Sustainable AI: Efficient Pruning of Large Language Models in Resource-Limited Environments

Anonymous Author(s)
Affiliation
Address
email

Abstract

The rapid growth and deployment of large language models (LLMs) like ChatGPT have revolutionized artificial intelligence, particularly in natural language processing, but they come with significant computational and environmental costs, including high energy consumption and carbon emissions. Addressing these challenges, our research introduces novel pruning techniques—"evolution of weights" and "smart pruning"—to enhance the efficiency of deep neural networks, especially on embedded devices. By systematically evaluating the importance of individual parameters during training, our methods achieve higher compression rates and faster computations while preserving accuracy, outperforming traditional pruning approaches. Extensive experiments with both scaled-down and larger multimodal LLMs demonstrate that moderate pruning can improve efficiency and reduce resource consumption with minimal accuracy loss, though excessive pruning can degrade performance. Our LLM experiment, available on GitHub, underscores the critical need for optimized AI models that balance technological advancement with ecological sustainability.

1 Introduction

Throughout their development, neural networks have witnessed significant advancements, beginning with the simple perceptron by [11] and expanding to the complex, multi-million-parameter Transformer-based models. The necessity for network optimization is highlighted by the rising computational expenses and the associated environmental concerns. Notably, the environmental toll of AI models, such as BERT and ChatGPT, is becoming increasingly apparent. BERT, with its 110 million parameters [15], has a carbon emission footprint akin to a transcontinental flight in the U.S. when trained on a GPU [12]. A heftier model, GPT-3, with approximately 137 billion parameters [4], accounts for carbon emissions equivalent to those of 13,483 Americans [10]. The daily operations of ChatGPT lead to the release of 3.8 tonnes of CO2, comparable to the carbon footprint of 93 Americans [10]. The environmental cost also encompasses water usage, which has drawn critical attention. For instance, GPT-3's training in top-tier U.S. data centers is associated with the consumption of 700,000 liters of water, sufficient to manufacture numerous cars [8]. A typical interaction involving 20-30 exchanges with ChatGPT uses about 500 ml of water [8]. While this might appear minimal, the rapid adoption rate of ChatGPT, with a surge of a million users in a mere five days [1], magnifies the overall environmental impact substantially. These points drive home the pressing need for optimizing networks to forge models that are not only efficient but also ecologically responsible.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

2 Related Work In pruning

This research presents an innovative approach to pruning deep neural networks, focusing on optimizing these models by removing less significant weights. Before exploring our method in detail, it's important to understand the prevalent techniques of structured and unstructured pruning that are commonly employed.

2.1 Unstructured and Structured Pruning

Unstructured pruning is a fundamental technique that involves setting individual parameters in a neural network to zero, effectively removing them from the model. This method, first introduced by Han et al. [5], has become a cornerstone for many subsequent pruning algorithms [2, 3]. The process typically begins after a model has been fully trained, where parameters that are close to zero are identified. A threshold is then established, and all weights below this threshold are zeroed out, enabling significant compression. Given that neural networks often contain millions of parameters, a large portion can be pruned without substantially impacting the model's performance. The disadvantage of this method is that it only provides theoretical compression since storing zeros still occupies memory. The actual storage costs are not reduced. Consequently, research has shifted towards pruning larger structural units, such as neurons in fully connected networks [6] and filters in convolutional networks [7, 9, 14], which helps in achieving more practical reductions in network size by thinning layers and decreasing the feature maps associated with the removed filters.

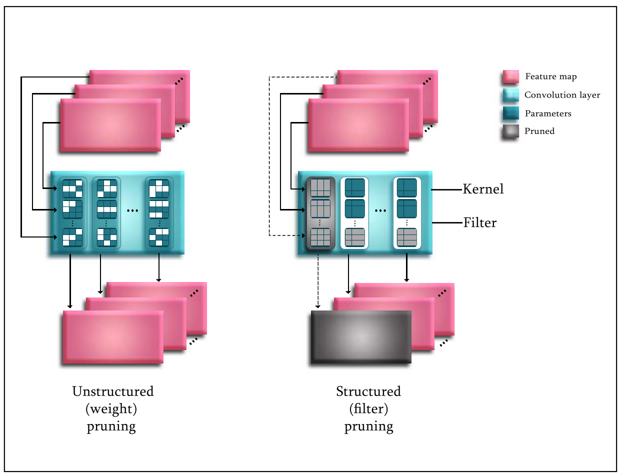

Figure 1: Structured Vs Unstructured Pruning. Adapted from [13]

Figure 1 gives examples of structured and unstructured pruning. Unstructured pruning offers more detailed granularity, but does not provide actual savings in storage cost. On the other hand, structured pruning, which involves removing parts of the layers, yields improvements in both time and memory efficiency at the cost of granularity.

