DisasterGPT: Democratizing Local Relevant Climate Disaster Information

Edrian Liao | ECE 495: Intro to Natural Language Processing Capstone

—

Abstract

To address the distribution of climate disaster information, this capstone project investigates the creation of a Retrieval-Augmented Generation (RAG) system that combines a GPT-2-based generator with a BERT-based retriever. Using semantic embeddings produced by Sentence-BERT and indexed with FAISS, the system extracts pertinent textual data from a well-curated archive of climate and environmental materials. GPT-2 uses the context provided by these retrieved chunks to produce responses that are both believable and instructive. Although there aren't many question-and-answer datasets specifically related to climate disasters, this work makes use of pre-existing datasets such as Climate Fever, SQuAD 2.0, and `environment_data`, which have been reorganized into context-question-answer formats, to successfully train the GPT-2 model.

The results show that while data limits hinder GPT-2's ability to provide highly coherent replies, performance can be enhanced by adding more data and fine-tuning model configurations. In particular, the performance of a 2-layer GPT-2 model with enhanced beam search was noticeably better than that of its 1-layer equivalent. However, problems still exist, such as verbosity, inattention, and the trade-off between sustainability and model complexity. This study emphasizes how crucial it is to select domain-specific datasets, strike a balance between model performance and energy efficiency, and encourage developing regions' access to cutting-edge technologies. To improve the efficiency of methods for disseminating climate information, future research should concentrate on growing high-quality datasets and investigating more sophisticated yet sustainable models.

Introduction

In this capstone project, I aim to explore the application of a **GPT-based model** on climate and environmental policy documents. The potential of such a tool lies in its ability to make critical and urgent climate information more accessible and understandable for individuals across the globe. While international and local policymakers, academics, and key stakeholders in climate mitigation and adaptation rely heavily on these insights, the importance of such knowledge extends to civilians as well. Informed individuals are better equipped to understand the risks posed by climate change and the measures necessary to mitigate its effects.

Timely and accurate dissemination of climate information can raise awareness and, in many cases, save lives. Coming from a country frequently impacted by typhoons, I have personally witnessed the devastating consequences of inadequate scientific communication. For example, the term **"storm surge"**—a temporary and significant rise in water levels resulting in unexpected flooding of coastal areas—was not widely understood by residents, leading to preventable casualties. Tools such as the one proposed in this project have the potential to bridge this gap in **science communication**, particularly in the context of climate-related disasters like typhoons, cyclones, and hurricanes.

Several recent projects have demonstrated impressive performance in democratizing access to climate knowledge. [1], [2], [3] Inspired by these advancements, my goal is to develop a proof of concept: a streamlined and relatively lightweight GPT-based system. While sophisticated large models have shown groundbreaking capabilities, their complexity often limits accessibility for individuals or communities with limited computational resources. I aim to address this challenge by focusing on building a simpler, easier-to-train, and faster-to-use model that can serve as an effective communication tool for climate-related information.

Due to the time constraints of this project, I will limit my scope to documents that cover the following topics:

- **The physics behind climate-related disasters** and how climate change exacerbates them.
- **Risk mitigation measures** to prevent or reduce disaster impacts.
- **Long-term adaptation planning** for communities vulnerable to climate disasters.
- **Emergency relief operations** for disaster-affected areas.

The curated **DisasterGPT Database** for this project contains **seven documents** that focus on these critical topics, including **climate adaptation planning** and disaster risk mitigation for extreme events such as heatwaves, flooding, drought, tropical cyclones, typhoons, and hurricanes.

Why is a RAG model better suited for this task?

The long-term objective of building a GPT model capable of supplying accurate and up-to-date information about climate disasters requires a timeline that extends beyond the duration of this course. Consequently, the scope of this final project must be further refined to align with the essential goals outlined in the submission rubric. To reiterate, the project prompts us to build a retrieval-augmented GPT (RAG) model capable of generating contextually relevant responses based on a curated database. This baseline goal serves as a foundation for demonstrating the feasibility and utility of such models in improving access to critical climate information.

