z. (2024). Large language model cascades with mixture of thought representations for cost-efficient reasoning. In *The Twelfth International Conference on Learning Representations*. Available at: https://openreview.net/forum?id=6okaSfANzh.

Zanotto, B. S., Beck da Silva Etges, A. P., Dal Bosco, A., Cortes, E. G., Ruschel, R., De Souza, A. C., Andrade, C. M., Viegas, F., Canuto, S., Luiz, W., *et al*. (2021). Stroke outcome measurements from electronic medical records: cross-sectional study on the effectiveness of neural and nonneural classifiers. *JMIR Medical Informatics*, 9(11):e29120. DOI: 10.2196/29120.

Zhang, W., Deng, Y., Liu, B., Pan, S., and Bing, L. (2024).

Sentiment analysis in the era of large language models: A reality check. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics. DOI: 10.18653/v1/2024.findings-naacl.246.

A Appendix - Evaluating Aggressive Reduction in the LLMs with Different Reduction Levels

As mentioned, fine-tuning is essential for LLM effectiveness. Here, we illustrate the impact of training data size on LLM effectiveness using the validation set and two sample datasets. The pattern of results is basically the same in all other datasets we experimented with.

Table 17 presents the effectiveness results when utilizing 30%, 50%, and 70% of the training data in Twitter and WebKB, two topic datasets in which CMBS performs very well. As we can see in the Table, 30% of training generally is not enough for achieving reasonable effectiveness, while the improvements of using 70% are either marginal or incur in higher costs.

As discussed in Section 5, the CMBS Partially-Tuned-IS (Aggressive) version we employed in our experiments uses 50% of the training data, randomly selected in a stratified manner, based on results of instance selection experiments Cunha *et al*. [2023a]. In all datasets, such a choice produced the best tradeoff between effectiveness and computational cost.

Table 17. Evaluate amount training LLM.

Dataset	**Portion Train**	**Macro-F1**
Twitter	30	66.2
Twitter	50	71.9
Twitter	70	76.1
Webkb	30	76.7
Webkb	50	83.4
Webkb	70	85.2

RoBERTa with DeepSeek

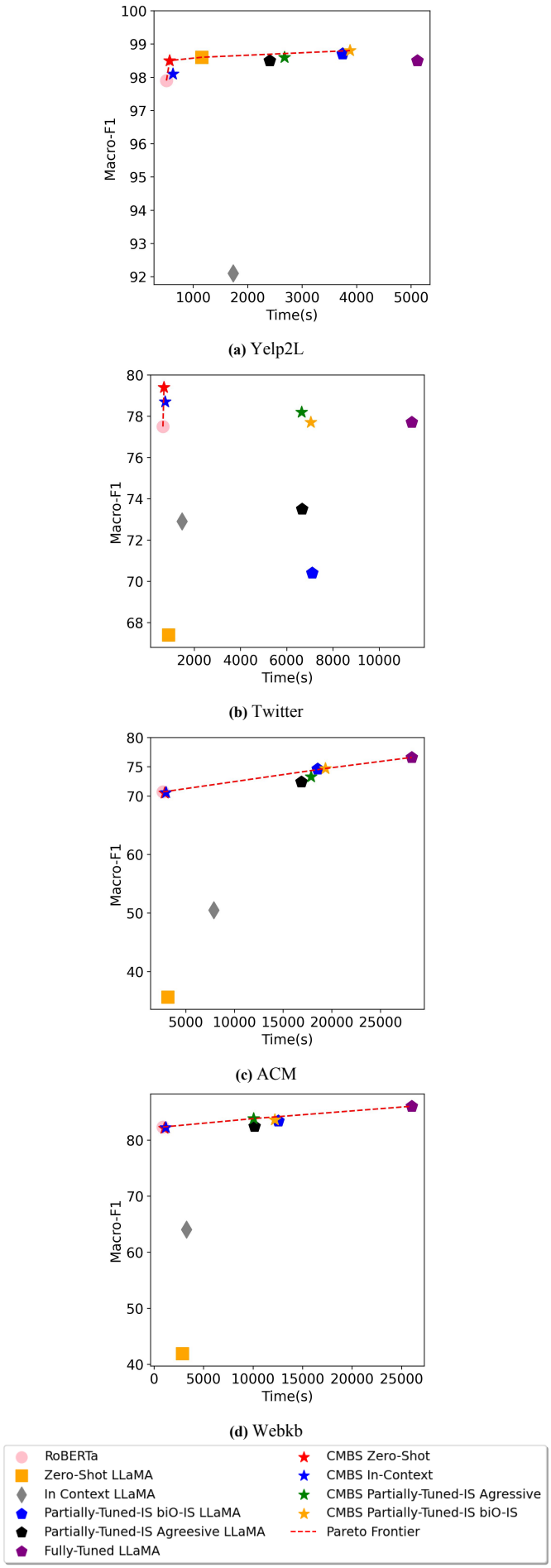
In addition to combining RoBERTa with LLaMA, we conducted a new experiment that integrates the SLM RoBERTa with the LLM DeepSeek 8B. Table 18 presents the results of this combination across two datasets. As observed, the combination with DeepSeek consistently yields lower average performance across all datasets, reinforcing LLaMA as the best choice among the evaluated LLMs.

Table 18. Average Macro-F1 and 95% confidence interval RoBERTa, CMBS Partially-Tuned-IS Agressive and Fully-Tuned LLaMA. Best results (including statistical ties) marked in **bold**.

Dataset	**RoBERTa**	**CMBS Partially-Tuned-IS Agressive DeepSeek**	**CMBS Partially-Tuned-IS Agressive LLaMA**	**Fully-Tuned LLaMA**
Sst	87.3±1.0	**90.4±0.8**	**90.9±0.9**	**91.1±1.0**
RottenT2024	93.7±1.1	**95.7±0.7**	**96.3±0.7**	**96.7±0.4**

Pareto-optimal

Figure 5 presents the graph of the effectiveness-efficiency trade-off across all methods evaluated in this study. The Pareto frontier is represented by the dashed red line, with all points overlapping this line being Pareto-optimal. We can observe that CMBS is the most frequent method on the Pareto frontier.



(a) Yelp2L

(b) Twitter

(c) ACM

(d) Webkb

Figure 5. Total time (seconds) and Macro-F1 for RoBERTa, different versions of LLaMA, and CMBS. The Pareto frontier is represented by a dashed red line, with all points overlapping this line being Pareto-optimal.