2.2 Global and Local Pruning

Pruning can also be classfifed by its application scope. When applied globally to the entire neural network, it's referred to as global pruning, which can lead to the removal of entire layers, potentially causing layer collapse. Conversely, it is advisable to implement pruning on a layer-by-layer basis, known as local pruning, to prevent the complete elimination of any single layer.

Figure 2 depicts the differences between local and global pruning. In global pruning, the threshold is applied on the entire model. As a result there is a risk that all weights in a layer with low values could be eliminated. This leads to layer collapse.

In this research we propose an alternate method of choosing the weights that can be pruned and show how it can be used to compress a Large Language Model, without effecting its efficacy till certain levels of compression.

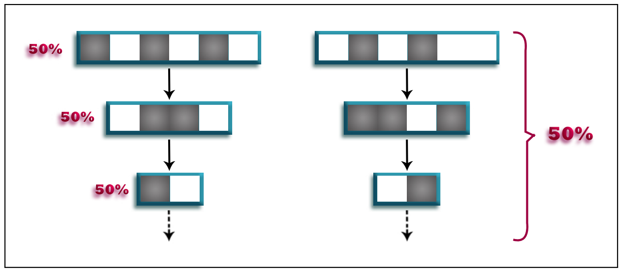

Figure 2: Global and Local Pruning. Adapted from [13]

3 Proposed Approach

The core aspect of the pruning method presented in this research involves tracking the evolution of parameters. This encompasses the regular observation of how parameter values change through the training epochs.

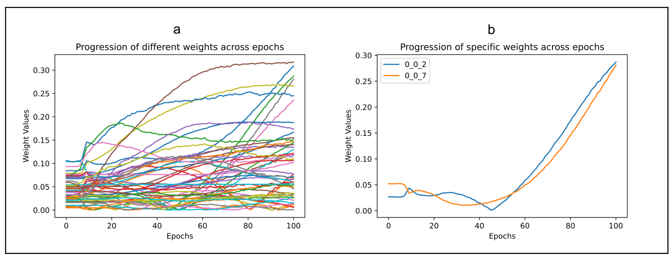

Figure 3: (a) Shows the development of randomly selected weights over 100 training epochs. (b) Demonstrates the progression of particular weights throughout the same 100 training epochs.

Figure 3 presents two graphs that depict the development of the network's weights: one graph tracks the change of weights selected at random across 100 training epochs, and the other graph focuses on the change trajectory of particular weights over an identical span. Within a neural network, the starting weight settings are chosen at random and then modified throughout training, with rates of change varying.

Our approach introduces a weighting system for the magnitudes of parameters, assigning more significance to those closer to the end of the epoch sequence but without neglecting earlier data. The importance score for each parameter is determined by multiplying its magnitude by a corresponding weight and averaging out these figures, which allows for the construction of an importance vector to clarify the parameter evolution through the epochs.

Table 1: To determine the significance of parameters over the course of training epochs, we track and record the value of each parameter at the conclusion of each epoch, organizing these figures into columns. The combined significance is obtained by performing a weighted summation of each weight's magnitude. The multipliers' values, displayed in the bottom row, indicate the extent to which each magnitude is factored into the calculation.

Weight #	Epoch 1	Epoch 2	Epoch ...	Epoch k	Aggregated Importance
1	4	6	...	3	17
2	8	9	...	5	15
3	6	8	...	8	5.66
4	2	5	...	9	4.66
Multiplier	$*1$	$*2$	...	$*k$	

As an example, to compute the weighted importance of a weight or filter, we compile a log of its magnitude values recorded at each epoch during training. This log aids in assessing the weighted significance according to the equation provided. The computed score reflects the significance of a parameter in terms of its magnitude and how it has changed over the entire training process. Table 1 lists magnitude recordings for weights over various epochs. When applying our method, it was observed that the most significant weights were Weight 1 (with a score of 17), Weight 2 (with a score of 15), and Weight 3 (with a score of 5.66).

For broader applicability of this calculation, we define a vector for every weight or filter ($val_i = [val_{i1}, val_{i2}, val_{i3}, ...val_{in}]$), with each entry corresponding to the weight's magnitude at a given epoch throughout the n epochs of training. This vector is the basis for computing the weighted significance, using the following equation which favors the most recent k epochs:

$$Imp_i = \frac{\sum_{L=0}^{k} val_{i(n-L)} * (n-L)}{\sum_{L=0}^{k} (n-L)} \tag{1}$$

Here, L varies from 0 to k, where 0 indicates the most recent epoch, and k counts back from the final epoch. The derived importance matrix thus becomes a pivotal tool for evaluating weight significance and informs the strategy for network pruning.

