American Academic Scientific Research Journal for Engineering, Technology, and Sciences
ISSN (Print) 2313-4410, ISSN (Online) 2313-4402
https://asrjetsjournal.org/index.php/American_Scientific_Journal/index

Bridging Zero-Shot and Fine-Tuned Performance in Text Classification through Retrieval-Augmented Prompting

Olesia Khrapunova*

Senior AI, ML Engineer, Paris, France
Email: khrapunova.ml@gmail.com

Abstract

Large Language Models (LLMs) have shown promise in zero-shot and few-shot classification. Yet, their performance often falls short of classic fine-tuned encoders, especially in fine-grained or domain-specific settings. This study compares fine-tuned BERT-family models with zero-shot and few-shot prompting of LLMs (GPT-4o, Llama 3.3 70B, and Mistral Small 3) on two benchmarks: AG News (coarse-grained topic classification) and BANKING77 (fine-grained intent classification). Baseline results confirm that fine-tuned models outperform zero-shot LLMs by ~10-25 points in accuracy, with a larger gap on the fine-grained task. We then test training-free methods to improve LLM performance, focusing on retrieval-augmented few-shot prompting, example ordering, and Chain-of-Thought (CoT) reasoning. Our results show that retrieval-augmented prompting consistently boosts accuracy, especially on the BANKING77 dataset with many semantically similar examples, where GPT-4o even slightly surpasses the best fine-tuned encoder. Ordering demonstrations from least to most similar further improves accuracy, reflecting the impact of recency bias in in-context learning. By contrast, CoT prompting decreases accuracy, suggesting that explanation-based prompting is not universally helpful for classification. These findings demonstrate that careful example selection and ordering can substantially narrow the gap between zero-shot LLMs and fine-tuned encoders, offering a practical, training-free alternative in data-scarce scenarios.

Keywords: Chain-of-Thought Prompting; In-Context Learning; Large Language Models; Prompt Engineering; Retrieval-Augmented Learning; Text Classification; Transformer Encoders; Zero-Shot Learning.

Received: 7/30/2025
Accepted: 9/30/2025
Published: 10/11/2025

---

* *Corresponding author.*

1. Introduction

1.1. Problem Description

Large Language Models (LLMs) have demonstrated impressive capabilities in zero-shot and few-shot settings, enabling practitioners to perform text classification without task-specific training data. By leveraging prompt engineering, an LLM can classify new inputs directly, bypassing the traditionally mandatory model fine-tuning step. This flexibility is especially valuable in scenarios where labeled data, essential for fine-tuning, is scarce and costly to acquire.

However, despite these advantages, zero-shot LLM performance often lags behind that of fine-tuned supervised models (such as BERT or RoBERTa), particularly in fine-grained or domain-specific classification tasks. This raises two key practical questions:

- How large is the performance gap between zero-shot LLMs and fine-tuned models in realistic classification tasks?
- Can we close that gap without fine-tuning the LLM, using only lightweight inference-time interventions?

In this work, we address these questions in two phases. First, we establish a clear, controlled baseline comparison between zero-shot LLM classification and fine-tuned transformer encoders across datasets of varying difficulty. We quantify the performance gap in terms of accuracy, highlighting where zero-shot methods fall short. Second, we explore a series of training-free upgrades designed to improve LLM classification performance without altering model weights. Specifically, we focus on retrieval-augmented in-context learning methods and their variations.

Our contributions are twofold:

- Gap quantification: A systematic head-to-head comparison of fine-tuned models and zero-shot LLMs across coarse- and fine-grained classification benchmarks.
- Training-free LLM lift: An evaluation of inference-time interventions that reduce the performance gap without LLM fine-tuning.

By evaluating the scale of zero-shot underperformance and testing targeted inference-time mitigations, this study aims to bridge the gap between the convenience of LLM prompting and the predictive power of fine-tuned supervised models.

1.2. Related Work

Early work on LLMs showed that increasing the size of the language model enables strong zero-, one-, and few-shot performance across NLP tasks, establishing prompting as a viable alternative to task-specific training. For example, GPT-3 demonstrated competitiveness in the few-shot scenario on multiple NLP tasks, occasionally even surpassing state-of-the-art fine-tuned models (e.g., in open-domain question-answering and translation into English) [1]. However, the result was not universal, with the LLM still struggling on many task types and datasets,

underscoring the need for task- and dataset-specific model evaluation.

