# CliMistral: Furthering Efficiency and Capabilities of Large Language Models in the Climate Domain

**Faye Titchenal**
UC Berkeley School of Information
fayetitchenal@berkeley.edu

## Abstract

Climate change poses one of the most significant threats to humanity, necessitating efficient methods to access and understand climate research. Large Language Models (LLMs) have shown promise in making climate data more accessible. However, the energy consumption associated with LLM training and inference can contribute to carbon emissions. This study introduces CliMistral, a 7-billion parameter model fine-tuned for climate-related question answering, aimed at enhancing performance while reducing energy consumption. We evaluate CliMistral's effectiveness in generating accurate responses to technical climate questions by benchmarking it against existing climate fine-tuned models. Our findings indicate that CliMistral achieves superior performance with less model complexity, highlighting the benefits of domain-specific fine-tuning and paving the way for carbon footprint reduction of LLMs in the climate domain.

## 1 Introduction

According to the World Health Organization, the greatest current threat to humanity is climate change. The effort to combat climate change requires comprehensive climate research and scientific recommendations to be broadly accessible and understandable. In many cases, this boils down to an ability to efficiently and accurately generate answers to technical questions based on the corpora of textual climate data.

In recent years, researchers and policymakers have successfully leveraged Large Language Models (LLMs) as a means to increase accessibility and understandability of climate research. While there are obvious benefits of leveraging LLMs for climate science tasks, LLM training and inference consumes a significant amount of energy which can result in significant carbon emissions (Budennyy et al., 2022).

Recent research demonstrates that domain fine-tuning can achieve superior model performance over baseline models on domain-specific tasks. In some cases, a smaller, fine-tuned model with fewer trainable parameters has achieved better performance to that of a larger, baseline model (Thulke et al., 2024). Furthermore, leveraging a pre-trained model removes the need for pre-training a new, domain-specific model from scratch. Reducing model complexity and the need for pre-training greatly reduces the energy demand and subsequent carbon footprint of the model.

## 2 Background

### 2.1 Task

We introduce CliMistral 7B, a fine-tuned Mistral model that generates answers to technical climate questions based on a passage of contextual input text. The aim of our research is threefold: 1) to determine if a fine-tuned Mistral model can outperform the foundational model on a climate QA task, 2) to compare the performance of CliMistral to existing climate fine-tuned models of varying model parameter size, and 3) to evaluate the impact of fine-tuning parameters on model performance and energy consumption.

### 2.2 Motivation

In 2021, researchers developed ClimateBERT, the first climate fine-tuned LLM for the purpose of processing and understanding climate-related texts (Webersinke et al., 2022). The authors fine-tuned the pre-trained DistilROBERTA model on a large corpus of climate-related documents in order to optimize the model for tasks like fact-checking. Their research demonstrated the ability of fine-tuned LLMs to learn the vocabulary and contextual complexities of textual climate data and opened the doors for further research and experimentation in this field.

ClimateQ&A was released in early 2023, a web-based chatbot used for question answering (QA) of over 30k climate-related questions since its release (De La Calzada et al., 2024). While the demand for this type of tool is evident, it was built on top of ChatGPT-3.5, a large model containing 175B trainable parameters. Therefore, there remains an opportunity to develop a less computationally expensive and less energy intensive solution for this task.

Another fine-tuned climate QA model, Mini-ClimateGPT leverages Vicuna as a baseline model, an instruction fine-tuned version of Meta's Llama-2 7B parameter model (Mullappilly et al., 2023). Mini-ClimateGPT is magnitudes smaller than ClimateQ&A, suggesting a reduction in energy required for training and inference. However, the generalizability of the model may be limited as synthetically generated data was used for fine-tuning.

In January 2024, researchers developed ClimateGPT, a family of LLMs for question-answering related to interdisciplinary research on climate change (Thulke et al., 2024). The Llama-2 LLMs were instruction and domain fine-tuned using a curated array of climate related data. ClimateGPT-7B achieved a higher accuracy than both the baseline Llama-2 7B and Llama-2 13B models. Interestingly, their research showed the out-of-the-box Mistral-7B model also achieved a higher accuracy than both the baseline Llama-2 7B and Llama-2 13B models.

The findings support that fine-tuning on climate specific data can yield superior model performance with a smaller, less complex model. However, using Mistral as the foundational pre-trained model may lead to further performance gains on climate-specific QA tasks. This finding motivates the foundational model selection for the research presented in this paper. To our knowledge, this is the first of a kind in fine-tuning Mistral for a climate question-answering task.

