Climate PAL: Climate Analysis through Conversational AI

Sonia Cromp
University of Wisconsin-Madison
sonic@cs.wisc.edu

Behrad Rabiei
University of California San Diego
brabiei@ucsd.edu

Maxwell Elling
University of Colorado Boulder
max.elling@colorado.edu

Alexander Herron
NASA Goddard Institute for Space Studies
alexander.herron@nasa.gov

Michael Hendrickson
NASA Goddard Institute for Space Studies
michael.hendrickson@nasa.gov

Abstract

To support climate change research and its communication to the public, we propose ***Climate Projection and Analysis with Language models (Climate PAL)***. Our system allows users to retrieve and analyze climate projection data through conversational English. Using a crowdsourced evaluation dataset, we demonstrate that Climate PAL's retrieved data are more relevant to user queries, with **over 20% higher accuracy** than baselines on several key metrics.

1 Introduction

Climate change research relies on staggering quantities of data. A prominent example is the 30-petabyte *Coupled Model Intercomparison Project (CMIP) 6*, which contains over 13,600,000 climate projection datasets [8]. Working with CMIP6 requires a range of technical knowledge, such as using specialized programming packages or understanding esoteric terms and abbreviations. These requirements pose challenges for less-experienced researchers, impede experts' ability to quickly evaluate new hypotheses, and deter non-technical stakeholders from engaging with climate data.

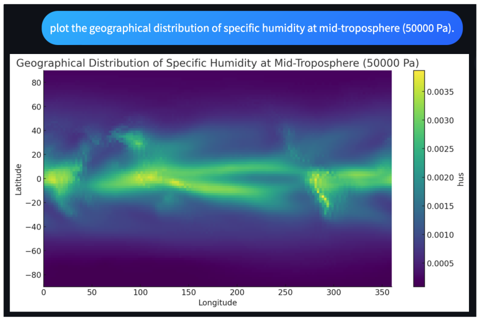
plot the geographical distribution of specific humidity at mid-troposphere (50000 Pa).

Figure 1: Climate PAL allows users to engage with CMIP6 climate data via conversational English through an intuitive graphical interface.

To address these issues, we present ***Climate Projection and Analysis with Language models (Climate PAL)***, a Large Language Model-based system to retrieve and analyze 343,119 CMIP6 datasets generated by NASA's Goddard Institute for Space Studies (GISS) using conversational English. Climate PAL will assist researchers and increase the accessibility of climate insights to the general public.

Climate PAL is, to our knowledge, the first general-use conversational system for retrieval and analysis of CMIP6 climate data. The interface is designed to be simple, intuitive and easily adaptable to a variety of device resources and screen sizes. Figure 1 demonstrates an interac-

Tackling Climate Change with Machine Learning workshop at NeurIPS 2024

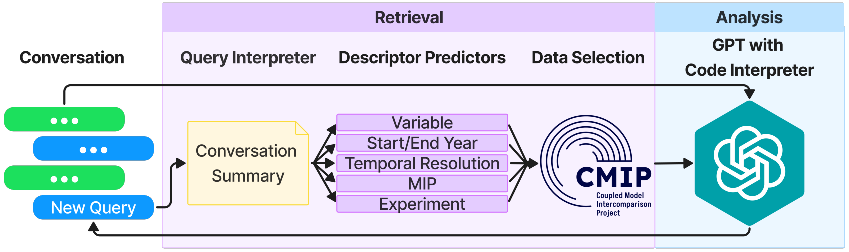

Figure 2: An overview of Climate PAL. The Retrieval Component is tasked with selecting datasets to provide to the Analysis Component, based on the user query and conversational history.

tion with Climate PAL. By relying on In-Context Learning (ICL) techniques [17], our system requires no fine-tuning and will allow for easy incorporation of new CMIP datasets as they are released. We summarize our contributions as follows:

- We propose Climate PAL for retrieval and analysis of CMIP6 datasets using conversational English.
- We crowdsource a dataset of CMIP6 analysis queries to evaluate Climate PAL and future systems.
- We demonstrate retrieved datasets are $> 20\%$ more accurate than baselines in several metrics.

2 Method

To motivate the design of Climate PAL, we provide additional information about the structure of GISS CMIP6. Next, we discuss each component of Climate PAL as summarized in Figure 2. Further details on Climate PAL's architecture are in Appendix A.