Follow-up studies on text classification offered a more nuanced picture. Bucher and Martini [2] compared major LLMs and fine-tuned encoders across several classification benchmarks and confirmed that zero-shot prompting alone is generally not sufficient, with smaller, fine-tuned models outperforming large LLMs. Domain-specific comparisons in political science reached similar conclusions and extended the analysis to few-shot settings, showing that while fine-tuned models consistently outperform LLMs in zero-shot mode, few-shot prompting narrows but rarely closes the gap [3,4]. Individual exceptions occur mainly in simpler tasks, where few-shot prompting can sometimes match the fine-tuned models. Large-scale analyses surfaced the same trend: Edwards and Camacho-Collados [5] showed that traditional fine-tuned models still surpass few-shot LLMs across 16 text classification datasets spanning multiple domains, while Zhang and his colleagues [6] reaffirmed that BERT-style models dominate in pattern-driven classification, with LLMs excelling mainly in knowledge-intensive tasks and those requiring deep semantic understanding.Importantly, the aforementioned studies paid little attention to prompting strategies and often relied on random selection of few-shot examples in their experiments. LLM research has now shown, however, that both the information included in a prompt and the way it is organized can cause significant variation in performance. A major line of work on in-context learning (ICL) highlights the importance of demonstration choice and ordering [7]. In this paradigm, LLMs learn a task by conditioning on a few labeled examples in the prompt, without updating their parameters. For example, Rubin and his colleagues [8] and Liu and his colleagues [9] find that adding semantically similar few-shot examples beats random selection, while Levy and his colleagues [10] demonstrate that incorporating not just semantically similar, but diverse examples is particularly beneficial for compositional generalization (the ability to combine known concepts in novel ways). Building on these insights, Qin and his colleagues [11] propose Iterative Demonstration Selection (IDS), which progressively refines the retrieval set to balance similarity and diversity, yielding accuracy gains across benchmarks. Ordering of examples also proves critical: Lu and his colleagues [12] show that it can make LLM performance range from state-of-the-art to almost random, and Zhao and his colleagues [13] report that models are biased towards the examples placed closer to the end of the prompt.In parallel, Chain-of-Thought (CoT) prompting has emerged as another powerful training-free strategy, particularly for reasoning-heavy tasks. Wei and his colleagues [14] showed that including just a few examples with explicit reasoning chains enables near state-of-the-art results on complex problems. Wang and his colleagues [15] further demonstrated that it is the presence and structure of reasoning, rather than its exact correctness, that drives these gains.

Taken together, the literature paints a consistent picture: (i) fine-tuned encoders remain strong reference points for text classification, outperforming zero-shot LLMs; and (ii) zero-shot LLMs can be substantially improved at inference time via training-free methods, like in-context learning and Chain-of-Thought reasoning. Our study builds on these threads by first quantifying the baseline gap between fine-tuned BERT-family models and zero-shot LLMs on the classification task of datasets of various difficulty and then attempting to close it with training-free upgrades. While prior work has typically examined these inference-time strategies in isolation, this study integrates them within a unified evaluation framework to assess their combined impact on LLM classification performance. In doing do, it connects the comparative analyses of model performance with recent advances in prompt optimization, exploring how established techniques can be systematically applied to bridge the long-standing gap between zero-shot LLMs and smaller fine-tuned encoder models.

2. Materials and methods

2.1. Data

To capture both coarse-grained and fine-grained text classification settings, we ran the experiments on two different benchmarks: AG News [16] and BANKING77 [17]. AG News is a large-scale news categorization dataset, consisting of news headlines and short descriptions labeled into four classes with a relatively clean separation: world, sports, business, and sci/tech. Constructed as a subset of a larger corpus of online news articles, it has a total of 120,000 training and 7,600 test samples. The BANKING77 is a fine-grained intent classification benchmark that consists of approximately 13,000 (10,000 train, 3,000 test) short customer service queries annotated into 77 categories. Many of these classes are semantically similar (e.g., card declined vs. card payment issue), making the class boundary more challenging to capture on this task.

2.2. Models

For our fine-tuned baselines, we focused on three widely adopted transformer encoders from the BERT family. BERT-base-uncased [18] serves as the canonical reference model for text classification and provides a historical baseline. RoBERTa-base [19] improves upon BERT with more robust pretraining and typically delivers stronger results across classification benchmarks. Finally, DeBERTa-v3-base [20] further enhances the pre-training efficiency of BERT and RoBERTa and outperforms them on several benchmarks. Together, these models offer a spectrum of increasingly powerful fine-tuned encoder models against which we compared zero-shot LLMs and training-free LLM lift techniques.

For our large language model baselines and further lift experiments, we selected three systems that span the spectrum from proprietary frontier models to efficient open-weight alternatives. GPT-4o [21] was included as the strongest closed-weight reference model. As an open-weight counterpart, we used Llama 3.3 70B Instruct [22], which has established itself as a leading open-source frontier model. Finally, we included Mistral Small 3 [23] to represent a smaller, efficiency-oriented class of models that still maintain competitive performance. Together, these three models provide a balanced view of current LLM capabilities across the accuracy-size spectrum.

2.3. Phase I: Baseline Comparison (Fine-Tuned BERT Models vs Zero-Shot LLMs)

We first established a clear, controlled baseline comparison between fine-tuned transformer encoders and zero-shot LLMs classification across the proposed datasets. Using accuracy as a consistent evaluation metric, we quantified the gap between fine-tuned and zero-shot performance. This phase provided a grounded understanding of how much zero-shot LLMs currently fall short and set a benchmark target for improvement.

