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

Anonymous ACL submission

Abstract

Transformers have achieved cutting-edge results, with Large Language Models (LLMs) considered SOTA in many NLP tasks. However, the literature has not yet fully demonstrated that LLMs are always superior to first-generation Transformers (a.k.a. Small Language Models (SLMs)) in all NLP tasks and scenarios. This study compares four SLMs (BERT, RoBERTa, Qwen, BART) with four open LLMs (LLaMA 3.1, Mistral, Falcon, DeepSeek) across 9 sentiment analysis and 4 topic classification datasets, totaling over 1000 results. Findings indicate that open LLMs can moderately outperform or tie with SLMs in all tested datasets, though only when fine-tuned, at a very high computational cost. To address the cost-effectiveness trade-off, we propose "Call My Big Sibling" (CMBS)[1], a novel confidence-based framework that efficiently integrates *calibrated SLMs with open LLMs using advanced instance selection techniques*. CMBS assigns high-confidence predictions to the cheaper SLM, while low-confidence cases are directed to LLMs in zero-shot, in-context, or partially-tuned (leveraging state-of-the-art **instance selection** approaches) modes, optimizing cost-effectiveness. Experiments show that CMBS significantly outperforms SLMs and achieves *LLM-level effectiveness at a fraction of the cost*, offering a practical alternative for cost-sensitive NLP applications.

1 Introduction

Automatic text classification (ATC) is essential in diverse contexts, ranging from organizing large data volumes to personalizing user experiences. ATC has experienced a huge revolution with the advent of semantically enriched Transformer models (Devlin et al., 2019) that have achieved state-of-the-art performance (de Andrade et al., 2023; Cunha et al., 2023a; Zanotto et al., 2021).

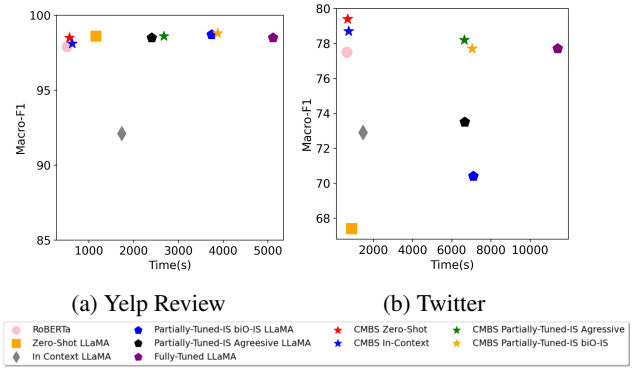


Figure 1: Total Time (seconds) and Macro-F1 in RoBERTa, version of LLaMA and CMBS. All CMBS proposals outperform the other baselines, being much cheaper.

More recently, Large Language Models (LLMs) emerged, built on top of the first generation of Transformers (a.k.a., small language models – SLMs). Studies have identified LLMs as the current SOTA for several NLP tasks (Liang et al., 2023). Although the literature reports LLMs' superiority for tasks such as summarization and translation, for others, such as sentiment analysis and topic classification, it is not yet clear whether LLMs' complexity and size (e.g., in terms of a number of parameters) translate into statistical and mainly *practical* gains. In fact, several studies point to the SLM RoBERTa as a very strong sentiment classifier (Cunha et al., 2023b) ranking prominently on leaderboards such as the GLUE benchmark[2].

Depending on the type of training (or its absence), LLM strategies can be categorized into four groups: *zero-shot, in-context, partially-, and fully-tuned*. In zero-shot, the model performs tasks without specific training. In in-context, a few examples are provided in the prompt for learning. Partially-tuned models use a significant part of the training set, while fully-tuned ones use the entire set for better optimization. As shown in (Table 1), open LLMs surpass SLMs mainly in the

[1] Code available at https://github.com/Anonymous

[2] https://gluebenchmark.com/leaderboard/

fully-tuned scenario, albeit at a **very high cost**, highlighting the need for cost-effective alternatives.

Accordingly, our first research question is **RQ1:** "Are (open) LLMs more effective than SLMs in sentiment and topic classification?" Recent work (Fields et al., 2024) shows no consensus on whether LLMs always excel in ATC tasks. To investigate, we compared four SLMs (BERT, RoBERTa, Qwen[3], BART) and four open LLMs (LLaMA 3.1 8B, Mistral 7B, Falcon 7B, DeepSeek 8B) on 9 sentiment and 4 topic datasets, including two post-LLM releases (IMDB2024, RottenT2024) to reduce data contamination (Liang et al., 2023). We focus on open LLMs, as closed ones (e.g., ChatGPT) lack transparency[4]. Our statistically rigorous experiments[5] show open LLMs can outperform SLMs, achieving up to 8.3% (avg. 3.1%) effectiveness gains, though mostly in fully-tuned mode.

Given the (much) higher computational costs associated with fully fine-tuning open LLMs (the most effective approach), a natural question we posited is **RQ2:** "How does the computational cost of using open LLMs for ATC compare to SLMs' cost?". To answer this question, we conducted a thorough analysis of our experimental results, considering zero-shot, in-context, partially-tuned, and fully-tuned strategies, to assess the trade-offs between effectiveness and costs in terms of computational cost[6] to train and employ the models and their impact on carbon emission. We found that LLMs are orders of magnitude more costly to fully fine-tune when compared to SLMs – fully fine-tuned LLMs are up to 1700% more expensive than SLMs. As current LLMs produce just moderate gains over SLMs and only through highly costly full fine-tuning processes, depending on the scenario, the benefits may not be worth the costs.

All this leads to our final research question **RQ3:** "Is it possible to perform a combination of SLM and (open) LLMs to achieve a better effectiveness/cost trade-off than using either SLM or LLM alone?" To answer this question, we proposed a novel confidence-based strategy called "Call My Big Sibling" (**CMBS**), which smoothly combines SLM and (open) LLMs based on *calibrated confidences*.