GISS CMIP6[1] contains the evaluation data and outputs of six climate models. Its 343,119 datasets simulate more than 400 different variables over 90,000 years of the Earth's past and future climate. CMIP6's modeling tasks are referred to as Model Intercomparison Projects (MIPs), each of which contain different sub-tasks, called Experiments.

Each CMIP6 dataset is described by a standardized set of attributes, which we refer to as *descriptors*. These descriptors include *Variable* (the dependent variable measured), *Start* and *End Year* (the range of years covered) and *Temporal Resolution* (whether Variable is measured hourly, monthly or yearly between Start and End Year), along with the *MIP* and *Experiment* to which the dataset belongs.

Climate PAL's Retrieval Component is responsible for selecting a dataset best-suited to answering the user query. Each time the user replies in a conversation, a GPT model [4, 18] is prompted to summarize the query and any conversational history into a few keywords, then determine if it is necessary to retrieve a new dataset to answer the query. If so, the retrieval component first builds a profile of an ideal dataset by predicting each of the dataset's descriptors. Next, a table of each GISS CMIP6 dataset is filtered to find the best match to the predicted descriptors.

Apart from Variable, each descriptor is predicted by prompting GPT with the conversational summary, plus information such as which descriptor to predict and the descriptor's set of possible values. Variable, however, is more challenging: there are 419 unique Variables in GISS CMIP6, each with a precise, technical definition. We use a three-step technique to predict Variable by further summarizing the conversation, then performing an embeddings-based search to find the top ten closest matches to the summary and providing this shortlist in an ICL GPT prompt similar to the other descriptors'.

The descriptor predictions form the profile of an ideal dataset for retrieval, but a dataset with this specific combination of descriptors may not exist in CMIP6. We filter all datasets by each descriptor's prediction sequentially, skipping descriptors that cannot be satisfied due to previous descriptor values.

The Analysis Component instantiates an OpenAI "Assistant" GPT model with the proprietary Code Interpreter Tool [1], allowing GPT to execute code for tasks such as data visualization and mathematical computations. This GPT agent is prompted with the full conversational history and all retrieved datasets, allowing the model to generate an informed response to the conversation using Retrieval-Augmented Generation (RAG) techniques [13].

[1] https://portal.nccs.nasa.gov/datashare/giss_cmip6/

CMIP6 datasets are stored in the specialized geospatial NetCDF format [19], which GPT cannot natively interpret. We use a specialized ICL prompt instructing GPT to install xarray [11], the Python library for interacting with NetCDF files.

3 Evaluation

We describe our crowdsourced evaluation dataset, then present our experimental setup.

Evaluation Dataset: We crowdsource a set of 35 GISS CMIP6-related queries from NASA scientists. Then, we manually annotate each query with the set of descriptor values necessary to the retrieved dataset. We perform semantic variation to augment our evaluation dataset to a full set of 210 queries. For further details and example queries, see Appendix B. We are in the process of gaining the rights to release our evaluation dataset publicly.

Setup: We evaluate Climate PAL retrievals in two phases. As described in Section 2, we first predict the values of each descriptor. We begin by assessing the accuracy of these predictions in Section 3.1. Next, in Section 3.2, we discuss the accuracy of the dataset that is ultimately retrieved. In all experiments, we include Climate PAL with both GPT-4o and GPT-3.5, which we refer to as Climate PAL-4o and Climate PAL-3.5. Furthermore, all experimental results are averaged across three runs for any non-deterministic approach. We focus here on the evaluation of dataset retrievals, with plans to evaluate the analysis component outlined in Appendix C.

3.1 Descriptor Prediction

Baselines: We compare Climate PAL's Variable predictions against two baselines: one embeddings-based and the other a simple Retrieval-Augmented Generation (RAG) pipeline [13] with GPT-3.5. The remaining descriptors use keyword and regular expression-based baselines. For instance, the Temporal Resolution baseline predicts "hour" if the query text contains "hr" or "hour". Refer to Appendix C for more details.

Results: In Variable prediction, *Climate PAL-4o is* 29.7% *more accurate on average than the best-performing baseline*. The RAG baseline using GPT-3.5 without Climate PAL reaches only 8.6% accuracy. Though RAG-3.5 and Climate PAL-3.5 rely on the same GPT-3.5 model, the very *low performance of RAG-3.5 demonstrates the importance of Climate PAL's Variable selection approach*.