During the fine-tuning process, for each dataset, we used the full available training set, fine-tuning the models for 3-5 epochs and using mainly default training settings. For zero-shot LLM classification, we created prompts closely based on those by Bucher and Martini [2] (see Appendix, Prompt 1) and provided the models only with a text to classify and a list of available labels.

We evaluated all the models exclusively on the test sets. Therefore, as the LLMs were not trained, they were not at all exposed to the training data in this step.

2.4. Phase II: LLM Performance Gains with Training-Free Methods

In the second phase, we systematically applied a set of inference-time techniques designed to improve zero-shot LLM classification performance without altering model weights. After exploring several learning-free approaches described in the Related Work section, we selected the following widely recognized techniques:

- Retrieval-augmented in-context learning (ICL), where task-relevant examples are dynamically selected and added to the LLM prompt;
- Chain-of-Thought (CoT) prompting, where the LLM is encouraged to provide reasoning before outputting the final answer.

To select few-shot examples for ICL, we prioritized semantic similarity over purely lexical overlap. Term-based retrieval methods such as BM25 [24] rely on surface-level word matching for candidate selection, whereas semantic retrieval captures the underlying meaning. This is an important factor in text classification, where subtle semantic distinctions often determine the correct label. For this purpose, we used a CrossEncoder model (ms-macro-MiniLM-L6-v2) from the Sentence-Transformers framework [25]. Unlike sentence embedding approaches, which embed queries and candidates separately and then compute similarity scores (e.g., via cosine similarity), the CrossEncoder jointly encodes each query-candidate pair and outputs a direct similarity score. This design provides higher retrieval accuracy at the cost of greater computational expense relative to embedding-based nearest-neighbor search, a trade-off we accepted to maximize the potential gains from strong example selection.

For each test instance, we first ranked candidate training examples within each target class using CrossEncoder similarity and retained the top five per class. For the BANKING77 dataset, we ranked all 10,003 training examples. For AG News, we limited the pool to 2,500 examples per class (10,000 total), as processing the full training set would have been prohibitively long and computationally expensive. At inference time, we then selected the ten most similar examples across all classes for inclusion in the prompt. This two-stage procedure ensured semantic relevance through similarity ranking while also preserving class diversity, since no class could contribute more than five candidates or exceed 50% of the demonstrations.

To study the effect of ordering, we compared two arrangements of demonstrations: most-to-least similar and least-to-most similar, motivated by the hypothesis that the latter would benefit from LLM recency bias [13].

For CoT prompting, we formatted few-shot examples (selected with the method above) as Query → Reasoning → Answer to elicit intermediate reasoning. Because the datasets lacked gold rationales, we generated reasoning steps automatically using Claude-3.5-Sonnet [26]. Prior work has shown that CoT does not require perfectly valid rationales to improve performance [15], so we accepted the trade-off between lower annotation cost and reduced explanation quality when using an LLM instead of a human annotator. To guide rationale generation, we included three hand-written justifications per dataset in the prompting process, each highlighting why the chosen label was

preferred over alternatives (see Appendix, Prompts 4 and 5).

We evaluated the proposed variations of ICL interventions in different combinations to better quantify their individual and joint contributions. As a result, we ran 4 different retrieval-based experiments of LLM training-free lift:

- **Few-Shot (best first):** 10 demonstrations added to prompt, ordered from most to least similar to query
- **Few-Shot (best last):** 10 demonstrations added to prompt, ordered from least to most similar to query
- **Few-Shot (best first) + CoT:** 10 demonstrations added to prompt, ordered from most to least similar to query, each demonstration includes an LLM-generated reasoning justifying label selection
- **Few-Shot (best last) + CoT:** 10 demonstrations added to prompt, ordered from least to most similar to query, each demonstration includes an LLM-generated reasoning justifying label selection

3. Results

3.1. Baseline Comparison: Fine-Tuned BERT Models vs Zero-Shot LLMs

We first share a comparison of fine-tuned transformer encoders and zero-shot LLMs on AG News and BANKING77. Across both datasets, fine-tuned encoders clearly dominated while remaining tightly clustered in performance. On AG News, BERT-base achieved the highest accuracy of 94.22%; on BANKING77, RoBERTa-base demonstrated the best performance, with an accuracy of 93.86%. However, the difference was minimal, with the difference between worst and best performance less than 0.5%, suggesting that even the simplest version (BERT-base) can reach high accuracy on these datasets with only a few epochs of training.

In contrast, zero-shot LLMs lagged substantially: on AG News, they trailed the best fine-tuned model by ~9-10% in accuracy, and on BANKING77 by ~19-26%. The much larger drop in performance on the BANKING77 dataset could be due to the higher label granularity, making it easier to make a mistake in the absence of training, especially when choosing between close labels. Notably, higher parameter count in LLMs did not automatically translate to better performance, with Llama 3.3 with 70B parameters falling behind Mistral Small 3 with 24B parameters on both tasks.

