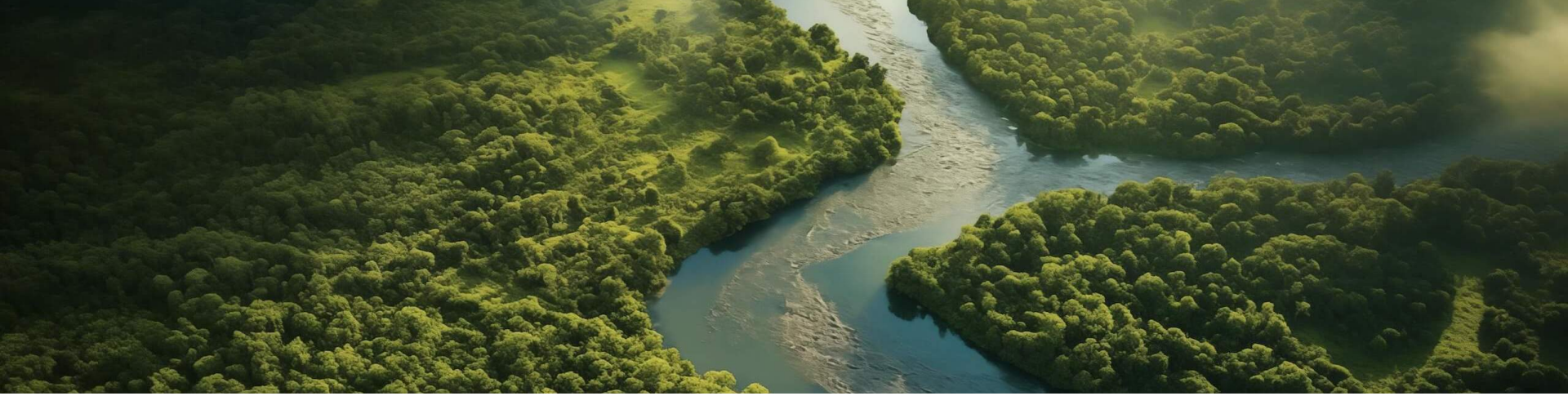


ClimateGPT: Towards Domain-Specific Large Language Models for Climate Change

David Thulke

Keynote Speech at the ClimateNLP Workshop at ACL 2024

Joint Work with:

AppTek
Y. Gao, R. Jalota, A. Nasir, H. Goldstein, T. Tragemann, K. Nguyen,
E. Tsymbalov, E. Matusov, M. Yaghi, M. Shihadah, H. Ney, C. Dugast

Erasmus.AI
P. Pelser, R. Brune, F. Fok, I. van Wyk, J. de Waal, D. Erasmus

EqtyLab
M. Ramos, A. Fowler, A. Stanco, J. Gabriel,
J. Taylor, D. Moro, J. Dotan

AppTek HLT | RWTH AACHEN UNIVERSITY

Motivation

A climate-specific large language model

- Climate change is an urgent, complex and interdisciplinary topic
- Applications of LLMs in the climate domain:
  - Make climate science more accessible - e.g. ClimateQ&A[1], ChatReport[2]
  - Identify climate misinformation - e.g. Climinator (Leippold et al., 2024)
  - Extract structured information from diverse documents - e.g. Climate Impact Events (Li et al., 2024)
- Mostly very large general-purpose models are used
- Other fields: domain-specific models can outperform general-purpose models
  - e.g. Science (Taylor et al., 2022), Medicine (Chen et al., 2023), Finance (Wu et al., 2023)

- ClimateGPT Project: Development of a climate-specific large language model
  - Main project time: September - November 2023
  - Prototype application: multilingual Q&A system demoed at COP28 in December
  - Partners: AppTek, Erasmus.AI, EqtyLab

[1] climateqa.com
[2] reports.chatclimate.ai

Outline

Domain-Specific Pre-Training

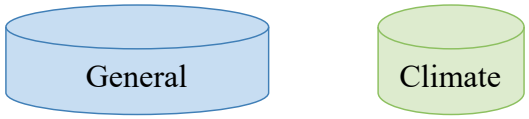
How to create a domain-specific language model?