In CMBS, we rely on *fully fine-tuned SLMs*, which have already attained effectiveness and efficiency and are calibrated[7] for ATC tasks. We then use the classification confidence to determine whether the LLM should classify a low-confidence document. In other words, the fully-tuned SLM classifies high-certainty documents (i.e., with high certainty calibrated scores), while low-confidence documents are sent to the *zero-shot, in-context or partially-tuned versions of the LLMs* for ATC. In the particular case of the partially-tuned LLMs, our CBMS solution introduces an innovation by exploiting recent SOTA **instance selection** methods (Cunha et al., 2025) to reduce the cost of tuning while maintaining effectiveness. Such combination with a cheaper LLM version (compared to the fully-tuned LLM) brings potential effectiveness gains to the SLM and is on par with fully-tuned LLM, being a very attractive, cost-effective option in basically all tested cases, as we shall see.

Our results show that for sentiment classification, combining an SLM with a zero-shot LLM (**CMBS Zero-shot**) improves effectiveness at the lowest cost, proving its practicality. Figure 1 illustrates the effectiveness (Macro-F1) and efficiency (Time(s)) of our approach versus baselines in two datasets, with our proposals marked by stars. In Figure 1a, CMBS Zero-Shot matches Fully-Tuned LLaMA's effectiveness at a fraction of the cost. Similarly, in Figure 1b, all CMBS methods outperform baselines while remaining much cheaper. Indeed, our experiments reveal that CBMS Zero-shot outperforms the SLM in 8 out of 9 sentiment datasets, tying with the remaining one, with an increase in computational cost over SLMs of only 8%. Moreover, compared to fully-tuned LLaMA, CBMS Zero-Shot delivers comparable effectiveness at a significantly lower cost. In 4 of the 13 datasets, CMBS Zero-Shot ties with the fully-tuned LLM, with minimal losses (on average, just 2%) in the other datasets, at $\frac{1}{10}$ of the cost. CMBS Partially-Tuned-IS (Aggressive), on turn, ties with fully-tuned LLaMA in *all* sentiment datasets at half of the cost.

For topic classification with several classes (up to 11) and uneven distributions, CBMS zero-shot and **CBMS In-Context** struggle with effectiveness. Only when combined with **CMBS Partially-Tuned-IS** can it surpass the SLM. Among 4 datasets, CMBS Partially-Tuned-IS (Aggressive)

[3]Among the multiple Qwen versions, we use the smallest one, with 0.5B parameters, which we consider as an SLM due its effectiveness comparable to the other SLMs used in this paper and its fast response time.

[4]Closed LLMs are irreproducible (Gao et al., 2024).

[5]Please see Section 4 for a full description of our experimental setup.

[6]Including both, time to train the model and predict test.

[7]The confidence of the SLM's softmax function is highly calibrated as we shall discuss.

outperforms RoBERTa in 2 and ties in the others. Against fully-tuned LLaMA, it achieves statistical parity in 3 datasets (with just a 2% loss in the 4th) while operating at half the computational cost.

In sum, the main contributions of this paper are:

- A comprehensive comparative evaluation of SLMs and (open) LLMs regarding cost/effectiveness trade-offs.
- The proposal of "Call My Big Sibling" (**CMBS**), a novel confidence-based strategy to combine calibrated SLMs and zero-shot or partially-tuned with *instance selection* LLMs, aimed at optimizing the effectiveness-cost trade-off.
- A thorough evaluation of our proposals considering 13 distinct datasets, in 2 tasks[8]: sentiment (binary) and topic (multi-class) classification tasks, 4 SLMs and zero-shot, in-context, partially-tuned and fully-tuned versions of 4 open LLMs.

2 Related Work

LLMs' computational costs have led to numerous studies highlighting their financial and environmental impacts. (Strubell et al., 2019) illustrates the substantial financial costs propelled by the continuous need for investment in specialized hardware to manage progressive LLMs. This trend not only limits access to these models but also escalates energy consumption, affecting the environment by increasing carbon dioxide (CO_2) emissions.

Among LLMs, proprietary and closed-source ones, such as *GPT*, operate as black boxes. This opacity poses challenges in comprehending their training and internal structures, obstructing reproducibility. Utilizing such LLMs often entails transmitting data through web platforms or APIs, a sensitive issues when privacy is a concern. As a result, numerous studies advocate restricting scientific evaluations to locally-run, open-source LLMs (Spirling, 2023).

(Xu et al., 2024) combines SLMs with LLMs aiming to improve effectiveness. In that work, computational costs are not evaluated, and LLMs are used to classify the entire test set, unlike ours, which only uses them to classify a subset of hard instances. On the other hand, we select only low-confidence documents to be forwarded to the LLM, a strategy that greatly reduces computational costs as the LLM is significantly more expensive. Moreover, a closed LLM is employed via an API, which provides no control over the computational structure or the model architecture. Finally, only a single sentiment dataset is used in the experiments.

Liang et al., 2023 investigate various LLMs across multiple tasks, prompts, metrics, and datasets. Like Liang et al., we investigate the effectiveness-cost tradeoff of LLMs. Unlike their study, which focuses on the breadth of evaluation with several domains (including only one sentiment dataset), our work is depth-oriented into the specific task of sentiment and topic classification, covering multiple datasets with diverse characteristics and domains. Moreover, although Liang et al. evaluates several models, they do not compare them with an SLM such as RoBERTa, considered SOTA in sentiment and topic classification (Bai et al., 2023a; Cunha et al., 2021a, 2020; França et al., 2024; Belém et al., 2024). Last, but not least, they do not provide any solution for the effectiveness-cost trade-off problem, which ***we do!***.

Yue et al., 2024 proposed a hybrid approach using GPT-3.5 (weak) and GPT-4 (strong), where the weak model is the first to respond, and a decision model checks response consistency across multiple prompts, calling the strong model if needed. This potentially improves effectiveness but does not reduce model costs – indeed, only financial costs based on the number of OpenAI API calls are considered in their work. In contrast, our approach considers both financial and computational costs (time) and uses SLM confidence, not response consistency, to decide when to invoke a cost-optimized LLM.

