Name: Joongeun Choi
Project Proposal: Fine-tuning ClimateGPT to Improve Climate Fact-Checking

Motivation:

In an ever-increasingly digital age, social media platforms have become a significant medium of communication, allowing information–regardless of its veracity–to be spread more easily than ever before. In the context of pressing climate issues, misinformation can give rise to doubts about the legitimacy of scientific research and mistrust in facts, slowing the necessary response to climate change mitigation or adaptation. Recently, a study conducted by the Center for Countering Digital Hate (CCDH) found that in 2023, 70% of climate change disinformation on YouTube concentrated on the infeasibility of climate solutions, the falsehood of climate science, and the harmlessness of global warming, compared to 35% in 2018.[1]

Previous Works:

In order to decelerate climate change by effecting regulatory policies, combating misinformation is crucial to increase trust for scientific solutions and environmentally progressive government action. Since manually fact-checking large amounts of data is difficult and time-consuming, there have been multiple attempts to automate the process.

Initially, research focused on using Natural Language Understanding (NLU) to conduct a three-way classification: labeling climate change-related claims as being supported, not supported, or having not enough information based on evidence texts. In 2021, the first such instance appeared in the paper discussing the new Climate FEVER dataset. The FEVER dataset is the largest dataset of artificially created claims; the Climate FEVER dataset followed a similar format but used real climate change-related claims from the internet that were manually categorized into the three classes mentioned above based on evidence texts. The paper discusses training the Albert model on the FEVER dataset and evaluating it on the Climate FEVER dataset, achieving an F1 score of 32.85%.[2] That same year, a research group that modified the model to Robustly Optimized BERT (RoBERTa) and fine-tuned it on a portion of the Climate FEVER dataset achieved an F1 score of 71.82% on unused Climate FEVER examples.[3] In 2022, an F1 score of 75.7% was achieved by ClimateBERT, which is a DistilRoBERTa model pre-trained on the general domain and subsequently on climate-related text (collected with web crawling), and finally, trained on a downstream task (fact checking).[4]

---

[1] Center for Countering Digital Hate, *The New Climate Denial*, (2024), https://counterhate.com/research/new-climate-denial/

[2] Thomas Diggelmann et al., "CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims," arXiv, 2021, https://doi.org/10.48550/arXiv.2012.00614.

[3] Wang, Gengyu, Lawrence Chillrud and Kathleen McKeown, "Evidence based Automatic Fact-Checking for Climate Change Misinformation." *ICWSM Workshops*, 2021, https://workshop-proceedings.icwsm.org/pdf/2021_39.pdf.

[4] Nicolas Webersinke et al. "Climatebert: A Pretrained Language Model for Climate-Related Text." arXiv, 2022, https://doi.org/10.48550/arXiv.2110.12010.

More recently, the sudden popularization of GPTs has prompted the use of generative language models to output a more detailed explanation to a user query. ChatClimate (2023)[5], ClimateGPT (2024)[6], and Climinator (2024)[7] are the latest developments in conversational climate chatbots. Additionally, there exist functional applications of both climate chatbots and fact checker bots: ChatNetZero[8] answers user queries by citing evidence from climate documents and reports, and Polestar Truths[9], a bot temporarily deployed on the social media platform X, responded to posts with misinformation. Finally, some large language models are tailored to a specific subset of climate change. For instance, WildfireGPT responds to user queries with insights on wildfire risks.[10] Only ChatClimate and ClimateGPT are open-source models. Notably, all studies implemented Retrieval-Augmented Generation (RAG) to elicit a more accurate response and avoid hallucinations.[11] The ClimateGPT study utilized 5-shot learning before evaluating on the Climate FEVER and Exeter Misinformation datasets. ClimateGPT-13B, the best-performing version of ClimateGPT for three-way classification on the Climate FEVER dataset achieved an accuracy of 77.6%.