In the context of this project, this means that the BERT model will fetch the most relevant chunks of information from the database that's also been embedded by the said model. The responses will then be fed to the GPT model as context and will hopefully spit out responses that are more plausible to humans and domain experts curating the information found on the database. This is what is referred to as the RAG model that has two main components: retrieval and generation. RAG systems are particularly useful in domains requiring evidence-based outputs, such as question-answering, summarization, and knowledge-driven tasks in fields like climate science, law, and healthcare. By leveraging external data dynamically without re-training the model, RAG systems improve scalability, adaptability, and performance, making them an essential tool for applications demanding reliable and domain-specific information generation. [4]

How does the BERT model retrieve relevant arguments from the database using the input query?

The BERT model retrieves relevant arguments from a database using an input query through two main steps: semantic encoding and similarity search. First, the query and all candidate arguments in the database are passed through the BERT model, which converts them into vector embeddings—numerical representations that capture the deeper semantic meaning of the text. These embeddings ensure that texts with similar meanings are positioned close to each other in a high-dimensional space. Once the query is encoded, its embedding is compared to the precomputed embeddings of the arguments in the database using cosine similarity or Euclidean (L2) distance. Tools like FAISS make this comparison efficient by ranking the arguments based on their proximity to the query in the vector space. The system then retrieves the top-k most relevant arguments as results. What makes BERT effective for this task is its ability to capture context, going beyond simple keyword matching to find arguments that are truly relevant based on meaning. This approach is particularly powerful when dealing with large databases or complex queries where traditional methods like TF-IDF or BM25 often fall short. [5]

Project Questions

In this project, we focused on the feasibility of realizing the RAG system using less complex models. Although counterintuitive, I wanted to stress the importance of energy conservation when training and deploying large language models. Estimates suggest that ChatGPT generates approximately 8.4 tons of carbon dioxide annually, more than double the 4 tons emitted by an average individual each year. This is apart from the 700,000 liters of freshwater used in training centers to stabilize the temperature within their surroundings. [6]

As such, one of the burning questions is whether a GPT-2 model is enough to capture the complexity of concepts within the climate space. In our previous homework, the model has consistently worked well on the TinyStories dataset with a sufficient amount of data, a number of embeddings, layers, and heads. However, it is obvious that the interplay of words within simple narratives is no way as complex as that in environmental policy reports.

Now the secondary questions are as follows:

- If that is not enough, what are techniques we can use both on the data and on the model to make the output produce more plausible responses?
- What do machine learning practitioners need to do to minimize carbon emissions while successfully completing the task of building a RAG system for information on climate disasters?
- What do the results mean to broader societal topics such as support of NLP research trained and conditioned on environmental data?

Experimental Setup

<u>Extracting Chunks of Information from the Document</u>

The first step in getting this project started is to extract the text from the PDF files. This code extracts text from specific page ranges of multiple PDF documents using the PyMuPDF library (`fitz`). The function `extract_text_from_pdf` takes a file path and a page range as input, ensuring the page numbers are converted to zero-based indexing and remain within valid bounds. It loops through the specified pages, extracting text using `get_text()` and handling any errors that might occur during extraction. The script iterates over a list of document paths and corresponding page ranges, printing progress messages for each document as it processes the pages. The extracted text is appended to a list called `text_data`, which stores the text for all the documents.

```
import fitz  # PyMuPDF

def extract_text_from_pdf(file_path, start, end):
   text = ""
   with fitz.open(file_path) as pdf:
```

```
        # Ensure the page range is within bounds
        start_index = max(start - 1, 0)  # Convert to zero-based index
        end_index = min(end, len(pdf))  # Ensure end is within bounds

        for page_num in range(start_index, end_index):
            try:
                text += pdf[page_num].get_text()
            except Exception as e:
                print(f"Error reading page {page_num+1} in {file_path}:
{e}")
    return text

# Paths to documents and page ranges
document_paths = ["doc1.pdf", "doc2.pdf", "doc3.pdf", "doc4.pdf",
"doc5.pdf", "doc6.pdf", "doc7.pdf"]
page_range = [[6, 87],
              [4, 116],
              [1, 15],
              [1, 38],
              [4, 78],
              [8, 51],
              [3, 82]]

# Extract text from the specified ranges
text_data = []
for i, doc in enumerate(document_paths):
    print(f"Extracting text from {doc} (pages {page_range[i][0]} to
{page_range[i][1]})")
    text = extract_text_from_pdf(doc, page_range[i][0], page_range[i][1])
    text_data.append(text)

print(f"Extraction completed for {len(document_paths)} documents.")
```