Finally, it is worth noticing that the recent success and impact, beyond the academic world, of DeepSeek (DeepSeek-AI et al., 2025), which matched or improved the effectiveness of state-of-the-art pre-trained LLMs while reducing computational demands, show how important the issue of trade-off effectiveness vs. cost is to the community. We investigate DeepSeek as one of the LLMs for our CBMS solution.

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

[8] An additional application of the CBMS on a different NLP task from the GLUE benchmark is discussed in Appendix J.

conjures up 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**. In this framework, we first employ *fully-tuned SLMs models*[9], which are already highly effective in some classification tasks (and faster to tune compared to LLMs). In our solution, (test) documents classified below a certain confidence threshold (a parameter) by the SLM are sent to an open LLM to be classified. We have 5 versions of LLMs: Zero-shot, In-Context, Partially-Tuned-IS biO-IS and Aggressive, and Fully-Tuned.

In particular, in the case of partially-tuned LLMs, we have two options. Partially-Tuned-IS biO-IS leverages the state-of-the-art just-realased instance selection method biO-IS (Cunha et al., 2025) 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., 2025). Partially-tuned-IS Aggressive 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. The procedure is illustrated in Figure 2.

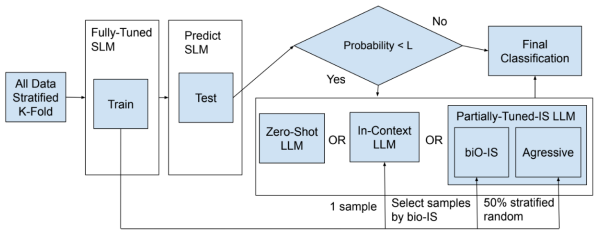


Figure 2: CBMS FlowChart.

For CBMS to properly work, we have to trust the probability outputs, or, in other words, the probabilities need to be calibrated[10]. 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 13 (Appendix G) presents the Brier score used to measure model calibration, in datasets used in our experiments, by 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. As we can observe, the table reinforces that RoBERTa is a very calibrated model (Brier score > 0.8), being as calibrated as well-calibrated classifiers such as Logistic Regression and Random Forests that obtain similar Brier scores for some of these benchmark datasets (Cunha et al., 2025). In addition to Table 13, Figure 5 presents the calibration curve for the BERT (red) and RoBERTa (blue) models. The dashed line along the main diagonal represents ideal calibration. As shown in Figure 5, RoBERTa is consistently closer to the ideal calibration across all datasets, further supporting the results presented in Table 13.

We select a document set for which the classifier is least confident about its classification (Probability < L) to send to an LLM for final prediction. Due to computational costs, we employ either the zero-shot, in-context, or partially-tuned strategies for this LLM. Finally, our final prediction set is built using the following procedure: 1) we evaluate the probability the model provides and compare it with the threshold parameter; 2) we decide whether the prediction will be made using an SLM or an LLM (zero-shot, in-context or partially-tuned).

In this proposal, the choice of confidence threshold L is essential to evaluate the documents that will be sent to the LLM. To illustrate this point, Figure 3 presents the effectiveness based on prediction confidence for the SST2 dataset. On the Y-axis, we have RoBERTa's effectiveness, and on the X-axis, we have confidence. We can observe the more confident RoBERTa is, the more effective its predictions are. The figure highlights the importance of selecting an appropriate confidence threshold, showing that it is more advantageous to forward low-confidence documents, as high-confident ones are classified with high accuracy by the cheaper SLM.

4 Experimental Methodology and Setup

4.1 Datasets

Our study draws on **thirteen** datasets developed for sentiment analysis and topic classification. The sentiment analysis datasets include **Finance**, **IMDB**, **PangMovie**, **SemEval17**, **SST**, **SST2** and

[9]Tuned with the full training data.

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

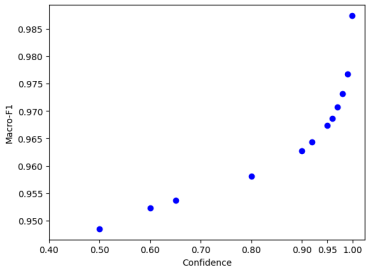


Figure 3: RoBERTa's Macro-F1 vs Confidence for SST2.

Yelp Review (Yelp2L), while prompt fication we used **ACM**, **DBLP**, **Twitter** and **Webkb**. With the significant amount of data used in building LLMs, several authors express concerns about contamination in evaluation data. To minimize this issue, we collected and curated two datasets with data post-LLMs release (**RottenT2024** from Jan-Nov 2024 and **IMDB2024** from Jan-May 2024), ensuring no contamination in the training of these LLMs. Further details about the datasets in Appendix B: domain, number of documents, density, and skewness (class imbalance). Our benchmark covers a wide variety of heterogeneous scenarios.

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 5) and topic classification (Table 6, Appendix A). 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 7) modifies this by replacing a generic example with the closest training document, determined via cosine similarity using RoBERTa embeddings.

4.3 Zero-shot, In-context, Partially-tuning or Fully-tuning for Text Classification

Applying SLM or LLM pre-trained models to ATC can be done within our CBMS solution through five strategies: zero-shot, in-context, partially-tuned-IS biO-IS, partially-tuned-IS Aggressive, and fully-tuned. **Zero-shot** strategy predicts text classes without using training examples or performing model fine-tuning. In an **In-context** approach, the model relies on a prompt containing the nearest neighbors of the evaluated example inserted into it to provide context for making predictions without adjusting its weights. In the **partially-tuned (IS)** strategies, a portion of labeled data is employed to adjust the model weights, simulating a scenario of data scarcity. As described in Section 3, we have two options for this strategy: using biO-IS instance selection for training set reduction, with variable reduction rates, depending on the dataset, and Aggressive, in which we fixed 50% of the training partition data for model training, selecting these instances in a stratified random manner. An evaluation using different fixed training data sizes was conducted in Appendix E, further justifying the 50% rate choice.

