Domain Specific Artificial Intelligence for Small Dateset

Anonymous ACL submission

Abstract

In the field of economics, analyzing market news for commodities like oil is crucial for forecasting trends and making informed decisions. The sheer volume of news data requires efficient methods for sentiment analysis. This thesis explores the use of language models for sentiment analysis within the oil commodity market, focusing on extracting information related to price, supply, and demand dynamics from daily news. The study investigates the efficacy of zero-shot and few-shot learning, along with the use of adapters for continuous training, in both small and large language models. It is hypothesized that few-shot prompt engineering offers a cost-effective and efficient solution for sentiment analysis in this context. The research examines the performance of various models, including those trained on domain-specific datasets and those continuously trained with adapters. The findings contribute to developing more accurate and efficient tools for economic analysis and forecasting, while also considering the environmental impact of different techniques.

1 Introduction

The rapid advancement of Domain-Specific Artificial Intelligence (DSAI) models has revolutionized complex tasks like economic analysis(Leck, 2022; Liu et al., 2020), promising unprecedented accuracy and efficiency. However, critical gap exists in the evaluation of these powerful models. Although research focuses primarily on the methodological efficacy of training and fine-tuning DSAI, the crucial perspectives of computational cost and energy consumption are often overlooked. This omission hinders the development of truly sustainable and scalable DSAI solutions, potentially leading to resource-intensive deployments that are neither economically nor environmentally viable.

To address this challenge, this research investigates the application of language models to the domain of economics, focusing on a multi-faceted evaluation framework that encompasses not only accuracy but also the computational and energy costs in a scenario where only a small labeled data exists. Domain specific data labeling requires subject matter expertise often at considerable cost. Traditionally, smaller models employing fine-tuning techniques have been favored for such tasks. However, the advent of larger language models (LLMs) has opened up new possibilities, particularly in scenarios with limited labeled data. This study explores the efficacy of these emerging paradigms, focusing on the potential of prompt engineering in LLMs to achieve comparable or superior performance with smaller models (SMLs).

This paper contrasts two distinct approaches. First, it investigates the performance of SMLs, such as RoBERTa-Base and RoBERTa-Large (Conneau et al., 2019), pre-trained on economic domain-specific datasets and enhanced with Quantized Low-Rank Adaptation (QLoRA) adapters (Hu et al., 2021; Dettmers et al., 2023) for supplementary task-specific training. Second, it explores the application of prompt engineering techniques, including zero-shot and few-shot prompting, in Llama3 (Dubey and et al., 2024) and MistralAI (Jiang et al., 2023) LLMs. In both cases, the models are specifically guided to activate economics-specific knowledge, ensuring domain relevance. For the few-shot approach, a limited number of examples are incorporated into the prompt to provide contextual guidance, while a system string is employed to prioritize the activation of relevant economic principles.

The central hypothesis underpinning this research posits that few-shot prompt engineering, employed within the framework of LLMs, presents a cost-effective and environmentally sustainable solution for sentiment analysis within commodity markets. This approach is hypothesized to attain performance levels that are either comparable to or surpassing those of fine-tuned SML models. The sentiment analysis focuses on three distinct categories pertinent to oil markets: price, supply, and demand. Each sentiment category is evaluated wherein "Positive" denotes a projected increase in the respective category, "Negative" signifies a projected decrease, and "Neutral" indicates an expectation of stasis or maintenance of the status quo. Specifically, within the context of oil markets, a positive sentiment regarding price forecasts an increase in oil prices, a negative sentiment concerning supply anticipates a reduction in oil availability, and a neutral sentiment towards demand suggests a stable level of oil consumption.

Rigorous evaluation of the performance and resource consumption of various language models provides valuable insights for developing more accurate, efficient, and sustainable tools for economic analysis and forecasting the economic dynamics of natural resources. The research contributes to a deeper understanding of how DSAI can be leveraged to enhance both the accuracy and the sustainability of economic insights. This, in turn, will empower stakeholders to make better-informed choices in a rapidly evolving economic landscape.