Defining the page range was crucial. In the Results and Discussion part of the paper, stark differences were seen in the top 5 chunks of information pulled from a database, given a query. After the text was attracted unnecessary symbols were removed to improve clarity and better model results. The code below was used to remove such symbols and to split the text with chunk sizes of around 200. Below, a histogram depicting the range of chunk sizes was also shown.

```python
def segment_text(text, chunk_size=200):
    sentences = text.split(". ")
    chunks = []
    current_chunk = []
    for sentence in sentences:
        current_chunk.append(sentence)
        if len(" ".join(current_chunk)) > chunk_size:
            chunks.append(" ".join(current_chunk))
            current_chunk = []
    if current_chunk:  # Add the last chunk
        chunks.append(" ".join(current_chunk))
    return chunks

# Segment all documents into chunks
all_chunks = []
for doc in text_data:
    doc = doc.replace("\n", " ")
    doc = doc.replace("\t", " ")
    doc = doc.replace("\r", " ")
    all_chunks.extend(segment_text(doc))
print(f"Number of chunks: {len(all_chunks)}")
```

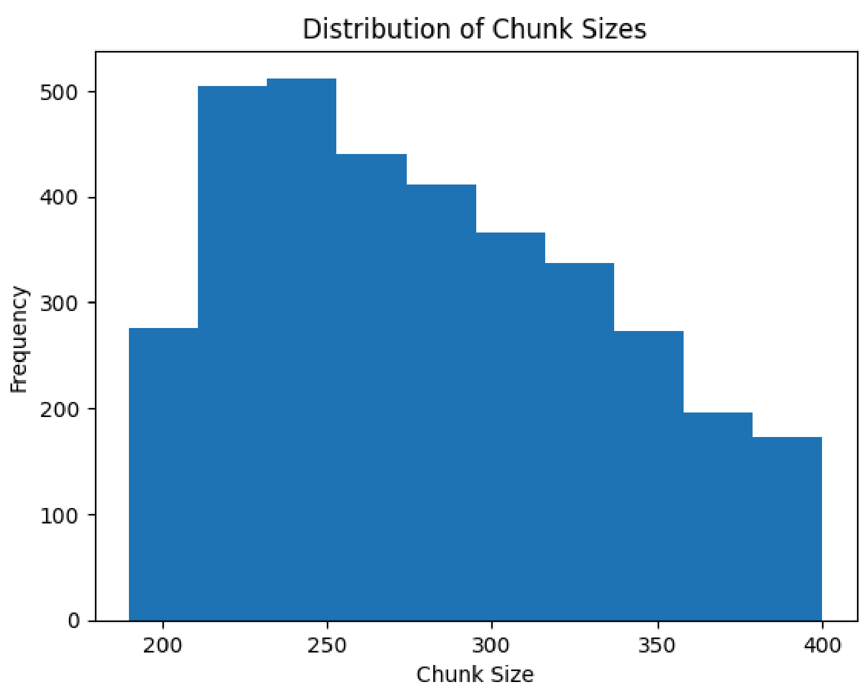

Figure 1. Distribution of chunk sizes

Retrieval System: Sentence Transformer

Because of the deeper abstractions present within scientific information, we need an augmented BERT model that not only masters the text on a word level but also on sentence and paragraph level. [7] This is precisely why we used a Sentence Transformer from the `sentence_transformers` package.