Lastly, the **fully-tuned** strategy utilizes all available labeled data in the model's training partition to maximize model adjustment for the task and data domain. While this strategy typically achieves better effectiveness, it has a very high computational cost. In our paper, the fully-tuned strategy is used as *baseline* to compare with our proposals.

We only employ the SLMs fully tuned, which is essential for effectiveness (de Andrade et al., 2023). Fully tuning SLMs involves fine-tuning the SLM's text representation (CLS token) and a fully connected layer that performs class prediction, utilizing all available training samples.

4.4 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, 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 hyper-parameterization, fixing the learning rate in 2×10^{-5}, the batch size with 64 documents, adjusted 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 12 (Appendix F) 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.

4.5 Metrics and Experimental Protocol

We evaluated SLMs and (open) LLMs regarding the effectiveness/cost tradeoff. All models were assessed on identical hardware configuration: a 4-core processor, 32GB of system memory, and an Nvidia Tesla P100 GPU. Classification effectiveness is assessed using Macro-F1 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 cost-effectiveness tradeoff, we evaluate each method's total model-building time, including learning and class prediction for the test set. For CMBS Zero-Shot, this includes tuning the SLM, predicting part of the test set with the SLM, and a smaller portion with the LLM. CMBS in-context also accounts for finding the k nearest neighbors, while CMBS partially-tuned includes training the LLM using biO-IS or 50% stratified random of the training for Aggressive.

5 Experimental Results and Analyses

5.1 SLMs vs. LLMs - Effectiveness

To address RQ1, we first evaluated several popular open-source LLMs, including Falcon 7B, Mistral 7B, DeepSeek 8B, and LLaMA 3.1 8B. We began by comparing the performance of these four LLMs on sentiment and topic tasks in the Zero-Shot setting. Table 9 in Appendix C presents the Macro-F1 scores, highlighting the best results in bold, including statistical ties. LLaMA 3.1 8B consistently achieves the best results (statistically) across most datasets in both sentiment and topic tasks. Due to the high computational cost of fully tuning LLMs, we selected LLaMA 3.1 8B for all subsequent tests.

Regarding SLMs, we did a similar experiment and compared four widely used SLMs – BERT, BART, RoBERTa and QWen. Results in Table 10 in Appendix D show that, among SLMs, RoBERTa achieves the highest effectiveness (or ties for it) in all cases, confirming findings reported in the literature (Cunha et al., 2023b; Bai et al., 2023b).

Still answering RQ1, we compare RoBERTa with five LLaMA versions in Table 1—zero-shot, in-context, partially-tuned (biO-IS and Aggressive), and fully-tuned—using the full training set. Zero-shot LLaMA 3.1 ties or underperforms compared to RoBERTa in sentiment datasets and performs much worse in topic classification. Similarly, In-Context LLaMA does not outperform RoBERTa in sentiment classification. Since Zero-Shot LLaMA lacked training labels, while RoBERTa was fully tuned, RoBERTa achieved better effectiveness in most cases. The performance gap is even larger in topic classification due to more classes and higher complexity of the ATC task.

Only partially-tunned-IS (in both versions, biO-IS and Aggressive, for most datasets) and fully-tuned LLaMA can outperform RoBERTa, with some advantage for fully-tuned, especially for topic classification. However, in some datasets, such as Finance and Yelp2L (sentiment), and Twitter (topic), fully-tuned and RoBERTa are statistically tied, mainly due to larger confidence intervals (meaning higher variability) for the RoBERTa results. In some other datasets, RoBERTa's and fully-tuned LLama's effectiveness are also close.

The instance selection 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 instance selection alternatives, we can see that biO-IS preserves effectiveness in more datasets, which is consistent with (Cunha et al., 2025), 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.

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±13	**98.6±0.1**	**98.7±1.6**
Imdb	93±0.5	93±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	87.9±0.7	88.5±1	**90.7±0.9**	**90.9±0.8**	**91.1±1**
Sst2	94.6±0.2	91.4±0.4	93.5±0.4	**95.8±0.3**	**95.7±0.2**	**96±0.1**
Yelp2L	**97.9±0.5**	**98.6±0.3**	92.1±1	**98.7±0.3**	**98.5±0.6**	**98.5±0.5**
IMDB2024	97.6±1	96.5±1	93.9±1	**98.4±0.7**	**98.6±0.7**	**98.7±0.7**
RottenT2024	93.7±1.1	95.2±1.4	95.3±1	**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	**86.7±1**	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±1.8	83.4±1	82.4±2.1	**86±1.3**

Table 1: 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	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**
RottenT2024	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 2: 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**.

5.2 SLMs vs. LLMs - Computational Cost

Table 3 presents 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 Partially-Tuned-IS, cost increases significantly due to weight adjustment performed via LLM backpropagation. Aggressive version is always around 9%-35% faster than biO-IS due to its larger training set reduction, which impacts the fine-tuning time. Of course, fully-tuned LLM is the most expensive solution, which is 1700% more expensive than RoBERTa. With an average improvement of 3.3% across all datasets (peaking at 8.3% in ACM), it is not clear whether such improvements justify the significant computational cost increases. Depending on the application scenario, costly solutions may not be ideal or even feasible. To address this, our proposed solution reduce costs associated with using LLMs while preserving effectiveness gains.

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

Table 3: Average Total Time for RoBERTa and versions of Llama3.1-8B.

