Comparing the performance of ChatGPT and state-of-the-art climate NLP models on climate-related text classification tasks

Dimitar Trajanov[1,2,*], *Gorgi* Lazarev[1], *Ljubomir* Chitkushev[2], *Irena* Vodenska[3,1]

[1]Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, 1000 Skopje, North Macedonia
[2]Computer Science Department, Metropolitan College, Boston University, Boston, MA, USA
[3]Administrative Sciences Department, Financial Management, Metropolitan College, Boston University, Boston, MA, USA

Abstract. Recently, there has been a surge in general-purpose language models, with ChatGPT being the most advanced model to date. These models are primarily used for generating text in response to user prompts on various topics. It needs to be validated how accurate and relevant the generated text from ChatGPT is on the specific topics, as it is designed for general conversation and not for context-specific purposes. This study explores how ChatGPT, as a general-purpose model, performs in the context of a real-world challenge such as climate change compared to ClimateBert, a state-of-the-art language model specifically trained on climate-related data from various sources, including texts, news, and papers. ClimateBert is fine-tuned on five different NLP classification tasks, making it a valuable benchmark for comparison with the ChatGPT on various NLP tasks. The main results show that for climate-specific NLP tasks, ClimateBert outperforms ChatGPT.

1 Introduction

The recent advancements in artificial intelligence, specifically in Natural Language Processing (NLP), have given rise to groundbreaking large language models that are capable of understanding natural human language [1]. These large language models (LLMs) are trained on diverse datasets, including books, articles, web pages, and more, enabling them to understand and generate human-like responses to a wide range of prompts [2].

The availability of publicly accessible language models has facilitated their integration into our daily lives, empowering individuals to automate repetitive tasks, improve writing skills, generate high-quality content, and even assist in information search and validation, surpassing traditional search engines. One prominent example is ChatGPT (Generative Pre-trained Transformer), developed by OpenAI, which has gained widespread adoption and found utility in various real-world applications such as chatbots, content generation for blogs and articles, code correction, language translation, and even medical diagnosis and treatment [3].

One significant advantage of ChatGPT is its potential to simplify and facilitate complex problem-solving tasks, impacting various real-world scenarios. Several studies have examined the challenges and opportunities of ChatGPT concerning environmental and climate-related issues. The authors of [4] present a comprehensive analysis of ChatGPT's application in environmental research. They provide ten real-world examples that highlight its versatility in explaining concepts related to PFAS, microplastics, life cycle assessment, and the circular economy. The authors also discuss how ChatGPT can offer personalized programming assistance, including code generation, syntax error identification, and clarification of complex syntax. While cautioning about the challenges of misinformation and biases in AI-generated responses, they advocate for responsible integration of AI in decision-making processes. They stress the importance of human judgment and discourage excessive reliance on AI, particularly in addressing public environmental concerns. The authors' balanced perspective encourages the responsible use of AI tools in research and beyond.

In [5], the authors present insights into the utilization of ChatGPT in biology and environmental science. They demonstrate how ChatGPT can simplify complex tasks in these fields by analyzing the answers to 100 important questions in biology and 100 important questions in environmental science. The authors acknowledge the potential benefits of ChatGPT for academic activities while highlighting the need to address potential risks and harms. Despite these limitations, they express optimism in fully harnessing recent technological advancements to push the boundaries of biology and environmental science.

In [6], the authors investigate ChatGPT's potential application in climate change research, highlighting its value in data analysis, interpretation, communication, outreach, decision-making support, and climate scenario generation. These tasks demonstrate ChatGPT's ability to simplify complex climate data and aid in policy decisions. However, also in this paper, the authors balance optimism by acknowledging ChatGPT's limitations. The authors

* Corresponding author: dimitar.trajanov@finki.ukim.mk

highlight how the quality and volume of training data impact the model's output quality, leading to diverse utility across various research questions. The authors also address significant challenges, including the model's potential difficulties in understanding complex scientific concepts, lack of contextual awareness, and possible biases in the training data.