Sentence Transformer is a model built on top of BERT and other Transformer-based architectures, specifically optimized to generate semantic embeddings for entire sentences rather than individual tokens. [8] While standard BERT produces contextualized embeddings for tokens, it does not directly output sentence-level embeddings suitable for comparing sentence meanings. Sentence Transformer addresses this by fine-tuning BERT using methods such as Siamese networks or triplet networks with a contrastive learning objective. This process enables the model to produce embeddings that capture the semantic meaning of sentences, making it highly effective for tasks like semantic similarity, retrieval, and clustering. Moreover, Sentence Transformer improves efficiency by allowing pre-computation of embeddings for sentences or documents, enabling fast similarity searches using tools like FAISS. This capability makes it ideal for use cases such as semantic search, text clustering, and retrieval-augmented generation (RAG) systems, where sentence-level understanding is critical.

Below is the code used to generate the embeddings associated with the chunks of information. There were 4,267 chunks/embeddings with dimension 384.

```
from sentence_transformers import SentenceTransformer

# Load a Sentence-BERT model
model = SentenceTransformer('all-MiniLM-L6-v2')

# Embed all text chunks
chunk_embeddings = model.encode(all_chunks, show_progress_bar=True)
print(f"Chunk Embedding shape: {chunk_embeddings.shape}")

# Embed predefined section titles
sections = [
   "Adaptation finance",
   "Flooding risks",
   "Extreme heatwave and drought preparations",
   "Tropical cyclone preparedness guidelines",
   "Developing countries and small island states"
]
section_embeddings = model.encode(sections)
print(f"Section Embedding shape: {section_embeddings.shape}")
```

Then, we created a FAISS index for efficient similarity search using the L2 distance (Euclidean distance) metric.

```
combined_index = faiss.IndexFlatL2(all_embeddings.shape[1])
```

<u>Training the GPT-2 Model</u>

We used GPT-2 not only as a good baseline but also as a path towards simplifying the technologies for the developing world to use.

We have only scraped two datasets that are not explicitly structured as question-and-answer datasets. Instead, they all have commonalities of having a column for claim and another one for evidence. These datasets are described below:

The Climate Fever dataset, developed by the University of Zurich's Department of Sustainable Finance, is a benchmark resource designed for fact verification in the context of climate change [9]. It contains 5,359 claims related to climate science, policy, and impacts, each accompanied by evidence sourced from credible scientific articles, reports, and trusted publications. Claims are annotated with one of three labels: SUPPORTS, REFUTES, or NOT ENOUGH INFO, indicating whether the provided evidence validates, contradicts, or fails to verify the claim. This dataset addresses the critical need for factually accurate information in climate-related discussions and serves as an essential tool for evaluating and training natural language processing (NLP) models for tasks like fact-checking, retrieval-based question answering, and misinformation detection.