2 Review

A common approach to creating a domain-specific expert is fine-tuning the layers of the Pre-trained Language Models (**PLM**) and then adding one or two output layers, known as prediction heads (Wolf et al., 2020). Typically, these are feed-forward layers for classification. The bulk of the computation is applied to fine-tuning the language model to produce the desired representation of the input(Min et al., 2021). For BERT(Devlin et al., 2018) sequence classification tasks, suggestions are to fine-tune the representation of the special $[CLS]$ token and follow with a single feed-forward layer that classifies it as one of the task labels. It is common practice to temporarily freeze the language model layers while initially training the feed-forward layers and then unfreeze the language model gradually for additional fine-tuning(Howard and Ruder, 2018; Yang et al., 2019), there are fewer benefits to training the feed-forward layer alone. An alternative strategy is to fine-tune a separate small network of the Natural Language Processing(NLP) model and select only a few weights of the model to keep. When using adapters, such as LoRA(Hu et al., 2021), only the weights of the adapter are fine tuned. Multiple adapters could be fine-tuned, each targeting and using the same frozen NLP model for each fine-tuned adapter.

Large Language Models (LLMs) demand substantial video random memory(VRAM) resources, making memory-efficient fine-tuning and continuous training techniques essential. Quantized LoRA (QLoRA) (Dettmers et al., 2023) addresses this challenge by enabling the training and adaptation of LLMs using quantized representations. During training, gradients are back-propagated (Rumelhart et al., 1986) through the quantized weights to update the model (Dettmers et al., 2023). QLoRA further enhances efficiency through double quantization by quantizing the quantization constants themselves, achieving additional memory savings of approximately 0.37 bits per parameter (Dettmers et al., 2023). By combining these quantization strategies, the model can be effectively reduced and converted to a 4-bit representation of the models weights, facilitating adapter-based training and fine-tuning. Furthermore, QLoRA introduces Paged Optimizers, a mechanism designed to mitigate memory spikes that occur during gradient checkpointing. This innovation prevents out-of-memory errors, which have historically hindered large model fine-tuning on single machines.

Prompting or Prompt Engineering refers to adding natural language text, often short phrases, to the input and output to encourage pre-trained models to perform specific tasks(Yuan et al., 2021). **Zero-shot learning** employs a pre-trained language model to perform a task without any further training or fine-tuning(Brown et al., 2020). The prompt gives only the question and context string or "system string" that is sent to the conversational model. Prompts also promote a better alignment of the new task formulation with the pre-training objective, leading to better use of knowledge captured in pre-training. The closer match in pretraining objectives facilitates a **Few-Shot Learning** approach(Liu et al., 2021), especially for tasks with small training datasets. A good prompt can be worth hundreds of labelled data points(Sanh et al., 2021). Few-Shot learning refers to a setting where the model is given one or few examples within the prompt for the task at inference time as conditioning(Alec et al., 2019), but no model weights are updated(Brown et al., 2020). This method employs in-context learning, providing the language model with K instances of context-completion pairs, concluding with the target question for evaluation and subsequent answer generation. Few-shot learning results in a major reduction in the need for task-specific data(Brown et al., 2020). Compared to an approach based solely on questioning, the ability of few-shot to improve classification even in the presence of misclassified examples, highlights the influence of four key factors on few-shot (Min et al., 2022): the input label mapping, the input text distribution, the label space, and the format of the examples.

Prompts can be used for specific tasks of interest, such as sentiment analysis and directly used in a zero-shot, unsupervised setting; or in fully supervised or few-shot settings where all or part of the specific-task training data is available. A fixed template-style prompt can perform tuning of the PLM. This method has consistently been shown to improve performance over fine-tuning which reduces the cost of computation and is a key factor to help reduce the environmental footprint when using AI.

3 Setup

3.1 Datasets

This section describes the creation of the two data sets. Separate datasets are used for model pre-training and fine-tuning.