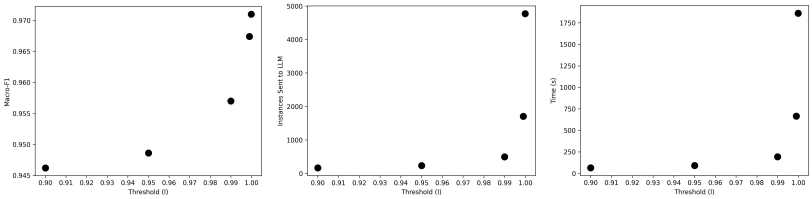
Dataset	RoBERTa	CMBS Zero-Shot	CMBS In-Context	CMBS Partially-Tuned-IS biO-IS	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLM
Finance	79	84	89	542	515	896
Imdb	2615	2930	3245	32813	25273	39257
PangMovie	934	994	1054	7386	6237	10921
SemEval17	2416	2574	2732	20004	16055	28087
Sst	1027	1089	1150	8149	6917	11791
Sst2	5817	6207	6597	48848	39508	65428
Yelp2L	510	568	626	3879	2676	5116
IMDB2024	681	762	844	8795	5921	12304
RottenT2024	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

Table 4: Average Total Time for RoBERTa and versions of CMBS and Fully-Tuned LLaMA.

5.3 RQ3: Evaluating the CBMS Approach

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 2 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 Finance. 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 4, 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 four topic datasets, both CBMS Partially-Tuned-IS



(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.

(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.

If choosing between the two partially-tuned-IS CBMS versions, we would prefer Aggressive. Both show similar effectiveness in sentiment and topic datasets, but Aggressive offers a better effectiveness-cost tradeoff. Its main cost reduction comes from randomly cutting the training set by 50%. Surprisingly, its effectiveness matches state-of-the-art biO-IS when used within CBMS. However, remember that partially-tuned LLaMA (Aggressive), used within CBMS, only needs to handle the hardest cases. Outside CBMS, Bio-IS performs better (Table 1), but within our solution, Aggressive achieves the best effectiveness-cost balance.

Summarizing, for sentiments, 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.

Finally, we calculated CO_2 emissions for obtaining final model predictions using Lacoste et al. (2019)'s methodology. Results in Table 14 (Appendix H) show emissions are significantly higher for LLMs, by orders of magnitude. Financial costs, analyzed in Appendix I, lead to similar conclusions.

5.4 Confidence Threshold Sensitivity Analysis

We analyze confidence threshold impact. Figure 4 shows IMDB results, where CMBS Partially-tuned-IS (Aggressive) achieved a strong cost-benefit tradeoff—outperforming RoBERTa and matching Fully-tuned LLaMA, at half the cost. Figures 4a, 4b, and 4c depict effectiveness gains, instances sent to LLM, and cost increases. Trends are similar across all graphs, despite different metrics. With appropriate thresholds, effectiveness can improve, though at a higher cost.

5.5 CBMS Applied to Other NLP Tasks- CoLA GLUE

To demonstrate the potential of our solution for other NLP tasks, Appendix J shows the results of applying CBMS-Partially-Tuned-IS (Aggressive) to CoLA (Corpus of Linguistic Acceptability), a GLUE task assessing a model's ability to classify English sentences as grammatically acceptable or unacceptable. As seen in Table 16, our solution outperforms RoBERTa and is statistically equivalent to Fully-Tuned at half of its computational cost.

6 Conclusion

We proposed Call-My-Big-Sibling (CMBS), an ATC solution that optimizes the effectiveness-cost trade-off by combining efficient, calibrated SLMs with more effective but costlier LLMs. CMBS uses LLMs only when the SLM is not confident, minimizing costs through zero-shot inference or instance selection when tuning is needed. Experiments on 13 sentiment and topic classification datasets confirmed its strong effectiveness-cost balance. CMBS Zero-Shot outperformed SLMs in 8 of 9 sentiment datasets with minimal overhead, while CMBS Partially-Tuned-IS matched fully tuned LLMs in sentiment tasks at half the cost. For topics, it surpassed partially tuned LLaMA and achieved near-equal effectiveness to fully tuned LLMs at twice the efficiency. Such efficiency gains are crucial in real-world scenarios. Future work includes extending CMBS to tasks like hate speech, irony detection, summarization, and Q&A, refining instance selection to reduce training needs, and exploring additional LLMs, configurations, and alternative prediction methodologies.

7 Limitations

Despite relevant contributions, our study has some limitations. Our current work covers only two classification tasks, which we have pursued to evaluate in depth. In this study, we used 13 datasets, 9 on sentiment analysis and 4 on topic classification, all with distinct characteristics. Additionally, we covered a single GLUE task on grammatical correctness to demonstrate generalization.

In spite of such limitations, there are only few studies - most of them covered in our related work - that analyze the complex issue of balancing effectiveness and the associated computational costs of NLP tasks in the Language Models context, particularly through the combination of LMs, as proposed in our paper. Indeed, this is a challenging trade-off to address: improving effectiveness often leads to higher costs. The recent success and real-world impact, including financial, of DeepSeek (DeepSeek-AI et al., 2025), which matched or surpassed the effectiveness of state-of-the-art pre-trained LLMs while reducing computational demands, highlights the importance of this topic (effectiveness vs. cost trade-off) to the community.

Although we have explored two (three if we consider GLUE CoLA) classification tasks - topic classification and sentiment analysis (and grammatical correctness) - we emphasize that ATC remains an open challenge in several NLP applications, such as hate speech detection, misinformation identification, and sentiment analysis—all of which can be framed as classification problems. Papers on these topics are recurrent in major ACL conferences, which attests to the relevance of this problem. Moreover, many classification tasks are imbalanced, posing significant challenges, particularly in multi-class classification. For instance, in the ACM dataset, which contains documents distributed across 11 imbalanced classes, the highest reported effectiveness is 76.6 Macro-F1, highlighting that this task remains challenging and far from being successfully solved.

We believe our proposal is comprehensive, as it evaluates four SLM methods, four LLMs - including four versions of LLaMA (zero-shot, in-context, partially tuned, and fully tuned) - and three versions of our solution across 13 different datasets. This extensive effort comprises more than 1000 experimental results (13 datasets x 16 methods x 5 runs). Nonetheless, more SLMs and LLMs, alone and in combination, will be tested in the future.

We focused our evaluation on open LLMs for the sake of reproducibility in subsequent research using our method. Hence we have refrained from using proprietary and closed-source LLMs, such as GPT, which operate as black boxes. This opacity poses challenges in understanding their training methodologies or internal structures, thereby obstructing reproducibility in research reliant on these models.