4 Experiment And Results

To check the consistency of our methods, two key experiments were conducted. These experiments focused on evaluating the effects of pruning, a process that reduces the number of parameters in a model, on model performance. The first experiment tested a scaled-down LLM trained from scratch, while the second involved a large pre-trained multimodal model. Both experiments aimed to determine how much compression could be applied to these models before significant performance degradation occurred. Before looking at the individual experiments, we take a look at the general procedure.

4.1 Record Weighted Average

In addition to directly training the model, a cloned version is maintained alongside it. The parameters of this clone are updated through a weighted average method that integrates historical parameter values across the training epochs. Initially, the cloned model's parameters are set to zero before the training starts. After each training step, both the original model's parameters and the corresponding parameters in the clone are updated. The updated values in the clone are computed as a weighted average, combining the existing parameters with the new ones from the original model, based on the current epoch. This approach ensures that recent updates are given more significance in the clone. The weighted average process, which gradually incorporates the model's parameter values over the epochs, is expressed as:

$$q_{\text{new}} = \frac{q_{\text{old}} \times S_{\text{prev}} + p \times (n+1)}{S} \tag{2}$$

Where:

- q_{new} are the updated parameters in the cloned model.
- q_{old} are the previous parameters in the cloned model.
- p are the current parameters in the original model.
- n is the current epoch number.
- S_{prev} is the sum of weights from epoch 1 to n (inclusive).
- S is the sum of weights from epoch 1 to $n+1$ (inclusive).

4.2 Model Training and Pruning

- Step I: The Transformer model is trained over 5000 epochs, with weight changes recorded throughout the process.

- Step II: After training, the model undergoes pruning based on the weighted parameters, followed by an additional 50 epochs of fine-tuning to maintain effective compression:
  - The importance of each parameter is assessed by evaluating its weighted absolute value:

$$W_{\text{abs}} = |W| \tag{3}$$

  - A pruning threshold is determined by scaling the standard deviation of these absolute values with a specific rate:

$$\text{Threshold} = \sigma(W_{\text{abs}}) \times \text{Prune Rate} \tag{4}$$

    Here, $\sigma(W_{\text{abs}})$ represents the standard deviation of W_{abs}, and "Prune Rate" is a constant that dictates the extent of pruning.
  - Parameters falling below the pruning threshold are set to zero:

$$P = \begin{cases} 0 & \text{if } W_{\text{abs}} < \text{Threshold} \\ P & \text{otherwise} \end{cases} \tag{5}$$

The effectiveness of Step II is evaluated by varying the pruning rates and observing the corresponding loss values.

4.3 Experiment and Model Details

The experiments summarized in Table 2 compare the performance of two different models under varying compression levels. Experiment 1 involved a scaled-down version of a ChatGPT-like Transformer-based LLM with 10.7 million parameters, trained on the complete works of Shakespeare. The model was subjected to pruning tests ranging from 0% to 94% compression, with performance tracked by the loss in the next-token prediction task. Experiment 2 used the Phi-3-vision model, a multimodal model with 4.2 billion parameters designed for both language and vision tasks, fine-tuned on a Burberry product dataset. The performance was evaluated by tracking the Mean Absolute Error (MAE) as the model underwent pruning at various compression levels.

Table 2: Summary of Experiments and Model Details

Exp #	Model	Model Type	#Parameters	Dataset	Training Procedure
1	GPT	Transformer based LLM	10.7M	Complete works of Shakespeare	Pruning from 0% to 94% compression; Performance tracked by loss
2	Phi 3 vision	Multimodal (Language + Vision)	4.2B	Burberry product dataset	Fine-tuning for 10 epochs; Pruning at various levels; Performance tracked by MAE

4.4 Results

In the first experiment, as shown in Figure 4 and Table 3, the scaled-down LLM demonstrated the ability to tolerate compression levels up to 60% without significant loss increases, reducing the loss to 1.656 from an initial 1.9. However, beyond 60%, there was a sharp escalation in loss, peaking at 3.098 at 94% compression, indicating that excessive pruning severely impacts model performance. The second experiment, depicted in Figure 5 and 4, involved the Phi-3-vision model and showed that initial pruning could enhance performance, reducing the Mean Absolute Error (MAE) from 439 to 374 at 10% compression. Nevertheless, aggressive pruning beyond 30% led to a dramatic rise in error, with the MAE surging to 11,041 at 48% compression. These results suggest that while moderate pruning can be beneficial, excessive pruning drastically deteriorates model performance in both cases.

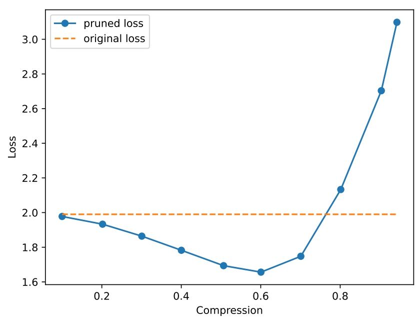

Figure 4: Loss as a function of compression levels, showing a decrease up to 60% compression, after which a sharp increase is observed.