The literature review indicates the potential and limitations of using ChatGPT for climate-related tasks. To provide quantitative and qualitative measures of ChatGPT's usefulness, we conducted a study comparing its performance with the state-of-the-art domain-specific language model, ClimateBERT [7][8]. ClimateBERT is trained on climate-related research paper abstracts, news, and company reports and fine-tuned for five distinct downstream NLP tasks: (1) climate detection classification, (2) climate sentiment analysis, (3) climate-related commitments and actions classification, (4) climate change specificity classification, and (5) climate change disclosure category classification. For the comparison, we selected all five downstream tasks and used the test dataset provided by ClimateBERT for these tasks. Employing prompt engineering best practices, we created multiple prompt variants to classify the text into relevant categories based on the selected downstream task. To compare the result of the ChatGPT, we compare its performance with the results obtained by the finetuned ClimateBERT models. The experiments that were conducted can be found in the GitHub[1] repository, along with detailed instructions on how to replicate the results.

The paper is structured into three primary sections: Methodology, Results and Discussion, and Conclusion. In the Methodology section, we detail the strategies and tools we utilized in our research, focusing specifically on how ChatGPT was utilized. In the Results and Discussion section, we present the outcomes of our study, and in the Conclusion, we summarize our research findings, consider their potential impact, and suggest possible future directions.

2 Methodology

The evaluation methodology consists of three main steps: (1) loading the necessary data from the datasets, (2) feeding the data into the models to obtain predicted labels for each paragraph, and (3) comparing the predicted labels with the actual labels to assess model accuracy and performance.

Although the ClimateBert and ChatGPT models have different operating mechanisms, the process of label prediction varied slightly. The ClimateBert model executed the task fine-tuned on the provided data without user interaction, while ChatGPT required a user prompt to generate a response.

In the following text, first, we will describe the used datasets, and then the details about ClimateBERT and ChatGPT models will be given.

2.1 Datasets

To test the ChatGPT performance as accurately as possible, we selected all five ClimateBERT downstream tasks. These tasks encompass several variations of the text classification, aiming to capture distinct aspects of climate-related content.

- Task 1: Climate Detection Classification - This task involves determining whether a given paragraph is climate-related or not, providing insights into the prevalence of climate-related content within a corpus.
- Task 2: Climate Sentiment Analysis - Here, the goal is to identify the sentiment expressed in climate-related paragraphs by categorizing paragraphs as representing neutral sentiment, opportunity sentiment, or risk sentiment towards the climate.
- Task 3: Climate-Related Commitments and Actions Classification - This task focuses on classifying climate-related paragraphs based on whether they pertain to commitments and actions or not. It helps analyze the translation of climate-related discussions into tangible commitments and actionable measures.
- Task 4: Climate Change Specificity Classification - This task assesses the specificity of climate-related paragraphs, determining whether they are specific or not. It provides insights into the level of detail and granularity in climate-related discussions, aiding in understanding the depth of information available for targeted climate action.
- Task 5: Climate change disclosure category classification. This is a multiclass classification task, which involves categorizing paragraphs into one of the four recommended categories outlined by the Task Force on Climate-related Financial Disclosures (TCFD). These categories include governance, strategy, risk management, and metrics and targets [9]. This classification allows for a comprehensive examination of how climate-related information is organized and reported within the financial sector, shedding light on areas of emphasis and potential gaps in reporting practices.

The datasets used for testing the tasks share a consistent structure, including the paragraph and a categorical label indicating the corresponding class.

2.2 ClimateBERT Baseline Models

The base ClimateBert model is pretrained on a corpus comprising climate-related and non-climate-related paragraphs from various sources such as news articles, Wikipedia articles, research papers, and climate reports. It is primarily fine-tuned for climate detection in paragraphs. Multiple variations of the ClimateBert model are available, each fine-tuned for one of the five tasks. These models can be accessed through the HuggingFace Library[2] [10]. Additionally, the datasets are uploaded to the HuggingFace Dataset Hub[3] for easy loading and utilization in the testing scripts.

1 https://github.com/gorgilazarev3/chatgpt-state-of-the-art-climate-models/

2 https://huggingface.co/

3 https://huggingface.co/datasets

For each of the five tasks, the corresponding model and dataset were loaded, and the results were evaluated.