LLMs have been made available for different purposes. Some of these LLMs have high execution costs, such as Falcon 180B (Penedo et al., 2023), which requires an expensive infrastructure to use it. In our work, we limited our study to the best evaluated LLMs in the Hugging Face platform[11], with around 7 billion parameters, which have a reasonable structure allowing us to evaluate zero-shot, in-context, partially and fully-tuned versions of our solutions.

Regarding, probability estimation, we employ a straightforward heuristic that exclusively considers the token with the maximum 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.

Finally, our work focused on applying our proposals with four open LLMs—Falcon 7B, Mistral 7B, DeepSeek 8B, and LLaMA 3.1 8B. However, new LLMs emerge all the time and it is hard to test all of them. In any case, we intend to explore new LLMs in future work. Given that these new LLMs tend to be increasingly more complex and costly, optimizing the cost-effectiveness of our combined approach using SLMs and LLMs will remain not only relevant but also an even more compelling objective.

References

Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vitor Silva, and Francesco Barbieri. 2022. Twitter Topic Classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

[11] https://huggingface.co/models

survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5369–5382, Singapore. Association for Computational Linguistics.

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

Fabiano Belém, Washington Cunha, Celso França, Claudio Andrade, Leonardo Rocha, and Marcos André Gonçalves. 2024. A novel two-step fine-tuning pipeline for cold-start active learning in text classification tasks. *arXiv preprint arXiv:2407.17284*.

GLENN W. BRIER. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Sérgio D. Canuto, Marcos André Gonçalves, and Fabrício Benevenuto. 2016. Exploiting new sentiment-based meta-level features for effective sentiment analysis. In *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.

Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, and Seán Slattery. 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.

Washington Cunha, Sérgio Canuto, Felipe Viegas, Thiago Salles, Christian Gomes, Vitor Mangaravite, Elaine Resende, Thierson Rosa, Marcos André Gonçalves, and Leonardo Rocha. 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.

Washington Cunha, Celso França, Guilherme Fonseca, Leonardo Rocha, and Marcos André Gonçalves. 2023a. An effective, efficient, and scalable confidence-based instance selection framework for transformer-based text classification. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 665–674.

Washington Cunha, Vítor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara M Almeida, et al. 2021a. 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.

Washington Cunha, Vitor Mangaravite, Christian Gomes, Sérgio Canuto, Elaine Resende, Cecilia Nascimento, Felipe Viegas, Celso França, Wellington Santos Martins, Jussara Almeida, Thierson Rosa, Leonardo Rocha, and Marcos A. Goncalves. 2021b. 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.

Washington Cunha, Alejandro Moreo, Andrea Esuli, Fabrizio Sebastiani, Leonardo Rocha, and Marcos André Gonçalves. 2025. A noise-oriented and redundancy-aware instance selection framework. *ACM Trans. Inf. Syst.*, 43(2).

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

Claudio M.V. de Andrade, Fabiano M. Belém, Washington Cunha, Celso França, Felipe Viegas, Leonardo Rocha, and Marcos André Gonçalves. 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.

DeepSeek-AI et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186.

John Fields, Kevin Chovanec, and Praveen Madiraju. 2024. A survey of text classification with transformers: How wide? how large? how long? how accurate? how expensive? how safe? *IEEE Access*, 12:6518–6531.

Celso França, Rennan C Lima, Claudio Andrade, Washington Cunha, Pedro OS Vaz de Melo, Berthier Ribeiro-Neto, Leonardo Rocha, Rodrygo LT Santos, Adriana Silvina Pagano, and Marcos André Gonçalves. 2024. On representation learning-based methods for effective, efficient, and scalable code retrieval. *Neurocomputing*, 600:128172.

Mingqi Gao, Xinyu Hu, Jie Ruan, Xiao Pu, and Xiaojun Wan. 2024. Llm-based nlg evaluation: Current status and challenges. *arXiv preprint arXiv:2402.01383*.

Tyler Griggs, Xiaoxuan Liu, Jiaxiang Yu, Doyoung Kim, Wei-Lin Chiang, Alvin Cheung, and Ion Stoica. 2024. M\'elange: Cost efficient large language model serving by exploiting GPU heterogeneity. *CoRR*, abs/2404.14527.

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

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 142–150. The Association for Computer Linguistics.

Pekka Malo, Ankur Sinha, Pekka J. Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4):782–796.

Luiz Felipe Mendes, Marcos Gonçalves, Washington Cunha, Leonardo Rocha, Thierson Couto-Rosa, and Wellington Martins. 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.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon LLM: outperforming curated corpora with web data, and web data only. *CoRR*, abs/2306.01116.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *CoRR*, abs/1912.00741.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 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.

Arthur Spirling. 2023. Why open-source generative ai models are an ethical way forward for science. *Nature*, 616(7957):413–413.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: Extraction and mining of academic social networks. KDD '08, page 990–998, New York, NY, USA. Association for Computing Machinery.

Felipe Viegas, Sergio Canuto, Washington Cunha, Celso França, Claudio Valiense, Leonardo Rocha, and Marcos André Gonçalves. 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.

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.

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2024. Small models are valuable plug-ins for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 283–294, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Bruna Stella Zanotto, Ana Paula Beck da Silva Etges, Avner Dal Bosco, Eduardo Gabriel Cortes, Renata Ruschel, Ana Claudia De Souza, Claudio MV Andrade, Felipe Viegas, Sergio Canuto, Washington Luiz, 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.

A Prompt Templates

Below are examples of the structure of the prompts we used for our experiments' Zero-Shot and In-Context versions of the LLMs. They all draw on (Liang et al., 2023). Table 5 provides the prompt used for sentiment classification, while Table 6 presents the prompt for topic classification. 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 exclusively considers the token with the maximum 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 7, 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.

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 5: Prompt template for sentiment 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**}

Table 6: Prompt template for topic classification.