Although three-way classification performance continues to improve, real-world applicability is limited by the vague and uninformative nature of resultant three-way labels. The use of GPT models in recent studies addressed the problem of informativeness by using text generation abilities to justify predicted labels. The vagueness can be addressed by utilizing the Science Feedback dataset, a gold-standard dataset with expert reviews and multiclass labels that more precisely identify and disprove nuanced claims. Climinator achieved an F1 score of 90.58% for three-way classification on Science Feedback after merging thirteen labels into three-way labels. However, the F1 score for claim verification shrunk to 43.84% when labels were condensed into five categories, highlighting the new challenge of identifying nuanced claims.

In summary, there have been significant advances in achieving a state-of-the-art fact-checking model within the past few years–in terms of both accuracy and reasoning capabilities. However, the low F1 scores for more precise classification of claims indicate that the fact-checkers for environmental claims are not yet fully ready to be deployed in the real world, where not all things are as clear-cut as ones and zeros. Doing so may result in misidentifying misinformation, defeating the purpose of fact-checking.

---

[5] Saeid Ashraf Vaghefi et al., "ChatClimate: Grounding conversational AI in climate science," *Commun Earth Environ* 4, 480 (2023). https://doi.org/10.1038/s43247-023-01084-x.
[6] David Thulke et al., "ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change," arXiv, 2024, https://doi.org/10.48550/arXiv.2401.09646.
[7] Markus Leippold et al., "Automated Fact-Checking of Climate Change Claims with Large Language Models," arXiv, 2024, https://doi.org/10.48550/arXiv.2401.12566.
[8] https://chatnetzero.ai/
[9] https://twitter.com/PolestarTruths
[10] Yangxinyu Xie et al., "WILDFIREGPT: TAILORED LARGE LANGUAGE MODEL FOR WILDFIRE ANALYSIS," arXiv, 2024, https://doi.org/10.48550/arXiv.2402.07877.
[11] Patrick Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," arXiv, 2021, https://doi.org/10.48550/arXiv.2005.11401.

Project Outline:

Currently, no open-source generative large language model has been fine-tuned and evaluated on the ground truth dataset Science Feedback. Hence, we aim to close this research gap by developing an open-source model trained on the multiclass (>3) labels of Science Feedback that can pick up on more nuanced claims while achieving satisfactory F1 scores. We propose further fine-tuning ClimateGPT–a Llama-2 model pre-trained on 4.2 billion climate tokens and instruction fine-tuned on question-answer pairs[12]–on the Science Feedback dataset to improve its fact-checking capabilities. After web scraping the 312 claims and their corresponding labels and expert reviews from the Science Feedback website, we would conform the labels into 5 classes according to the Science Feedback organization's metrics for credibility. For instance, sources with very high credibility are labeled as "accurate" or "correct," high credibility with "mostly accurate" or "mostly correct," neutral credibility with "lacks context," "imprecise," or "partially correct," low credibility, with "unsupported" or "misleading," and very low credibility with "inaccurate," "incorrect," or "flawed reasoning." We would split the dataset into training, validation, and testing sets, concatenating the claims to fit the ClimateGPT and Climinator claim-verification prompt templates. First, we would evaluate the baseline ClimateGPT-7B and ClimateGPT-13B model's performance on the testing set. Next, we would further fine-tune ClimateGPT-7B and ClimateGPT-13B through supervised learning to provide multiclass labels and evidence-backed justification using the training and validation set. Finally, we would determine the best-performing model's combination of model type, training procedure, and prompt type, and compare its F1 score with the baseline. The optimal setup–along with the web-scraped Science Feedback data–will be uploaded on the open-source Huggingface model hub for public use. Once more claim reviews accumulate, future work could conform the labels into 7 classes introduced by the Climinator study and test ClimateGPT's performance on an even more nuanced dataset than the 5 class task. Furthermore, a practical application would be implementing the fine-tuned ClimateGPT model as a Chrome extension that alerts users of misinformation on social media in real time.

[12] Thulke et al., "ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change."