ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change

David Thulke[1,4] **Yingbo Gao**[1] **Petrus Pelser**[3] **Rein Brune**[3]
Rricha Jalota[1] **Floris Fok**[3] **Michael Ramos**[2] **Ian van Wyk**[3]
Abdallah Nasir[1] **Hayden Goldstein**[3] **Taylor Tragemann**[1] **Katie Nguyen**[1]
Ariana Fowler[2] **Andrew Stanco**[2] **Jon Gabriel**[2] **Jordan Taylor**[2]
Dean Moro[2] **Evgenii Tsymbalov**[1] **Juliette de Waal**[3] **Evgeny Matusov**[1]
Mudar Yaghi[1] **Mohammad Shihadah**[1] **Hermann Ney**[1,4]
Christian Dugast[1] **Jonathan Dotan**[2] **Daniel Erasmus**[3]

[1]AppTek [2]EQTY Lab [3]Erasmus.AI [4]RWTH Aachen University
dthulke@apptek.com climategpt@dtn.net
eci.io

Abstract

This paper introduces ClimateGPT, a model family of domain-specific large language models that synthesize interdisciplinary research on climate change. We trained two 7B models from scratch on a science-oriented dataset of 300B tokens. For the first model, the 4.2B domain-specific tokens were included during pre-training and the second was adapted to the climate domain after pre-training. Additionally, ClimateGPT-7B, 13B and 70B are continuously pre-trained from Llama 2 on a domain-specific dataset of 4.2B tokens. Each model is instruction fine-tuned on a high-quality and human-generated domain-specific dataset that has been created in close cooperation with climate scientists. To reduce the number of hallucinations, we optimize the model for retrieval augmentation and propose a hierarchical retrieval strategy. To increase the accessibility of our model to non-English speakers, we propose to make use of cascaded machine translation and show that this approach can perform comparably to natively multilingual models while being easier to scale to a large number of languages. Further, to address the intrinsic interdisciplinary aspect of climate change we consider different research perspectives. Therefore, the model can produce in-depth answers focusing on different perspectives in addition to an overall answer. We propose a suite of automatic climate-specific benchmarks to evaluate LLMs. On these benchmarks, ClimateGPT-7B performs on par with the ten times larger Llama-2-70B Chat model while not degrading results on general domain benchmarks. Our human evaluation confirms the trends we saw in our benchmarks. All models were trained and evaluated using renewable energy and are released publicly[1].

[1]https://huggingface.co/eci-io/

Preprint.

Contents

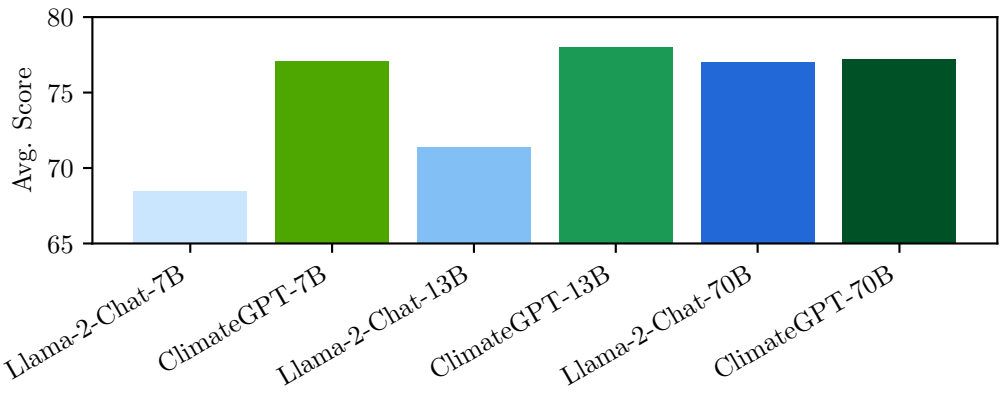

Figure 1: Overview of automatic evaluation results on climate-specific benchmarks.

1 Introduction

Large Language Models (LLMs) have the exceptional ability to comprehend and generate human-like text that empowers them to address a wide array of tasks with Claude-2[2], GPT-4 (OpenAI, 2023), Llama-2 (Touvron et al., 2023b) or Gemini (Gemini Team, Google, 2023) to cite a few. They have been trained on diverse large datasets (from hundreds of billions to trillions of tokens) covering a wide range of topics and domains. The universality of these general-purpose models has made them accessible for a broad spectrum of applications: from text comprehension, over content generation and summarization up to chatbots and much more. Recent research has pointed to the potential of LLMs trained on domain-specific data, e.g. Biomedical sciences (Lee et al., 2020), Finance (Wu et al., 2023) and Medicine (Peng et al., 2023; Luo et al., 2022). These models, while being smaller, have outperformed general-purpose models in their respective domains. The work reported in this paper continues this line of research, addressing one of the most pressing and complex challenges this time: climate change.

Climate change stands out as a multifaceted discipline, covering climate science (the natural science behind modeling climate and the development of the earth's atmosphere) as well as human issues related to climate that impact our environment, our economies, our societies, public health and biodiversity. Right now, we are moving to the brink of multiple risk tipping points (UNU, 2023). Efforts are underway to avoid getting at these potentially irreversible changes in the climate system. Accelerating this process requires global climate awareness and collective knowledge, that we call "climate social intelligence". Building an LLM that addresses climate questions requires access to this collective knowledge, understanding, and decision-making capacity of the human population to harness the collective climate social intelligence available.

We propose an LLM on climate change, called ClimateGPT, which should help the diverse science communities involved to exchange information and knowledge along the three major multi-disciplinary dimensions it covers at large: environmental and natural science, economics, and social science. As opposed to other work done around climate-related LLMs, e.g. ClimateBERT (Webersinke et al., 2022), ClimateGPT-2 (Vaghefi et al., 2022), MBZUAI Arabic Mini-ClimateGPT (Mullappilly et al., 2023), ChatClimate (Vaghefi et al., 2023), the focus of our work was to develop high quality in-domain Instruction Fine-Tuning (IFT) data as well as to train our model with as much climate data as possible, specifically technical reports from the Intergovernmental Panel on Climate Change (IPCC) as well as top papers from climate change research and related fields, such as the UN Sustainable Development Goals. Further, we developed a multi-domain Large Language Model, which can give four types of answers for each request: a natural science-related answer, an answer about the

[2] https://www.anthropic.com/index/claude-2

economic aspects of climate change, as well as an answer about social impacts. The fourth answer, the main one, gives a general high-level overview, addressing all of these sub-fields.

This paper introduces a large language model that seeks to be used across domains by people learning from and collaborating with other specialists in the realm of climate information, rather than merely acting as a chatbot. We are looking at it as a climate intelligence platform that can assist governments, organizations, and individuals in making informed decisions and that contributes to a global social intelligence related to climate.

1.1 Technical Approach

This section outlines our technical approach and the different steps we have taken to develop ClimateGPT.

Language Modeling is done with a large decoder-only Transformer (Vaswani et al., 2017b; Liu et al., 2018) architecture, which is in line with most of the recent literature on large language models (LLMs) (Radford et al., 2019; Brown et al., 2020; Scao et al., 2022; Biderman et al., 2023). The model represents tokens as continuous-valued hidden vectors and makes use of the attention mechanism (Bahdanau et al., 2014) to model inter-token dependencies. The training criterion is cross-entropy, which rewards the model for high probability mass on the correct target token.

From-Scratch (FS) training is done to obtain a foundation language model in the climate domain, the training data for which is cleaned with the climate domain in mind. We train a climate foundation model as well as a general domain model with a focus on scientific content to study the effect of up-sampling domain-specific data during foundation model training. We follow closely the training hyper-parameters that were documented in the Llama-2 paper (Touvron et al., 2023b).

Continued Pre-Training (CPT) is a common alternative to training a new foundation model from scratch (Gupta et al., 2023; Chen et al., 2023). The goal is to adapt an existing LLM trained on a large set of general domain data to the target domain by continuing the pre-training on a smaller set of in-domain data. After an initial evaluation, we focus on the Llama-2 model series as well as for our general not climate-specific from-scratch model. During CPT, we keep the training criterion of the pre-trained model.

Instruction Fine-Tuning (IFT) is an important step to inject instruction-following capabilities into the model. In the literature, this is also often referred to as Supervised Fine-Tuning (SFT). We prefer IFT to make the distinction to domain adaptation via CPT or task-specific supervised fine-tuning approaches. Instruction-completion pairs both from the general domain and climate domain are prepared and gear the model towards following user instructions. We collaborate with climate experts to create a high-quality human-generated dataset. During the data collection, we follow standard approaches (Ouyang et al., 2022a) and also tune the distribution among our and different public instruction-tuning datasets. Although our model is capable of chatting, we focus on its question-answering and instruction-following aspects, which also greatly simplify the instruction fine-tuning data creation and retrieval steps.

Retrieval Augmented Generation (RAG) is implemented with high-quality climate resources to increase factuality as well as to extend the system with new knowledge. We crawl text from manually curated sources and process these sources into smaller chunks. To retrieve relevant documents for a user query, we use a bi-encoder model to calculate embeddings and make use of efficient nearest-neighbor search. During the generation phase, the user instruction is concatenated with the most relevant text chunks for the model to come up with more reliable and stable answers. As the sources of retrieved documents are known, RAG also gives the possibility to provide citations for parts of the generated output.

Cascaded Machine Translation (MT) is included at the system level to enable support for multiple languages. Specifically, non-English queries are first translated to English for our underlying LLM and retrieval engine to generate an English answer. Finally, this answer is translated back to the original language for display.

Model	Base Model	Tokens	LR	GPU Hours
ClimateGPT-70B	Llama-2 70B	4.2B	$1 \cdot 10^{-5}$	2,182
ClimateGPT-13B	Llama-2 13B	4.2B	$1 \cdot 10^{-5}$	301
ClimateGPT-7B	Llama-2 7B	4.2B	$1 \cdot 10^{-5}$	157
ClimateGPT-FSC-7B	-	319.5B	$3 \cdot 10^{-4}$	14,131
ClimateGPT-FSG-7B	-	323.7B	$3 \cdot 10^{-4}$	14,288

Table 1: ClimateGPT model variants.

Benchmarking and Evaluation is done both automatically and with human experts. For automatic evaluation, we evaluate the model both on climate-domain tasks and general-domain tasks. Furthermore, we describe our approach to human evaluation with domain experts, hoping to address the limitations that come with the automatic evaluation of LLMs. We present the results of an initial human evaluation comparing our main model variants.

Responsible AI is an important aspect of our work because as LLMs become stronger, we strongly believe that the models should "do good." To this end, we include instruction fine-tuning data that teaches the model to avoid answering unwanted or even malicious user queries. During the whole development process, we carefully considered and actively worked on reducing the environmental impact of our work. Finally, the models and evaluation protocols are released publicly to increase the reproducibility of our work.

2 Domain-Specific Pre-Training

Foundation models are pre-trained on vast datasets encompassing a wide array of domains (Brown et al., 2020; Touvron et al., 2023a,b). These domains range from general knowledge and common sense reasoning to more specialized areas like science, technology, and literature (Gao et al., 2020; Penedo et al., 2023). Training on a large-scale dataset enables the models to exhibit impressive zero-shot and few-shot (in-context learning) learning capabilities (Brown et al., 2020; Kojima et al., 2022; Wei et al., 2021, 2022), allowing them to perform reasonably well on tasks they are not explicitly trained for. However, despite their versatility, foundation models are not intrinsically designed to possess deep expertise in specific domains. Therefore, recent efforts focused on training domain-specific language models that are either significantly smaller or outperform their general domain counterparts on domains like finance (Wu et al., 2023), science (Taylor et al., 2022) or medicine (Singhal et al., 2023; Chen et al., 2023). To create such a model one can either perform domain adaptation on an existing general domain model (Singhal et al., 2023; Chen et al., 2023) or train a new model from scratch (Wu et al., 2023; Taylor et al., 2022). Which approach is preferable depends on various factors, like the total compute budget, the amount of available domain-specific pre-training data and how well the target domain is represented in the general domain data. To gain insights into these tradeoffs for the climate change domain, we compare both approaches.

In this section, we first describe the general model architecture we used for ClimateGPT. Then, we describe how we curated and collected our climate change and science-specific pre-training dataset. Next, we make use of continued pre-training as a domain adaptation technique to adapt a strong general domain model to the target domain. Finally, we describe how we train a climate-specific model from scratch.

2.1 Model Architecture

We follow Llama-2 (Touvron et al., 2023b) closely in terms of the model architecture. Specifically, the model is a decoder-only Transformer (Vaswani et al., 2017b; Liu et al., 2018) network with word embedding layers sandwiching a stack of self-attention layers. Key components, such as pre-normalization (Xiong et al., 2020) with RMSNorm (Zhang and Sennrich, 2019), SwiGLU (Shazeer, 2020) activation function, and rotary positional embeddings (RoPE) (Su et al., 2023) are retained in this work. Improvements on top of Llama-1 (Touvron et al., 2023a), such as increased context length (4096) and the introduction

Subset	Tokens	Weight	Tokens in model		Percentage of data	
			FSG	FSC	FSG	FSC
news	193.9	1	125.1	120.0	39.1%	37.5%
publications	23.1	4	59.6	57.1	18.6%	17.9%
modern books	28.4	3	55.0	52.7	17.2%	16.5%
patents	19.5	4	50.2	48.1	15.7%	15.1%
wikipedia	6.3	5	20.4	19.6	6.4%	6.1%
policy and finance	3.7	3	7.1	6.8	2.2%	2.1%
science	0.7	5	2.2	2.1	0.7%	0.6%
climate change	4.2	5	0	13.0	0.0%	4.1%
Total	279.7	-	319.5	100%	319.5	100%

Table 2: Subset breakdown of the 300B-token from scratch pre-training dataset.

of grouped-query attention (GQA) (Ainslie et al., 2023) for larger model variants were also kept.

2.2 Pre-Training Dataset

The preparation of high-quality in-domain data is important for the success of the model. Therefore, we started with a corpus of roughly 300B tokens from curated sources compiled by Erasmus.AI. While the corpus spans a wide range of domains sources were evaluated and selected based on their relevance to the topic climate, humanitarian issues and science. The upper part of Table 2 shows the different subsets of the dataset and the corresponding weight for model training. The last columns indicate the effective number of tokens each of the from-scratch models has seen during from-scratch pre-training and the resulting data distribution (see Section 2.4).

The *news* subset is a web crawl with a focus on relevant news and blog articles. It also contains data from an internal extreme weather index. *Publications* is a collection of abstract and full-text papers. The *Modern books* set consists of fiction and non-fiction books and should help to model long-range context. *Patents* are collected mostly from the United States Patent and Trademark Office. *Wikipedia* is a recent dump of the English Wikipedia website. *Policy and finance* is a collection of text related to law, finance and companies and stocks in the climate sector. Finally, *science* covers other science and climate-related texts like EPA documents and ESG reports.

From this dataset, we identified high-quality sources such as scientific papers, and further included primary sources, such as reports from the Intergovernmental Panel on Climate Change (IPCC), and applied cleaning and filtering using keywords and topic classification. In addition, we included our manually curated climate-specific data. More details on these datasets are described in Appendix A.3. In total, we arrive at a corpus of 4.2B climate-specific tokens which is used for continued pre-training.