2.3 ChatGPT Models

Given that ChatGPT is a universal language model, it requires carefully curated prompts to understand both the task at hand and the desired format of the output. Our assessment of ChatGPT was two-fold: we leveraged our own prompts, adhering to best practices for prompt construction, and alternatively, utilized an existing library.

2.3.1 Prompt engineering-based approach

In our first approach, we examined how effectively ChatGPT could perform text classification tasks using manually designed prompts. To complete this task, we called the OpenAI API in batches and then collected the corresponding responses. Each request to ChatGPT comprised a specific prompt that began by stating that ChatGPT is an expert in the respective field, followed by the paragraph that needed to be classified, and finally, guidelines regarding the expected format of the response. After data collection, each response was mapped to a categorical label based on the output instructions provided in the prompt. Responses that deviated from the anticipated output format, which couldn't be automatically mapped, were manually labeled. This was achieved by examining the content of the response and associating it with the relevant class. After labeling, we compared these assigned labels against the ground truth to determine the results.

Additional efforts were made to make sure that the prompts that we send to ChatGPT are as descriptive as possible and follow patterns defined in [11]. The prompt patterns from the catalog that were implemented in our prompts include "*The Persona Pattern*", which enables the model to take a certain point of view or role, in our case, a climate, sustainability, and environmental expert; *"The Fact Check List Pattern*", which instructs the model to output the most important points of a text and then use those points as the input in a follow-up prompt and the *"Reflection Pattern*" in which the model is asked to explain the reasoning behind its response. In all techniques for enhancing prompt engineering, a coherent chain of thought processes comprising a series of intermediate reasoning steps [12] is followed. The utilization of CoT (Chain-of-thought) enables models to generate more comprehensive reasoning processes. However, due to its emphasis on intermediate reasoning steps, there is a potential risk of introducing hallucinations and accumulated errors, thereby constraining the models' effectiveness in addressing complex reasoning tasks. Motivated by the natural carefulness and deductive reasoning processes of humans when completing complex tasks, the authors of [13] aim to empower language models to perform explicit and rigorous deductive reasoning and ensure the reliability of their reasoning process through self-verification.

2.3.2 Classification using off-the-shelf library

The alternative approach involved the use of an off-the-shelf library, Scikit-LLM[4], which allows using ChatGPT as any other conventional Classifier that would be found in the Scikit-Learn library. Scikit-learn [5] library is a machine learning package that includes a large list of machine learning methods like Classifiers and Regressors that can be conventionally used for a selection of data science tasks [14]. The Scikit-LLM library already has its prompts set up in the Classifiers, and all that is needed is to provide the labels for the classes and optionally give a few samples to train ChatGPT in the context of the data.

We have used the ZeroShotClassifier from the library in our evaluation of ChatGPT in this approach which is a classifier that only requires the labels for the classes in order to be able to classify the provided text. The responses are returned as predictions in the format of the labels provided to the Classifier as input. In the same manner, as with the other approach, the responses were compared against the actual labels to get the results.

3 Results and Discussion

This section provides a detailed examination of the results. We utilize two metrics, accuracy [15] and f1-score [16], to evaluate the generalization ability of classifiers and consider class balance. These metrics are presented in Table 1.

3.1 ClimateBERT Performance

The results show that the individual fine-tuned models, based on ClimateBert, generally achieve satisfactory performance across all five tasks.

The base ClimateBert model is primarily fine-tuned for text classification to determine whether paragraphs are climate-related. As expected, the model performs exceptionally well on task 1, achieving high accuracy (0.97) and an f1-score of 0.9572. While the other tasks also involve classification, their performances are acceptable but less outstanding than the climate detection task. The sentiment analysis task (task 2) and the classification of paragraphs as commitments and actions (task 3) exhibit similar accuracy and f1-scores, with slightly better results observed for the commitments and actions classification. The specificity classification (task 4) shows lower accuracy and f1-scores, though still satisfactory to some extent. The last task for classifying paragraphs into the TCFD recommended categories (task 5) is the least performing task, which can be attributed to its complexity and multiple classes.