- Climate-specific data
  - 4.2B tokens from a private corpus from Erasmus.AI
  - Pipeline identifying climate-relevant documents from news, papers and reports

Domain-Specific Pre-Training

How to create a domain-specific language model?

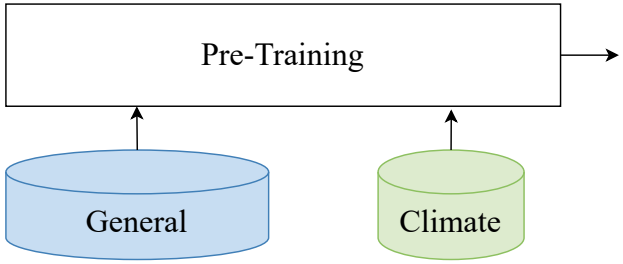
- Climate-specific data
  - 4.2B tokens from a private corpus from Erasmus.AI
  - Pipeline identifying climate-relevant documents from news, papers and reports



Domain-Specific Pre-Training

How to create a domain-specific language model?

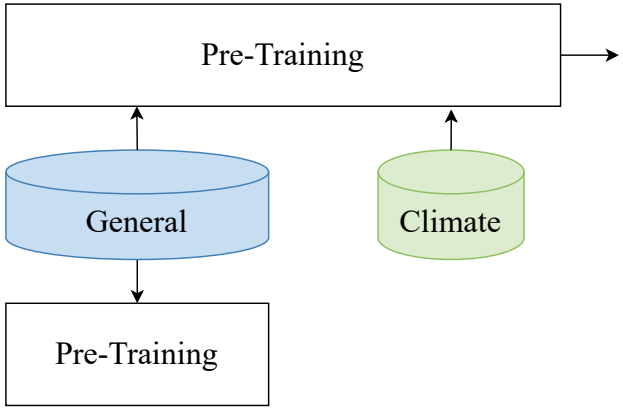
- Climate-specific data
  - 4.2B tokens from a private corpus from Erasmus.AI
  - Pipeline identifying climate-relevant documents from news, papers and reports



Domain-Specific Pre-Training

How to create a domain-specific language model?

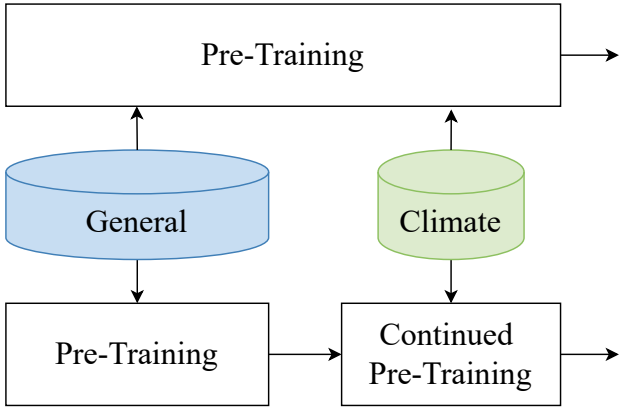
- Climate-specific data
  - 4.2B tokens from a private corpus from Erasmus.AI
  - Pipeline identifying climate-relevant documents from news, papers and reports



Domain-Specific Pre-Training

How to create a domain-specific language model?

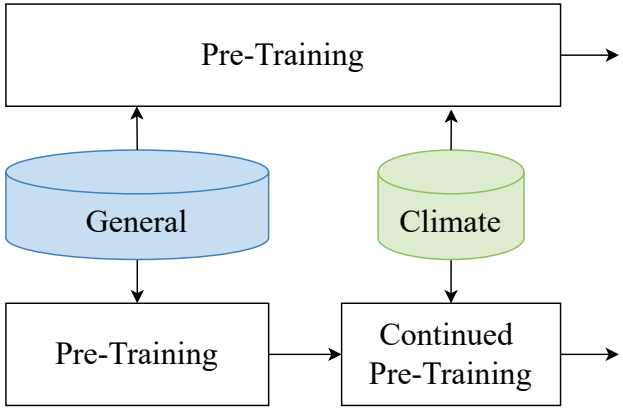
- Climate-specific data
  - 4.2B tokens from a private corpus from Erasmus.AI
  - Pipeline identifying climate-relevant documents from news, papers and reports



Domain-Specific Pre-Training

How to create a domain-specific language model?