To improve the quality of the training data, a set of cleaning, filtering, and pre-processing steps was done, which included:

- filtering of sources from unrelated domains, such as sport and entertainment, politics and crime, as well as fiction. By doing so, we hope to limit the number of opinion pieces and information irrelevant to the climate domain;
- personal identifiable information reduction, such as email addresses, telephone numbers, URLs etc.;
- keeping sentences with a Flesch reading score (Kincaid et al., 1975) between 5 and 120;
- handling of errors related to character encoding and special symbols;
- elimination of documents by text length;

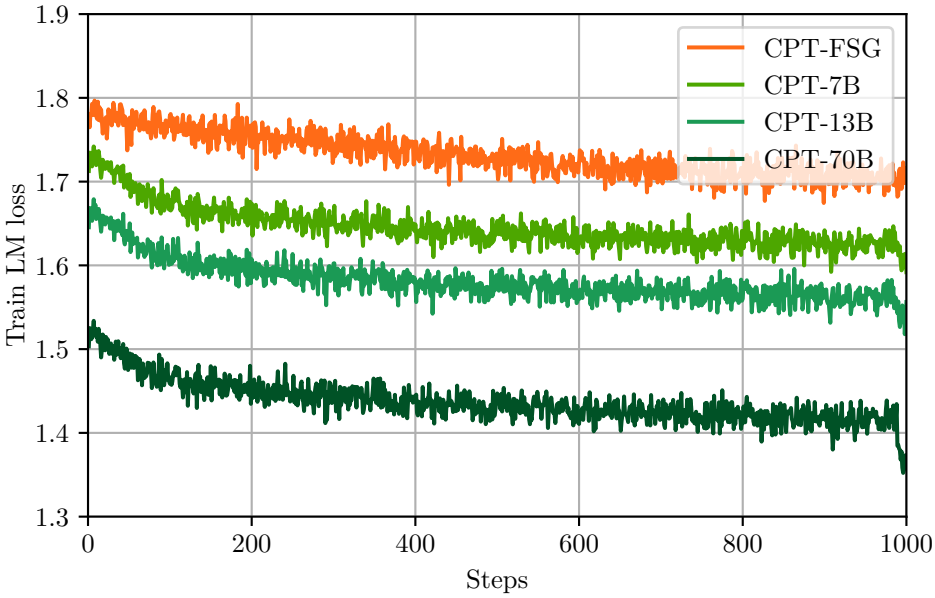

Figure 2: Training loss of the CPT models.

- focusing on sources from the past eight years;
- aggregating themes concerning climate, humanitarian issues and science;
- discovering key sentences and entities that are associated with climate;
- filtering based on symbol distribution, i.e. removing documents that contain at least 80% non-symbols;
- filtering based on language identification, i.e. removing documents that do not score above 85% to be in English;
- removing double spaces, consecutive empty newlines, lines containing long repeating characters such as "======" and "+++++", etc.;
- deduplication based on MinHash (Broder, 1997) with proprietary extensions by Erasmus.AI as well as removal of duplication identifiable in the source metadata.

2.3 Continued Pre-Training

We employ domain adaptation methods, as we aim to develop a model that is specialized for climate change and possesses understanding and domain-specific knowledge. Domain adaptation can help tailor the foundation model to the climate domain, ensuring that it not only retains its broad knowledge base but also develops a more refined understanding of climate-specific concepts, terminologies, and contextual nuances.

Domain adaptation (Ben-David et al., 2010), while not new, remains a cornerstone in the evolution of machine learning systems, e.g. in language modeling (Karouzos et al., 2021), machine translation (Kim et al., 2019), and automatic speech recognition (Baevski et al., 2020). Fundamentally, the method involves the continued training of a baseline model on specific, in-domain data to enhance its performance within that domain. This approach has been widely recognized for its ability to significantly boost a model's proficiency on in-domain test data, while still maintaining robust performance on general tasks. In the context of our work, we adopt this principle to further refine foundation models for climate change applications. We prefer to term this process as "continued pre-training" (CPT), rather than the more commonly used "fine-tuning", to differentiate our approach from other methods like "supervised fine-tuning" (Ouyang et al., 2022a) and to highlight the similarity to the initial pre-training stage. We deliberately apply this CPT step before proceeding to instruction

fine-tuning. If instruction fine-tuning is done before domain adaptation, there is a risk that the model might lose some of its newly acquired instruction-following capabilities. By first adapting the model to the intricacies of the climate domain, we lay a solid foundation upon which instruction fine-tuning can then be built, ensuring a more effective and domain-savvy instruction-following model.

Domain adaptation, while offering significant benefits, presents two primary challenges. The first challenge lies in preparing high-quality data for the in-domain training. The effectiveness of domain adaptation is largely contingent on how closely the distribution of this training data aligns with that of the in-domain test data. The closer the match, the more we can anticipate enhanced performance in domain-specific tasks. To address this we make use of the curated dataset described in Section 2.2. The second challenge is the prevention of degradation in the model's performance on general domain tasks, a phenomenon often referred to as "catastrophic forgetting" (Kirkpatrick et al., 2017) in the literature. This occurs when a model, upon being further trained on specific data, loses its proficiency in tasks it was previously capable of handling. To mitigate this, we carefully tune the batch size, learning rate, learning rate schedule, and data composition for the 7B model variants. Due to time and compute-budget constraints during the project, we did not have time to tune the hyperparameters for the 13B and 70B models and just chose the same values as for the 7B models.

To choose a foundation model, we considered multiple candidates and chose the one that performed best on our climate-specific benchmarks (discussed in Section 6.1). Candidates that we considered were Llama-2 (Touvron et al., 2023b), Falcon (Almazrouei et al., 2023), Pythia (Biderman et al., 2023) and Jais (Sengupta et al., 2023). From these models, we achieved the best results with Llama-2 (see Tables 11 and 12), and thus we continued this model. Redoing these experiments today, we would also consider Mistral-7B (Jiang et al., 2023) and Mixtral (Jiang et al., 2024), but these models were not available at this time.

For training, we use a fork of NVIDIA's Megatron-LM (Narayanan et al., 2021) by the EPFL LLM Team (Cano et al., 2023; Chen et al., 2023). The main modifications to the original version from Nvidia are support for Llama and other recent models. We use a cosine learning rate schedule with a peak learning rate of 10^{-5}, a warm-up of 100 steps and decay to a learning rate of $5 \cdot 10^{-6}$. The batch size is set to 1024 and we use the full sequence length of 4096 tokens. For regularization, we use weight decay of 10^{-2}. All models are trained for 1,000 steps which corresponds to one epoch on the 4.2B climate dataset. The training loss curves for the models are shown in Figure 2.

While we observed that higher learning rates resulted in better training and validation losses, we observed a degradation on our downstream benchmarks. Therefore, we settled with this learning rate as a trade-off between domain adaptation and avoiding overfitting.

2.4 From-Scratch Pre-Training

In contrast to the continued pre-training approach, an alternative strategy involves departing from the use of pre-trained models such as Llama-2 (Touvron et al., 2023a) or Falcon (Penedo et al., 2023). Instead, we initiate the weights of a domain-specific foundation model entirely from scratch and directly train it on domain-specific data.

Adopting the approach of training a model from scratch comes with two significant implications. On the one hand, by choosing not to utilize a pre-trained foundation model, we inherently forego the advantages that come from training on the vast corpus of trillions of tokens that such models have been exposed to. These pre-trained models, despite not having fully disclosed datasets, are likely to have been trained on a diverse range of information, some of which could be beneficial for our purposes. On the other hand, initializing the model from scratch offers us complete control over the training data, which is particularly crucial in a field like climate change that is prone to misinformation and bias (Coan et al., 2021). By carefully selecting and curating the data, we can ensure that the model is trained on accurate, reliable, and scientifically valid information. This level of control allows us to mitigate the risk of perpetuating biases or inaccuracies that might be present in larger, less controlled datasets. While we can expect better performance from training on more data (Kaplan et al., 2020; Hoffmann et al., 2022), projects developing domain-specific models

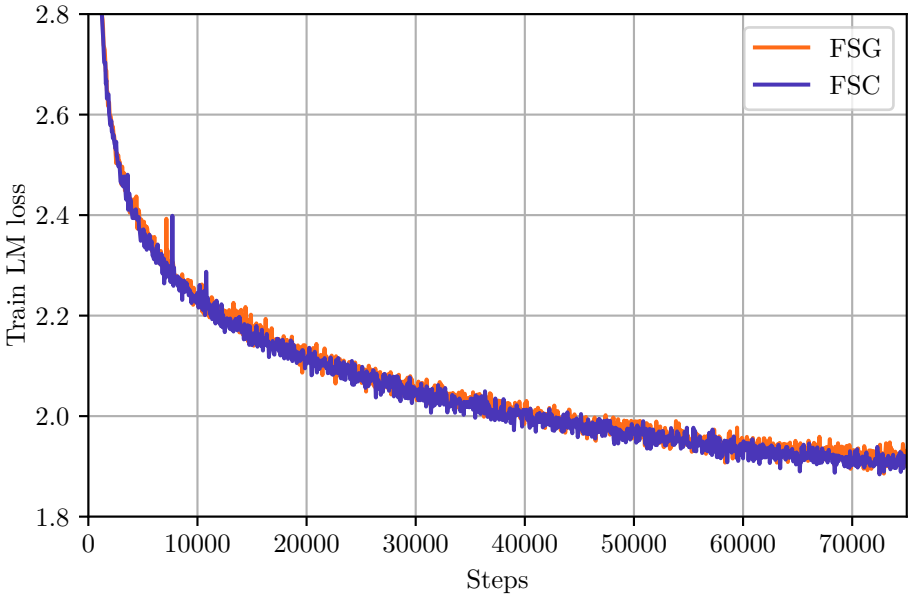

Figure 3: Training loss of the from-scratch general (orange) and from-scratch climate (purple) models.

often have lower compute budgets compared to general-purpose models with a broader range of applications. Thus, training on less but more relevant and high-quality data could still result in better performance.

In our setup, we align our model architecture closely with the Llama-2-7B model developed by Meta, also utilizing the Llama tokenizer (Touvron et al., 2023a), which employs the Byte Pair Encoding (BPE) algorithm (Sennrich et al., 2015). We recognize that developing our own tokenizer, tailored specifically to climate-related terminology, could potentially give better vocabulary compression for domain-specific terms. However, due to time constraints within the project, we left this for future work. Nonetheless, our experience here provides a data point to judge the impact of different training datasets on model performance, while keeping other variables constant. For this reason, we continue with the rest of the development steps, such as instruction fine-tuning, with the from-scratch pre-trained model.

For from-scratch training, we use the same setup as for CPT training. We use a cosine learning rate schedule with a peak learning rate of $3 \cdot 10^{-4}$, a warm-up of 100 steps and decay to 10% of the peak learning rate, i.e. to $3 \cdot 10^{-5}$. The batch size is set to 1040 and we use the full sequence length of 4096 tokens. For regularization, we use weight decay of 10^{-1}. Both models are trained for 75,000 steps the resulting effective tokens seen per subset are shown in Table 2. The training loss curves for the models are shown in Figure 3. To train both models we use the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and $\epsilon = 10^{-5}$. While these values are commonly used to train large language models (Brown et al., 2020; Biderman et al., 2023; Touvron et al., 2023b), we want to highlight that decreasing the β_2 momentum from the common default value of 0.999 decreases training instabilities and loss spikes caused by large batch sizes (Zhai et al., 2023).

2.5 Training Hardware

Given the computationally intensive nature of training a foundational model from scratch (Hoffmann et al., 2022), there are significant environmental considerations, especially pertinent in the context of our work in the climate domain. Therefore, we choose to utilize a high-performance computing cluster that is entirely powered by hydropower (24g CO_2eq / kWH (Schloemer et al., 2014)), provided by MLFoundry. The cluster has 32 nodes, each equipped with 8 H100-SXM GPUs. These nodes are interconnected through InfiniBand, ensuring

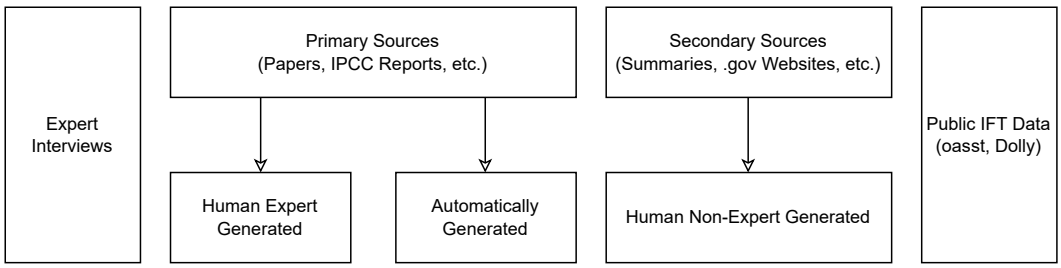

Figure 4: Instruction Fine-Tuning Tracks

high-speed data transfer and communication across nodes. Additionally, within each node, the GPUs are connected via NvLink, facilitating efficient intra-node GPU communications. Leveraging Megatron's (Shoeybi et al., 2019; Cano et al., 2023) efficient implementations of data parallelism, tensor parallelism, and pipeline parallelism, we achieved an average training speed of 250 TFLOPS per GPU, and the training run took 3.7 days on 20 nodes. When fully utilized, we assume a power consumption of 775W per GPU (including CPU).

3 Instruction Fine-Tuning

After pre-training, we expect that the resulting domain-specific language models have a deeper understanding and knowledge of the target domain than comparable foundation models. Since these models were merely trained to predict the next token in our pre-training dataset, using them for specific downstream tasks requires careful prompting or providing the model with few-shot examples. Adapting these models to follow users' instructions formulated in natural language and generate answers in a style appropriate for our use case requires Instruction Fine-Tuning (IFT) (Ouyang et al., 2022a). In the literature, this is also often referred to as Supervised Fine-Tuning (SFT), but we use this terminology as a clearer distinction to other fine-tuning steps (e.g. CPT or task-specific fine-tuning). To do this, the model is trained on instruction and corresponding completion pairs. In the following, we are also referring to these pairs as demonstrations. To limit the complexity of the required data, we decided to only focus on prompt and completion pairs and not to collect any multi-turn chat interactions.

To adapt the style of completions to be appropriate for our envisioned use case, we require a sufficiently large amount of in-domain data for IFT. However, collecting such a dataset is challenging, as it requires a certain level of expertise in the target domain. During the project, we had the privilege to be able to work with a small team of climate experts as well as a larger team of non-experts with limited domain knowledge.

Figure 4 shows the different tracks we followed to collect IFT data. The first track of our IFT data consists of demonstrations (i.e. instruction and completion pairs) collected through *interviews with senior climate experts* (i.e. experienced researchers in the field like professors or other leading experts). During the interviews, the main questions in the field of study of the expert were discussed, implications on other fields as well as different use cases for a climate-specific LLM.