For pre-training, the Economic Abstracts (EA) dataset is comprised of 2.8 million abstracts of published Economic papers from 1900 to 2022, compiled from three repositories. The Research Papers in Economics (RePEc) repository abstracts were copied using RSync and then filtered to remove duplicate abstracts. Non-English abstracts were assessed using the Python *lang-detect* library and then removed from the dataset. Abstracts from the National Bureau of Economic Research (NBER) and the American Economic Association (AEA) repository were scraped using the Python libraries *Mechanize* and *BeautifulSoup*. All abstracts from this repository are in English. Abstracts from NBER and AEA were compared to those from RePEc to remove duplicates from the dataset. The combined dataset results in 2.8 million unique English abstracts.

The Oil Markets (OM) dataset is used to fine-tune the sentiment analysis of the different models. All source information comes from news articles in "The Oil Bulletin." From the gathered articles, 1000 paragraphs were randomly chosen split into 500 for training and 500 for evaluation. Economic experts from Canada's central bank labeled the dataset. The economists provided hierarchical labels for these sampled paragraphs. At the first level, paragraphs were labeled "Relevant" or "Not Relevant" to oil commodities. The next level subset "Relevant" paragraphs into one or more combinations of {"Prices", "Supply","Demand"} × {"Positive", "Neutral", "Negative,"}. Table 1 shows the counts of the labeled dataset categories. One sentence can be tagged with multiple labels because of the independence of prices, demand, and supply.

Label	Total Count
Relevant	300
Not Relevant	200
Prices Positive	25
Prices Negative	32
Prices Neutral	243
Supply Positive	60
Supply Negative	52
Supply Neutral	188
Demand Positive	16
Demand Negative	22
Demand Neutral	262

Table 1: The number of paragraphs containing the specified labels in the Oil Markets Data set. One sentence can contain multiple labels from Supply, Demand and Prices. If Not Relevant, then it is its only label. Each number is the total count of the label in the 500 sentence training dataset.

3.2 Zero Shot Learning

This study employs a zero-shot learning approach to evaluate the performance of two prominent language models, FinBERT and RoBERTa-Large, in establishing a baseline for the sentiment analysis of oil market articles. The objective is to assess the inherent capacity of these models to generalize to unseen data and tasks, thereby providing a benchmark for future research in this area. The energy cost for zero shot learning is the Tesla V3 GPU with 16GB of memory required for inference.

This study further extends the zero-shot learning paradigm to Mistral and LLama LLMs. To obtain economically relevant insights, we employ engineered prompts designed in consultation with economists. These prompts are structured to extract information pertaining to the impact of news articles on oil markets, specifically focusing on price, demand, and supply dynamics. The energy cost in this case is the inference cost using a A100GPU with 80GB of video memory.

To ensure accurate interpretation and response formatting, the GPT-4 engine was instrumental in developing the system message. This initial prompt gauges the relevance of an article to the oil market, effectively filtering out irrelevant information and streamlining the analysis. Articles deemed relevant are then subjected to three further prompts, each designed to assess the sentiment within each of price, demand, and supply categories. This structured approach allows for a granular analysis of market sentiment.

Outputs generated by these LLMs are evaluated against a validation set annotated by economists using the same labels as the training set for OM dataset. This comparative analysis will provide insights into the efficacy of LLMs in zero-shot economic analysis. The specific LLMs and SLMs employed for this zero-shot learning task are detailed in Table 2

3.3 Training Methodologies

This section details the training procedures employed to develop LoRA adapters for both RoBERTa-Large and Mistral-7B models to enhance their sentiment analysis within the oil market domain. Furthermore, we discuss the energy consumption associated with each training phase.

3.3.1 Adapter Training for RoBERTa-Large and Mistral-7B

To further refine the models' understanding of economic language for sentiment analysis in the context of oil market news, we employed LoRA (Hu et al., 2021) to train adapters for both Mistral-7B (Jiang et al., 2023) and RoBERTa-Large (Conneau et al., 2019). The EA dataset was used to impart domain-specific knowledge to these adapters. To mitigate the computational demands of LoRA, we incorporated quantization techniques (Dettmers et al., 2023), thereby reducing memory requirements and enabling training on the available hardware. This quantization also reduced the energy required compared to full fine-tuning.