[4] https://github.com/iryna-kondr/scikit-llm

[5] https://scikit-learn.org/stable/

Table 1. ClimateBert vs. ChatGPT-based Models performance comparison on five climate-related NLP tasks.

Task	Model									
	ClimateBert Model (trained on the task)		GPT-3.5 with simple single-stage prompts		GPT-4 with simple single-stage prompts		ChatGPT 3.5 with Scikit LLM Library		ChatGPT 3.5 complex multiple-stage prompts	
	accuracy	f1-score	accuracy	f1-score	accuracy	f1-score	accuracy	f1-score	accuracy	f1-score
Task 1: Climate detection classification	0.97	0.9572	0.86	0.8150	0.80	0.7502	**0.89**	**0.8465**	**0.89**	0.8418
Task 2: climate sentiment analysis	0.80	0.7837	**0.48**	**0.4958**	0.47	0.4626	0.46	0.4732	**0.48**	0.4879
Task 3: climate-related commitments and actions classification	0.81	0.7979	**0.53**	**0.5252**	0.47	0.4681	0.44	0.4374	0.42	0.4183
Task 4: climate change specificity classification	0.71	0.6509	0.37	0.3245	0.37	0.2876	0.35	0.3040	**0.39**	**0.3650**
Task 5: climate change disclosure category classification	0.62	0.4983	0.39	0.3994	**0.50**	**0.4865**	0.40	0.4063	0.42	0.4198

3.2 ChatGPT

3.2.1 Simple single-stage prompts

In the single-stage approach, we utilize manually designed prompts using a simple Persona Pattern [11]. In this method, we asked ChatGPT to assume the role of an expert in climate, sustainability, and the environment. We then submitted the paragraphs to be assessed, explicitly outlining the task and the required output format. The prompts provided to ChatGPT for each task were as follows:

- Task 1: Climate detection in paragraphs: "*You are the sustainability, environment, and climate change expert. Is the following text about sustainability, the environment, or climate change? Answer only with 1 if the text is sustainability, environment, or climate change related or 0 if not:*"
- Task 2: Climate sentiment analysis: "*You are the sustainability, environment, and climate change expert. Does the following text indicate risk, is neutral, or indicates an opportunity about sustainability, the environment, or climate change? Answer only with 0 if the text indicates risk, answer only with 1 if the text is neutral, or with 2 if the text indicates an opportunity:*"
- Task 3: Climate classification on commitments and actions: "*You are the sustainability, environment, and climate change expert. Is the following text about climate commitments and actions or not? Answer only with 0 if the text is not about climate commitments and actions and with 1 if the text is about climate commitments and actions*:"
- Task 4: Climate specificity classification: "*You are the sustainability, environment, and climate change expert. Is the following text specific about sustainability, environment, and climate change or not? Answer only with 0 if the text is not specific and with 1 if the text is specific:*"
- Task 5: Climate change disclosure Category classification task: "*You are the sustainability, environment, and climate change expert. Is the following text about metrics, strategy, risk, governance or is not climate change-related? Answer only with 0 if the text is not climate-related, answer only with 1 if the text is about metrics for sustainability, environment, and climate change, and answer only with 2 if the text is about strategy for sustainability, environment, and climate change, answer only with 3 if the text is about risk for sustainability, environment, and climate change and answer only with 4 if the text is about governance for sustainability, environment, and climate change*:"

The results follow a similar pattern to the distribution of performance of the ClimateBert models, with the best-performing task being climate detection, with climate commitments and actions following along with the sentiment analysis what is interesting here is that the classification of disclosure categories (task 5) is performing better than the specificity classification (task 4) even though it is a more complex task and involves multiple classes. The reason for this is most likely that the multiple classes are more separable, and ChatGPT can distinguish between them, leading to better classification scores.

3.2.2 Simple single-stage prompts with ChatGPT 4

As a second experiment we wanted to assess the impact of different versions of ChatGPT on performance. This investigation involved a basic, single-stage experiment with ChatGPT 4, with comparative analysis against ChatGPT 3.5.