For the remainder of the collected climate-specific IFT data, we provided annotators with existing documents as the basis for the demonstrations they generate. We identified that coming up with new topics can be a limiting factor in the data creation process for the annotators, and this approach can help to concentrate their mental load to write a good completion. Additionally, having control over these "seed" documents means that we can increase the diversity of topics to be covered in the IFT corpus. For the *human expert generated* part of our IFT data, non-senior climate experts (i.e. graduate or PhD students or other early career researchers) created data based on primary sources (like research papers and technical sections of the IPCC reports). As the time of climate experts is valuable and limited we additionally worked with a larger team of non-expert data annotators. Most primary sources, like climate change papers or technical sections of the IPCC reports, were

Domain	Name	Total Size	Training Samples
Climate	Senior Expert Interviews	74	1,332
	Grounded Expert Demonstration	403	7,254
	Grounded Non-Expert Demonstrations	9,663	146,871
	Synthetically Generated Demonstrations	57,609	0
	StackExchange	3,282	9,846
General	AppTek General	700	2,100
	OASST-1	3,783	11,349
	Dolly	15,001	45,003
	Llama-2 Safety	939	2,817
	FLAN	38,909	30,000
	CoT	448,439	15,000
Total			271,572

Table 3: Details about the Instruction Fine-Tuning datasets.

not completely comprehensible by the non-experts. Therefore, for this team, we decided to focus on secondary sources such as governmental websites (e.g. from the EPA, NASA, or European Parliament) and summary sections of the IPCC reports. As we were still concerned that we might not be able to collect a sufficient amount of in-domain data, we simultaneously experimented with *synthetically generated* demonstrations from the documents using a general-purpose LLM as opposed to the manual IFT creation from above.

Finally, to increase the amount and diversity of instruction-tuning data and to be able to reuse well-developed non-climate domain-specific instructions, we made use of existing *general domain IFT data*. Table 3 gives an overview of the different IFT datasets that were used to train ClimateGPT. The rest of this section is devoted to providing more details on these datasets and how they are used to train ClimateGPT.

3.1 Senior Expert Interview Demonstrations

Our vision for the model is for everyone to have a personal climate expert at their fingertips breaking down questions and concepts to the level of expertise of the user. Interviews with climate experts most closely resemble this goal and thus IFT data created in this process is the most valuable data source for ClimateGPT. We started the interviews by defining foundational concepts in the area of expertise of the interviewee and the role of climate change. Second, we discussed current trends in the field and the expected developments in the future. Next, we discussed pivotal findings and research papers in the field and extracted key arguments. Finally, we brainstorm ways in which a climate-specific LLM could be helpful for stakeholders involved in this specific field. As the time of the corresponding experts is very limited, instruction and completion pairs were developed afterward by the interviewer.

For the first version of ClimateGPT, we conducted a series of interviews with the agricultural ecologist Dr. David Lobell. He is the Director of the Center on Food Security and the Environment at Stanford University and also served as lead author for the food chapter on the IPCC Fifth Assessment Report (AR5). The result of this process was a high-quality IFT dataset of 74 demonstrations. Based on these promising results, we plan to refine our methodology and conduct additional interviews.

3.2 Grounded Expert Demonstrations

In addition to the non-expert annotators, we collaborated with nine climate scientists (graduate or PhD level) from different European universities. For the data collection, AppTek's data annotation tool Workbench[3] was used. The team worked in close collaboration with the authors to improve the style of the generated data.

[3] https://www.apptek.com/technology/workbench-data-annotation

Task Category	%
Open Ended QA	26.9
Open Ended Generate	48.0
Open Ended Classification	1.2
Open Ended Chat	4.7
Open Ended Chain of Thought	0.1
Open Ended Brainstorm	7.5
Closed Ended Summarize	4.7
Closed Ended Rewrite	0.2
Closed Ended QA	3.8
Closed Ended Extract	1.6
Closed Ended Classification	1.3

Table 4: Task distribution for the non-expert data collection.

As a first step, we asked the nine climate scientists to think themselves about five to ten questions relevant to climate change that they find important to address and feel comfortable answering. We proposed to them to organize their answers with a scientific mindset (the style we want ClimateGPT to use) by first making a summarizing statement followed by a list of bullet points explaining or developing elements of the summary. Each answer should refer to a scientific source, from which the experts should extract a couple of paragraphs relevant to the answer. The retrieved paragraphs were stored so that they could be used later on. This first exercise was a first test used to evaluate the writing skills of our experts. At the end of this first phase, we continued with seven of the nine experts with the IFT creation task.

In the second step, we provided the experts with references to primary sources and questions related to these sources that have been generated by our synthetic IFT pipeline (Section 3.4), as a source of inspiration. The synthetically generated questions can be very specific reading comprehension questions with respect to the reference source that often has no relevance outside of the source document used. Such questions would either need to be generalized or skipped by the expert. Also, to be time-efficient while producing an answer, we proposed to our experts to choose those questions that relate to their domain of expertise (e.g. city climate, tropical climate, etc.). At the end, we gave each expert a set of 1,000 question-answer pairs, from which 50 to 250 have been selected. In contrast to the non-expert data collection effort (Section 3.3), we did not suggest specific task categories to the experts and instead let them decide on relevant instructions.

As addressed previously, we want the model to be able to generate different in-depth responses addressing the different dimensions of climate change, namely natural science, economics, and social aspects. To collect IFT data for this feature, we asked the expert annotators to create four responses to the same prompt one giving a general answer and three focusing on one of these dimensions.

3.3 Grounded Non-Expert Demonstrations

For the non-expert data collection, we worked with a team of 99 annotators employed by external contractors from six different countries and three continents. Annotators were selected based on their educational background, domain-specific expertise and interests, strong communication skills, and writing skills. More details on the demographics of the annotators are provided in Appendix A.4. All annotators were trained by the corresponding project managers on the project scope, guidelines and requirements. The team used the same tool as the expert annotators (Section 3.2).

To ensure a certain level of diversity of types of instruction, we provided annotators with a task category. The set of tasks and their distribution is based on the use case categories reported in (Ouyang et al., 2022b). The resulting task distribution is shown in Table 4. For each category textual guidelines were provided to the annotators.

For the initial phase of the project, we did not provide annotators with any specific topic to work on in addition to the general climate topic. However, we observed that this resulted in too many simplistic and overlapping prompt and completion pairs. Providing annotators with a specific topic to work on resulted in more diverse and interesting data. Topics were selected based on interviews with climate experts to cover the climate impact across various real-life situations and elements. Table 19 in Appendix A.4 shows the full list and distribution of topics.

The recommended way of data creation was to find content from approved data sources to develop ideas for prompt and completion pairs. Initially, we provided annotators with primary sources, such as research papers and technical sections of the IPCC reports. However, initial feedback showed that our annotators struggled with these documents. Therefore, we decided to switch to secondary sources, such as governmental websites (e.g. from the EPA, NASA, or the European Parliament) and summary sections of the IPCC reports. Besides the trustworthiness of the content, data sources were approved to avoid copyright issues.

Annotators were instructed to give in-text citations to sources they were using in the completion. We instructed annotators to give citations in MLA style (i.e. author name, title, and source in brackets) but noticed that this resulted in inconsistencies that had to be corrected in post-processing. Later we switched to IEEE style (i.e. reference number in square brackets). The data annotation tool allows storing additional details for each citation, such as the URL or the cited text, as additional metadata for the prompt-completion pair. At the beginning of the data annotation process, we decided to instruct annotators to only store the URL of the cited source and not the cited text itself. While the latter would have been useful to improve the retrieval augmented generation capabilities of the model, we decided against it in concern that annotators would restrict themselves to the referenced text (instead of making use of all information in the document) and to avoid increasing the complexity of the annotations process, and, thus, the volume of data we can collect. As an alternative, we can make use of the URLs to crawl the complete document and reconstruct the cited paragraph automatically. Section 4.3 discusses this process in more detail.

3.4 Synthetically Generated Demonstrations

As access to experts who can make use of primary sources is limited (and we initially were concerned that we may not be able to collect enough human-generated IFT data), we were also investigating synthetically generating demonstrations from primary sources. To achieve that we prompted an existing general-purpose LLM with few-shot examples and a document and instructed the model to first generate a question and the corresponding system completion. The prompts were carefully designed to increase the diversity of the generated data. Further, we applied multiple post-processing steps to ensure that the generated data is of high quality. These steps included verifying that there is not too much and not too little overlap to the reference documents and prompting general-purpose LLM again to check whether the generated completion is plausible. Further, we filter out questions or responses that mention figures or specific sections from papers and try to detect other text generation artifacts like repeating sequences. This process was initially designed with a multi-turn model in mind. Therefore, completions were intentionally kept shorter with the intent that the user might ask follow-up questions. The decision not to allow multi-turn interactions in this initial version and that more comprehensive answers are preferable came later in the project.

While initial experiments showed promising results, we did not observe consistent improvements in our automatic benchmarks for later versions of the models when using this data. Thus, due to this and due to the lenght mismatch this data is not directly included in our final IFT data mixture.

3.5 General Domain Data

As the last track of our IFT training dataset, we make use of existing human-written IFT datasets that are available to us. The first is an internal high-quality set of prompt-completion pairs originally collected by AppTek. We are referring to this dataset as *AppTek General*.

The data collection methodology was similar to the one described for the non-expert data collection.

Further, we make use of two openly available crowd-sourced IFT datasets. First, *Databricks Dolly* (Conover et al., 2023) was the first openly available human-generated IFT dataset with a permissive license. The dataset consists of 15,001 prompt and completion pairs across 7 task categories and was generated over two months by over 5,000 employees at Databricks. Second, *OpenAssistant Conversations 1* (OASST-1) (Köpf et al., 2023) is a dataset consisting of 161,443 messages in 35 different languages. The corpus is the result of a worldwide multilingual crowd-sourcing effort involving over 13,500 annotators. In contrast to all previously mentioned IFT datasets, this dataset does not only contain instruction and completion pairs but also multi-turn conversations. For ClimateGPT, we only make use of English conversation and only include the best-rated messages in each conversation tree, resulting in a total of 3,783 conversations.

As an additional source of data, we included 3,282 question-and-answer pairs from domain-relevant *StackExchange* communities (earth science, sustainability and economics). Another common approach to curating IFT datasets is to format existing NLP datasets as instruction and completion pairs using task-specific templates (Wang et al., 2022; Longpre et al., 2023). While training on this type of data alone is not sufficient to achieve good performance (Ouyang et al., 2022b), combining this type of data can be beneficial (Wang et al., 2023a). A possible explanation for this is that this way the model is exposed to a larger variety of tasks and more examples of in-context learning. At the same time, this type of data is closer to our evaluation tasks than human-written pairs, which might explain improvements in automatic evaluation that might not translate to improvements under realistic use. We decided to include 15,000 examples per epoch from *FLAN v2 and CoT* as described by Wang et al. (2023a) into our training data.

Most recently published instruction fine-tuning datasets were created by distillation from large proprietary LLMs like GPT-4. Examples of these include Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023) or WizardLM (Xu et al., 2023). We intentionally decided not to make use of these datasets. First, recent research has shown that training on these approaches can successfully transfer the style of the models but not their factuality (Gudibande et al., 2023). Second, the biases of the teacher model may be transferred to the student. And finally, the licensing terms of commercial LLM providers often limit the use of their API to train models that potentially compete with them. Due to this, the usage of this type of data in models intended for commercial use is problematic (Taori et al., 2023).

3.6 Safety Data

One missing component in our IFT dataset is examples to align the model to be safe and harmless. While both Dolly and OASST-1 contain a few examples of refusing to answer intentionally harmful prompts, we observed that this was not enough to make the model safe. To evaluate this, we analyzed completions of initial versions of the model on a subset of the Do-Not-Answer dataset (Wang et al., 2023b). This dataset consists of around 1,000 prompts that are intentionally designed to invoke harmful or offensive model outputs. As expected, the initial model produced multiple unsafe and potentially harmful outputs, which suggests that additional demonstrations of expected model behavior are required. As writing safe completions to these types of prompts can be especially stressful for annotators and new approaches to safety are not the center of this work, we decided to make use of an already safe model to generate the completions synthetically. Specifically, we generated completions for each prompt in the dataset using Llama-2-Chat-70B (Touvron et al., 2023a) and included this data in our IFT set. We are referring to this dataset as *Llama-2 Safety*. The design considerations around safety are discussed in more detail in Section 8.1.

3.7 Data Preparation

We use a mix of different sources for our IFT data to enable alignment with the different aspects outlined in the previous sections. Table 3 shows the mixing ratios of the different subsets in our final model training. We just train for a single epoch on the general domain data and up-sample the climate-specific data.

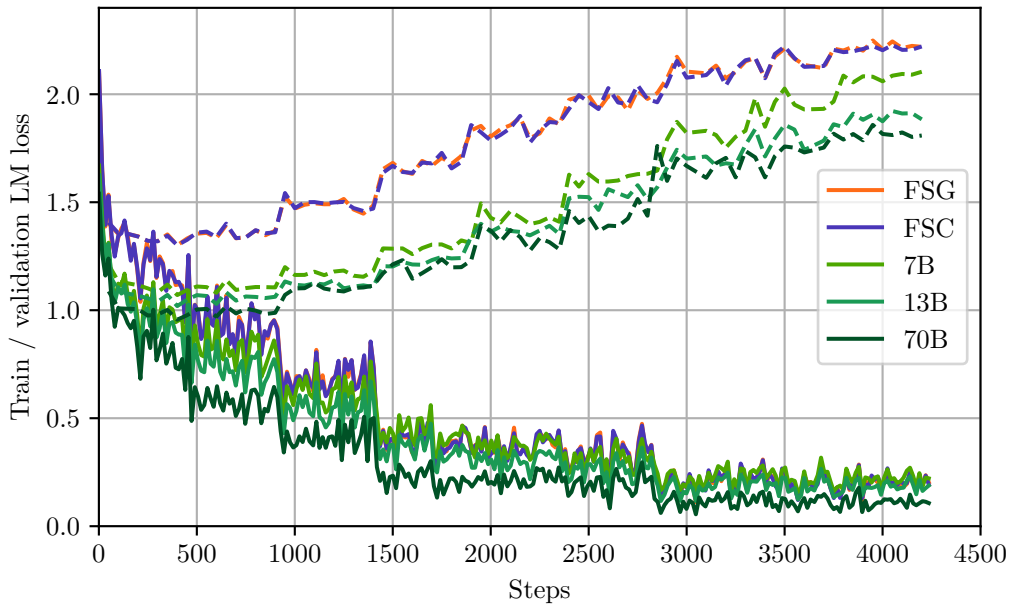

Figure 5: Training (solid) and validation (dashed) loss of the final IFT models.

During inference, we want the model to generate text that is as close as possible to our expert-generated data. However, since the majority of the IFT data comes from other sources, we need some mechanism to counteract this. Our solution to this is to use different system prompts for each data source to condition the model. By using the system prompt corresponding to the expert IFT data we can control its style at inference time. Further, this also allows us to train on data that, e.g. does not make use of all the features of the model (e.g. does not provide citations to references as discussed in Section 4.3). The system prompts for each of the subsets are listed in Appendix A.6.