Climate PAL-4o exceeds the baseline by significant margins on all descriptors except Experiment, while Climate PAL-3.5 achieves the second-best performance for all descriptors. The keyword baselines perform worse than uniform random guessing on Temporal Resolution, which takes four unique values, and MIP, which takes three unique values. Results are summarized in Table 1.

3.2 Dataset Selection

Baselines: We implement three baselines for this task.

Two baselines are combinations of the descriptor prediction baselines introduced in Section 3.1, along with Climate PAL's process of using these predictions to select the retrieved dataset. The first baseline, called E+K, uses the embedding approach to predict the Variable. The second, called 3.5+K, predicts Variable using the GPT-3.5 RAG baseline. These baselines both rely on the keyword-based approaches to predicting Start/End Year, Temporal Resolution, MIP and Experiment.

The third baseline, called 3.5, is a single-step RAG approach. We provide GPT-3.5 with a table of all 343,119 GISS CMIP6 datasets and prompt the model to choose a dataset appropriate to the query.

Results: Table 2 compares the accuracy of the dataset selection methods. *Climate PAL-4o and Climate PAL-3.5 are more accurate in the Variable descriptor than all baselines, by large margins.*

Start and End Year see lower performance. The 3.5 baseline achieves the highest accuracy, at 62.7%. While this baseline chooses its dataset in one step, the other methods are constrained by their choice of Variable before attempting to select their predicted Start Year. When the predicted combination of Variable and Start Year does not exist in the GISS CMIP6 datasets, Climate PAL, 3.5+K and E+K opt for their predicted Variable instead of their predicted Start Year. We refer to this effect of degraded performance due to constraints from prior descriptors as *prior descriptor limitation*.

Table 1: Descriptor prediction accuracy. Climate PAL (denoted CP) outperforms baselines on all descriptors except Experiment. Baselines are keyword-based (Keywords) except for Variable, which has an embeddings-based (Embeds) and RAG (3.5) baselines.

	Variable
CP 4o (Ours)	$\mathbf{66.7 \pm 3.3}$
CP 3.5 (Ours)	62.9 ± 4.9
Embeds	36.2 ± 1.6
3.5	08.6 ± 0.0

	Year	Temporal Resolution	MIP	Experiment
CP 4o (Ours)	$\mathbf{94.6 \pm 0.1}$	$\mathbf{83.8 \pm 0.8}$	$\mathbf{88.6 \pm 1.3}$	57.1 ± 1.7
CP 3.5 (Ours)	86.0 ± 0.5	68.3 ± 0.5	57.8 ± 0.3	62.1 ± 0.5
Keywords	64.3	12.9	14.3	$\mathbf{63.3}$

Table 2: Accuracy of retrieved dataset. Climate PAL (denoted CP) achieves top performance on 4/6 descriptors. Baselines are, in order: keywords with embedding-based or RAG Variable prediction and a single-step RAG dataset selection approach.

	Variable	Start Year	End Year	Temporal Resolution	MIP	Experiment
CP 4o (Ours)	$\mathbf{65.4 \pm 1.0}$	26.7 ± 0.5	30.6 ± 0.7	80.0 ± 0.5	$\mathbf{94.0 \pm 0.3}$	77.9 ± 0.7
CP 3.5 (Ours)	61.0 ± 0.5	31.9 ± 1.0	$\mathbf{38.6 \pm 0.0}$	67.0 ± 0.3	91.7 ± 0.3	$\mathbf{86.0 \pm 0.7}$
E+K	8.6 ± 0.0	0.0 ± 0.0	5.7 ± 0.0	77.1 ± 0.0	42.9 ± 0.0	40.0 ± 0.0
3.5+K	35.7 ± 0.5	16.0 ± 0.3	35.1 ± 0.3	$\mathbf{81.3 \pm 0.3}$	74.6 ± 0.3	67.9 ± 0.3
3.5	11.4 ± 0.0	$\mathbf{62.7 \pm 0.3}$	37.0 ± 0.3	32.9 ± 0.0	85.6 ± 0.3	69.5 ± 0.0

Despite its freedom from prior descriptor limitation, the 3.5 baseline struggles at selecting relevant datasets. In fact, this method sees the second-lowest accuracy for the Variable descriptor, and is *outperformed by Climate PAL-3.5 or Climate PAL-4o in all descriptors but Start Year*.