The findings revealed that the performances were quite similar. ChatGPT 3.5 exhibited slightly better results for tasks 1 through 4, while ChatGPT 4 demonstrated better performance in task 5. Notably, task 5 is the most complex one, and ChatGPT 4's results in this task were the best among all the experiments with the different versions of ChatGPT.

Furthermore, it is important to highlight the cost differences in conducting these experiments. The total expense for running all tests with ChatGPT 4 was around $30. In contrast, the cost with ChatGPT 3.5 was around 3 USD. This cost variation should be considered in light of the observed differences in performance between the two versions.

3.2.3 Classification using off-the-shelf Slikit-LLM library

In the second approach, we used the Scikit-LLM library that provides the ability to use ChatGPT as a Zero Shot Classifier. This approach was easier to evaluate since we did not have to write our own prompts. We just need to set the base model (gpt-3.5-turbo in our case) and the labels to ZeroShotGPTClassifier class. Then we simply call the model to predict the test data, and the model will take care of sending the data to ChatGPT and receiving the response in the correct format of the predicted label.

The results using this approach, compared to the approach where we provide our own prompts, varied across tasks. The model performed better in climate-related text detection (task 1) and classification into disclosure categories (task 5), but it yielded poorer results in other tasks, such as specificity (task 4), commitments and actions (task 3), and sentiment analysis (task 2).

3.2.4 Complex multiple-stage prompts.

In addition to the approach described in 3.2.1, we explored the use of multiple prompt engineering techniques to assess the extent of performance improvement compared to the simple single-stage prompts. We used the gpt-3.5-turbo model for this experiment.

The first pattern used is the same one that was used in the approach described in 3.2.1, the Persona Pattern [11]. In addition to this pattern, we have used the Fact Check List Pattern [11] that instructs ChatGPT to extract the most important points from the text and then use these generated points as the input in the follow-up prompt instead of the original text. The model was also instructed to provide explanations and reasoning behind its answers, implementing the Reflection Pattern [11]. We have noticed that when ChatGPT explains the reasoning behind its response, the information is more correct, and the response corresponds to the actual situation. By using these patterns, we have created a simple three-step chain prompt pipeline in which we first introduce ChatGPT to the topic by having it take the role of an expert in the field, then we instruct him to extract the most important points from the text provided, and then using those points as the input to the actual classification task. This approach was successful in achieving a better performance in most of the tasks but was not better in all tasks; specifically, this approach has the best accuracy in tasks 1, 2, 3, and 5 and the best f1-score for tasks 4 and 5. For all other tasks, it has slightly worse results compared to the other two approaches.

The prompts in which the patterns were applied were the following:

- Task 1: Climate detection task: "*You are the sustainability, environment, and climate change expert. Read the following text about sustainability, the environment, or climate change and analyze it as an expert:*" followed by "*Now you need to classify the analyzed text whether the text is about sustainability, the environment, or climate change. Answer only with 1 if the text is sustainability, environment or climate change related or 0 if not:*"
- Task 2: Sentiment analysis task: "*You are the sustainability, environment, and climate change expert. Read the following paragraph and extract the most important points from the text and return only the points and their explanations:*" followed by "*Read the following points and answer only with the overall sentiment of all points summarized without any explanations. Answer only with 0 if the sentiment is risk related to sustainability, environment and climate change, answer only with 1 if the sentiment is neutral related to sustainability, environment and climate change and answer only with 2 if the sentiment is opportunity related to sustainability, environment and climate change:*"
- Task 3: Climate classification on commitments and actions task: "*You are the sustainability, environment, and climate change expert. Read the following paragraph and extract the most important points from the text and return only the points and their explanations:*" followed by "*Read the following points and answer only with one number that is the overall class of all points summarized without any explanations. Answer only with 0 if the points are not about climate commitments and actions, and answer only with 1 if the points are about climate commitments and actions:*"
- Task 4: Climate specificity classification task: "*You are the sustainability, environment, and climate change expert. Read the following paragraph and extract the most important points from the text and return only the points and their explanations:"* followed by "*Read the following points and answer only with one number that is the overall specificity of all points summarized without any explanations. Answer only with 0 if the points are not specific about sustainability, environment, and climate change, and answer only with 1 if the points are specific about sustainability, environment and climate change:*"
- Task 5: Climate disclosure category classification task: "*You are the sustainability, environment, and climate change expert. Read the following paragraph and extract the most important points from the text and return only the points and their explanations:*" followed by "*Read the following points and answer only with the overall class of which all points are summarized without any explanations. Answer only with 0 if the text is not climate-related, answer only with 1 if the text is about metrics for sustainability, environment, and climate change, answer only with 2 if the text is about strategy for sustainability, environment, and climate change, answer only with 3 if the text is about risk for sustainability, environment, and climate change*

and answer only with 4 if the text is about governance for sustainability, environment, and climate change:"