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**}

Table 7: Prompt template for sentiment classification for In-Context Llama and CMBS In-Context.

B Datasets

Our study draws on **thirteen** datasets developed for topic and sentiment classification. Our choice was strategically purposeful due to the effort to perform an in-depth analysis of this task. The datasets include **Finance** (Malo et al., 2014) focusing on economic news, **IMDB** (Maas et al., 2011)[12] compiling movie reviews as well as **PangMovie** (Pang and Lee, 2005) including Rotten Tomatoes[13] 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., 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

[12] https://www.imdb.com/
[13] https://www.rottentomatoes.com/

(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., 2021b), DBLP (Tang et al., 2008), Twitter Topic (Antypas et al., 2022) and WebKB (Craven et al., 1998).

	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
	RottenT2024	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

Table 8: Datasets Statistics.

As detailed in Table 8, 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.

C Evaluating LLMs

We evaluate four LLMs in Zero-shot mode: Falcon 7B, Mistral 7B, LLaMA 3.1-8B and DeepSeek 8B. Table 9 presents the results regarding Macro-F1, with the best outcomes highlighted in bold. As observed, LLaMA is consistently the best performer, either alone or tied with Mistral, across **all** datasets, except WebKB.

Dataset	Falcon 7B	Mistral 7B	Llama 3.1 8B	DeepSeek 8B
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±0.3**	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±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**	94.6±0.8
RottenT2024	65.8±1.3	93.8±1.2	**95.3**±1	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±0.3	**62.2±2.1**	**63.5±1.7**	**63.6±1.2**
Webkb	3.8±0.3	**42.1±0.6**	37±2.1	24.1±3.5

Table 9: Effectiveness in 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) marked in **bold**.

D Evaluating SLMs

We evaluate four fully-tuned SLMs: BART, BERT, Qwen 0.5B and RoBERTa. Table 10 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.

Dataset	BERT	BART	Qwen 0.5B	RoBERTa
Finance	94.1±3.8	**97±1.7**	67.8±14.2	**98.1±1.9**
Imdb	91.7±0.4	**92.8±0.4**	81.6±1.3	**93±0.5**
PangMovie	87.5±0.7	**88.4±1**	72.5±1.8	**88.7±0.9**
SemEval17	90.3±0.3	**91±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**
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**
RottenT2024	92.5±1	**93.5±0.5**	84.6±0.8	**93.7±1.1**
ACM	**69.8±1.8**	68±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 10: Average Macro-F1 and 95% confidence interval for SLMs . Best results (including statistical ties) marked in **bold**.

E 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 the LLM effectiveness using the validation set and a sample of two datasets. The pattern of results is basically the same in all other datasets we experimented with.

Table 11 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.

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

Table 11: Evaluate amount training LLM.

F Evaluating Threshold L

We evaluate the impact of the parameter L, which determines the number of documents sent to the LLM. The higher the value of L, the more documents fall below the threshold, increasing the number of documents forwarded to the LLM. Table 12 presents this evaluation for a sample of four datasets, showing the dataset name, the percentage of instances sent to the LLM relative to the total of test instances, and the effectiveness of both the SLM and LLM on this subset of instances. It is interesting to notice that the choice of L that maximizes the cost-effectiveness threshold (around 0.9) is similar in all datasets, and that the LLM effectiveness in these hard-to-classify instances is better than the SLM, which justifies the CBMS Partially-Tuned-IS Aggressive gains.

Dataset	Percentage of Instances	SLM Macro-F1	LLM Macro-F1	Threshould (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

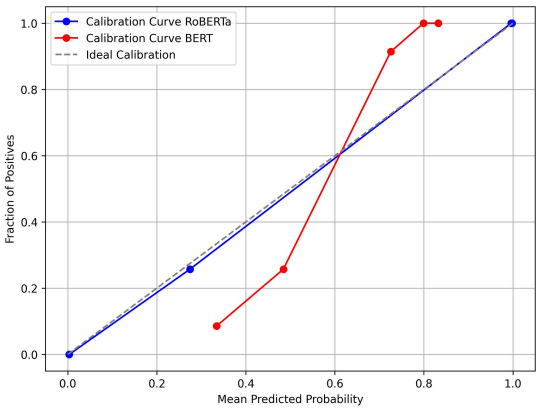
Table 12: Evaluation Threshold L.

G Calibration Model

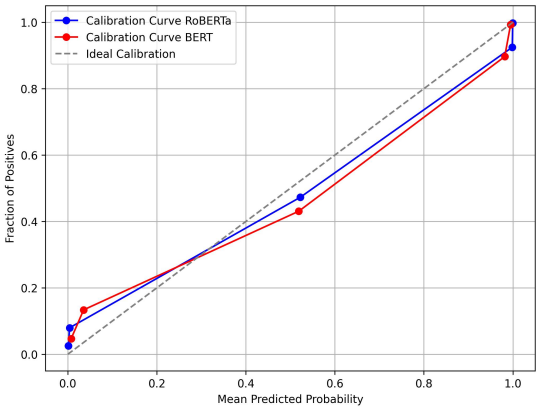
We analyze the behavior of the Transformer calibrations using Brier Score (BS) (BRIER, 1950), a scoring rule applied to measure the accuracy of probabilistic predictions. Brier (BRIER, 1950) defines $BS = \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}(P(Y = y_c|x_i) - o_{ci})^2$, where o_{ci} is the one-hot vector with a value 1 in the index of the true class of x_i, and 0 otherwise. BS ranges from 0 (worst) to 1 (best) – the closer to one, the better in achieving more calibrated probability estimations. Table 13 presents the obtained Brier Score averaged across the datasets by applying BERT and RoBERTa classifiers on all datasets considered in our experiments. Meanwhile, Figure 5 presents the calibration curves for BERT and RoBERTa, reinforcing that RoBERTa is a better-calibrated model compared to BERT.