To prepare the IFT data for training, we make use of the codebase from Open Assistant[4]. During training the IFT data is formatted using the ChatML prompt template[5], following other recent open source models like Open Assistant[6] or Meditron (Chen et al., 2023). Standardizing prompt templates in open-source models results in greater compatibility with existing tools and libraries.

3.8 Training

As for pre-training, we use a fork of NVIDIA's Megatron-LM (Narayanan et al., 2021) by the EPFL LLM Team (Cano et al., 2023; Chen et al., 2023) for IFT training. We use a cosine learning rate schedule with a peak learning rate of 10^{-5} and a warm-up of 100 steps. The batch size is set to 64 and we use the full sequence length of 4096 tokens. For regularization, we use weight decay of 10^{-2} and dropout as used for LIMA (Zhou et al., 2023). The training and validation loss curves for our models are shown in Figure 5. As the validation set, we used a held-out set of 400 prompt and completion pairs from our non-expert climate data.

As was observed by previous work (Ouyang et al., 2022b; Zhou et al., 2023), the perplexity on the validation set first decreases for the first steps and then increases. Typically, this is a clear sign of over-fitting, but, as in other works, we observe that the quality of the model still improves. This is measured by evaluating the model on our automatic evaluation benchmark (Section 6).

[4] https://github.com/laion-ai/open-assistant/tree/main/model/pretokenizer

[5] https://github.com/openai/openai-python/blob/120d225b91a8453e15240a49fb1c6794d8119326/chatml.md

[6] https://huggingface.co/OpenAssistant/llama2-70b-oasst-sft-v10

4 Retrieval Augmented Generation

While pre-training and instruction fine-tuning on climate-specific data improve the climate knowledge of the model, large language models still tend to hallucinate information, especially facts that are not well represented in the training data. An example, where this is especially problematic, is specific numeric figures, for example, CO_2 emissions of a country in a specific year. In addition, the model's knowledge is frozen and it is not possible to incorporate new facts or knowledge without additional training or other approaches to modify the model's weights (Mitchell et al., 2022). In the domain of climate change, these issues are critical. The field is constantly evolving and having access to the latest findings is crucial to draw relevant conclusions.

Retrieval augmented generation (RAG) addresses both of these issues by retrieving relevant documents from external databases and providing these documents as additional context to the model. While the approach was originally proposed for question-answering (Chen et al., 2017; Guu et al., 2020), it has been successfully applied to other tasks like machine translation (Khandelwal et al., 2021), task-oriented dialog (Thulke et al., 2021) and recently in the context of instruction-tuned LLMs (Shi et al., 2023b). The general approach for RAG is to have a separate retriever module which given the user query returns a list of relevant documents. Then, in the language model – often referred to as reader in this context – both user query and retrieved documents are given as context to the model to produce a response. Due to the limited sequence length of transformers and efficiency concerns, it is often not feasible to provide full documents (like full research papers) to the model. Instead, shorter excerpts of a few sentences or paragraphs are typically used as the unit for retrieval and input to the model. Systematic studies show that RAG can reduce the number of hallucinations in language models (Shuster et al., 2021).

Nevertheless, the approach still suffers from limitations. Models usually suffer from noise or irrelevant documents in the retrieval context (Shi et al., 2023a; Cho et al., 2023). We address this issue by including distractor documents during IFT training to allow the model to learn to ignore irrelevant documents.

While RAG is an obvious approach to increase the factuality of generated responses, it also suffers from an inherent trade-off between factuality and abstractiveness (Daheim et al., 2022, 2023). With current RAG approaches, generated responses are often limited to the information provided in the retrieved documents and do not provide a broader perspective. For climate communication, it's especially important to provide an interdisciplinary perspective integrating different viewpoints in the response. To address this in our approach, we propose to make use of distinct sets of documents to generate different answers each covering one of the main perspectives.

We would like to note that RAG is widely used in literature to improve LLMs in the domain of climate change and communication First, ChatClimate (Vaghefi et al., 2023) makes use of GPT-4 as LLM and follows the standard approach outlined above. As a document source, the IPCC AR6 reports are used. The reports are converted to raw text and split into smaller chunks. For retrieval, OpenAI's `text-embedding-ada-002`[7] embedding model is used. In contrast to our approach the retrieval database is just limited to IPCC reports and the base LLM was not adapted to the domain. The system explicitly refuses to generate completions for prompts for which no relevant passages can be found in the IPCC reports. Secondly, for Arabic Mini-ClimateGPT (Mullappilly et al., 2023) the authors fine-tune Vicuna-7B (Chiang et al., 2023) which is based on the first version of Llama-7B (Touvron et al., 2023a). Similar to this work, they dynamically retrieve both English and Arabic climate-specific documents but do not specify the source of the documents in more detail. In contrast to us, they do not perform continued pre-training to adapt the model to their domain. Furthermore, their IFT data was synthetically, generated using ChatGPT, while our climate-specific IFT data was manually curated by humans in close cooperation with multiple climate scientists and experts.

Source	# Docs	# 512 Chunks
IPCC Reports	16	17897
Potsdam Papers	390	8539
Earth4All	14	235
73 other (open access)	336	8648

Table 5: Statistics of the different data sources of the primary retrieval dataset.

4.1 Retrieval Dataset

The dataset for retrieval consists of a manually curated collection of scientific reports and papers. We used the IPCC reports as a starting point and then manually extended the dataset with additional trusted sources in collaboration with climate experts. Therefore, we focused on recent documents to avoid including outdated research. During the data collection, we carefully evaluated the license of each document and only included content with open access or Creative Commons licenses allowing commercial use. To reduce the complexity of the text extraction pipeline, we only considered digitally native PDF documents (i.e. documents where the content can be directly extracted without requiring OCR or similar approaches).

After collecting the PDF documents, we first split the documents into separate PDF pages and for each page, the text is extracted using PyMuPDF[8]. While with this approach we might split relevant paragraphs in the middle of a sentence and loose cross-page context, it greatly simplifies our data processing pipeline. Then, the text on each page is split into chunks of 115 tokens. Next, we observed that many pages in these documents do not contain any relevant information for RAG and potentially degrade performance. These include, for example, tables of content or pages with references. These pages have a high density of superficially relevant content and thus are likely to be retrieved. On the other hand, in most cases, these pages do not provide the full information required to generate a response. To remove these pages, we use a combination of manual data cleaning as well as heuristics to detect problematic content. We deployed a custom tool to iterate, filter, and manually edit data and end with a final set of curated and clean data. The resulting pages are converted to sub-word tokens and then split into chunks of length 115 with stride 10. As the last step, we filter chunks that do not contain enough information (e.g., chunks only containing numbers from tables).

4.2 Retrieval Approach

For retrieval, we follow the common approach of using a transformer bi-encoder model (Mazaré et al., 2018; Reimers and Gurevych, 2019). Here, both document and query are passed separately through a transformer encoder to produce embedding vectors for both sequences of tokens. The similarity between the query and the document is then measured by calculating the dot product or cosine similarity between corresponding vectors. The main advantage of this approach is that embeddings for all documents can be pre-computed and only the query has to be passed through the model at inference time. Other retrieval methods, like cross-encoders (Reimers and Gurevych, 2019), pass query and document through the model simultaneously. While this results in better retrieval performance, the inference cost becomes prohibitively expensive if the document database exceeds more than a few hundred documents; as in the case of our approach, where we want to access a broad range of content. Therefore, the cross-encoder approach is commonly used only to re-rank results from other more efficient methods.

As training our own retrieval model was out-of-scope for this project, we evaluated several existing embedding models. We only considered bi-encoder models and did not integrate an additional model for reranking. As an initial set of models, we considered the best-performing models on the MTEB benchmark (Muennighoff et al., 2023). To be able to do our own

[7] https://openai.com/blog/new-and-improved-embedding-model
[8] https://pymupdf.readthedocs.io/en/latest/

Model	Params	Question		Answer	
		R@1	R@5	R@1	R@5
bge-base-en-v1.5	0.1B	54.8	71.5	81.8	92.1
bge-large-en-v1.5	0.3B	55.8	73.6	83.3	93.1
gtr-t5-large	0.3B	48.8	67.4	79.6	90.1
gtr-t5-xxl	4.8B	47.6	66.3	79.2	89.7
gte-large	0.3B	50.7	68.2	80.9	91.4
ember-v1	0.3B	49.5	68.6	79.7	91.1
instructor-large	0.3B	50.0	68.2	81.7	91.8
instructor-xl	1.2B	53.3	69.7	83.3	92.1

Table 6: Recall (R@1 and R@5) of retrieving the correct document given the question or the answer from the synthetically generated IFT dataset.

in-domain and use-case-specific evaluation, we selected a subset of the synthetically generated IFT data (Section 3.4) as a test set. The generated question and source paragraph are considered positive pairs and all other paragraphs as negatives. We run the benchmark both with the question and with the answer as a query. The results of this evaluation and the models under consideration are shown in Table 6. We got the best results using `bge-large-en-v1.5` (Xiao et al., 2023) and decided to use this embedding model for retrieval.

To increase the relevance of retrieved chunks, we use a hierarchical retrieval approach. Therefore, we first retrieve the top-k PDF pages. We use the full text on the page to calculate the embedding. If the length of the text on a page exceeds the maximum sequence length of the embedding model (512 in our case with variable stride length), we use a sliding window over the sequence and average the embeddings to get a single embedding for the whole page. After selecting the top-k ($k = 60$) pages, as a last step, we retrieve the top-k ($k = 5$) most relevant chunks corresponding to these pages (note that the number of chunks per page varies). This ensures that the page context of retrieved chunks is also relevant to the query.

HuggingFace Transformers (Wolf et al., 2020) and Sentence Transformers[9] (Reimers and Gurevych, 2019) was used to deploy the embedding models and to embed text in preprocessing and inference stages. For efficient nearest-neighbor search we use ScaNN[10] (Guo et al., 2020).

4.3 Grounding

To improve the model's capabilities to make use of reference paragraphs provided in the context, we train the model with IFT examples that already include reference paragraphs. For the IFT data collected from climate experts (Section 3.2) and from interviews (Section 3.1) we asked the annotators to provide the reference paragraphs as additional metadata. For the non-expert data (Section 3.3), we used the URLs of the cited sources to crawl corresponding documents and extract cited paragraphs. Most of the cited sources were websites; therefore, we constructed a separate pipeline to crawl these websites, extract the text using Mozilla's readability[11] and inscriptis[12], and split the content into smaller chunks.

As also shown in Table 6, it is easier to retrieve the correct paragraph using the completion as a query than using the prompt. The reason for this is that the completion contains all the relevant information from the document and thus has high semantic similarity. In contrast, the question just asks for the corresponding information which may not necessarily imply semantic similarity. In our case, we can make use of this fact by not only considering the relevance of the prompt to the potential reference chunks, but also the relevance of the answer. Furthermore, as discussed at the beginning of this section, numbers are a common source of hallucinations in LLMs. Thus, we want to make sure, that the numbers that

[9] https://github.com/UKPLab/sentence-transformers
[10] https://github.com/google-research/google-research/tree/master/scann
[11] https://github.com/mozilla/readability
[12] https://github.com/weblyzard/inscriptis

annotators extracted from the provided reference are part of the selected reference chunk. To achieve this we add the overlap in numbers between the completion and potential reference paragraph as an additional scoring factor. Based on these three scores, we select the best matching chunk from the reference document as source chunk which is given as context during IFT training.

To increase the robustness of the model to irrelevant retrieval results, we added distractor chunks as additional context. These chunks are selected from the total set of chunks produced by the pipeline above. To select chunks not relevant to the answer, we use the opposite of the scoring function outlined above, i.e. we select chunks with high similarity to the prompt but with low similarity to the answer and with low numerical overlap. Additionally, some tasks for the model do not require information from any reference paragraph. Examples include all closed-ended task categories described in Table 4. For these tasks, the prompt already includes all relevant context, so everything that is retrieved is just unnecessary noise. Instead of detecting this at inference time and disabling the retrieval augmentation, we also include distractor paragraphs during the training of IFT examples of these categories to make the model more robust.

4.3.1 Citations

In a scientific context, it is important that provided information is attributable. In RAG, the retrieved chunks are attributable since their source is known. However, the generated completion is not necessarily explicitly related to any of the chunks. It would be helpful to know which chunks exactly were used to generate which part of the response. As discussed in Section 3.3, annotators were asked to provide in-text citations for each reference. We converted these citations to special tokens `[[0]]`, `[[1]]`, etc. which are prepended to the corresponding chunk in the context as well as the token for the citation. While the general approach seems to work, we observed that citation tokens are hallucinated in some cases. We attribute this to the automatic extraction of reference chunks that might not cover all relevant details of the source documents and, thus, introduce noise. For this reason, we removed citations from the IFT data for the final ClimateGPT models and plan to reconsider this in the future.

4.4 Three Dimensions

Climate change is inherently an interdisciplinary field. Therefore, to effectively serve our intended audience, including policymakers, scientists, and journalists, we aim to enhance our model to adeptly address queries from three critical perspectives: natural, economic, and social science aspects. Our goal is to have the model's outputs tailored to the multifaceted nature of climate change, thereby providing comprehensive insights essential for informed discussion and decision-making in the field.

To this end, we devise a three-step approach. First, we utilize the ChatGPT `gpt-3.5-turbo` API to tag our retrieval database with labels corresponding to the natural science, economic, and social aspects. Our preliminary experiments demonstrate that the quality of tags generated through few-shot prompts with `gpt-3.5-turbo` are satisfactory, and we do not further use `gpt-4`. The detailed prompt can be found in Appendix A.8. Second, during inference, we retrieve the most relevant documents for each dimension using the tags above. These sets of documents are then fed separately to the model to generate three distinct completions. The final step involves modifying the system prompt to instruct the model to focus on the specific dimension. These special system prompts were already used during IFT training, for the examples where we had demonstrations focusing on these dimensions.

5 Multilinguality

To achieve our goal of making climate science accessible to a broader range of users, it is important that the model is not only accessible in English but is multilingual. To support multilinguality in LLMs, there are two options. The first is to include multilingual data in the pre-training and IFT data. The other is to rely on Machine Translation (MT) to translate the user input into English, and to translate the generated text back into the

Method	Supported Languages		
native in LLM	English		
cascaded MT	Arabic	Bengali	Chinese (simplified)
	Dutch	Finnish	French
	German	Greek	Hebrew
	Indonesian	Japanese	Korean
	Lithuanian	Pashto	Persian
	Portuguese	Russian	Spanish
	Thai	Turkish	Vietnamese

Table 7: Supported languages. English is native to the underlying LLM, while the support for other languages is achieved via a cascaded translation approach, i.e. `xx`→`en` at the input and `en`→`xx` at the output.

user's language, i.e. a cascaded approach. We chose the latter because there is a lack of multilingual data in the climate domain and we wanted to maintain consistent quality in multiple languages. In Table 7, we list the supported languages.

To enable higher translation quality for climate text, we performed several domain adaptation experiments. Building on a generic base NMT model, we continued fine-tuning the model on parallel sentence pairs extracted from climate data only. As the initial version of the model was presented in December 2023 at the 28th United Nations Climate Change conference in Dubai, we focussed our experiments on Arabic.