The code below was used to store the content of the dataset as something that is digestible for the GPT-2 model. This is just a horizontal concatenation of the context, the question, and the answer.

```
import json

# Load the Climate Fever dataset (assuming it's in JSONL format)
file_path = "climate-fever-dataset-r1.jsonl"
climate_data = []

# Read the JSONL file
with open(file_path, "r") as f:
   for line in f:
       climate_data.append(json.loads(line))

# Generate context and answers for each claim
contextual_data = []
for entry in climate_data:
```

```
    claim = entry["claim"]
    evidences = entry.get("evidences", [])

    # Combine evidence from supporting articles into a context
    context = " ".join([ev["evidence"] for ev in evidences if ev["evidence_label"] == "SUPPORTS"])

    # If no supporting evidence, use all available evidence
    if not context:
        context = " ".join([ev["evidence"] for ev in evidences])

    # Generate the answer (use supporting evidence as answer)
    answer = " ".join([ev["evidence"] for ev in evidences if ev["evidence_label"] == "SUPPORTS"])
    if not answer:
        answer = " ".join([ev["evidence"] for ev in evidences])  # Fallback to all evidence if no support

    # Append the claim, context, and answer to the dataset
    if context and answer:  # Ensure context and answer are not empty
        contextual_data.append({"claim": claim, "context": context, "answer": answer})

# Save the contextualized data with answers in JSON format
with open("climate_contextual_data.json", "w") as f:
    json.dump(contextual_data, f, indent=4)

print(f"Generated {len(contextual_data)} contextualized claims with answers.")

# Generate a text file for GPT training
output_file = "climate_context_qna.txt"

with open(output_file, "w") as f:
    for entry in contextual_data:
        claim = entry["claim"]
        context = entry["context"]
        answer = entry["answer"]
        # Format the entry as context, question, and answer
```

```
        formatted_entry = f"Context: {context}\nQ: {claim}\nA:
{answer}\n---\n"
        f.write(formatted_entry)

print(f"Context, question, and answer pairs saved to '{output_file}'.")
```

The second dataset I used is named "The Stanford Question Answering Dataset (SQuAD) 2.0". It is a widely used benchmark dataset for reading comprehension and question-answering tasks. Specifically, SQuAD 2.0 combines the original SQuAD dataset with additional unanswerable questions, making the task more challenging and realistic. [10] It contains over 150,000 questions posed on passages sourced from Wikipedia articles. Each question is designed to either have an answerable span of text directly within the passage or be unanswerable, requiring models to distinguish between the two cases. We filtered out the topics that are only about climate change using the code below. We can see that only selected topics were extracted, so only around 4k triads are utilized from the dataset.

```
from datasets import load_dataset, concatenate_datasets
import re

# Load SQuAD 2.0 dataset
squad_dataset = load_dataset("squad_v2")

# Keywords to filter IPCC-related questions or contexts
keywords = ["IPCC", "Intergovernmental Panel on Climate Change", "climate
report", "climate assessment"]

# Function to filter examples based on keywords
def filter_ipcc_reports(example):
   context = example["context"]
   question = example["question"]
   # Check if any keyword is in the context or question (case insensitive)
   for keyword in keywords:
       if re.search(keyword, context, re.IGNORECASE) or re.search(keyword,
question, re.IGNORECASE):
           return True
   return False

# Apply the filtering function to train and validation splits
ipcc_train = squad_dataset["train"].filter(filter_ipcc_reports)
ipcc_validation = squad_dataset["validation"].filter(filter_ipcc_reports)
```

```
# Combine the filtered splits using concatenate_datasets
from datasets import concatenate_datasets
filtered_ipcc = concatenate_datasets([ipcc_train, ipcc_validation])

# Function to format context, question, and answer into the desired text
format
def format_context_qna(example):
   context = example["context"]
   question = example["question"]
   # Select the first answer (SQuAD usually has multiple possible answers)
   answer = example["answers"]["text"][0] if example["answers"]["text"]
else "No answer provided"

   # Return the formatted text
   return f"Context: {context}\nQ: {question}\nA: {answer}\n---\n"

# Write the formatted data to a text file
output_file = "ipcc_context_qna.txt"
with open(output_file, "w") as f:
   for example in filtered_ipcc:
       formatted_entry = format_context_qna(example)
       f.write(formatted_entry)

print(f"Filtered Context-QA pairs saved to '{output_file}'.")
```

For both datasets, we used the following hyperparameters for the GPT-2 model:

```
config = GPT2Config(
   vocab_size=len(tokenizer),
   n_positions = CONTEXT_SIZE,
   n_embd = 512,
   n_layer = 2,
   n_head = 8
)
```

<u>Addressing the Lack of Question-and-Answer Datasets regarding Climate Disasters</u>

We have observed a significant lack of question-and-answer (QnA) datasets specifically focused on climate disasters. To address this gap, I have combined datasets scraped from Hugging Face, resulting in a total of just over 9,000 triads of context, questions, and answers related to climate disasters. While this is a valuable starting point, it is clear that more data is needed to enhance the performance of the model.