### 2.3 Foundational Model

Mistral 7B Instruct is a decoder-only transformer LLM developed and released open-source by Mistral AI in 2023 (Jiang et al., 2023). Mistral 7B Instruct outperforms both the Llama-2 7B and Llama-2 13B chat models on instruction benchmark datasets. The techniques and datasets used during pre-training of the model have not been publicly disclosed making it difficult to draw definitive conclusions about its performance over existing, larger LLMs on certain benchmark datasets.

## 3 Methodology

### 3.1 Data

Pirá 2.0 is a reading comprehension dataset about the ocean, Brazilian coastline, and climate change which was curated from a collection of scientific reports (Pirozelli et al., 2023). The dataset includes roughly 2k context paragraph, question, and answer sets. Despite its limited size, this high-quality dataset is well suited for evaluating a model's ability to learn highly technical and context specific climate concepts and vocabulary. A random subset, accounting for 10% of the total dataset, was used for benchmarking of the baseline and fine-tuned models. The dataset was originally designed for a multiple-choice QA task. To better address the need for generative QA in the climate sector, that data were modified for generative question-answering.

### 3.2 Baseline

Three pre-trained LLMs were selected as baseline models and evaluated on the benchmark dataset: Mistral Instruct 7B, ClimateGPT 7B, and ClimateGPT 13B. Mistral Instruct 7B was tested to establish model performance prior to domain and task fine-tuning. The ClimateGPT models were used to establish a benchmark for comparison to climate fine-tuned models on the Pirá 2.0 dataset. Additionally, the two sizes of the ClimateGPT model were used to study the relationship between fine-tuning and model complexity.

### 3.3 Metrics

Model inference was measured using the average ROUGE (1-gram, 2-gram, and L-gram) F1 scores calculated on the benchmark dataset. ROUGE is a standard metric for natural language generation tasks and is well-suited for evaluating exact word matches. The rigidity of this metric is appropriate in the context of climate science, where there are few appropriate synonyms for technical concepts. However, ROUGE does not account for semantic similarities and overall coherence of the generated text. To balance the limitations of ROUGE, BERTScore was used as a supplementary metric to evaluate sentence structure and contextual similarity between the model output and the label.

Measuring energy consumption and carbon emissions during LLM training and inference is an active area of research amongst ML and climate scientists. However, there is currently no industry standard for measuring and analyzing the carbon footprint of LLMs. Therefore, the Weights & Biases[1] API was used for estimating energy consumption during model fine-tuning due to its ease of use and automated data capture capabilities.

### 3.4 Training

Fine-tuning all 7B model weights of the Mistral Instruct model was not feasible given compute constraints. Therefore, QLoRA (quantized low-rank adapters), a parameter efficient fine-tuning method, was used due to these limited compute resources and to prevent overfitting as a result of the limited dataset size. A 40GB A100 GPU was used for both fine-tuning and model inference.

QLoRA combines two fine-tuning efficiency techniques: model quantization and trainable parameter reduction (Dettmers et al., 2023). Quantization reduces model complexity by representing model weights with lower precision data types. LoRA reduces the trainable parameters by approximating the original weight matrix with two, lower rank matrices. These decomposed matrices represent a fair approximation of the original weight matrix while consuming significantly less memory and requiring significantly fewer resources to train (Hu et al., 2021).

The LoRA hyperparameters were adjusted across three fine-tuned models to evaluate changes in model performance and energy consumption during training. The matrix rank (r), which determines the size of the decomposed weight matrices, was scaled linearly across each of the three fine-tuned models. Per existing guidance, the LoRA alpha ($\alpha$) parameter was scaled linearly with r to maintain a scaling factor of 2. Additionally, to target the most relevant trainable weights, only attention and linear layer weights were included in the matrix decomposition (Dettmers et al., 2023).

## 4 Results

### 4.1 Performance & Model Output

While ClimateGPT 13B performed the best of the three baseline models, CliMistral with 1% fine-tuned parameters performed the best overall on the test dataset (Table 1); providing evidence that domain fine-tuning improves performance with reduced model complexity on domain specific tasks. Fine-tuned model performance, however, is highly dependent on the parameters used for fine-tuning. Fine-tuning too few parameters leads to underfitting to the data and even some "memory loss" as compared to the original pre-trained model. On the other hand, fine-tuning too many parameters in conjunction with a small training dataset, leads to overfitting and an inability to generalize to new data.