4 Conclusion

Our study indicates that context-specific models, such as ClimateBert, which have been fine-tuned for specific tasks, tend to outperform general-purpose language models, like ChatGPT, for classification tasks. Particularly for tasks involving climate-related texts, ClimateBert emerged as a more efficient alternative. This suggests that its application could be highly beneficial in tackling real-world climate issues.

Our research uncovered intriguing findings when tackling the most complex tasks, specifically the classification of climate change disclosure categories (Task 5). Here, the performance of ChatGPT closely matched that of the ClimateBert model. This suggests that with well-structured labels, a Zero-shot approach utilizing ChatGPT could yield noteworthy results.

It's important not to underestimate the versatility and adaptability of general-purpose models. These models demonstrated above-average performance even when handling tasks they were not specifically fine-tuned for or data they were encountering for the first time. This contrasts with conventional classifier models, which require pre-training and fine-tuning with appropriate data to discern and perform the tasks assigned.

Our findings also highlight the benefits of prompt engineering. Simple adjustments to the prompts could significantly enhance the performance of ChatGPT without necessitating further training. Given that these models are conversationally trained with diverse user-provided data, their capacity for learning and adaptation to new information is continually improving.

Despite current limitations, we anticipate that the strengths and potential of general-purpose models will continue to grow over time. Their integration into everyday life is likely to extend beyond general tasks to more complex and scientific endeavors, including analyses, diagnoses, and treatments.

For future research, it would be interesting to explore other strategies for enhancing performance, similar to our approach with prompt engineering. This could potentially include task-specific context considerations and developing automated pipelines that leverage the strengths of both context-specific and general-purpose models.

References

1. K. Tirumala, A. Markosyan, L. Zettlemoyer, A. Aghajanyan. Advances in Neural Information Processing Systems, **35**, 38274-38290 (2022)
2. W.X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du. A survey of large language models. *arXiv preprint arXiv:2303.18223.* (2023)
3. W. Hariri. ArXiv.2304.02017 (2023)
4. J.J. Zhu, J. Jiang, M. Yang, Z.J. Ren. Environ. Sci. Technol. (2023)
5. E. Agathokleous, C.J. Saitanis, C. Fang, Z. Yu, Sci. Total Environ. **888**, 164154 (2023)
6. S.S. Biswas, Ann. Biomed. Eng. **51**, 1126–1127 (2023)
7. N. Webersinke, M. Kraus, J.A. Bingler, M. Leippold. ArXiv.2110.12010 (2021)
8. J.A. Bingler, M. Kraus, M. Leippold, N. Webersinke. Finance Res. Lett. **47**, 102776 (2022)
9. D.A. Nisanci. *World Scientific Encyclopedia of Climate Change: Case Studies of Climate Risk, Action, and Opportunity* **3**, 3-8 (2021)
10. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, arXiv.1910.03771 (2019)
11. J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, D.C. Schmidt, ArXiv. 2302.11382 (2023)
12. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q.V. Le, D. Zhou, Adv. Neural Inf. Process. Syst. **35**, (2022)
13. Z. Ling, Y. Fang, X. Li, Z. Huang, M. Lee, R. Memisevic, H. Su. ArXiv.2306.03872 (2023)
14. J. Hao, T.K. Ho. J. Educ. Behav. Stat. (2019)
15. M. Hossin, M.N. Sulaiman. Int. J. data Min. Knowl. Manag. Process, **5**(2), 1 (2015)
16. J. Lever, Nat. methods, **13**(8), 603-605 (2016)