- Climate-specific data
  - 4.2B tokens from a private corpus from Erasmus.AI
  - Pipeline identifying climate-relevant documents from news, papers and reports

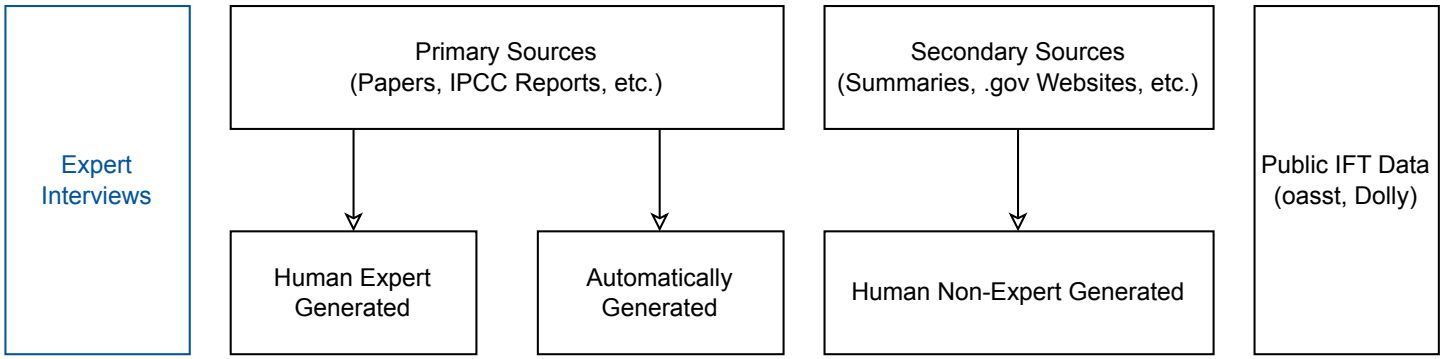


- ClimateGPT Model Variants (300B tokens general data)
  1. FSC (From-Scratch Climate): Pre-Training on climate and general data
  2. FSG (From-Scratch General): Pre-Training on general data + CPT on climate data
  3. CPT (Continued Pre-Training): Llama-2 (trained on 2,000B tokens) + CPT on climate data

Instruction Fine-Tuning

Expert Interviews

- Participants: 1 Professor working on food security
- Detailed interview with an expert on their field of expertise
- Separate person transforms interview notes into prompt and completion pairs
- Idea: high-quality responses from experts with a good high-level understanding of the field

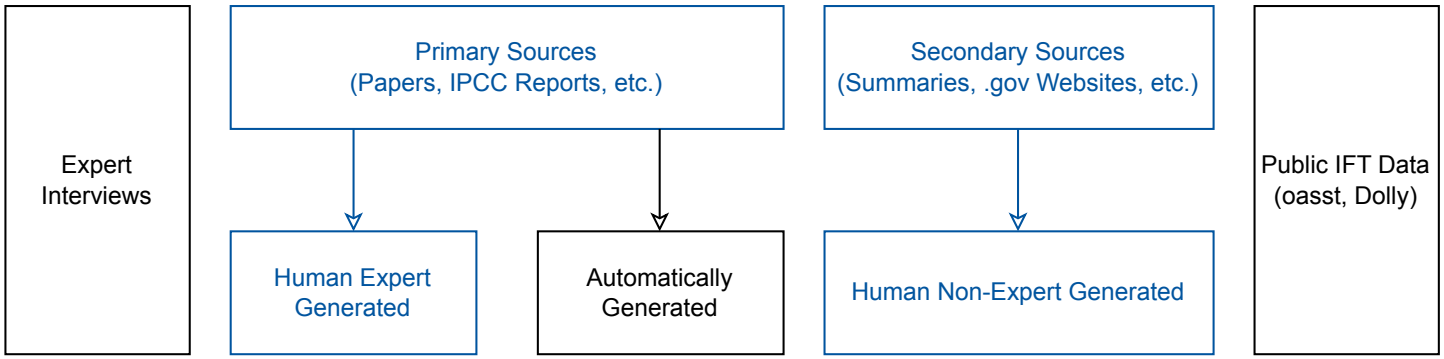


Instruction Fine-Tuning Tracks

Instruction Fine-Tuning

Grounded Demonstrations

- Participants:
  - 9 PhD or graduate student in climate-related fields
  - 99 non-expert annotators from AppTek's data workforce
- Start with a topic and reference document (depending on the expertise)
- Documents can be used as grounding for RAG