Table 3 provides a detailed summary of the training hyper-parameters, including the specific low-rank matrices employed for each model. Notably, two RoBERTa-Large adapters were trained with distinct r-values (8 and 32) to investigate the relationship between adapter capacity and performance. The training durations for these adapters were 182 and 185 hours, respectively, which, assuming a similar 300 Wh power draw as pre-training, correspond to estimated energy consumption of 54.6 kWh and 55.5 kWh, respectively.

3.3.2 Adapter Integration and Model Evaluation

Following training, the adapters were integrated into their corresponding base models. For RoBERTa-Large, the adapter was incorporated into the pre-existing sentiment analysis model to augment its economic language comprehension. For Mistral-7B, the adapter was merged with the base model, enabling prompt-based interaction and allowing the model to process and respond to natural language prompts. This adapter enhanced Mistral-7B model, referred to as EconMistral AI, was evaluated using the same prompts employed in the zero-shot and few-shot learning experiments, ensuring consistency across evaluation methodologies. The training of EconMistral AI took 370 hours, corresponding to an estimated energy consumption of 111 kWh, again assuming a 300W power draw.

3.4 Fine-Tuning & Few-Shot

In conjunction with fine-tuning(Min et al., 2021) the SLMs, we explore the effectiveness of few-shot(Brown et al., 2020; Alec et al., 2019), learning employing LLMs. We utilize prompt engineering techniques to construct prompts that incorporate a single example of each sentiment label for each classification level. For the relevance classification stage, the prompt includes randomly selected examples of both "Relevant" and "Not Relevant" articles from the OM dataset. If an article is classified as relevant, three distinct prompts are subsequently activated. Each of these prompts contains examples of "Positive," "Negative," and "Neutral" sentiments specific to Price, Supply, and Demand. This few-shot learning approach is evaluated using Llama 3.1, Mistral AI, and EconMistral, with a temperature setting of 1 and a maximum output token limit of 1000.

Table 4 details the 5 fine-tuned models used in this study, all trained on the OM dataset. The LoRA-trained RoBERTa-Large models, which is trained on the EA dataset and further trained on the OM dataset, is referred to as EconRoBERTa. For the LLMs, fine-tuning was implemented on a single model using the QLoRA technique (Dettmers et al., 2023) to mitigate GPU memory constraints, and hierarchical prompts were used during the fine-tuning process. To establish performance baselines,

Name	Parameters(Millions)	temperature	Max Return Tokens
Llama3.1	8,000	1	1000
Mistral AI-v.02	7,600	1	1000
FinBERT	355	N/A	N/A
RoBERTa	355	N/A	N/A

Table 2: Zero Shot learning engines

Model	Epochs	Learning Rate	Rank (r)	Compute Time (Hours)	kWh
RoBERTa-Large	3	1e-4	32	182	54.6
RoBERTa-Large	3	1e-4	8	185	55.5
MistralAI-7B	3	1e-4	8	370	111

Table 3: Adapter Trained LoRAs, Hyperparamaters

we also fine-tuned RoBERTa-Large (Conneau et al., 2019) and FinBERT (Liu et al., 2020) as non-LLM benchmarks, while zero-shot results were used to benchmark LLM performance. The fine-tuning process demonstrated high energy efficiency, consuming less than 300 Wh of energy, equivalent to under one hour of computation on one A100 GPU.

3.5 Evaluation Dataset

Model performance is evaluated using accuracy in Table 5 and F1 score vs energy consumption in Figure 1 based on the 500 evaluation paragraphs from OM. The master labeling dataset were labeled by the same experts who created training OM dataset.