For all models, the ROUGE and BERTScores varied linearly with one another, confirming accuracy of the content and semantics of the output (Figure 1). Due to the rigidity of ROUGE scoring, some model outputs received a very low ROUGE score and a high BERTScore.

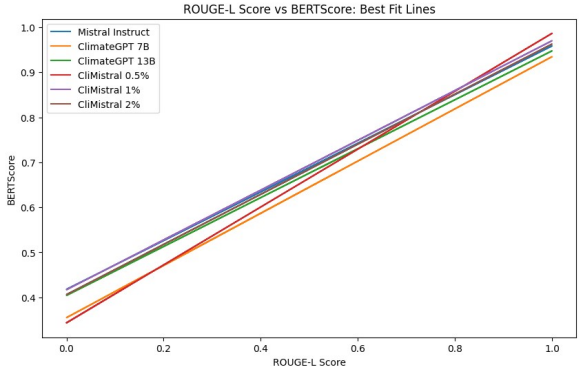


Figure 1: Linear relationship between F1 ROUGE-L and F1 BERTScores on the test dataset.

This was particularly evident for questions with exact numerical answers. ROUGE penalized imprecise answers, while BERTScore rewarded answers with approximate or similar numbers (Table 3). Due to the need for accuracy in climate QA generation, ROUGE is better suited for evaluation of numerical output.

### 4.2 Inference Speed & Energy Consumption

To evaluate resource-use during inference, the inference speed was measured for each of the baseline models and the highest performing CliMistral model. Of the baseline models, Mistral Instruct achieved marginally faster inference than the same sized ClimateGPT model, while the larger ClimateGPT model was roughly 20% slower.

Mistral has two defining architectural features that could explain this superior performance and efficiency as compared to Llama 2. The first is that Mistral leverages Grouped Query Attention (GQA)

[1] https://wandb.ai/site

| Model | ROUGE-1 | ROUGE-2 | ROUGE-L | BERTScore | Energy Consumption (kWh/epoch) |
|---|---|---|---|---|---|
| Mistral Instruct 7B | 0.45 | 0.27 | 0.42 | 0.64 | - |
| ClimateGPT 7B | 0.47 | 0.32 | 0.45 | 0.62 | - |
| ClimateGPT 13B | 0.54 | 0.37 | 0.51 | 0.68 | - |
| CliMistral 0.5% | 0.22 | 0.10 | 0.19 | 0.47 | 0.098 |
| CliMistral 1% | **0.55** | **0.40** | **0.52** | **0.71** | 0.097 |
| CliMistral 2% | 0.50 | 0.34 | 0.47 | 0.67 | 0.100 |

Table 1: Average F1 metrics for the baseline and fine-tuned models.

instead of Multi-Head Attention (MHA). GQA reduces the quantity of key and value attention heads in the architecture, reducing the computational steps required to perform inference (Ainslie et al., 2023). Secondly, the Mistral architecture uses Sliding Window Attention (SWA) which, instead of attending to all token pairs within a sequence, limits the attention mechanism to a fixed size window around each token. Both GQA and SWA reduce the computational complexity of the model (Jiang et al., 2023).

| Model | Inference Speed (s) |
|---|---|
| Mistral Instruct 7B | 3.15 |
| ClimateGPT 7B | 3.21 |
| ClimateGPT 13B | 3.90 |
| CliMistral 1% | **1.99** |

Table 2: Inference speed per query.

Furthermore, the CliMistral inference was approximately 60% faster than the baseline Mistral Instruct (Table 2). A comparison of the output lengths generated by each of the models showed that, on average, the output generated by the Mistral Instruct model was approximately three times longer than that of the ClimateGPT and CliMistral models (Figure 2). As transformers generate text sequentially, longer sequences increase inference time. The pre-training used for Mistral Instruct likely predisposed the model to longer generated outputs; an attribute that was less ideal for this specific task. The baseline Mistral Instruct model was clearly able to learn both domain knowledge and the appropriate output length for this task as a result of fine-tuning.

Finally, there was minimal difference in the estimated energy consumption during training across the three fine-tuned models. Additional fine-tuning trials are needed to determine a statistically significant difference between fine-tune training of these parameter sizes.