Table 1: Baseline BERT-family outperforms zero-shot LLM performance on AG News and BANKING77

Model group	*Model*	*Params*	AG News *Accuracy*	BANKING77 *Accuracy*
Fine-tuned BERT models	BERT-base	110M	**0.9422**	**0.9347**
	RoBERTa-base	125M	**0.9374**	**0.9386**
	DeBERTa-base	185M	**0.9411**	**0.9373**
Zero-shot LLMs	Mistral Small 3	24B	0.8457	0.7032
	Llama 3 70B	70B	0.8408	0.6805
	GPT-4o	> 175B*	0.8526	0.7461

*no official data, assumed based on GPT-3 size

3.2. LLM Performance Gains with Training-Free Methods

In-context learning reliably improves performance. On both datasets, retrieval-augmented few-shot prompting yielded consistent gains over zero-shot prompting for all LLMs, with much larger improvements on BANKING77 (fine-grained intents) than on AG News (coarse topics). On AG News, best configurations increased accuracy from 84-85% to 88-90%, which was still ~4-6% below the best fine-tuned BERT-family baseline. On BANKING77, the same techniques closed almost the entire gap, with GPT-4o even slightly surpassing the best fine-tuned encoder.

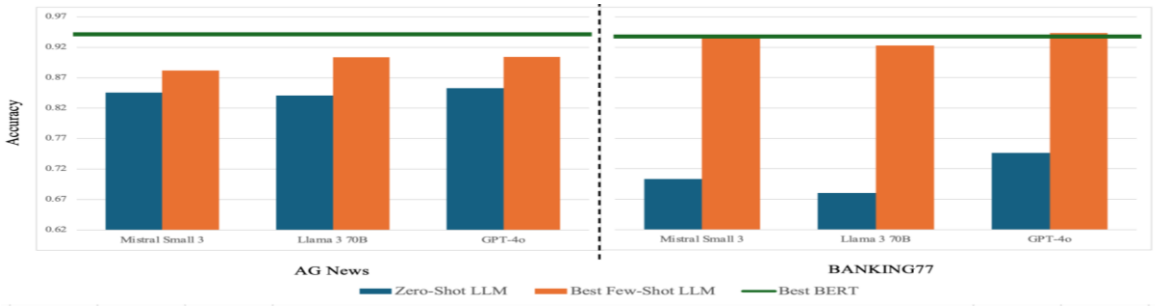

Figure 1: In-context learning boosts LLM performance, reaching near parity with fine-tuned encoders

Analysis of few-shot examples revealed that demonstrations in the BANKING77 dataset were semantically closer to the queries being classified, which may explain the larger gains from in-context learning on this dataset. When provided with highly similar correctly classified examples, LLMs are more likely to assign the accurate label to a new query. This suggests that datasets with many close few-shot examples naturally yield better performance, underscoring the importance of careful demonstration selection.

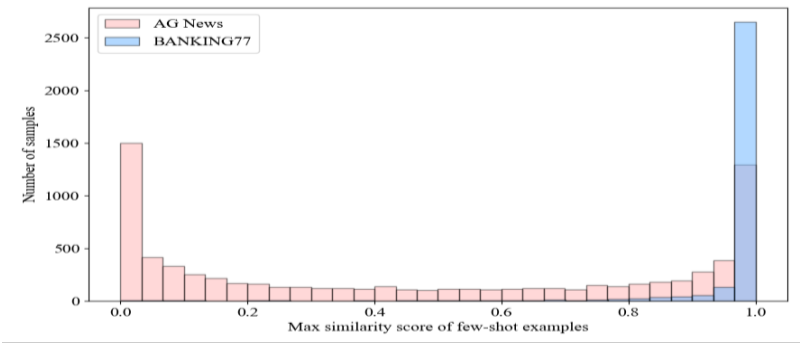

Figure 2: Few-shot examples in BANKING77 were more semantically similar to query than those in AG News

Example order matters. Across almost all model and dataset combinations, placing the most similar examples last ("best last" strategy) outperformed placing them first ("best first" strategy) by ~0.2-1.7%. A likely explanation is the effect of recency bias: demonstrations at the end of the prompt exert greater influence on the model's prediction. When these last positions are occupied by the examples most semantically similar to the query, the model is nudged toward the label that is more likely to be correct according to the retrieval (semantic similarity)

model. Notably, among the LLMs tested, Llama was the least responsive to the ordering of demonstrations, with performance barely changing across the ordering strategies.