To supplement this, I turned to existing datasets with broader context on climate change and environmental topics. One such dataset is `environment_data` from `ESGBERT`, a comprehensive collection of textual data centered around environmental, social, and governance (ESG) issues. It contains claims, contextual evidence, relevance labels, and categorical classifications that span climate policies, environmental risks, mitigation strategies, and sustainability efforts. For the generation model within the RAG system, the goal is not necessarily to answer questions directly but to generate plausible, informative texts that reflect the patterns seen in articles, news reports, and research papers on climate and environmental topics. This approach aligns well with the scope of our project and partially mitigates the issue of limited open-source QnA datasets for environmental sciences and climate disaster research.

The said dataset has 2,100,586 data points consisting of 2-3 sentences about the abovementioned dataset. Because of the large volume of data, we restricted the GPT-2 model to only train on a tenth and a twentieth of the dataset using the following hyperparameters outlined below.

Experiment 1: 210058 instances of the environmental dataset were used to train a GPT-2 model with the configuration:

```
config = GPT2Config(
    vocab_size=len(tokenizer),
    n_positions = CONTEXT_SIZE,
    n_embd = 512,
    n_layer = 1,
    n_head = 8
)
```

Experiment 2: 105029 instances of the environmental dataset were used to train a more complex GPT-2 model.

```
config = GPT2Config(
    vocab_size=len(tokenizer),
    n_positions = CONTEXT_SIZE,
    n_embd = 512,
    n_layer = 2,
    n_head = 8
)
```

RAG: Combining the BERT model and GPT-2 model

The process begins with a user query, such as "What is a storm surge?". The BERT model is used to retrieve relevant text chunks from a pre-indexed database of embeddings. The query is first embedded using BERT, and then its embedding is compared to the precomputed embeddings of the text chunks using a similarity metric, like cosine similarity or L2 distance. This allows the system to identify and return the top k most relevant chunks that serve as contextual information. These retrieved chunks are then passed to the GPT-2 model along with the original query. The query and retrieved text chunks are concatenated to form a contextual prompt for GPT-2, enabling it to generate a coherent and informative response. GPT-2 then generates an output like "A storm surge is a sudden and abnormal rise in sea level caused by strong winds and low atmospheric pressure during storms, leading to coastal flooding.". To demonstrate the system, two queries can be tested, and for each, the retrieved chunks are displayed before passing them to the GPT-2 model. This setup ensures that GPT-2 generates responses that are both contextually grounded and relevant by leveraging the information retrieved by the BERT model.

Results and Discussion

BERT Models' Retrieval of Information

Below, we can see an example of the top 5 chunks of information returned by the BERT model when we extract all text from the documents. Only 2 out of the 5 chunks actually relate to relevant information, and the rest are only references usually found at the end of the document.

```
query = "How can we help developing countries adapt to climate?"
```

```
Retrieved Chunk: Such support can help these countries  improve the level
of ambition in the national commitments  through the implementation of
climate-friendly technologies,  to both mitigate and adapt to the
inevitable impacts of  climate change
Retrieved Chunk:  Casado-Asensio, J., Blaquier, D and Sedemund, J (2022)
Strengthening Capacity for Climate Action in Developing Countries:
Overview and Recommendations Paris: OECD Publishing
https://doi.org/10.1787/22220518
Retrieved Chunk: npj Climate Action,  3(40)
https://doi.org/10.1038/s44168-024-00128-y Casado-Asensio, J., Blaquier, D
and Sedemund, J (2022) Strengthening capacity for climate action in
developing countries:  Overview and recommendations
Retrieved Chunk:  Sovacool, B.K., Linnér, B.-O and Klein, R.J.T (2017)
Climate change adaptation and the Least Developed Countries Fund  (LDCF):
Qualitative insights from policy implementation in the Asia-Pacific
Climatic Change 140, 209-226
```