We see the effect of prior descriptor limitation even more clearly in Temporal Resolution. As presented in Section 3.1, the keyword baseline for predicting Temporal Resolution achieved only 12.9% accuracy (worse than random guessing), versus 83.8% and 68.3% for Climate PAL-4o and Climate PAL-3.5. Despite the low performance of the keyword Temporal Resolution predictions, which are used identically by both E+K and 3.5+K, we see that datasets selected by E+K and 3.5+K perform similarly to datasets selected by Climate PAL on the accuracy of Temporal Resolution.

Due to the prior descriptor limitation effect, we find a need to adjust assessments of Climate PAL, E+K and 3.5+K's retrieved datasets' descriptors on the basis of how limited each method is by its prior descriptor choices. The design of a clearer metric for evaluating the retrieved datasets is a priority for future research. Despite these challenges, *the competitive performance of Climate PAL is demonstrated in its high Variable accuracy, along with its highest or near-highest accuracy on four of the five other descriptors*: End Year, Temporal Resolution, MIP and Experiment.

4 Related Work

Information retrieval has been a focus of AI for decades [20], with several recent approaches relying on Large Language Models (LLMs) [15, 21, 9]. RAG is a related task, where the LLM accesses an external knowledge base to better-inform its outputs [13, 14]. However, many of these works are focused on retrievals from natural-language datasets, as opposed to specialized modalities such as geospatial data. Two exceptions are Chen et al. [6] and Zhang et al. [22], which are for road design and satellite control respectively. Unfortunately, such works are highly fitted to their specific applications and offer no clear method for adaptation to CMIP6.

LLM researchers have given much attention lately to the task of automated data analysis [16, 10], with the GPT Data Analysis tool [2] particularly relevant to these efforts. Similar to the problem of RAG, however, these works have primarily focused on domains such as natural language and pure mathematics. In-Context Learning (ICL) is also a popular topic of LLM research [17, 7], allowing

LLMs to perform novel tasks by following instructions in a text prompt instead of undergoing further training/finetuning.

5 Conclusion

We present Climate PAL, a system for the retrieval and analysis of CMIP6 data using conversational English. We hope Climate PAL will be useful for accelerating climate research, and providing greater exposure of this field to the general public.

Acknowledgements

We are grateful to the NASA Office of STEM Engagement (OSTEM) for their support of Sonia Cromp and Behrad Rabiei, who completed this work during internships at the NASA Goddard Institute for Space Studies. We additionally thank Ellen Salmon and the NASA Center for Climate Simulation staff for hosting and providing convenient access to the GISS CMIP6 model output[2].

References

[1] Openai platform. URL `https://platform.openai.com`.

[2] Improvements to data analysis in chatgpt. URL `https://openai.com/index/improvements-to-data-analysis-in-chatgpt/`.

[3] New embedding models and api updates. URL `https://openai.com/index/new-embedding-models-and-api-updates/`.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[5] I. Beltagy, K. Lo, and A. Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[6] J. Chen, W. Xu, H. Cao, Z. Xu, Y. Zhang, Z. Zhang, and S. Zhang. Multimodal road network generation based on large language model. *arXiv preprint arXiv:2404.06227*, 2024.

[7] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[8] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[9] J. Gavilanes, Y. Bozhilov, U. Dodeja, G. Valtas, and A. Badrajan. Use of llm for methods of information retrieval.

[10] S. Hong, Y. Lin, B. Liu, B. Wu, D. Li, J. Chen, J. Zhang, J. Wang, L. Zhang, M. Zhuge, et al. Data interpreter: An llm agent for data science. *arXiv preprint arXiv:2402.18679*, 2024.

[11] S. Hoyer and J. Hamman. xarray: Nd labeled arrays and datasets in python. *Journal of Open Research Software*, 5(1):10–10, 2017.

[12] M. Khorasani, M. Abdou, and J. H. Fernández. Web application development with streamlit. *Software Development*, pages 498–507, 2022.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[14] H. Li, Y. Su, D. Cai, Y. Wang, and L. Liu. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*, 2022.

[2] https://portal.nccs.nasa.gov/datashare/giss_cmip6/

[15] Z. Liu, Y. Zhou, Y. Zhu, J. Lian, C. Li, Z. Dou, D. Lian, and J.-Y. Nie. Information retrieval meets large language models. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1586–1589, 2024.

[16] P. Ma, R. Ding, S. Wang, S. Han, and D. Zhang. Demonstration of insightpilot: An llm-empowered automated data exploration system. *arXiv preprint arXiv:2304.00477*, 2023.