Table 2: ICL with 'best-last' few-shot ordering has the largest LLM lift on AG news and BANKING77

	AG News				
Model	Baseline *Accuracy*	Few-Shot (best first) *Accuracy*	Few-Shot (best last) *Accuracy*	Few-Shot (best first) + CoT *Accuracy*	Few-Shot (best last) + CoT *Accuracy*
Mistral Small 3	0.8457	0.8658	**0.882**	0.8404	0.8689
Llama 3 70B	0.8408	0.9012	**0.9036**	0.8847	0.8867
GPT-4o	0.8526	0.887	**0.9042**	0.8722	0.8932
	BANKING77				
Model	Baseline *Accuracy*	Few-Shot (best first) *Accuracy*	Few-Shot (best last) *Accuracy*	Few-Shot (best first) + CoT *Accuracy*	Few-Shot (best last) + CoT *Accuracy*
Mistral Small 3	0.7032	0.926	**0.9373**	0.9032	0.9104
Llama 3 70B	0.6805	**0.9231**	0.9218	0.9104	0.9068
GPT-4o	0.7461	0.9383	**0.9432**	0.9195	0.9338

Reasoning hurts here. Adding a Chain-of-Thought reasoning consistently reduced accuracy relative to the no-reasoning prompts (by ~0.9-2.7%, with a larger drop on BANKING77). A closer analysis revealed that reasoning improved the classification of some examples, but at the dataset level, it generally hurt performance. For instance, on AG News classification with GPT-4o, adding reasoning changed the predicted label in 4.6% of examples: it degraded performance in 2.7% (flipping correct labels to incorrect), improved it in 1.6% (flipping incorrect labels to correct), and produced different but still incorrect labels in the remainder.

Inspection of model outputs suggests this decline stemmed from a mismatch between task demands and reasoning behavior. The classification tasks did not require deep reasoning but rather sensitivity to how annotators drew fine-grained class boundaries. LLM-generated rationales failed to capture this nuance, occasionally confusing the model's final prediction. For instance, the most noticeable effect for GPT-4o on the AG News dataset was an increased tendency to predict the 'business' class. While this shift sometimes aligned with the ground truth, it more often resulted in misclassifications. For example, the text *'Novell reshuffles biz for Linux focus Novell is reorganising its business to focus on two key areas - Linux and identity management…'* had the true label 'sci/tech'. GPT-4o instead labeled it as 'business', reasoning that *"this text discusses Novell's business reorganization to focus on Linux and identity management. While it involves technology elements, the primary focus is on the company's strategic business decisions and restructuring efforts, which align more with business news than purely technology updates."* This justification is coherent and well-argued but diverges from the dataset's labeling scheme, illustrating how CoT can encourage plausible reasoning that nonetheless conflicts with annotation guidelines.

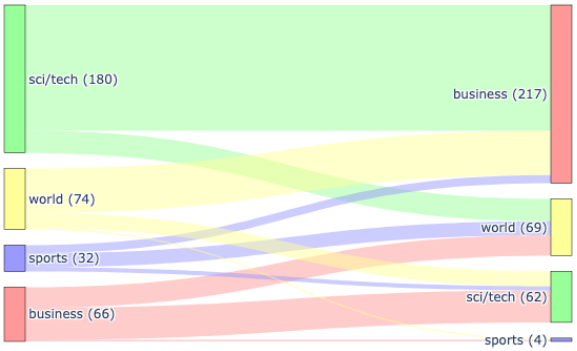

Figure 3: Adding reasoning led GPT-4o to produce more 'business' predictions on the AG News dataset

It is important to note that the exact label-shift patterns from reasoning varied across models, indicating that the effect was not driven by a bias in the few-shot rationales. Rather, a consistent theme emerged: after reasoning, each LLM produced a coherent justification, but one that often diverged from the implicit reasoning applied by human annotators. This mismatch led to system-level misclassifications.

4. Discussion

4.1. Discussion of the Results

This work set out to (i) quantify the accuracy gap between fine-tuned encoders and zero-shot LLM prompting, and (ii) test whether training-free, inference-time interventions could close that gap. The results directly addressed both aims.

First, our baselines showed a substantial gap on two datasets of different complexity (~10% on AG News and ~20-25% on BANKING77), reinforcing the finding that zero-shot prompting alone is typically not competitive with fine-tuned BERT-style models for classification. These results align with prior findings by Bucher and Martini [2] and Edwards and Camacho-Collados [5], while extending their analyses to newer LLMs and showing that, despite recent progress, the performance gap with fine-tuned encoders remains.Second, and most importantly, retrieval-augmented few-shot prompting substantially closed this gap. The improvement was particularly striking for fine-grained intent classification (BANKING77 dataset), where zero-shot LLMs typically struggle due to subtle label distinctions. By dynamically selecting semantically relevant examples, the retrieval process provided and implicit form of label disambiguation, allowing GPT-4o and Mistral Small 3 to perform on par with, or in some cases exceed, the best fine-tuned encoders. In contrast, gains on AG News were more modest (though still significant). One contributing factor was that BANKING77 contained many semantically similar examples within its fine-grained classes, making it easier for retrieval to supply demonstrations that closely matched the query. AG News, by comparison, featured fewer examples that align this tightly, which limited the benefit of retrieval. This underscores that the effectiveness of retrieval depends strongly on the availability of semantically close examples within the dataset. The analysis of prompt structure revealed another practical design factor – ordering. Demonstrations arranged from least to most similar consistently outperformed the reverse, corroborating the findings of Lu and his colleagues [12] and Zhao and his colleagues [13] regarding the importance

of the right sample order in in-context learning. This suggests that LLMs, when reading prompts sequentially, assign more weight to later examples, effectively treating them as stronger evidence for label prediction. Thus, even simple manipulations in example sequencing can translate into measurable performance gains, offering low-cost optimization technique for practitioners. However, this effect assumes that the few-shot examples themselves are of high semantic quality and correctly labeled. Poor or noisy demonstrations would be equally amplified by ordering, potentially degrading rather than improving performance.