5.1 In-Domain Data

To extract parallel climate-related data from our large background collection of public and proprietary datasets, we explore two methods.

For the first method, we used around 2K climate terms and their human translations. We filtered the parallel data based on exact matches of these terms. We then fine-tuned our base model on these parallel sentences. We use Exact Match (EM) to denote this model.

The second method is based on sentence embedding similarity. For this method, we take climate-related monolingual text, mainly a subset from our LLM pre-training data, as seed data. First, we use a weighted average over the word embeddings of a sentence to generate a fixed-size sentence embedding. To obtain a sentence pair embedding, we concatenate the source and target sentence embedding of each bilingual sentence pair. Afterwards, we employ k-Means clustering in the sentence pair embedding space. After obtaining a set of clusters, we use the in-domain seed data to determine which clusters should be used for training. This is done by selecting all clusters that contain a non-negligible portion of the in-domain data using a fixed threshold. For details, refer to AppTek's submission to the shared task of the IWSLT[13] evaluation (Bahar et al., 2020). The resulting parallel corpus is then used for fine-tuning the baseline translation model. We call this approach Embeddings Clustering (EC).

In Table 8, data statistics related to the machine translation adaptation are shown. Initially, we took a subset of our training data, excluding those corpora that most likely will not benefit the conversation about climate. For example, we excluded subtitling data, transcriptions, and some others. The remaining data is the *Filtered Base* data in Table 8. We filtered this data again to extract EM and EC data.

5.2 Training

We used Transformer big architecture and parameters for the MT model (Vaswani et al., 2017a). The fine-tuning process is done for a fixed number of steps for both data setups. The training stopped after 15M parallel sentence pairs with a learning rate of $8.0 \cdot 10^{-5}$. This

[13]International Workshop (Conference) on Spoken Language Translation.

Data	Line count	Word count
Filtered base	88.4M	1.9B
Exact Match	2.5M	63M
Embedding Clustering	22.5M	0.5B
IPCC Testing	5.7K	68K

Table 8: Data statistics related to machine translation adaptation.

	IPCC	FLORES	OpenSubtitles
Base	28.1	40.3	29.1
EM	29.5	39.7	26.9
EC	28.8	40.0	28.1

Table 9: BLEU scores in % of the baseline and adapted Arabic→English models, using in-domain climate data or out-of-domain data as held-out evaluation sets.

means that the EM data were seen multiple times, while the training did not go through all of the EC data. For model adaptation usually the data size is much lower, and it does not make sense to train on a huge amount of data.

5.3 MT Evaluation

To construct a climate-domain test set, we used a translated IPCC report (Pachauri and Meyer, 2014). These reports have professional translations into multiple languages. We converted the PDF reports into text and aligned the sentences using vecalign (Thompson and Koehn, 2020).

Apart from evaluating our adapted MT models on this in-domain test set, we also computed automatic MT evaluation measures on a general-domain FLORES test set (Team et al., 2022) and a test set extracted from OpenSubtitles[14]. Since climate-related questions and answers may have a broad range of styles and topics, we wanted to make sure that the adapted models did not over-fit to the style of the IPCC reports, but were still able to translate more general climate-related content well.

Table 9 presents the BLEU scores of the EM and EC models on the three evaluation sets. Here, we notice an improvement on the in-domain test data as expected, while degraded performance on the out-of-domain test set especially for the EM model can also be seen. This decrease in performance is not as large for the EC model. One probable explanation is that data filtering on exact matches is more strict than similarity-based filtering. From these results, we decided to go with the EC approach for the reverse translation directions, i.e. from English to Arabic.

5.4 Glossary Adaptation

To make sure that climate-specific terms are almost always translated correctly, we use the climate term base that has formed the seed data for parallel data selection for the EM model also at inference time. We convert it to a glossary and use a glossary override method in which the desired glossary-based translation (a word or a phrase) is encoded together with the source term in the source sentence, both in training and at inference time. Our approach is similar to the glossary override method suggested by (Dinu et al., 2019), with a difference that in training, randomly selected bilingual phrase pairs, determined via statistical word alignment, are inserted as artificial glossary entries. We also use special translation factors to mark the source and target glossary terms in the sentence, based on our factored NMT architecture (Wilken and Matusov, 2019). With the successful adaptation of the MT models to the climate domain, the models are then used in a cascaded approach to enable

[14]http://www.opensubtitles.org/

Aspect	Datasets	Test Set Size	Eval. Metric
Climate-Specific	ClimaText	1.6K	Accuracy
	ClimateStance	0.3K	Accuracy
	ClimateEng	0.3K	Accuracy
	ClimateFever	1.5K	Accuracy
	CDP-QA	1.1K	Accuracy
	Pira 2.0 MCQ	0.2K	Accuracy
	Exeter Misinformation	2.9K	F1-Macro
General Domain	PIQA	1.8K	Accuracy
	WinoGrande	1.2K	Accuracy
	MMLU	14.0K	Accuracy
	HellaSwag	10.0K	Normalized Accuracy
	OpenBookQA	0.5K	Normalized Accuracy

Table 10: Benchmarks in Climate and General Domains for downstream task evaluation and their corresponding metrics.

multilinguality of the system, i.e. `xx` → `en` on the user query and `en` → `xx` on the LLM response.

6 Automatic Evaluation

Automatic evaluation of LLMs presents significant challenges that stem not only from the inherent complexities of natural language tasks but also from the difficulty in accurately capturing the multifaceted capabilities of LLMs with limited metrics. Despite these challenges, automatic evaluations, with their simplicity, interpretability, availability, and cost-effectiveness can serve as valuable proxies for assessing model performance.

In our evaluation process, we utilize both established tasks and custom tasks integrated into the LM Evaluation Harness (Gao et al., 2021), a widely adopted resource in large language model evaluations, as seen in platforms like the Open LLM[15]. Our primary evaluation format involves text classification and multiple-choice questions (MCQs), structured as log probability ranking tasks. In the case of MCQs, the question and its corresponding choices are concatenated for model assessment. For a comprehensive overview of these tasks, please refer to Table 10. We publish all prompt templates and instructions on how to reproduce the evaluation[16].

6.1 Climate-Specific Benchmarks

ClimaBench (Spokoyny et al., 2023) consists of a collection of diverse climate-related datasets designed to systematically evaluate model performance across a range of classification tasks. We employ the following 5 datasets from ClimaBench for evaluation: (i) **ClimaText** (Leippold and Varini, 2020): This is a binary classification dataset containing sentences from the web, Wikipedia and the section of US public companies' 10-K reports that address climate-related risks. The task is to predict whether a given sentence is relevant to climate change or not. (ii) **ClimateStance** (Vaid et al., 2022) contains climate change-related tweets that were posted during the 2019 United Nations Framework Convention on Climate Change. The tweets have been manually categorized into three groups for the purpose of stance detection: those expressing support for climate change prevention, those opposing it, and those with an ambiguous stance. (iii) **ClimateEng** (Vaid et al., 2022) is also a climate change related Twitter dataset, collected in the same manner as ClimateStance, for the task of fine-grained classification into topics such as: Disaster, Ocean/Water, Agriculture/forestry, Politics, General. (iv) **ClimateFever** (Leippold and Diggelmann, 2020) is a fact-verification dataset of climate change-related claims. Consisting of 1,535 claims obtained from the

[15] https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard
[16] https://github.com/eci-io/climategpt-evaluation

Internet, each claim is paired with five pertinent evidence passages extracted from Wikipedia. Each claim-evidence pair is labelled into one of three categories: Supports, Refutes, or Not Enough Info. We use this dataset in two ways for fact-verification: with evidence and without. In the first case (dubbed Fever-Evidence), the task is to detect whether certain evidence supports or refutes the claim or neither. The second task (dubbed Fever-Boolean) is to classify if a certain claim is true or false (without providing any evidence). (v) **CDP-QA** (Spokoyny et al., 2023) is a dataset compiled from the questionnaires of the Carbon Disclosure Project, where cities, corporations, and states disclose their environmental information. The dataset presents pairs of questions and answers, and the objective is to predict whether a given answer is valid for the corresponding question.

Pira 2.0 MCQ (Pirozelli et al., 2023) is constructed using a compilation of scientific abstracts and United Nations reports focusing on the ocean, the Brazilian coast, and climate change. The objective involves choosing the correct answer from a set of five candidates in response to a given question, with or without supporting text. The candidate answers are carefully crafted to exhibit substantial lexical similarity with the supporting text and closely resemble the correct answer. This deliberate design adds an extra layer of complexity to the task, demanding a more profound comprehension of the question at hand.

Exeter Misinformation (Coan et al., 2021) dataset contains text from 33 influential climate contrarian blogs and climate change-related content from 20 conservative think tanks spanning the years 1998 to 2020. Annotation of the dataset was done manually using a thorough three-layer taxonomy of contrarian claims related to climate change, developed by the authors. We utilize this dataset specifically for the binary classification task of discerning whether a given text contains a contrarian claim on climate change or not.

6.2 General Domain Benchmarks

Besides climate-specific benchmarks, we also conduct evaluations in the general domain. The goal is to maintain proficiency across general benchmarks while enhancing performance on climate-specific tests. The evaluation of models in the general domain focuses on two key aspects: commonsense reasoning and world knowledge.
The reasoning capabilities of models within the general domain are examined using four datasets: (i) **HellaSwag** (Zellers et al., 2019): comprising multiple-choice questions derived from ActivityNet or wikiHow and challenging models to predict the next event in grounded situations; (ii) **PIQA** (Bisk et al., 2020): containing binary-choice questions that require understanding of real-world object interactions in physical scenarios, (iii) **OpenBookQA** (O.B.QA) (Mihaylov et al., 2018): containing multiple-choice elementary-level science questions that evaluate the understanding of scientific facts and the ability to apply them to novel scenarios; (iv) **WinoGrande** (Sakaguchi et al., 2021): containing sentences in fill-in-the-blank format for the task of resolving ambiguous pronouns, given two options for completion.
The evaluation of models' world knowledge is carried out using the **MMLU** dataset (Hendrycks et al., 2021), which consists of 57 subjects spanning STEM, humanities, social sciences, and beyond, with varying difficulty levels from elementary to advanced professional.

6.3 Results

We perform an automatic evaluation on all of our final ClimateGPT models as well as a set of baselines and other publicly available foundation models. As baselines, we consider the Llama-2 Chat models by Meta that were instruction fine-tuned on general domain data as well as further tuned using reinforcement learning from human feedback.

Table 11 shows the 5-shot results of all models on the set of climate-specific benchmarks. To increase readability, we here just report the weighted average of ClimaBench results, the weights[17] and results for individual tasks are shown in Ap-

[17]Weights are assigned based on the nature of the task and its relevance to the practical application of the model.

Model	ClimaBench	Pira 2.0 MCQ	Exeter Misinf.	Weight. Avg.
Stability-3B	71.4	48.7	52.6	62.8
Pythia-6.9B	63.6	22.9	48.9	50.8
Falcon-7B	62.9	19.8	39.9	48.3
Mistral-7B	73.1	80.0	63.7	73.7
Llama-2-7B	68.5	51.1	59.4	62.6
Jais-13B	66.9	26.4	54.2	54.4
Jais-13B-Chat	65.8	66.3	61.3	65.3
Llama-2-Chat-7B	67.8	72.0	64.3	68.5
Llama-2-Chat-13B	68.6	79.3	68.6	71.4
Llama-2-Chat-70B	72.7	88.8	72.5	77.0
ClimateGPT-7B	75.3	86.6	65.9	77.1
ClimateGPT-13B	75.0	89.0	70.0	78.0
ClimateGPT-70B	72.4	89.9	73.4	77.2
ClimateGPT-FSC-7B	59.3	17.2	45.1	46.2
ClimateGPT-FSG-7B	53.1	17.4	41.5	42.1

Table 11: Results on the climate benchmarks.

pendix A.9. The first half of the table shows a comparison among foundation models: Stability-3B (`stabilityai/stablelm-3b-4e1t`), Pythia-6.9B (`EleutherAI/pythia-6.9b`), Falcon-7B (`tiiuae/falcon-7b`, Mistral-7B (`mistralai/Mistral-7B-v0.1`), Llama-2-7B (`meta-llama/Llama-2-7b-hf`) and Jais-13B (`core42/jais-13b`). We observe that Mistral-7B shows the best performance, followed by Stability-3B and LLama-2-7B. As mentioned in Section 2.3, due to the unavailability of Mistral-7B at the time of development, we continued with LLama-2-7B as our base model.

Next, we compare Llama-2 chat models with Llama-2-based ClimateGPT models and observe that all Llama-2-based ClimateGPT models outperform the corresponding Llama-2 Chat variant. ClimateGPT-7B even outperforms the two times larger Llama-2 13B Chat model and performs on par with the 70B Chat model. The two from-scratch models significantly underperform the Llama-2-based models. While this was expected as Llama-2 was trained on significantly more data (2T compared to 0.3T tokens), we hoped that the potentially higher data quality of our corpus could counteract this.

We note that ClimateGPT-70B performs worse than expected on the climate-specific benchmarks and even worse than ClimateGPT-13B. As discussed in Section 2.3, we did not have enough time at the end of the project to optimize hyper-parameters for the 70B models, so we assume the results can be significantly improved by additional optimization (e.g. lower learning rates).

Further, we observe that the FSC model outperforms the FSG model on climate-specific tasks. While this gives an indicator that including domain-specific data already during pre-training could result in better results than the CPT of a general domain model, the difference is not large enough to justify training domain-specific models from scratch. On the contrary, when considering the resource requirements of from-scratch training, this confirms the CPT approach used for the main ClimateGPT models.

Table 12 shows the results of the general domain benchmarks for our baselines and the ClimateGPT models. We report 10-shot results on HellaSwag and 5-shot on all other benchmarks. We used these benchmarks to verify that our models do not over-fit on the climate domain and still perform on par with comparable models on non-climate tasks. Comparing the main ClimateGPT models to the Llama-2 Chat models, we observe that we not only do not degrade but even outperform the baseline models. Most of our general benchmarks are still science focused so we assume that the additional climate data also benefits these benchmarks.

Model	PIQA	WinoGrande	HellaSwag	O.B.QA	MMLU	Avg.
Stability-3B	79.4	66.3	76.1	40.8	44.6	61.5
Pythia 6.9B	76.6	61.2	65.5	36.2	25.9	53.1
Falcon-7B	80.3	67.3	78.1	43.6	27.1	59.3
Mistral-7B	81.8	73.9	83.4	48.0	63.5	70.1
Llama-2-7B	79.0	69.1	79.0	45.2	47.0	63.9
Jais-13B	76.5	68.4	73.1	38.6	35.0	58.3
Jais-13B-Chat	76.5	68.4	73.1	38.6	68.4	63.1
Llama-2-7B-Chat	79.0	69.1	79.0	45.2	69.1	62.9
Llama-2-13B-Chat	79.9	72.3	82.3	48.2	72.3	66.4
Llama-2-70B-Chat	83.9	78.0	87.0	52.2	78.0	68.6
ClimateGPT-7B	79.8	70.3	78.4	47.6	68.6	65.1
ClimateGPT-13B	80.7	73.4	82.0	51.8	73.1	68.8
ClimateGPT-70B	83.6	79.4	85.8	53.0	66.6	73.7
ClimateGPT-FSC-7B	72.9	53.4	58.9	36.0	23.0	48.8
ClimateGPT-FSG-7B	72.5	54.5	58.7	38.6	25.1	49.9

Table 12: Results on the general benchmarks.