[17] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[18] OpenAI. Introducing chatgpt. URL `https://openai.com/index/chatgpt/`.

[19] R. Rew and G. Davis. Netcdf: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4):76–82, 1990.

[20] A. Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4): 35–43, 2001.

[21] C. Zhai. Large language models and future of information retrieval: Opportunities and challenges. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 481–490, 2024.

[22] R. Zhang, H. Du, Y. Liu, D. Niyato, J. Kang, Z. Xiong, A. Jamalipour, and D. I. Kim. Interactive generative ai agents for satellite networks through a mixture of experts transmission. *arXiv preprint arXiv:2404.09134*, 2024.

A Further Methodological Details

A.1 Retrieval Component

We provide additional information on the design of each descriptor predictor.

Variable: The Variable predictor selects the most relevant Variable from a set of 419 possible values. Examples of possible Variables include “tasmax", the maximum air temperature at the Earth’s surface, and “sithick", the average thickness of sea ice. Climate PAL has access to a short, natural-language description of each variable. These descriptions are publicly available online[3].

To select a Variable, Climate PAL employs a three-step process:

1. Although Climate PAL initially creates a general-purpose summary of the user’s query, this summary might contain analysis-related words like “plot", or words relevant to other descriptors, such as year ranges. To help the Variable predictor focus on Variable-relevant information, we prompt GPT to write a description specifically of a CMIP6 Variable relevant to the conversational summary.
2. Each Variable’s description and the description produced in step (1) are embedded using OpenAI’s text-3-embedding-large model [3]. We determine the set of 10 Variables with descriptions of smallest cosine distance to the step (1) description.
3. This shortlist is provided in a second call to GPT, along with the original user query and an ICL prompt to choose the Variable from the list that is most relevant to answering the query.

Other descriptors: All of the remaining descriptors follow a similar format to each other: a GPT model is provided with the conversational summary, as well as an ICL prompt. Each of these ICL prompts is listed in Table 3.

A.2 Analysis

In order to create a GPT Assistant using the OpenAI Assistant API, we must provide a client and our retrieved CMIP6 datasets. The client handles HTTP requests, manages API authentication, processes user input, and integrates responses into the application. The client should manage conversation context, handle errors, and allow customization of API parameters.

With all necessary sub-components set up, each time a user provides a new analysis query, the Analysis Component will append the query to the conversational history and feed this history into the Assistant. The Assistant then generates code to perform analysis using the Code Interpreter Tool and displays the results to the user through our custom graphical interface.

This process repeats until the user is satisfied or a new dataset must be retrieved. Each time that a new dataset is added to the conversation, a dataset summary is created and displayed to the user, including information such as the dataset name, size and features.

A.3 Graphical User Interface

To abstract away the underlying components of Climate PAL, we have created a custom user interface using the Streamlit library [12], a Python library that allows developers to quickly create and share custom web apps. The interface mimics a conversational text message format. See Figure 1 for an example of the interface.

[3] https://github.com/PCMDI/cmip6-cmor-tables/tree/main

Table 3: The ICL prompts used by Climate PAL's Retrieval Component, edited slightly for brevity.

Descriptor	ICL Prompt
Variable (1)	You are a climate scientist and expert on CMIP6. Given a colleague's query, describe what CMIP6 variable you would use to answer the query. For instance, you might want a rainfall-related variable for questions about drought. For a query about days below freezing, you might want a variable describing minimum temperature instead of average temperature. Formulate your response as a detailed list of keywords. Be specific because a lot of CMIP6 variables are very similar but there is only one correct answer to these queries.
Variable (2)	RETURN A ONE-WORD RESPONSE: You are an expert climate scientist working with the CMIP6. Following, is a colleague's climate analysis query and a list of 10 CMIP6 variables with their descriptions. From these 10 variables, choose the variable best-suited to answer the colleague's query. Return ONLY the variable's name and nothing else. For instance, return 'tas', or 'zostoga', or 'sithick' alone, no explanation, no alternative answer, nothing else.
Start/End Year	You are an expert climate scientist. Does the following CMIP6 query require or specify a year range for the data required to answer the query? If yes, provide the year range in format START-END, for instance 1960-1970 or 2100-3100. If no, respond NA-NA. If only the start or end is specified, provide just that year in format START-NA (eg 2100-NA) or NA-END (eg NA-1900). Provide only the year range in this format and nothing else.
Temporal Resolution	You are an expert climate scientist. Is the following CMIP6-related query best answered using data gathered at which of the following resolutions? A. hour B. day C. month D. not applicable, none of the above, or unclear Respond with only the one letter corresponding to your choice and nothing else. If a query does not specify any given temporal resolution, like the query "plot average temperature", then choose option D.
MIP	You are an expert climate scientist working with CMIP6. Here is a list of the MIPs you work with: CMIP, ScenarioMIP To answer the following query, which of the above experiments would you use? Return JUST the name of the experiment and nothing else. If the choice of experimentdoes not matter, return 'None'.
Experiment	You are an expert climate scientist working with CMIP6. Here is a list of the experiments you work with: To answer the following query, which of the above experiments would you use? Return JUST the name of the experiment and nothing else. If the choice of experimentdoes not matter, return 'None'.