4 Results

4.1 Baseline Model and Continued Training

The RoBERTa-Large model establishes a robust baseline for the study, exhibiting a high degree of precision in the initial task of relevance labeling. However, its performance falters when extended to the more complex task of downstream label prediction, highlighting a limitation in its generalizability. Subsequent models developed through the continued training of RoBERTa largely fail to surpass this baseline performance, indicating diminishing returns from further training. A notable exception to this trend is the EconRoBERTa-Large-R8 model, which demonstrates a marked improvement specifically in the domain of downstream label prediction. This anomaly underscores the potential for targeted model enhancements to yield significant performance gains in specific tasks, even when broader continued training approaches prove ineffective.

Nevertheless, the broader context of these findings necessitates a critical evaluation of the trade-offs between performance optimization and resource expenditure. The observed marginal gains achieved through continued training, including the development of EconRoBERTa-Large-R8, must be weighed against the substantial environmental and computational costs associated with such endeavors, consuming an addittional 185 kWh of GPU resources. The relatively modest improvements across most models suggest that, in many cases, the resource investment may not be justified by the resulting performance enhancements. This highlights the increasing importance of considering the efficiency and sustainability of model development alongside traditional performance metrics in the field of natural language processing.

4.2 Model Performance

Table 5 presents a comparative analysis of the accuracy of various models in assigning sentiment labels to text related to oil markets. The models were evaluated on their ability to classify both the relevance of a given text to the oil market and the sentiment expressed regarding prices, supply, and demand. Accuracy, in this context, represents the proportion of correctly assigned labels to the 500 of the evaluated instances within the OM evaluation dataset, which served as the ground truth.

To illustrate the interpretation of the table, consider the performance of the Mistral AI models. The "Mistral AI Few Shot" model achieved a Relevance accuracy of 80.4%, indicating that it correctly identified texts relevant to the oil market in approximately four out of five instances. Furthermore, when focusing on texts deemed relevant to oil, this model demonstrated a 67.4% accuracy in correctly classifying sentiment related to oil prices. Similar interpretations apply to the "Supply" and

Model	Epochs	Learning Rate	R-value
EconRoBERTa-Large	3	$1e-4$	32
EconRoBERTa-Large	3	$1e-4$	8
RoBERTa-Large	3	$1e-4$	8
$EconRoBERTa_{pre}$	3	$1e-4$	8
FinBERT	3	$1e-4$	8

Table 4: Fine-tuned Models. The EconRoBERTa Models refer to the ones that received pre-training and continuous training.

"Demand" columns, where the model achieved accuracies of 65% and 63.4%, respectively. Comparing these figures to the "Mistral AI Zero Shot" model, we observe a slight improvement in relevance detection and supply sentiment classification when using a few-shot approach, while price and demand sentiment classification accuracy remained relatively consistent. These nuances highlight the impact of different prompting techniques on model performance across various facets of oil market sentiment analysis. Analogous interpretations can be applied to analyze the performance of the "EconMistral" variants and other models listed in the table.

5 Discussion

Large Language Models (LLMs) have exhibited remarkable capabilities across a variety of natural language processing tasks, particularly when augmented by techniques such as zero-shot and few-shot learning. This is exemplified in the comparative analysis of the EconMistral model, a Mistral AI model fine-tuned with a Low-Rank Adaptation (LoRA) adapter, and the standard Mistral AI model (version 0.2-Instruct). Both models demonstrate comparable performance gains when employing zero-shot prompts, suggesting that the LoRA adaptation, in this specific scenario, fails to confer a discernible advantage. This finding raises questions regarding the cost-effectiveness of the LoRA adapter, given the substantial computational resources required for its training, specifically the estimated 111 kWhs use of energy. The investment of such resources appears unjustified when the resulting performance improvement is negligible in the zero-shot learning context.