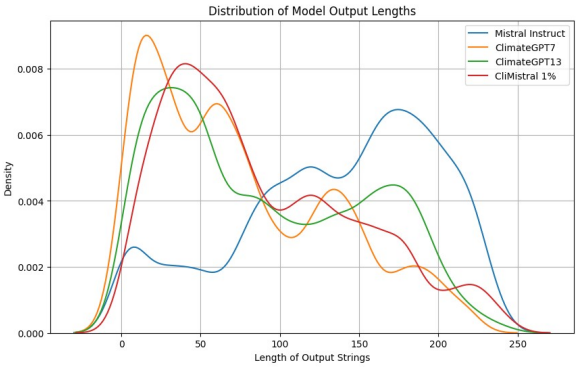


Figure 2: Distribution of length of output strings for each baseline model and the highest performing CliMistral model.

## 5 Conclusion

CliMistral demonstrates the value of fine-tuning pre-trained LLMs to improve model performance on domain specific tasks while reducing the required model complexity. However, due to the limited size and the narrow focus of the Pirá 2.0 dataset, CliMistral likely will not generalize well to a broader corpus data or climate related language tasks. ClimateGPT was fine-tuned on a larger and more varied corpus of climate data. Therefore, while it had lower performance in this case, it is likely a better tool for generalized climate QA tasks.

To encourage further research in this area, researchers should focus on curation of a high-quality, broad, and public corpus of climate related texts and standardization techniques for estimating energy consumption and carbon emissions of model training and inference.

| **Question:** | How much more CO2 is being emitted in 2019 compared to 2010, in gigatons? | | |
|---|---|---|---|
| **Label:** | 2.9 | | |
| **Model** | **Output** | **ROUGE** | **BERTScore** |
| Mistral Instruct 7B | 33.3 gigatons in 2019, which is an increase of approximately 3 gigatons compared to 2010. | 0.00 | 0.36 |
| ClimateGPT 7B | 30.7 | 0.00 | 0.69 |
| ClimateGPT 13B | 10.3 gigatons | 0.00 | 0.57 |
| CliMistral 0.5% | *No Answer* | 0.00 | 0.00 |
| CliMistral 1% | 3.3 gigatons | 0.00 | 0.51 |
| CliMistral 2% | 3.3 gigatons | 0.00 | 0.51 |

Table 3: Example output for precise numerical extraction prompt.

## References


Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. ArXiv:2305.13245 [cs].

Semen Budennyy, Vladimir Lazarev, Nikita Zakharenko, Alexey Korovin, Olga Plosskaya, Denis Dimitrov, Vladimir Arkhipkin, Ivan Oseledets, Ivan Barsola, Ilya Egorov, Aleksandra Kosterina, and Leonid Zhukov. 2022. Eco2AI: carbon emissions tracking of machine learning models as the first step towards sustainable AI. ArXiv:2208.00406 [cs].

Natalia De La Calzada, Théo Alves Da Costa, Annabelle Blangero, and Nicolas Chesneau. 2024. ClimateQ&A: Bridging the gap between climate scientists and the general public. ArXiv:2403.14709 [cs].

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. ArXiv:2305.14314 [cs].

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. ArXiv:2106.09685 [cs].

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. ArXiv:2310.06825 [cs].

Sahal Shaji Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. 2023. Arabic Mini-ClimateGPT : A Climate Change and Sustainability Tailored Arabic LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14126–14136. ArXiv:2312.09366 [cs].

Paulo Pirozelli, Marcos M. José, Igor Silveira, Flávio Nakasato, Sarajane M. Peres, Anarosa A. F. Brandão, Anna H. R. Costa, and Fabio G. Cozman. 2023. Benchmarks for Pir\'a 2.0, a Reading Comprehension Dataset about the Ocean, the Brazilian Coast, and Climate Change. ArXiv:2309.10945 [cs].

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian van Wyk, Abdallah Nasir, Hayden Goldstein, Taylor Tragemann, Katie Nguyen, Ariana Fowler, Andrew Stanco, Jon Gabriel, Jordan Taylor, Dean Moro, Evgenii Tsymbalov, Juliette de Waal, Evgeny Matusov, Mudar Yaghi, Mohammad Shihadah, Hermann Ney, Christian Dugast, Jonathan Dotan, and Daniel Erasmus. 2024. ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change. ArXiv:2401.09646 [cs].

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2022. ClimateBert: A Pretrained Language Model for Climate-Related Text. ArXiv:2110.12010 [cs].