B Evaluation Dataset

We provide a small sample of our crowdsourced evaluation dataset in Table 4. Each query is a standalone question to be answered by Climate PAL, rather than a multi-turn conversation that must first be summarized. These single-turn queries are easier to crowdsource from volunteers than multi-turn conversations. Despite this, we still pass each query through the retrieval component's conversational summarization step in order to shorten the query and allow for greater similarity to Climate PAL's intended use-case of multi-turn conversations.

After manually annotating each of the 35 queries, we perform semantic variation to augment our evaluation dataset. For each query, we ask GPT-4o to rephrase the query five different ways for a total of set of 210 queries (35 original plus 175 augmented). As such, the manual annotations for each original crowdsourced query can be used for the semantic variation queries as well.

We use this semantic variation method to augment our dataset because of its simplicity, and also because of the quality of the generated queries as measured by cosine distance. We refer to the original, crowdsourced queries as “parents" and their rephrased queries as “children". As demonstrated by the plots of SciBERT [5] and OpenAI embedding cosine distances between each child query and its parent in Figures 3 and 4, the augmented queries have small– but non-zero– distances to their parent query. As a result, the augmented queries tend to have similar meanings to the crowdsourced queries, without being identical.

To help foster future research in this area and a culture of reproducibility, we are currently in the process of gaining the rights to release our evaluation dataset publicly.

Table 4: Example queries in the evaluation dataset.

Query	Variable	Start Year	End Year	Temporal Resolution	MIP	Experiment
Are there going to be increased heatwaves in South America under SSP370 for 2085?	tasmax	2085	2085	day	ScenarioMIP	ssp370
Show me in the future, all the suitable places that wheat could grow	clt, pr, etc	2025			ScenarioMIP	
Show me the expected average winter ice coverage for Lake Ontario is 2050?	sblIs, sftgif	2050	2050	month	ScenarioMIP	
Plot the change in cloud cover from 1930 to 2015	clt	1930	2015	month	CMIP	historical
What are the projected changes in global ocean salinity by 2050 under SSP126?	so	2025	2050	month	ScenarioMIP	ssp126

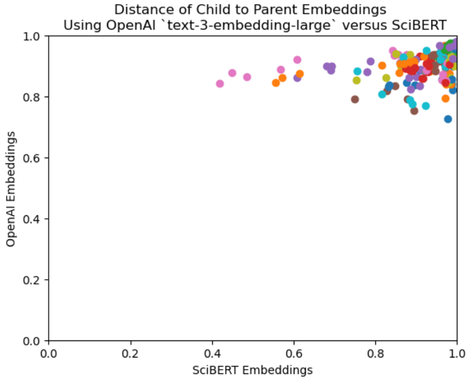

Figure 3: A comparison of child queries' embedding cosine distances to their parents' embeddings, using SciBERT [5] or OpenAI's text-3-embedding-large model. Each color corresponds to one parent query. We observe that distances produced by the OpenAI embeddings are generally closer to 1 than the SciBERT embedding distances.

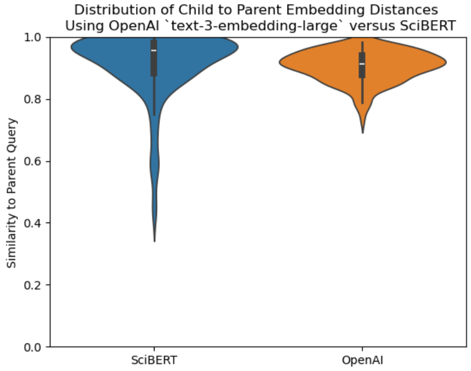

Figure 4: The distributions of cosine distances between child and parent queries, using SciBERT or OpenAI's text-3-embedding-large model. The range of distances for OpenAI (right) are closer to 1, despite a lower *average* distance than SciBERT embeddings. The SciBERT distribution exhibits a longer tail of lower-similarity embeddings as well.