	EXAMS (Acc [%])	
Model	Arabic	MT Ar-En
Llama-2-13B-Chat	25.0	38.0
Llama-2-70B-Chat	28.7	*38.6*
Jais-13B-Chat	**40.9**	36.0

Table 13: Arabic EXAMS evaluation results of the English-only Llama-2 Chat models and compared with the bilingual model Jais, taking the exam directly in *Arabic* and on an *Ar-En* machine translation of the dataset.

6.4 Cascaded Machine Translation

To evaluate the cascaded machine translation approach we evaluate on the Arabic subset of the EXAMS dataset (Hardalov et al., 2020). EXAMS is a multiple choice question answering collected from high school examinations. The Arabic subset covers questions from biology, physics, science, social science and Islamic studies.

For this evaluation we compare Llama-2 Chat which was primarily trained on English text to the bilingual Arabic-English Jais Chat model (`core42/jais-13b-chat`) (Sengupta et al., 2023). All models are first evaluated directly on the original *Arabic* version of the datasets. Second, we translate the questions and answer options using our adapted *Ar-En* MT system (see section 5) from Arabic to English. The results are shown in Table 13. We first observe that when evaluating directly on Arabic, as expected, the Llama-2 Chat models significantly underperform Jais Chat. But, when making use of cascaded machine translation, the Llama-2 Chat models recover most of the performance gap compared to Jais. At the same time with Jais, we observe a degradation in the results when using the cascaded MT approach from 40.9% in Arabic to 36.0% in English. An explanation for that is, as shown in table 12, Jais performs worse than Llama-2 Chat on English benchmarks (i.e. 65.3 for Jais compared to 71.4 for Llama-2-Chat-13B). Another explanation could be errors or potential ambiguities in the automatic translation.

Overall, this shows that cascaded MT is a promising direction to scale LLMs to a large number of languages. While truly native LLMs for each specific language seem to perform better, training LLMs for each specific language or a truly multilingual LLM would require a large amount of resources.

7 Human Evaluation

The value of automatic evaluations is their very strict definition as well as their repeatability which allows precise comparisons of systems and models. Automatic evaluations need a set of clearly defined references and a metric that allows one to measure the difference between the system output and the reference. For tasks like speech recognition the reference is clear (the sequence of words that have been spoken) and counting substitutions, insertions and deletions is fairly easy. For machine translation (respectively summarisation), because, given a source sentence, several equivalent translations having the same meaning are possible, flexible metrics have been developed like BLEU (Papineni et al., 2002) (respectively ROUGE (Lin, 2004) for summarisation). For question answering, (e.g. answering questions on climate change) the variability in the formulation of responses is much higher than that for translation or summarisation of text where the original text serves as base. Domain knowledge, factuality and the bringing of arguments in a certain order (reasoning) are key to the quality of an answer. While the tasks used for our automatic evaluation do cover in some respects climate change and climate science knowledge they do not give us indications of how well an answer is formulated, essentially they do not give us insights on how well the arguments (facts) of the answer are brought together, how good the reasoning supported by the arguments do lead to an easy to understand conclusion. Multiple-choice questions (used for automatic evaluations as they allow to limit the number of valid outputs) are good at verifying the understanding of atomic knowledge. ClimateGPT being targeted at answering complex questions has to generate the dots between the acquired atomic knowledge, bringing in the reasoning needed to make the answer self-explainable. It is probably necessary to master the knowledge behind a domain like climate science to answer complex questions on climate, but it is not enough. A good answer depends on the ability to reason on this knowledge and on the formulation clarity of this reasoning. As of today, quality evaluation of answers to questions is best done with humans having a comprehensive overview of the field and who can judge whether a response is adequate, if it covers all relevant aspects to the question and if the reasoning supported by the information provided is well formulated/expressed.

For our human evaluation, a set of 7 climate change post-docs, PhD students and master students has been asked to provide feedback on the output of 3 different models by ranking them against each other, Also they were asked to tell us if some claims in the generated outputs have been hallucinated or not. For the ranking, human evaluators had positive and negative points to distribute according to the following principles:

- Evaluate the quality of each answer by ranking them within each other and also qualifying the goodness of each answer.
- Positive numbers are good, negative numbers are bad, the zero is neutral.
- Refer to the sheet "Quality Dimensions" for your evaluation.
- If answers differ only in their syntactical form, please consider them equal.
- The answers in each column have been randomly taken from one of the 3 system outputs so that you shall not be tempted to find a pattern per system or that you do not develop a preference for a system.
- Edit columns B C and D from the "Ranking" sheet according to the following principles:
- You will have the following numbers at your disposition:
  - 2 / 1 / 0 / -1 / -2
- The negative numbers are bad grades.
  - 2 is best
  - 0 is average
  - -2 is the worst
- As we want to rank the answers, try to avoid giving the same rank to 2 system outputs.

Model	Average Rank	# Hallucinations
ClimateGPT-70B	1.0	2
ClimateGPT-7B	0.6	4
ClimateGPT-FSC-7B	0.2	5

Table 14: Human evaluation comparing the answers of 50 questions from 3 different systems. An average rank around 0 means the system has been evaluated half of the time as good and half of the time as not good, independently of its rank.

- If all are of the same quality, all get the same grade, between 2 if all are very good and -2 if all very bad.
- If 2 are similarly good and 1 is bad, the 2 good ones get a positive number (e.g. 1) and the 3rd a negative number (e.g. -1)
- If 2 answers are good and one is better, the better one gets a 2 and the less better one a 1

The "Quality dimensions" referred to in the above guidelines are those from (Bulian et al., 2023).

We noticed that human evaluators tended to try to find patterns for each system. To increase the neutrality of their judgment, we decided not to name each answer by the model name and to randomly order the answers of each system so that the annotators were not tempted to try to guess from which system each answer comes.

7.1 Results

For the human evaluations, we asked the evaluators to compare and rank three versions of our ClimateGPT models: the from-scratch model ClimateGPT-FSC-7B, the CPT models ClimateGPT-7B and ClimateGPT-70B. This evaluation shows us that CPT models compare positively against the from-scratch model (1.0 points vs. 0.2), and the 70B CPT model performs better than the 7B CPT model (see table 14 first block). While the ranking between from-scratch and CPT models correlates well with the automatic evaluation, it is not the case when comparing the 7B and 70B within the CPT model family. We need to investigate this further. Another outcome of this evaluation is the observation that the lower the rank of a model, the higher the number of hallucinations.

8 Responsible AI

The pursuit of responsible AI systems is a critical aspect as important as, if not more than, the model performance itself. In this work, we aim to follow closely the standard approaches in the field. Of course, as an active and evolving field of study, the definition and scope of "responsible AI" continue to develop in tandem with the advancement of more sophisticated AI systems.

8.1 Content Moderation

Our perspective on responsible AI encompasses two fundamental aspects: maximizing benefit and minimizing harm. This reflects an inherent trade-off between a model's usefulness and its safety. In the realm of LLMs, for instance, a system that refrains from answering any question minimizes risk but offers limited utility, whereas a system that responds to all queries increases usefulness but may be prone to misuse like generation of misinformation. A pertinent example is content moderation. Simple methods like keyword block lists, as used in the Jais model (Sengupta et al., 2023), can be effective: a safe refusal message is triggered by a regex check against a predefined word list. However, such approaches risk having too many false positives, for example, the keyword 'sex', though potentially problematic, can be a part of legitimate biological discussions. This illustrates how surface-level safety measures might inadvertently constrain a model's utility.

A more elegant solution is to fine-tune the model on data that gracefully answers unintended contents. In our case, we adopt the Do-Not-Answer dataset (Wang et al., 2023b), and manually check many responses from the baseline Llama-2-Chat 70B model (Touvron et al., 2023a). The model responses are often satisfactory, i.e. not only refusals but also include helpful explanations and suggestions. Encouraged by this, we decide to augment the Do-Not-Answer dataset with these model completions and include it in our IFT dataset (see Section 3.5). Additionally, this automated approach to the curation of content moderation examples spares human annotators from the stress of handling toxic data just to replicate Meta's existing efforts, further aligning with our responsible AI principles. While this approach effectively helps to reduce undesired outputs and reduces the potential for misuse of the model, we want to note that these fine-tuning approaches can be easily circumvented if an attacker has access to the model (Yang et al., 2023; Zhan et al., 2023).

8.2 Transparency

Transparency is a cornerstone of responsible AI, fostering reproducibility, facilitating communication, and revealing potential issues. The recent introduction of the Foundation Model Transparency Index (FMTI) (Bommasani et al., 2023) offers a framework to assess the transparency of foundation models.

Although we agree that the specific questions and their weightings in FMTI may be subject to debate (Lambert et al., 2023), it represents a significant step towards standardizing disclosure practices in LLM research. In our work, we nonetheless reference FMTI to self-assess and achieve an FMTI score of 69 and also self-assessed using the revised methodology and achieved a score of 62.[18]

These self-evaluations underscore our commitment to sustainability and the ongoing discourse towards transparent model development.

8.3 Environmental Impact

The environmental footprint is a critical consideration in responsible AI, especially for projects in the climate domain. Recognizing the substantial economic and computational resources required for training LLMs, we prioritized the use of sustainable energy sources. In collaboration with MLFoundry, we accessed a high-performance computing cluster powered exclusively by hydroenergy. Although securing high-end GPU computing resources, especially those powered by green energy, is challenging, our decision to partner with a provider committed to clean energy sources reflects our dedication to minimizing the environmental impact of our work.

Using a framework developed by Hershcovich et al. (2022), we have published a sustainability scorecard that details the energy usage and emissions associated with training experiments, as well as the final models. You can find the scorecard in the appendix in Table 16. This data is currently being evaluated by Filecoin Green to establish a Green Score[19] and will be published once finalized.

9 Conclusion

This paper has introduced ClimateGPT, a domain-specific large language model (LLM) that gives access to interdisciplinary research information on climate change. ClimateGPT is therefore the first family of LLMs to generate not only one answer but four different answers, three answers each along a different perspective (science/economic/social which we also call dimension) plus a fourth answer summarising the answer of the three perspectives provided to the user. We have compared five different ClimateGPT versions. The first two models are from scratch (FS) foundational models trained on our own 300B tokens corpus, both with a 7B parameters transformer architecture similar to that of Llama-2. The training corpus of the first FS model contains 4B climate change data (Climate-FSC-7B), the other do not

[18] Detailed results: https://github.com/eci-io/climategpt-fmti

[19] https://www.greenscores.xyz/

have these 4B tokens (Climate-FSG-7B) which allows us to compare the value of climate change data within a foundational model. The next three models are based on a Llama-2 foundational model (7B, 13B and 70B, all pre-trained on 2T tokens). These models as well as the ClimateGPT-FSG-7B are fine-tuned on 4B climate tokens (Continued Pre-Training, CPT). We did not fine-tune the ClimateGPT-FSC-7B model as it had these 4B tokens in the from-scratch training. All models have been further trained on a manually defined Instruction Fine Tuning (IFT) climate-specific prompt/completion pair corpus that has been produced by experts (climate consultants and climate scientists) and by non-experts. We have been benchmarking our models on two different sets of tasks, one set specific to climate (ClimaBench, Pira and Exeter), and the other one on standard non-climate tasks. We have shown that, while adapting our models to climate change we do not lose performance on general tasks (MMLU, HellaSwag, PIQA and WinoGrande) and that our Llama-2-based ClimateGPT-7B outperforms Llama-2-Chat-13B (77.1% resp. 71.4%) on climate tasks with two times fewer parameters and is on par with the Llama-2-Chat-70B results (77.1% resp. 77.0%) with 10 times less parameters.

The quality of the IFT data plays an important role. An interesting question that we did not have time to address is whether a general science IFT dataset would also have contributed and by how much.

In this paper, we show also the value of cascaded machine translation as opposed to using a general one-system fits-all-languages approach. The comparison made on the Arabic subset of EXAMS between the mono-lingual system Llama-2-13B-Chat with that of Jais-13B-Chat (which has been also trained on Arabic data) shows that Machine Translation (translating the EXAMS from Arabic to English, so translating the query and the answer into English) allows Llama2-13b-Chat to improve from 25.0 to 38.0 very near the performance of Jais-13b-Chat (40.9) which was evaluated directly in Arabic. We did not fine-tune our MT for this specific task.

Another important aspect of our work is related to the sustainability of domain-specific models: fine-tuning our ClimateGPT-7B has been done with a tiny fraction (so small that it is apparent to a rounding error) of the CO2 production needed to produce the complete Llama-2-Chat-70B model. Further, at inference time, our system answers questions producing 12 times less CO2 (needing 12 times less energy) than Llama2-70B would do, for the same result.

Finally, our human evaluations show some correlation with the set of automated tasks used for benchmarking.

10 Limitations

Like any LLM, ClimateGPT is subject to hallucinations. Retrieving relevant documents for grounding before calling ClimateGPT can help control hallucinations. While subjectively our proposed RAG approach seems to reduce hallucinations, we have not yet performed a systematic evaluation.

This interdisciplinary project involving multiple partners is challenging in itself and inevitably has many limitations. Here, we acknowledge the drawbacks by following standard practices in the NLP community and hope to inspire future work.

Firstly, while Reinforcement Learning from Human Feedback (RLHF) is an effective method for enhancing model performance, we do not employ it due to time and resource constraints. We notice the model gives decent performance without RLHF, and thus we focus our efforts elsewhere.

Secondly, LLMs are shown to exhibit strong native multilinguality and can perform the machine translation task very well. However, we apply a cascaded approach to make use of domain fine-tuned existing MT systems at AppTek because fully training a multilingual LLM would inevitably be much more costly.

We also acknowledge the limitation of not using a domain-specific tokenizer which could have improved the model's representation of climate-related vocabulary as discussed in Section 2.4.

For retrieval, we relied on a standard bi-encoder model and did not investigate domain adaptation or more advanced retrieval techniques like reranking or verification of the relevance of retrieved documents.

Finally, another limitation is in evaluation. Automatic evaluation is limited in reliability and what they can evaluate and our human evaluation is limited in scale, completeness, and breadth. Therefore, a lot of design decisions still need to be validated with systematic evaluations.

Acknowledgments

The model development and evaluation was completed as independent research in advance of the COP28 Conference from August to December 2023 through a grant provided by ADQ, TAQA, Masdar, Etihad Rail, ADNEC Group, and Hedera.

We would like to thank The Club of Rome for their partnership and unwavering support, specifically Sandrine Dixson-Declève, Paul Shrivastava, Mmampele Rampele, Peter Blom, Carlos Alvarez Pereira, Wouter van Dieren, Per Espen Stoknes, Jorgen Randers, Gunter Pauli; Nature Finance, specifically Simon Zadek; Goals House specifically Matthew Freud, Arlo Brady, Anna Biles; Info.nl specifically Jann de Waal, Dominik Vrbic, Anandita Punj, Jorrit Tinholt, Paul Domen; The Internet Archive, Brewster Kahle, Mark Graham, Wendy Hanamura, Jamie Joyce, as well as the support from Babiche Veenendaal-Westerbrink.