Conversely, the Chain-of-Thought (CoT) prompting strategy reduced accuracy across all the considered datasets and models. This result underscores that reasoning-based prompting, though effective for tasks requiring explicit logical inference or multi-step problem solving [14,15], may be counterproductive in the classification context that is driven by pattern recognition rather than reasoning. The LLMs' rationales, although sound, often diverged from dataset-specific annotation logic, moving the model further away from rather than closer to the desired label.

All in all, this study addressed a well-documented limitation of LLMs – their underperformance in zero-shot classification compared to smaller fine-tuned encoders – by integrating known inference-time techniques that have individually shown promise for improving LLM accuracy. By systematically combining these techniques and testing them on both coarse- and fine-grained tasks, we demonstrated that the long-standing performance gap can be largely resolved through careful prompt design and examples selection rather than model retraining. Specifically, (1) retrieval-augmented prompting reaffirmed its value as a strong training-free alternative to fine-tuning, while highlighting that performance gains depend critically on the semantic similarity between retrieved examples and the text being classified; (2) the ordering of examples proved to be a deciding factor, corroborating earlier observation of demonstration-order sensitivity while offering clearer guidance on how recency bias can be leveraged in practice; and (3) reasoning-based prompting, though widely celebrated, was shown to be task-dependent and potentially detrimental when label definitions are not reasoning-driven. Collectively, these findings illustrate that the persistent zero-shot gap can be mitigated through strategic prompt engineering, combining retrieval, example ordering, and task-aware reasoning choices into a coherent, practical framework for text classification with modern LLMs.Practically, the significance of the study is twofold. First, organizations with limited labeling budgets can now feel confident pursuing text classification use cases they may have previously avoided due to the prohibitive cost of data annotation. Our results showed that, with newer LLMs and thoughtful prompt design, it is possible to achieve performance comparable to smaller fine-tuned models without extensive labeled data. Second, the study demonstrated what that "thoughtful" design entails in practice, offering a clear, reproducible recipe for applied practitioners: retrieve semantically similar examples, order them from least to most similar to the query, and avoid CoT prompting. This combination of design choices provides a low-cost pathway to high-performing LLM classifiers that approach fine-tuned accuracy without additional training.

4.2. Limitations and Future Work

Despite these promising results, several limitations of the study should be acknowledged. First, both datasets and models were limited in scope. The experiments were conducted on two English-language benchmarks (AG News and BANKING77) that capture coarse- and fine-grained classification. While the design offered a valuable contrast, the findings may not fully generalize to other domains or languages. Likewise, the analysis focused on

a small subset of large language models (GPT-4o, Llama 3.3 70B, and Mistral Small 3) and fine-tuned encoders from the BERT family (BERT, RoBERTa, DeBERTa). Although these represented a variety of models, results may differ on other LLMs, and traditional encoders not included in this study. Future work should expand the evaluation to a broader range of datasets and model families to assess the generality of the observed trends.

Second, the CoT rationales used in this work were automatically generated rather than human produced. Due to time and resource constraints, it was not feasible to create human-written rationales for thousands of examples in the datasets. While automatic generation helped overcome these limitations, it may have introduced reasoning artifacts that failed to capture underlying classification patterns and, as a result, reduced classification accuracy when CoT prompting was used. Future studies should examine whether human-verified rationales produce different outcomes.

Finally, the retrieval-augmented prompting setup of this study employed only a few methods, including semantic similarity ranking with a CrossEncoder and ordering by increasing or decreasing similarity. While this configuration provided a strong foundation for improving LLM performance, it represents only one of many possible designs. Additional techniques - such as calibration methods, automatic prompt optimization, or hybrid retrieval strategies that combine similarity with diversity – were not explored but may further enhance LLM performance in text classification and warrant investigation in future research.

5. Conclusion

This study establishes a clear baseline where fine-tuned encoders (BERT/RoBERTa/DeBERTa) outperform zero-shot LLMs and remain the most reliable approach on both coarse- and fine-grained text classification. We also show that training-free, retrieval-augmented few-shot prompting with a simple least to most similar ordering closes most of that gap, while Chain-of-Thought rationales consistently degrade performance. The practical recipe that emerges is straightforward: retrieve a handful of semantically similar demonstrations, order them from least to most similar, and skip CoT. This makes few-shot prompting a viable substitute when fine-tuning is infeasible.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal et al. “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[2] M. J. J. Bucher and M. Martini. “Fine-tuned ’small’ LLMs (still) significantly outperform zero-shot generative AI models in text classification,” arXiv:2406.08660v2 [cs.CL], Aug. 2024.