Instruction Fine-Tuning Tracks

Grounded Demonstrations

Document

Title: Assessing the Impacts of Urbanization on Albedo in Jing-Jin-Ji Region of China

Excerpt: As an indicative parameter that represents the ability of the Earth's surface to reflect solar radiation, albedo determines the allocation of solar energy between the Earth's surface and the atmosphere, which plays an important role in both global and local climate change. Urbanization is a complicated progress that greatly affects urban albedo via land cover change, human heat, aerosol, and other human activities. Although many studies have been conducted to identify the effects of these various factors on albedo separately [...]

Prompt

What is the effect of urbanization on the albedo of roofs and roads in the Jing-Jin-Ji region?

Completion

In the Jing-Jin-Ji region, the effects of urbanisation on the albedo of roofs and streets have been identified. Albedo, or the ability of the Earth's surface to reflect solar radiation, plays a [...]

Summary IFT Datasets

Domain	Name	Total Size	Epochs
Climate	Senior Expert Interviews	74	18
	Grounded Expert Demonstr.	403	18
	Grounded Non-Expert Demonstr.	9,663	15
	StackExchange	3,282	3
General	OASST-1	3,783	3
	Dolly	15,001	3
	AppTek General	700	3
Safety	Llama-2 Safety	939	3
NLP Tasks	FLAN (Subset)	30,000	1
	CoT (Subset)	15,000	1

Automatic Evaluation

Task Overview

- Few-shot evaluation of climate-specific NLP tasks

Aspect	Datasets		Task Type
Climate-Specific	ClimaBench	ClimaText	Topic Classification
		ClimateStance	Sentiment Analysis
		ClimateEng	Topic Classification
		ClimateFever	Fact Checking
		CDP-QA	QA-Pair Prediction
	Pira 2.0	MCQ	Multiple-Choice QA
	CARDS	Binary	Misinformation Classification

- Limitation: most of these tasks do not require detailed domain knowledge

\+ Commonly used general benchmarks (e.g. MMLU)

Automatic Evaluation

Results on climate-specific benchmarks

Model	ClimaBench	Pira 2.0 MCQ	CARDS	Weight. Avg.
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

Accuracies on the climate benchmarks.

- No degradations on general benchmarks

Automatic Evaluation

Ablation Experiments

Model	ClimaBench	Pira 2.0 MCQ	CARDS	Weight. Avg.
ClimateGPT-7B	75.3	86.6	65.9	77.1
w/o CPT	72.7	85.7	66.5	75.3
only general IFT	72.6	57.0	63.0	67.2
only climate IFT	70.5	50.2	59.8	63.7
only NLP Task IFT	74.5	83.9	64.3	75.6

Accuracies on the climate benchmarks.

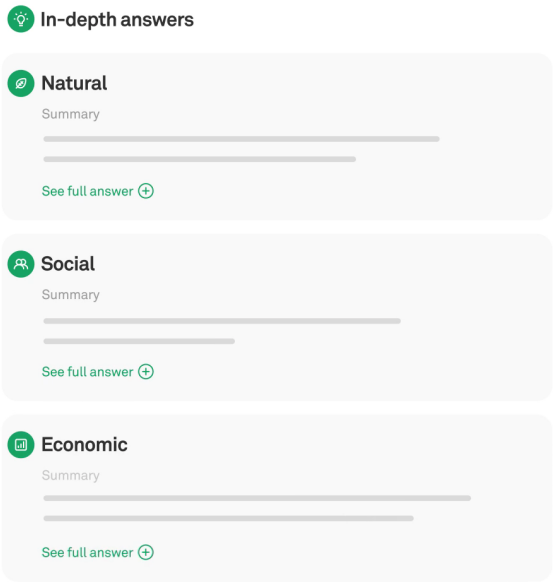
Q&A System

RAG

- Goal: reduce hallucinations by grounding outputs in retrieved documents
- Document base: IPCC reports, scientific papers, etc.
- Hierarchical Retrieval approach
- Instruction fine-tuning data adapted for RAG