Compression (%)	**Loss**
0	1.900
0.1	1.977
0.2	1.932
0.3	1.864
0.4	1.782
0.5	1.693
0.6	**1.656**
0.7	1.747
0.8	2.133
0.9	2.703
0.94	3.098

Table 3: The table details compression loss observed in Experiment 1, with a significant loss increase beyond 70% compression, consistent with trends in Figure 4.

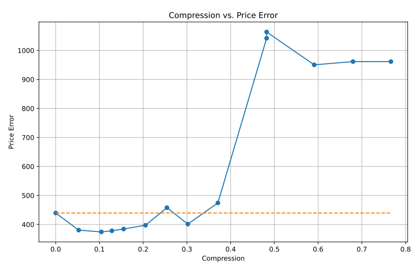

Figure 5: Price error as a function of compression levels. The figure demonstrates that while the model maintains a relatively low error up to moderate compression levels, the error escalates sharply beyond 30% compression, consistent with the MAE trends observed in Table 4.

Compression (%)	**MAE**
0	439
5	380
10	374
12	378
15	384
20	397
25	457
30	**401**
37	474
48	11041
59	950
67	961
76	961

Table 4: The table details the MAE observed across different compression levels, showing a significant increase in error beyond 30%, particularly at 48% compression, which aligns with the trends illustrated in Figure 5.

5 Limitations And Future Work

The approaches presented in this research offer a robust strategy for reducing the size of large-scale models, particularly large language models, without compromising performance. However, several limitations must be acknowledged. Fine-tuning LLMs for specialized use cases may restrict their applicability across diverse tasks, necessitating more adaptable solutions. As models grow in size, the proportion of parameters that can be effectively pruned diminishes, highlighting the need for more advanced techniques to handle large-scale models efficiently. Additionally, managing memory requirements for models with millions or billions of parameters remains a significant challenge, requiring memory-efficient strategies. Future work will focus on optimizing LLMs more efficiently to achieve tangible energy savings and sustainability, exploring smarter pruning methods to enable deeper compression while maintaining model accuracy and generalization capabilities. Balancing innovation with environmental responsibility will be crucial as the research community continues to advance AI technology.

References

[1] David Baidoo-Anu and Leticia Owusu Ansah. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and

Learning. *SSRN*, 2023.

[2] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *7th International Conference on Learning Representations, ICLR 2019*, pages 1–42, 2019.

[3] Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

[4] Matthew Gooding. Google takes on ChatGPT with new Bard chatbot and AI-powered search, 2023.

[5] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, pages 1–14, 2016.

[6] J K Kruschke and Javier Movellan. Benefits of Gain: Speeding Learning and Minimal Hidden Layers in Back-Propagation Networks. *Systems, Man and Cybernetics, IEEE Transactions on*, 21:273–280, 1991.

[7] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

[8] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. Making ai less "thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv*, 2023.

[9] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pages 2736–2744, 2017.

[10] Kasper Groes Albin Ludvigsen. The Carbon Footprint of ChatGPT, 2022.

[11] F Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[12] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, (1):1393–13696, 2020.

[13] Hugo Tessier. Neural Network Pruning 101: All you need to know not to get lost, 2021.

[14] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[15] Ryle Zhou. Question Answering Models for SQuAD 2 . 0. *Stanford Web*, (1):1–7.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, the paper accurately reflects its main claims by introducing novel pruning techniques to enhance the efficiency of large language models, focusing on reducing computational and environmental costs. The research effectively addresses the need for optimization by demonstrating how these techniques can achieve significant compression while maintaining performance, aligning well with the paper's stated goals of balancing technological advancement with ecological sustainability. The experiments and results presented support the paper's contributions, making the claims made at the outset credible and well-substantiated.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, the paper discusses the limitations of the work performed, acknowledging challenges such as the restricted applicability of fine-tuning for specialized use cases, the decreasing effectiveness of pruning as model size increases, and the significant memory requirements for managing large-scale models. The paper also emphasizes the need for more advanced techniques to address these limitations and outlines potential directions for future research to optimize large language models more effectively while maintaining performance and sustainability.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: nswerNA

   Justification: The paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, the paper provides sufficient information to reproduce the main experimental results, detailing the models used, the pruning techniques applied, and the evaluation metrics. It mentions placeholders for github link where code will be placed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Necessary details are provided to access the data and code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper explains how the data was split and used to fine tune models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Necessary evaluation metrics have been provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The paper mentions the resources used to train and test the models

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The authors have gone through the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines` to make sure that the research conforms to the same.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: It is a technical paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The model is under MIT License. `https://huggingface.co/microsoft/Phi-3-vision-128k-instruct`

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Citation has been made to the resource from which the model was used.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.