Special thanks to David Lobell for participating in our expert interviews which gave invaluable insights in his work and are the foundation of our IFT data. Next, we want to thank Acheampong Baafi-Adomako, Hamidreza Mosaffa, Pan Hao, Qinghua Yu, Ralf Liebermann, Thomas Kreuzwig and Yurong Gao for creating the expert IFT dataset as well as for having participated in our human evaluations. We further want to thank Eugen Beck, Nico Daheim, Nils Hilgers and Ege Beysel for helpful discussions on this work.

We thank ML Foundry for the opportunity to train ClimateGPT using renewable energy, giving us early access to their H100 machines and for all the support across different time zones during the main training phase.

In addition, we'd like to acknowledge Faisal Al Hammadi and the entire team at Further Ventures, who helped coordinate the sponsorship funding and supported our vision to bring the model to COP28.

We would also like to thank the engineering and cryptography team at EQTY Lab that worked on the integrity of the AI lifecycle: Benedict Lau, Yurko Jaremko, Paul Dowman, Cameron Fyfe, Tyler Brink, Mauve Signweaver, Tucker McCoy and Ziv Weissman. And a special thanks to Dan Boneh.

Further, the team at Gladeye for designing and creating our website: Tarver Graham, Conrad Blight, Nathan Walker, Antony Zouch, Kate Forsythe, Alastair Gray, Michael Cannon, Giuliana Aliotti, Daniel Bonham, Joris Rotteveel and William Hamlin.

And all the people who worked on the responsible AI pipeline and pilots: Judie Muhrez, Alex Feerst, Chris DiBona, Monica Granados and the entire Creative Commons team, Marc Johson (Filecoin Green), Dr. Regina Stanback Stroud (RSSC), Anton Blewett, Travis Coan (Exeter), John Cook (University of Melbourne), Nathan Schneider, Joshua Tan, Connor Spelliscy, Scott Moore and Khalifa University, Students: Benhur Tekeste, Divora Yemane, Maryam Alblooshi, Maryam Alraeesi and Noof Alhammadi.

Finally, this work would not have been possible without various contributions from the open-source community. We want to highlight the Open Assistant project for crowd-sourcing a high-quality multi-turn IFT dataset, Meta for sharing Llama-2, the EPFL LLM Team for their work on Megatron-LLM and finally Databricks for sharing Dolly.

References

2023. *Interconnected Disaster Risks 2023: Risk Tipping Points*. United Nations University - Institute for Environment and Human Security (UNU-EHS), Bonn. 96 pp.

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 preprint arXiv:2305.13245*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Parnia Bahar, Patrick Wilken, Tamer Alkhouli, Andreas Guta, Pavel Golik, Evgeny Matusov, and Christian Herold. 2020. Start-Before-End and End-to-End: Neural Speech Translation by AppTek and RWTH Aachen University. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 44–54, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning*, 79:151–175.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The Foundation Model Transparency Index. *arXiv preprint arXiv:2310.12941*.

A.Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels Mede, Markus Leippold, and Nadine Strauss. 2023. Assessing Large Language Models on Climate Information.

Alejandro Hernández Cano, Matteo Pagliardini, Andreas Köpf, Kyle Matoba, Amirkeivan Mohtashami, Olivia Simin Fan, Axel Marmet, Deniz Bayazit, Igor Krawczuk, Zeming Chen, Francesco Salvi, Antoine Bosselut, and Martin Jaggi. 2023. epfLLM Megatron-LM.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. Blog post.

Sukmin Cho, Jeongyeon Seo, Soyeong Jeong, and Jong C. Park. 2023. Improving Zero-shot Reader by Reducing Distractions from Irrelevant Documents in Open-Domain Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore. Association for Computational Linguistics.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M. Ponti. 2023. Elastic Weight Removal for Faithful and Abstractive Dialogue Generation.

Nico Daheim, David Thulke, Christian Dugast, and Hermann Ney. 2022. Controllable Factuality in Document-Grounded Dialog Systems Using a Noisy Channel Model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1365–1381, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training Neural Machine Translation to Apply Terminology Constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

O. Edenhofer, R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel, and J.C. Minx. 2014. *Climate Change 2014: Mitigation of Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Gemini Team, Google. 2023. Gemini: A Family of Highly Capable Multimodal Models.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The False Promise of Imitating Proprietary LLMs.

Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual Pre-Training of Large Language Models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A Multi-subject High School Examinations Dataset for Cross-lingual and Multilingual Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Daniel Hershcovich, Nicolas Webersinke, Mathias Kraus, Julia Bingler, and Markus Leippold. 2022. Towards Climate Awareness in NLP Research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2480–2494, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. UDALM: Unsupervised domain adaptation through language modeling. *arXiv preprint arXiv:2104.07078*.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest Neighbor Machine Translation. In *International Conference on Learning Representations*.

Yunsu Kim, Yingbo Gao, and Hermann Ney. 2019. Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1246–1257, Florence, Italy. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Minh Nguyen, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Alexandrovich Glushkov, Arnav Varma Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Julian Mattick. 2023. OpenAssistant Conversations - Democratizing Large Language Model Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nathan Lambert, SE Gyges, Stella Biderman, and Aviya Skowron. 2023. How the Foundation Model Transparency Index Distorts Transparency. `blog.eleuther.ai/`.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text minings. *Bioinformatics, 36(4), 1234-1240.*

Markus Leippold and Thomas Diggelmann. 2020. Climate-FEVER: A Dataset for Verification of Real-World Climate Claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Markus Leippold and Francesco Saverio Varini. 2020. ClimaText: A Dataset for Climate Change Topic Detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. In *International Conference on Learning Representations*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. *arXiv preprint arXiv:2301.13688*.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics 23, no. 6*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022. Memory-Based Model Editing at Scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Sahal Shaji Mullappilly, Abdelrahman Shaker, Omkar Thawkar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. 2023. Arabic Mini-ClimateGPT : A Climate Change and Sustainability Tailored Arabic LLM. In *EMNLP 2023*.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '21, New York, NY, USA. Association for Computing Machinery.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022a. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022b. Training language models to follow instructions with human feedback. In *NeurIPS*.

R.K. Pachauri and L.A. Meyer, editors. 2014. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. IPCC, Geneva, Switzerland. 151 pp.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, and Cheryl Martin. 2023. A Study of Generative Large Language Model for Medical Research and Healthcare. *arXiv preprint arXiv:2305.13523*.

Paulo Pirozelli, Marcos M José, Igor Silveira, Flávio Nakasato, Sarajane M Peres, Anarosa AF Brandão, Anna HR 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 preprint arXiv:2309.10945*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Commun. ACM*, 64(9):99–106.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Steffen Schloemer, Thomas Bruckner, Lew Fulton, Edgar Hertwich, Alan McKinnon, Daniel Perczyk, Joyashree Roy, Roberto Schaeffer, Ralph Sims, Pete Smith, and Ryan Wiser. 2014. *Annex III: Technology-specific cost and performance parameters*, pages 1329–1356. Cambridge University Press, United Kingdom. This annex should be cited as: Schlömer S., T. Bruckner, L. Fulton, E. Hertwich, A. McKinnon, D. Perczyk, J. Roy, R. Schaeffer, R. Sims, P. Smith, and R. Wiser, 2014: Annex III: Technology-specific cost and performance parameters. In: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Edenhofer, O., R. Pichs-Madruga, Y. Sokona, E. Farahani, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel and J.C. Minx (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023b. REPLUG: Retrieval-Augmented Black-Box Language Models.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models.

Daniel Spokoyny, Tanmay Laud, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. Towards Answering Climate Questionnaires from Unstructured Climate Reports.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2023. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, page 127063.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. GitHub repository.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Brian Thompson and Philipp Koehn. 2020. Exploiting Sentence Order in Document Alignment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5997–6007, Online. Association for Computational Linguistics.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient Retrieval Augmented Generation from Unstructured Knowledge for Task-Oriented Dialog. In *AAAI 2021, Workshop on DSTC9*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Saeid Vaghefi, Veruska Muccione, Christian Huggel, Hamed Khashehchi, and Markus Leippold. 2022. Deep Climate Change: A Dataset and Adaptive domain pre-trained Language Models for Climate Change Related Tasks. In *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*.

Saeid Vaghefi, Qian Wang, Veruska Muccione, Jingwei Ni, Mathias Kraus, Julia Bingler, Tobias Schimanski, Chiara Colesanti Senni, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding Conversational AI in Climate Science. *Swiss Finance Institute Research Paper No. 23-88*.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards Fine-grained Classification of Climate Change related Social Media Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. *Advances in neural information processing systems*, 30.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023a. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023b. Do-Not-Answer: A Dataset for Evaluating Safeguards in LLMs.

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

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Patrick Wilken and Evgeny Matusov. 2019. Novel applications of factored neural machine translation. *arXiv preprint arXiv:1910.03912*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *ArXiv preprint: https://arxiv. org/pdf/2303.17564*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. WizardLM: Empowering Large Language Models to Follow Complex Instructions.

Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. 2023. Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training.

Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. 2023. Removing RLHF Protections in GPT-4 via Fine-Tuning.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment.

A Appendix

A.1 Model Card

Table 15 presents a model card (Mitchell et al., 2019) that summarizes details of the models.

Model Details	
Model Developers	AppTek, EQTYLab, Erasmus AI
Variations	ClimateGPT comes in a range of parameter sizes: 7B, 13B, and 70B. Additionally, there are two 7B model variants trained from scratch.
Input	Models input text only.
Output	Models generate text only.
Model Architecture	ClimateGPT is an auto-regressive language model that uses an optimized transformer architecture. After pre-training, instruction fine-tuning (IFT) is used to align the models to the expected output format.
Model Dates	ClimateGPT was trained between September 2023 and November 2023.
Status	This is a static model trained on an offline dataset and intended to dynamically include new knowledge via RAG.
License	ClimateGPT Community License
Where to send comments	Feedback can be given by creating a discussion thread on the model's Huggingface page (`https://huggingface.co/eci-io/`).
Intended Use	
Intended Use Cases	ClimateGPT is intended to be directly used as a question-answering model that is specialized in the climate domain. It is built to provide useful feedback for decision-makers, scientists and journalists involved in climate discussions.
Out-of-Scope Uses	Use in any manner that violates applicable laws or regulations.
Hardware and Software (Section 2)	
Training Factors	We used a fork of the Megatron-LLM Repository by the EPFL LLM Team (`http://github.com/epfLLM/Megatron-LLM`). A cluster provided by MLFoundry was used for pre-training, instruction fine-tuning and evaluation.
Carbon Footprint	Pre-training utilized a cumulative 31,059 GPU hours of computation on hardware of type H100 SXM (including CPU TDP of 775W). The cluster for training and evaluation was powered using 100% hydropower (24g CO_2eq/KWh (Schloemer et al., 2014)) which resulted in the emission of 577.7kg CO_2eq.
Training Data (Sections 2.2 and 3)	
Overview	ClimateGPT was continuously pre-trained on a dataset of 4.2B climate-specific tokens. The from-scratch models were trained on 300B tokens. All models were instruction fine-tuned on a dataset consisting of public IFT data as well as IFT data collected in cooperation with climate experts during the project.
Data Freshness	The pretraining data contains documents up to October 2023.
Evaluation Results	
See automatic evaluation in Section 6 and human evaluation in Section 7	
Ethical Considerations and Limitations (Section 10)	
Despite the efforts from the development team to eliminate them, as with every other chat-capable LLM, this model may generate biased, offensive or inaccurate responses. Testing done to date has been mostly in English and no extensive red-teaming was conducted. Therefore, for all downstream applications users should be made aware of these limitations and should be incentivised to double check model outputs.	

Table 15: Model card for ClimateGPT.

A.2 Sustainability Scorecard

Model Publicly Available	Yes
Time to train final models	31,059 GPU Hours
Location for computations for final models	United States (WA)
Energy mix at location for final models	24 gCO2eq/kWh
Power of GPU and CPU for final models	0.775 kW
CO2eq for final models	577.70 kgCO2eq
Time for all experiments	1,535 GPU Hours (Canada, ON) 2,150 GPU Hours (United States, CA)
Power of GPU and CPU for experiments	0.55 kW
Location for computations for experiments	Canada, ON & United States, CA
Energy mix at location for experiments	134 gCO2eq/kWh & 186 gCO2eq/kWh
CO2eq for all experiments	113.13 kgCO2eq & 219.95 kgCO2eq
Average CO2eq for inference per sample	24.5 mgCO2eq

Table 16: Sustainability scorecard for ClimateGPT.

A.3 Curated Climate-Specific Pre-Training Data Details

The following section gives additional details on the high-quality and manually curated climate-specific datasets that are part of our pre-training data.

Extreme Weather Events A corpora built out of the the most recent decade (2023-2013) of extreme weather news reports, in excess of on average 1M articles per year (slightly less in the earlier years and more in the latter). The intent of the corpora is to build a collection of human-centered effects of climate change, and how extreme weather events impact human activity systems.

The articles were categorized using a custom-built classifier from Erasmus.AI which from its daily planetary scale web crawl organized the articles into 19 categories (Drought, Sandstorm, Extreme Heat Wave, Forest Fire, Wildfire, etc.) and one not-relevant class. Low certainty scoring articles were eliminated from training corpora.

Candidate articles were translated into English from 19 languages. Extreme weather events in certain areas such as West Africa, Latin America, and parts of China do not have a great deal of reporting in English. For example in 2023 in Columbia, 78% of the articles were reported in Spanish. In Peru, 97% of the articles were non-English and were collected from 128 unique websites and associated with 442 Peruvian cities.

The candidate articles were geographically assigned using named entity recognition and a proprietary framework from Erasmus.AI to ensure higher accuracy in assigning events to specific locations as well as to check the events against each other as it would be highly unlikely to have single reports of extreme weather events.

The events were matched to human timelines, and reporting about expected future weather events was eliminated, as much of climate reporting relates to future events that may or may not happen.

The overall goal of the corpora was to train the model family on how, at a human activity systems level, the changing climate connects with geographic knowledge as well as deeper knowledge on the causal effects of climate events (e.g. Snowstorm leads to electricity outages, stay at home orders, supply disruptions; Drought in the Horn of Africa leads to increased civil conflict, etc.).

Technical Game-Changing Breakthroughs For Europe's largest technology company Erasmus.AI in partnership with the Digital Thinking Network researched and identified 153 game-changing breakthroughs in Energy, Climate Change, Food Security, Health, etc.. The process encompassed 500+ pages of technical documentation, and 153 themes set up in a proprietary interface NewsConsole run by Erasmus.AI. The interface enables graph-based curation of large bodies of articles through multiple views (narrative analysis, temporal, etc. views). A theme might present breakthroughs in super-capacitors, desalination technologies, multiple approaches to batteries, novel bacteria that convert sunlight directly into animal feed, saltwater-based agriculture, etc.. Each of these themes presents a forward-looking approach to addressing climate challenges with technology. This includes, for example, not just experience curves in Solar PV and Battery technology, but the viewpoint that these experience curves will continue to make Solar PV (and wind) the cheapest forms of energy in most locations or breakthroughs in animal feed. A selection of the top-ranked few thousand articles per theme was used, where ranking was a combination of human feedback, automated systems, and curation on rich visual interfaces.