H CO_2 emissions

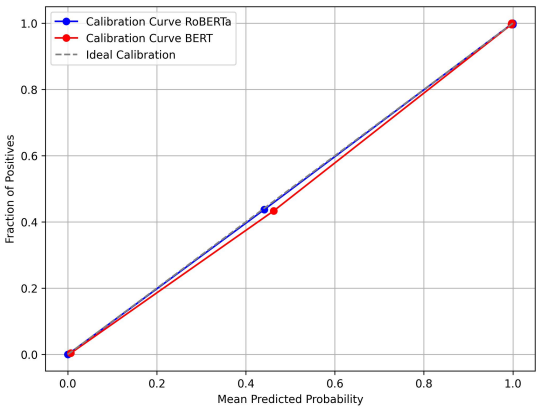
We calculated the CO_2 emissions associated with the execution of the model using the methodology developed by Lacoste et al. (2019). It is possible to associate the value of emission 0.14 kg of CO_2eq per hour with a machine of similar structure to



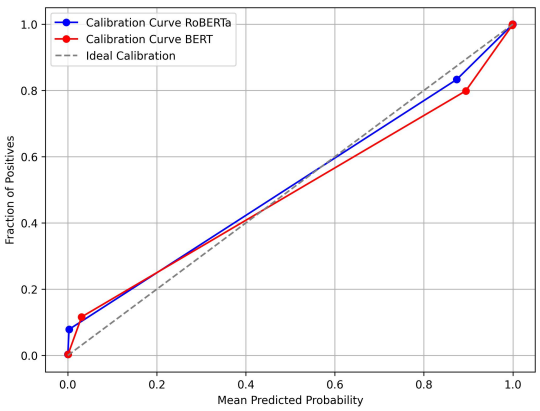
(a) Finance



(b) Pang Movie



(c) IMDB2024



(d) RottenT2024

Figure 5: Calibration curve 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

Table 13: Brier score for BERT and RoBERTa.

the one used in our experiments[14]. The emission values are presented in Table 14.

I Financial Cost

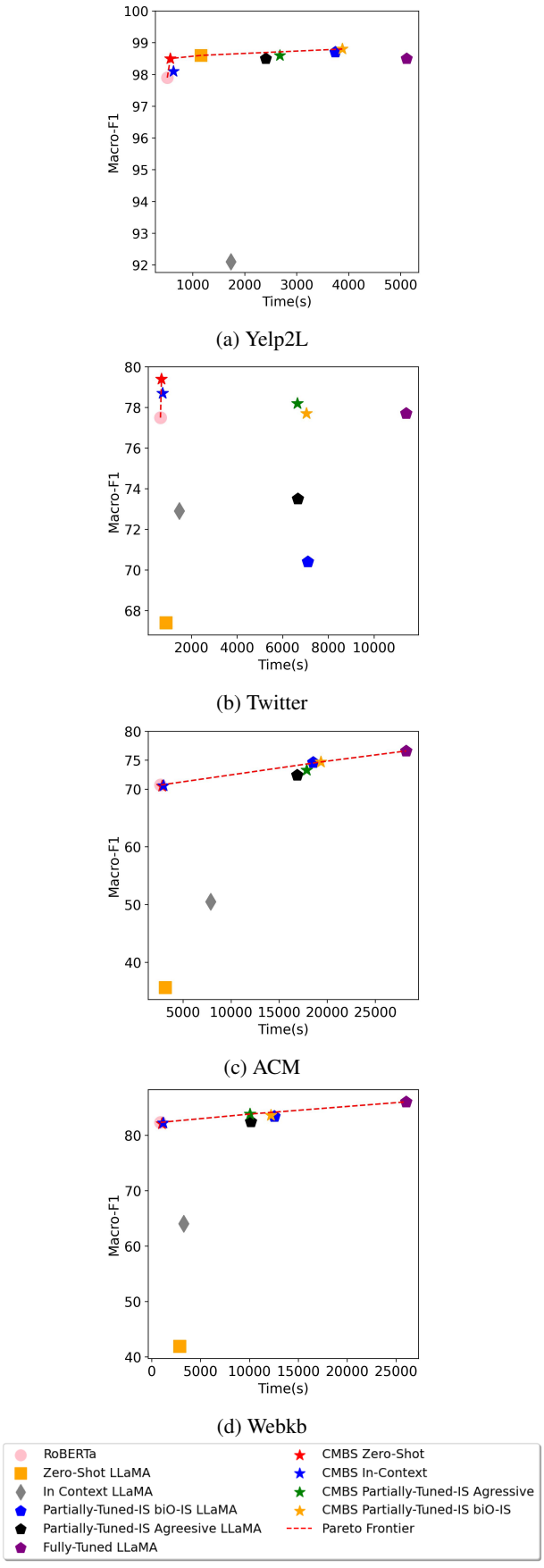
Following (Griggs et al., 2024), Table 15 presents the financial cost in dollars for executing the main methods discussed in this paper. As a reference, we used the hourly price of a setup similar to the one used in this research [15], offered by a large cloud company, which currently costs $0.752 per hour.

J GLUE CoLA

To demonstrate the potential of our solutions for other NLP tasks, we present the application of the best version of CBMS (CBMS Partially-Tuned (Aggressive) to CoLA (Corpus of Linguistic Acceptability), a GLUE task that evaluates a model's ability to predict the grammaticality of an English sentence by identifying it as either grammatically acceptable or unacceptable. For example, the sentence "The cat sat on the mat" is grammatically correct, whereas the sentence "On the mat sat cat the" is grammatically incorrect. Table 16 presents the results. As it can be observed, our solution outperforms RoBERTa and is statistically equivalent to Fully-Tuned at approximately half the computational cost.

K Pareto-optimal

Figure 6 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 6: 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.

[14] https://mlco2.github.io/impact/#co2eq
[15] aws.amazon.com/ec2/instance-types/g4/

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: 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	0.08	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

Table 15: 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	CMBS Partially-Tuned-IS Agressive	Fully-Tuned LLM
GlueCola	80.4±1.8	**83.4±2.1**	**84.5±0.4**

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**.