Furthermore, while the application of few-shot prompts does lead to observable improvements in the accuracy of relevance labeling, these gains do not significantly translate into enhanced downstream label prediction performance. This limited impact suggests a potential deficiency in the current methodology for selecting exemplars for few-shot learning. Specifically, the reliance on a uniform set of examples across different articles, irrespective of their thematic nuances or specific content, may be hindering the effectiveness of the approach. It is, therefore, plausible that a more nuanced strategy for exemplar selection, potentially one that dynamically tailors the examples to the specific context of the article being analyzed, is required to unlock the full potential of few-shot learning for downstream label prediction tasks. Future research should thus focus on investigating context-aware exemplar selection strategies to optimize the performance of LLMs in such applications.

5.1 LLM Performance and Prompt Engineering

The efficacy of large language models (LLMs) in specialized tasks, such as commodity sentiment analysis, is demonstrably influenced by the underlying engine architecture. A comparative analysis reveals that the Llama3.1 engine exhibits a lower degree of accuracy relative to the Mistral AI engine. This discrepancy in performance can likely be attributed to fundamental differences in the training methodologies and information retrieval mechanisms employed by each engine. The implication is that the inherent biases and strengths developed during the training process manifest as varying proficiencies in handling specific types of input and generating appropriate outputs.

Consequently, prompts that are finely tuned and highly effective for one engine may not yield equivalent results when applied directly to another. This observation underscores the critical role of prompt engineering as a pivotal technique in optimizing the performance of LLMs for specific applications. Tailoring prompts to the unique characteristics of a given engine, taking into account its training data, architecture, and retrieval mechanisms, is essential for maximizing accuracy and achieving desired

Model	Relevance	Prices	Supply	Demand
Mistral AI Zero Shot	76.9%	67.6%	63.6%	65.6%
Mistral AI Few Shot	80.4%	67.4%	65%	63.4%
EconMistral Zero Shot	75.3%	62.3%	60.1%	61.1%
EconMistral Few Shot	79.8%	68%	64.2%	65.6%
Llama3.1 Zero shot	60%	42.3%	45.7%	41.5%
Llama3.1 Few shot	63%	49%	43.9%	48.6%
EconRoBERTa-Large-R32	48.3%	6.1%	21.9%	3.8%
EconRoBERTa-Large-R8	60.3%	31.6%	21.5%	35.8%
FinBert	60.3%	30.5%	21.1%	33%
RoBERTa-Large (baseline)	73.5%	28.1%	19.2%	32.2%

Table 5: Accuracy table for each of the models evaluated on the 500 OM labeled dataset for evaluation. Each percentage represent the accuracy for the proper tested label.

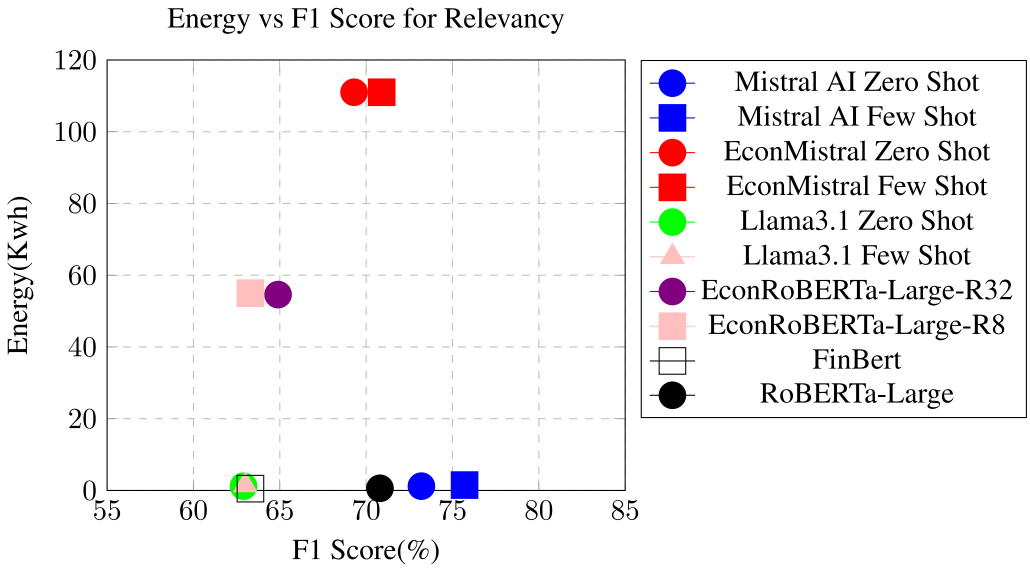

Figure 1: Llama3.1 and FinBert use roughly the same amount of energy with a difference of less than one Kwh.