[3] M. Bosley, M. Jacobs-Harukawa, H. Licht, and A. Hoyle. “Do we still need BERT in the age of GPT? Comparing the benefits of domain-adaptation and in-context-learning approaches to using LLMs for Political Science Research,” presented at the Annual Meeting of the Midwest Political Science Association, Chicago, IL, USA, Apr. 2023.

[4] Y. Wang, W. Qu, and X. Ye. “Selecting between BERT and GPT for text classification in political science

research," arXiv:2411.05050v1 [cs.CL], Nov. 2024.

[5] A. Edwards and J. Camacho-Collados. "Language models for text classification: Is in-context learning enough?" in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* May 2024, pp. 10058-10072.

[6] J. Zhang, Y. Huang, S. Liu, Y. Gao, and X. Hu. "Do BERT-like bidirectional models still perform better on text classification in the era of LLMs?" arXiv:2505.18215v1 [cs.CL], May 2025.

[7] M. Luo, X. Xu, Y. Liu, P. Pasupat, and M. Kazemi. "In-context learning with retrieved demonstrations for language models: A survey," in *Transactions on Machine Learning Research*, Oct. 2024.

[8] O. Rubin, J. Herzig, and J. Berant. "Learning to retrieve prompts for in-context learning," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jul. 2022, pp. 2655–2671.

[9] J. Liu, D. Shen, Y. Zhang, B. Dolan, L. Carin, and W. Chen. "What makes good in-context examples for GPT-3?" in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, May 2022, pp. 100–114.

[10] I. Levy, B. Bogin, and J. Berant. "Diverse demonstrations improve in-context compositional generalization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jul. 2023, pp. 1401–1422.

[11] C. Qin, A. Zhang, C. Chen, A. Dagar, and W. Ye. "In-context learning with iterative demonstration selection," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Nov. 2024, pp. 7441–7455.

[12] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, May 2022, pp. 8086–8098.

[13] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. "Calibrate before use: Improving few-shot performance of language models," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139 PMLR, Jul. 2021, pp. 12697-12706.

[14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia et al. "Chain-of-thought prompting elicits reasoning in large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Nov. 2022, pp. 24824-24837.

[15] B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. "Towards understanding chain-of-thought prompting: An empirical study of what matters," in *Proceedings of the 61st Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), Jul. 2023, pp. 2717–2739.

[16] X. Zhang, J. Zhao, and Y. LeCun. "Character-level convolutional networks for text classification," in *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, Dec. 2015, pp. 649–657.

[17] I. Casanueva, T. Temčinas, D. Gerz, M. Henderson, and I. Vulić. "Efficient intent detection with dual sentence encoders," in *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, Jul. 2020, pp. 38–45.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jun. 2019, pp. 4171–4186.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen et al. "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:19007.11692v1 [cs.CL], Jul. 2019.

[20] P. He, J. Gao, and W. Chen. "DeBERTaV3: Improving DeBERTa using electra-style pre-training with gradient-disentangled embedding sharing," arXiv:2111.09543v4 [cs.CL], Mar. 2023.

[21] OpenAI. "GPT-4o system card." Internet: https://openai.com/index/gpt-4o-system-card/, Aug. 2024 [Aug. 7, 2025].

[22] HuggingFace. "Meta-Llama-3.3-70b-Instruct model card." Internet: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct, Dec. 2024 [Aug. 7, 2025].

[23] Mistral AI. "Mistral Small 3," Internet: https://mistral.ai/news/mistral-small-3, Jan. 2025, [Aug. 7, 2025].

[24] S. Robertson and H. Zaragoza. *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and Trends® in Information Retrieval, 2009. pp. 333-389.

[25] N. Reimers and I. Gurevych. "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019, pp. 3982–3992.

[26] Anthropic. "Introducing Claude 3.5 Sonnet." Internet: https://www.anthropic.com/news/claude-3-5-sonnet, Jun. 2024 [Aug. 7, 2025].