Perspectives

- Goal: provide detailed responses from different perspectives
- Approach:
  - Prompting
  - Special instruction fine-tuning data
  - RAG-specialisation



Q&A System

Multilinguality

- In addition to English the demo should be available in Arabic (+ ideally all other UN languages)
- Options for multilinguality:
  1. Multilingual LLMs
     - At the time no open multilingual LLMs
     - Requires language-specific instruction fine-tuning data
  2. **Cascaded Machine Translation**
     - No additional work on the LLM is required
     - Easy to extend to new languages
     - Additional latency due to additional translation step

- AppTek's Machine Translation adapted to climate
- The final Q&A system was deployed in 22 languages

Human Evaluation

Methodology

- Evaluation dimensions based on work by (Bulian et al., 2023)
  - **Style:** Avoid overly informal language; balance text length to maintain engagement without oversimplifying.
  - **Clarity:** Use clear, simple language; avoid jargon to enhance understanding, especially for those with lower numeracy skills.
  - **Correctness:** Adhere to linguistic conventions such as proper punctuation, spelling, and grammar to maintain credibility.
  - **Tone:** Maintain a neutral tone; avoid overly positive or negative language to effectively convey factual information.
  - **Accuracy:** Ensure scientific information is correct, with no out-of-context or contradictory statements.
  - **Specificity:** Provide region and time-specific info, tailoring to the audience for relevance.
  - **Completeness:** Address all parts of the question fully, reflecting the depth of scientific knowledge.
  - **Uncertainty:** Communicate the certainty/agreement of findings, and if unknown, share the uncertainty.
- Overall assessment on scale from -2 to 2

Human Evaluation

Results

- 50 prompts from our heldout IFT data
- 7 academics in climate-related fields (graduate, PhD or postdoc)

Model	Climate Avg.	General Avg.	Average Rank	# Hallucinations
ClimateGPT-FSC-7B	46.2	48.8	0.2	5
ClimateGPT-7B	77.1	65.1	0.6	4
ClimateGPT-70B	77.2	73.7	1.0	2
ClimateGPT-70B	77.2	73.7	0.8	2
GPT-3.5-Turbo	-	-	0.9	3

Conclusion

Plans for the next version of the model:

- Based on Llama 3
- Extended climate pre-training data (4.2 $\rightarrow$ 10.4B tokens)
- Extend model capabilities by additional instruction fine-tuning data
- Extend instruction fine-tuning by additional *climate-related* NLP tasks

A few open challenges:

- Increase faithfulness and robustness in RAG
- More complex benchmarks are needed for evaluating foundation models for climate tasks
- Rigorous evaluation in real-world use-cases

Paper
`https://arxiv.org/abs/2401.09646`

Models on Hugging Face
`https://huggingface.co/eci-io`

Bulian, J., Schäfer, M. S., Amini, A., Lam, H., Ciaramita, M., Gaiarin, B., Huebscher, M. C., Buck, C., Mede, N., Leippold, M., and Strauss, N. (2023). Assessing large language models on climate information.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., Pagliardini, M., Fan, S., Köpf, A., Mohtashami, A., Sallinen, A., Sakhaeirad, A., Swamy, V., Krawczuk, I., Bayazit, D., Marmet, A., Montariol, S., Hartley, M.-A., Jaggi, M., and Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models.

Leippold, M., Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Luterbacher, J., and Huggel, C. (2024). Automated fact-checking of climate change claims with large language models.

Li, N., Zahra, S., de Brito, M. M., Flynn, C. M., Görnerup, O., Koffi, W., Kurfali, M., Meng, C., Thiery, W., Zscheischler, J., Messori, G., and Nivre, J. (2024). Using LLMs to build a database of climate extreme impacts. In *Natural Language Processing meets Climate Change @ ACL 2024*.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. (2022). Galactica: A large language model for science.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., and Mann, G. (2023). Bloomberggpt: A large language model for finance. *ArXiv preprint: https://arxiv. org/pdf/2303.17564*.