Sustainable Development Goals For the Club of Rome as pre-work into the Earth4All process, Erasmus.AI prepared a breakdown of the 17 Sustainable Development Goals (SDGs) of the United Nations into sub-goals and set up a framework of themes on the platform described above that tracks these subgoals of the SDGs. The intent here is to provide a more holistic human-scale view of climate change, and its effects, where action on carbon reduction using cutting edge Solar PV and Battery technologies is constrained by for example poverty. Action on climate change is not just simply the rational allocation of resources to enable the best long-term returns for a healthy planet and humanity. The climate change action discussion is deeply political with countries in the Global South making the case that they

will be bearing the brunt of the climate liabilities in which historically polluting countries in the Global North have accrued the benefits. Capturing these nuances in terms of the 17 SDGs and a holistic outcomes-driven discussion seemed prudent in data selection to ensure the model has some degree of the human development challenges inherent in the climate change discussion.

Climate Change News Next to the Extreme Weather Corpora, Erasmus.AI searched and curated corpora based on a set of climate semantic concepts. The semantic concepts were built and curated initially from themes inside Erasmus.AI's NewsConsole which displays machine- and human-curated visualisations of narrative analysis of vast amounts of articles. Once the set of these themes was deemed of sufficient quality through human inspection, these concepts were used for larger-scale searches through the Erasmus.AI corpora.

Climate Change Specific Corpora International development organizations, treaty organizations, and the broad NGO community (World Bank, OECD, IPPC, UN, EU, TCFD, US Gov, Nation State Governments, NASA, ESA, WRI, IREA, WEF, Nature Finance, etc.) together publish significant well-researched work on climate change and its impacts on financial systems, countries, ecologies, etc. Erasmus.AI built a collection of these reports from a combination of existing collections and performed a set of custom crawls.

Climate Academic Research A set of academic publications of open access and open web academic articles were collected on climate change. It was decided to limit this corpus to open access and open web full articles (and not just abstracts) to ensure that the model represented logical arguments, inherent in full academic publications not just conclusions taken from abstracts.

A.4 AppTek Non-Expert IFT Data Details

Age	Count
18-29	72
30-49	25
50-69	2

Table 17: Age distribution of non-expert IFT data annotators.

Country	Num. Annotators	Num. prompts
India	51	4970
USA	35	1901
Germany	6	560
Mexico	3	3090
CAN	3	107
Costa Rica	1	200

Table 18: Geographic distribution of non-expert IFT data annotators.

Topic	%
Climate	9.2
From Experts Interviews	
Central Bank Policies	2.3
Extreme Weather	2.4
Geo-engineering	2.3
Industrial Systems	2.0
Natural Systems and Services	2.3
Reducing Carbon Emissions	2.6
Regenerative Agriculture	0.2
Other Topics	
Agriculture	2.3
Animals	3.9
Culture	4.0
Ecosystems	4.1
Energy	4.1
Environment	5.5
Health	3.7
History	3.9
Legal	4.1
News	3.6
Politics	4.0
Technology	4.1
Travel	3.9
Weather	24.0

Table 19: Distribution of topics provided to non-expert IFT data annotators during the data collection process.

A.5 Retrieval Augmentation Example

	Rank	Text
user query	-	What policies should India implement in Kolkata to mitigate flooding?
retrieved chunk	1	land use and land cover of ekw between 2000 and 2019 when engaging with officials and engineers from the kolkata municipal corporation and the i & wd, it becomes evident that wastewater regulation is influenced by various pressures that must be considered in urban environmental planning, management, and governance. officials utilize the bantala lock gate to lower the wastewater level in the main canal, especially during monsoons, to accelerate sewage and stormwater flow and mitigate the risk of urban flooding. the dense informal settlements in the deltaic city of kolkata heighten the threat of waterlogging, and
	2	heighten the threat of waterlogging, and the deterioration of both natural and human - made blue infrastructure amplifies the city's vulnerability to floods. officials are cautious about attracting media attention and criticism if flood risks are not managed effectively. however, opening the bantala lock gate hinders wastewater supply to the inlet canals and, consequently, the bheris. this puts kolkata municipal corporation and the i & wd in a dilemma, torn between addressing the needs of low - lying urban residents and fishers during the monsoons. the media portrayal suggesting a deliberate
	3	the monsoons. the media portrayal suggesting a deliberate jeopardizing of fishing livelihoods to facilitate real estate development is likely an exaggeration. decisions made by other government agencies and departments emphasize that there is no systematic effort to convert the wetlands into built - up areas.
	4	these floods is likely to increase as the climate changes, particularly due to storm surges, sea level rise and more intense precipitation. " future proofing " kolkata against climate change, population growth and economic development is an immense challenge, particularly considering the scale of poverty and informality in the city. iii. methods this paper evaluates the implications of " business - as - usual " modes of development for kolkata ' s energy use, energy bills and greenhouse gas emissions in the period to 2025. it also evaluates a wide range of energy efficiency, renewable energy and other mit
	5	kolkata and other indian cities are experimenting with more inclusive forms of urban planning and policymaking. kolkata has a tropical climate, with monthly mean temperatures varying from 19 to 30°c. most rainfall occurs during the monsoon season between june and september. the city frequently experiences flooding during this time due to the inadequate drainage and sewer networks, which do not serve the city ' s whole population. where this infrastructure exists, it is often a century old and lacks the capacity to meet the current population ' s needs. the frequency and severity of these floods is likely to increase as the climate changes

Table 20: Examples retrieval results.

A.6 System Prompts

Subset	System Prompt
Senior Expert Interviews	You're ClimateGPT a large language model synthesizing inter-disciplinary research on climate change. Always answer as helpfully and professional as possible, while being safe. Avoid colloquial language. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased.
Grounded Expert Demonstrations	You're an expert in climate science. Always answer as helpfully and professional as possible, while being safe.
Grounded Non-Expert Demonstrations	You're a helpful assistant supporting users with their questions on climate change.\n Cite the documents provided in the context.
StackExchange	You're an AI assistant generating answers to questions on the website stackexchange on the topic {source}.
AppTek General	You're a helpful and harmless AI assistant.
OASST-1	You're Open Assistant, an AI language model, developed by Laion AI together with an open source community and trained using crowdsourced data.
Dolly	You're an AI language model trained on data generated by employees of databricks.
Llama-2 Safety	You're a helpful assistant supporting users with their questions on climate change.
FLAN	You're a multi-task model solving a variety of NLP tasks. Give short responses only and follow the format of the user query.
CoT	You're a multi-task model solving a variety of NLP tasks. Give short responses only and follow the format of the user query.

Table 21: System prompts used for IFT training for each of the different subsets.

A.7 Prompts used in Climate-Specific Automatic Evaluation Tasks

Task	Prompt
ClimaText	Given the following statement, is it relevant to climate change or not:\n{}\nAnswer:
ClimaStance	Is the following statement in-favor, against, or ambiguous about climate change prevention:\n{}\nAnswer:
ClimateEng	Given the five categories: 'general', 'politics', 'ocean/water', 'agriculture/forestry', 'disaster', assign the following statement to one of the categories: '{}'. \nAnswer:
CDP-QA	Given a question and an answer, examine if the answer addresses the question.\nQuestion: {}\nAnswer: {}\n\nOutput:
Fever-Boolean	Is the following statement on climate correct or misinformation:\n{}\nAnswer:
Fever-Evidence	Given the following documents:\n{}\n\nIs the following claim:\n{}\nSupported or Refuted?\n
Pira 2.0 MCQ (no context)	Answer the following question with the correct alternative. GIVE ONLY THE CORRECT LETTER. \nQuestion: {}. A: {}. B: {}. C: {}. D: {}. E: {}\nANSWER:
Pira 2.0 MCQ (with context)	Based on the following context: \n\n {}. \n Answer the following question with the correct alternative. GIVE ONLY THE CORRECT LETTER \n Question: {{question}} \n\n A: {}. B: {}. C: {}. D: {}. E: {}\nANSWER:
Exeter Misinformation	This is a climate-misinformation classification task. Your task is that of telling whether the given text presents a contrarian claim regarding climate change. Your reply should be: 1: contains a contrarian claim; 0: does not contain a contrarian claim. Your reply should contain only the corresponding number and nothing else (i.e., 0 or 1).\nTEXT: {} ANSWER:

Table 22: Prompts used in each of the climate-specific automatic evaluation tasks.

A.8 Prompt for Retrieval Database Tagging

In Figure 6, we present the prompt we used with the OpenAI `gpt-3.5-turbo` text completion API to tag text chunks from our retrieval database. The text chunks in the few-shot example sections in the prompt all come from the IPCC Climate Change 2014 Mitigation of Climate Change report (Edenhofer et al., 2014). Our initial runs with the prompt proved to be satisfactory, but not perfect. For example, despite adding "Please generate comma-separated, plain-text tags, e.g. natural,social (no need to add space after the comma separators and do NOT repeat your tags).", the model sometimes did not follow this specific instruction exactly. Nonetheless, such imperfections are easily fixable with post-processing scripts, and we stuck with the prompt.

Your task is to tag a chunk of text with labels from ("natural", "economic", "social"), depending on which science discipline the text is closest to. A chunk of text can have multiple tags. Please generate comma-separated, plain-text tags, e.g. natural,social (no need to add space after the comma separators and do NOT repeat your tags). Below, you will be presented with some examples. Each example is formatted as:
text
<text to be tagged>
tags
<natural|economic|social>...
Examples will follow "=== example ===" and the text to be tagged will follow "=== to be tagged ===".

=== example ===
text
5.2.3.5 Sulphur dioxide and aerosols Uncertainties in SO2 and carbonaceous aerosol (BC and OC) emissions have been estimated by Smith et al. (2011) and Bond et al. (2004, 2007). Sulphur dioxide emissions uncertainty at the global level is relatively low because uncertainties in fuel sulphur content are not well correlated between regions. Uncertainty at the regional level ranges up to 35%. Uncertainties in carbonaceous aerosol emissions, in contrast, are high at both regional and global scales due to fundamental uncertainty in emission factors. Carbonaceous aerosol emissions are highly state-dependent, with emissions factors that can vary by over an order of magnitude depending on combustion conditions and emission controls. A recent assessment indicated that BC emissions may be substantially underestimated (Bond et al., 2013), supporting the literature estimates of high uncertainty for these emissions.
tags
natural

=== example ===
text
The energy intensity of the industry sector could be directly reduced by about 25% compared to the current level through the wide-scale upgrading, replacement and deployment of best available technologies, particularly in countries where these are not in use and in non-energy intensive industries (high agreement, robust evidence). Additional energy intensity reductions of about 20% may potentially be realized through innovation (limited evidence, medium agreement). Barriers to implementing energy efficiency relate largely to initial investment costs and lack of information. Information programmes are a prevalent approach for promoting energy efficiency, followed by economic instruments, regulatory approaches and voluntary actions. [10.7, 10.9, 10.11]
tags
economic

=== example ===
text
Reduction of subsidies to fossil energy can achieve significant emission reductions at negative social cost (very high confidence). Although political economy barriers are substantial, many countries have reformed their tax and budget systems to reduce fuel subsidies that actually accrue to the relatively wealthy, and utilized lump-sum cash transfers or other mechanisms that are more targeted to the poor. [15.5.3]
tags
social

=== example ===
text
No single factor explains variations in per-capita emissions across cities, and there are significant differences in per capita GHG emissions between cities within a single country (robust evidence, high agreement). Urban GHG emissions are influenced by a variety of physical, economic and social factors, development levels, and urbanization histories specific to each city. Key influences on urban GHG emissions include income, population dynamics, urban form, locational factors, economic structure, and market failures. Per capita final energy use and CO2 emissions in cities of Annex I countries tend to be lower than national averages, in cities of non-Annex I countries they tend to be higher. [12.3]
tags
economic,social

=== to be tagged ===
text
<Text Chunk To Be Tagged>
tags

Figure 6: Prompt used for retrieval database tagging with the OpenAI `gpt-3.5-turbo` text completion API.

A.9 Full Automatic Evaluation Results

Table 23 shows the performance of the models on the individual climate-specific tasks grouped under ClimaBench and Pira benchmarks.

Models	CDP-QA	Clima-Text	Climate-Eng	Climate-Stance	Fever-Boolean	Fever-Evidence	Pira-MCQ (no ctx)	Pira-MCQ (with ctx)
Weights	1.0	0.5	0.5	0.5	1.0	1.0	1.0	1.0
Stability-3B	74.7	57.1	61.1	80.6	72.2	74.9	40.5	56.8
Pythia-6.9B	67.8	52.8	36.6	78.0	63.1	71.6	21.6	24.2
Falcon-7B	78.7	57.7	48.2	75.8	50.5	63.0	21.1	18.5
Mistral-7B	79.3	66.7	65.4	78.0	71.7	72.9	67.0	93.0
Llama-2-7B	73.1	55.8	61.1	72.7	66.3	74.0	45.8	56.4
Jais-13B	67.1	62.9	60.8	56.9	70.6	73.0	19.8	33.0
Jais-13B-Chat	71.3	72.8	35.8	40.0	70.5	80.0	58.1	74.4
Llama-2-Chat-7B	77.4	72.1	60.8	70.1	62.0	64.2	63.0	81.1
Llama-2-Chat-13B	72.2	74.7	52.7	62.3	69.7	72.0	68.3	90.3
Llama-2-Chat-70B	77.0	78.1	59.4	69.9	70.8	75.7	83.3	94.3
ClimateGPT-7B	81.2	70.5	65.1	59.4	73.5	81.0	81.1	93.0
ClimateGPT-13B	83.0	76.3	68.5	56.6	77.6	76.1	82.4	95.6
ClimateGPT-70B	83.2	78.3	68.7	50.1	69.9	74.3	85.5	94.3
ClimateGPT-FSC-7B	43.7	46.5	53.0	77.7	63.5	70.9	18.1	16.3
ClimateGPT-FSG-7B	35.1	50.0	45.4	77.7	45.2	72.3	20.3	14.5

Table 23: Five-shot performance on climate-specific automatic evaluation tasks. Task-specific weights are used to compute the weighted-average score in Table 11.

A.10 MT Glossary Examples

In Table 24, we share a few examples of Glossary entries that are were used during the MT inference.

Original	Adjustment/Correction (if applicable)
Adaptation Research	أبحاث التكيف
Paris Agreement	اتفاقية باريس
Obligate Species	أجناس شبه محددة الموطن
Water Stress	إجهاد المياه
Carbon Capture and Sequestration	احتجاز الكربون وامتصاصه
Carbon capture and storage	احتجاز الكربون وتخزينه
Bond event	أحداث بوند
Earth's energy imbalance	اختلال توازن الطاقة
Eustatic Sea-Level Rise	ارتفاع مستوى سطح البحر
Sea Level Rise	ارتفاع مستوى سطح البحر

Table 24: Example climate-related glossary used during machine translation inference.