6.Appendix

We include prompts used in the experiments. Variables that changed from one query to another are included in curly brackets. Their names indicate the expected content.

```
You have been assigned the task of zero-shot text classification. Your objective is to classify a given snippet into one of several possible class labels, based on the content in the text. Your output should consist of a single class label that best matches the given text. Choose ONLY from the given class labels below and ONLY output the label without any other characters.

Text: ```{{ text }}```

Class Labels:
{% for label in labels %}
-           {{ label }}
{% endfor %}

Provide the prediction after an `Answer:` tag.
```

Figure 4: Used for zero-shot LLM prompting

```
You have been assigned the task of text classification. Your objective is to classify a given snippet into one of several possible class labels, based on the content in the text. Your output should consist of a single class label that best matches the given text. Choose ONLY from the given class labels below and ONLY output the label without any other characters.

Class Labels:
{% for label in labels %}
-           {{ label }}
{% endfor %}

{% for example in few_shot_examples %}
Text: ```{{ example.text }} ```
Answer: {{ example.label }}
{% endfor %}

CLASSIFY THE FOLLOWING TEXT:
Text: ```{{ text }}```

Provide the prediction after an `Answer:` tag.
```

Figure 5: Used for few-shot LLM prompting

```
You have been assigned the task of text classification. Your objective is to classify a given snippet into one of several possible class labels, based on the content in the text. Your output should consist of a single class label that best matches the given text. Think about it step by step - first provide the reasoning, then the final answer. Choose ONLY from the given class labels below and ONLY output the label without any other characters.

Class Labels:
{% for label in labels %}
-           {{ label }}
{% endfor %}

{% for example in few_shot_examples %}
Text: ```{{ example.text }} ```
Reasoning: ``` {{example.reasoning }}```
Answer: {{ example.label }}
{% endfor %}

CLASSIFY THE FOLLOWING TEXT:
Text: ```{{ text }}```

Provide the reasoning after a `Reasoning:` tag, and the prediction after an `Answer:` tag.
```

Figure 6: Used for few-shot LLM prompting with CoT reasoning

```
You have been assigned the task of providing reasoning for text classification. Your objective is to explain why, out of all the possible labels, the specific one was assigned to the text, based on the content of the text. Make sure that your reasoning matches only the assigned label and makes a clear distinction from other potential label choices.
Your output should consist ONLY of a short (30 words max) reasoning for why the text was chosen.

Possible Class Labels:
{% for label in possible_labels %}
-           {{ label }}
{% endfor %}


Text: ```How can my boss pay me directly to the card?```
Reasoning: ```Asks about how they could receive payment to the card. The main concern is receiving money to their account.```
Label: receiving_money

Text:  ```Someone might have access to my card!  There are some strange payments showing up on my account.  What do I do?```
Reasoning: ```Alerts about payments on their account that they did not make and asks for help. The main concern is an unrecognized payment.```
Label: card_payment_not_recognised

Text:  ```Won't let me top my card up!! Who can get it sorted out for me? It's urgent! How do I find out what the problem is with my card and how else to top it up?```
Reasoning: ```Alerts about inability to top up their card and asks for help resolving. The main concern is unsuccessful top up.```
Label: top_up_failed


PROVIDE REASONING FOR THE FOLLOWING TEXT CLASSIFICATION, MAKING A DIFFERENTIATION FROM OTHER CLOSE OPTIONS:
Text: ```{{ text }}```
Label: {{ label }}

Provide the prediction after a `Reasoning:` tag.
```

Figure 7: Used to generate reasoning for BANKING77 few-shot examples (with dataset-specific demonstrations)

You have been assigned the task of providing reasoning for text classification. Your objective is to explain why, out of all the possible labels, the specific one was assigned to the text, based on the content of the text. Make sure that your reasoning matches only the assigned label and makes a clear distinction from other potential label choices.
Your output should consist ONLY of a short (30 words max) reasoning for why the text was chosen.

Possible Class Labels:
{% for label in possible_labels %}
- {{ label }}
{% endfor %}

Text: ```Carlyle Looks Toward Commercial Aerospace (Reuters) Reuters - Private investment firm Carlyle Group,\\which has a reputation for making well-timed and occasionally controversial plays in the defense industry, has quietly placed its bets on another part of the market.```
Reasoning: ```This article focuses on news related to a specific investment company and its next business move.```
Label: business

Text: ```Kerry Campaign Helping With Fla. Recovery (Reuters) Reuters - Democratic presidential candidate John Kerry does not plan to visit Florida in the aftermath of Hurricane Charley because he's concerned his campaign entourage could distract from recovery efforts, he said Saturday.```
Reasoning: ```This article discusses news related to a presidential campaign in the context of a natural disaster. It focuses on politics and society.```
Label: world

Text: ```Oracle Sales Data Seen Being Released (Reuters) Reuters - Oracle Corp. sales documents detailing highly confidential information, such as which companies receive discounts on Oracle's business software products and the size of the discounts, are likely to be made public, a federal judge said on Friday.```
Reasoning: ```This article focuses on releasing sales documents of Oracle. Although it concerns a business, Oracle is a technology company selling software, so this is related more to tech news.```
Label: sci/tech

PROVIDE REASONING FOR THE FOLLOWING TEXT CLASSIFICATION, MAKING A DIFFERENTIATION FROM OTHER CLOSE OPTIONS:
Text: ```{{ text }}```
Label: {{ label }}

Provide the prediction after a `Reasoning:` tag.

Figure 8: Used to generate reasoning for AG News few-shot examples (with dataset-specific demonstrations)