C Evaluation

C.1 Retrieval Evaluation

For the Variable, Temporal Resolution, MIP and Experiment descriptors, accuracy is calculated as the percentage of queries for which the descriptor prediction or the retrieved dataset matches the descriptor's gold label. As the Variable descriptor may be annotated to have multiple acceptable values on a given query, a prediction or retrieved dataset is considered to be accurate if the Variable equals to any of the manually-annotated Variables. For the Year descriptor, 50% accuracy is awarded for a correct start- *or* end-year prediction, while 100% accuracy is awarded for correct start- *and* end-year predictions.

We implement two Variable prediction baselines for the descriptor prediction evaluation in Section 3.1. The first is embedding-based: we use OpenAI's text-3-embedding-large model to embed the natural-language descriptions of the 419 Variables, as well as the conversational summary for each evaluation query. A variable with an embedded description of minimum cosine distance to the embedded conversational summary is chosen as this baseline's description. Our second Variable baseline is a simple RAG pipeline: we provide GPT 3.5 with a table of all 419 variables and their natural-language descriptions, along with an ICL prompt to return the variable best-suited to answer the evaluation query.

C.2 Analysis Evaluation

To evaluate the effectiveness of the analysis component in Climate PAL, we will conduct a series of assessments. These evaluations are designed to measure the accuracy and usability of our method across different contexts. Please note that we are still in the process of conducting these evaluations and currently do not have any results to report.

C.2.1 Variable Identification

The first evaluation focuses on the component's ability to correctly identify relevant descriptors from a retrieved dataset based on a given analysis query. To achieve this, we will curate a dataset comprising queries paired with the correct descriptors as ground truth. The performance of our method will be compared across GPT-4o, GPT-3.5, and the standard ChatGPT interface [1]. We will use Accuracy, Precision and Recall as performance metrics, defined as follows:

- **Accuracy** is the proportion of correct variable identifications out of all identifications made by the system:
$$Accuracy = \frac{nCorrectIdentifications}{nIdentifications}$$
- **Precision** is the proportion of correct variable identifications out of all identifications predicted as relevant:
$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$
- **Recall** is the proportion of correct variable identifications out of all actual relevant variable identifications:
$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

C.2.2 Temporal Window and Resolution Identification

The second evaluation examines the component's ability to accurately identify the Start/End Years and Temporal Resolution specified by the user in an analysis query. We will create a dataset that includes queries with the correct temporal window and resolution, similar to the Variable Identification evaluation. The same model variations (GPT-4o, GPT-3.5, and the ChatGPT interface) will be tested. Accuracy, Precision, and Recall will also be used as metrics for this evaluation.

C.2.3 Plot Generation Capability

This evaluation measures the component's capability to generate a plot when appropriate, regardless of the plot's correctness. We will compile a dataset containing queries and corresponding boolean values indicating whether a plot is needed. The same model variations as in previous evaluations will be tested. Performance will be evaluated using Accuracy, defined as:

$$Accuracy = \frac{\sum_{i=1}^{N} \mathbb{1}[P_i = R_i]}{N}$$

where:

- N is the total number of queries in the dataset,
- P_i is the system's prediction for query i (1 if a plot is generated, 0 otherwise) and
- R_i is the ground truth for query i (1 if a plot is required, 0 otherwise).

C.2.4 User Satisfaction

The fourth evaluation is a user study. We will assess user satisfaction with Climate PAL's responses to a fixed set of n queries for U participants. This evaluation will involve a diverse group of users, from novices to experts, who will use the system and provide satisfaction ratings on a scale from 1 to 5. The metric for this evaluation will be the average satisfaction score, calculated as:

$$Average\_Score = \frac{1}{n}\sum_{q=1}^{n}\frac{\sum_{u=1}^{U} satisfaction(q,u)}{U},$$

where n is the total number of queries and $satisfaction(q,u)$ is the satisfaction of the u-th user on Climate PAL's response to the q-th query.