outcomes. In the context of commodity sentiment analysis, these findings highlight the necessity of a nuanced, engine-aware approach to prompt design in order to fully leverage the capabilities of different LLMs and achieve optimal performance across diverse model architectures.

One of the key findings of this study is the substantial compute time and environmental impact associated with continuous training of both small and large language models. Observed performance gains, when present, proved negligible compared to less computationally expensive techniques. These findings underscore the critical need for energy-efficient training methodologies to mitigate the environmental and financial footprint associated with these computationally demanding processes.

The limitations in the performance of smaller language models when applied to the complex task of sentiment analysis in the commodity market. These models, even when fine-tuned with a small number of expert-labeled datasets, struggled to achieve adequate accuracy in predicting market sentiment. The financial and time burden associated with dataset augmentation presents a significant obstacle to this avenue of inquiry.

Large language models, especially when optimized through prompt engineering techniques, demonstrated superior performance in sentiment analysis. Few-shot learning emerged as a particularly promising approach, offering a cost-effective and efficient solution for sentiment analysis in this context.

This research is of significant value for economists specializing in the study of oil markets. Using the daily aggregate of news sentiment, generated by the few-shot models, they can gain a deeper understanding of past market trends and dynamics.

Future research should further explore the potential of techniques such as Adaptive Few Shot prompting and the use of Agentic knowledge graphs to improve the accuracy and robustness of sentiment analysis in commodity markets.

6 Limitations

Replication of this work necessitates utilization of the Azure Cloud platform for training the LoRA adapters and fine tuning wihtout a full refactor of the code. The economic abstracts were sourced exclusively from publicly available articles within the RePEc repository; while duplicate entries were largely eliminated, comprehensive verification of data uniqueness was not performed. Access to the proprietary, expert-annotated oil markets dataset, integral to this research, requires a specific license for the oil bulletin. Consequently, full reproduction would require the creation of a comparable dataset, comprising oil market data annotated by domain experts, correlated with publicly accessible economic publications.

While the master evaluation dataset was labeled by domain experts, the process was facilitated by startng from the Zero-Shot Learning results from MistralAI(Jiang et al., 2023) and then corrected by experts. This may add a bias when deciding if the label is correct, but is required to speed up the process.

7 Impact Statement

This paper aims to advance the field of Machine Learning while prioritizing responsible stewardship of environmental resources. The work specifically addresses the societal and environmental implications of high computational costs associated with certain machine learning tasks, using sentiment analysis as a case study. A key finding suggests that smaller datasets, coupled with existing generative AI models below 10 billion parameters, can offer a more environmentally sustainable approach to leveraging artificial intelligence technologies. This method potentially mitigates the significant energy consumption and carbon footprint associated with training and deploying large-scale models, thereby offering a path towards democratizing access to AI while minimizing its negative environmental impact. The research therefore not only contributes to the technical advancement of the field but also emphasizes the importance of considering the broader societal and environmental consequences of emerging AI applications.

References

Radford Alec, Wu Jeffrey, Child Rewon, Luan David, Amodei Dario, and Sutskever Ilya. 2019. Language models are unsupervised multitask learners | enhanced reader. *OpenAI Blog*, 1.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 2020-December.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

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

Abhimanyu Dubey and et al. 2024. The llama 3 herd of models.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. volume 1.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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

Sebastian Leck. 2022. *Macroeconomics, The Canadian Encyclopedia*. Historica Canada.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. volume 2021-January.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE International Conference on Computer Vision*.