Retrieved Chunk: Capacity-building, which includes climate education and training, is the most important need for this group of countries, which make the least contribution to the problem of climate change, but are most affected by its increasingly runaway impacts

The highlighted retrieved chunks provide great context for the GPT-2 model to use. According to ChatGPT, the relevance of both chunks can be rated as 8/10 and 8.5/10, respectively. Moving forward, I decided to filter out the pages consisting of the Table of Contents, References, and Appendices. Here is a sample response from the updated BERT model using the same query.

```
Top results:
Result 1 (Chunk):   ▶ Water, food and agriculture are the sectors most
frequently cited by developing countries as key  sectors for investment in
both capacity-building and technology transfer for addressing climate
change adaptation
Result 2 (Chunk): Yet considering the pace  of climate change, a boost in
support of adaptation  implementation is urgently needed Greater focus on
and  support provided for adaptation in the next round of NDCs  could give
credence to strengthened country ambitions  and actions
Result 3 (Chunk): Benefits for other types of strategies might be linked
to economic development,  public health, air and water quality, and
improved housing In some situations, climate adaptation  strategies could
also contribute to reductions in GHG emissions
Result 4 (Chunk): Put simply, those with  the fewest resources to cope
with climate change are often the most vulnerable to its impacts  Keeping
equity at the center of climate adaptation planning is key to ensuring the
entire  community – not just a small subset – has access to the resources
needed to adapt to climate  change
Result 5 (Chunk): Countries need to ramp up  their ambitions to prepare
for increasing climate risks Across different data sources, information on
the  implementation of adaptation actions shows large annual  fluctuations
but they ultimately result in a slight upward  trend over time (figure
ES.2)
```

ChatGPT rated them consistently high, ranging from 6-9 out of 10. Because it specifically identified important sectors (agricultural, food, and water) for investment in adaptation and capacity-building, Result 1 was given the highest rating of 9 out of 10. Result 2 received an 8 out of 10, highlighting the importance of Nationally Determined Contributions (NDCs) and the necessity of adaptation, although it lacked details on what poor nations should do. Although it was still conceptual, Result 4 came in second with an 8.5/10 for emphasizing equity in climate adaptation and the necessity of equitable resource allocation. Although Result 5 acknowledged the need to raise adaptation aspirations, it lacked practical insights, earning it a 7 out of 10. Because it covered the wider advantages of adaptation techniques without specifically addressing activities for developing nations, Result 3 had the lowest relevance (6/10).

This makes filtering important in developing the retrieval model for the RAG system. This could be improved further by eliminating the headings and captions that clearly do not constitute either a whole sentence or a whole thought.

Training the GPT-2 Model on the QnA Datasets

Here is a sample output of the GPT-2 model that was trained on the Climate Fever dataset.

```
Question: What is global warming?
Answer:,, the of the, of of,.', ', and, in,., to the the and of in of.' of
' of and the.' the ' the. the in the to, with,-, as, is, a of to of. of as
of that,0 the a, (, an, ",\, global, on, that and and in '.' in. and.. '
in in and ' ' and.'.' and as the is the- playlist,The,) of\ of with of a
the global ofThe the\.'- of is of- the ( the as Devices, human theThe of
an the an and- correlate, atmosphere, gases the with the that the0 of0 '.
in to ' a.' ' ( to in.'..' that of CO, for, be, climate, are, CO of ( and
a ' toENS, or the " of gases,C, greenhouse,xa, have of global photograp
the)grass, warming,° PV,– of " in\ and to induces the atmosphere of on of
are- ' that ' global the CO the are of) Blocks, was, from, other,'s the on
the have, ice, 2007, increased the for.' a and gases of have the be and
that in ( in a- and Bloomberg, carbon, only the climate of atmospherefish,
by Beth,.",
```

Terminologies:
While some relevant words like "global," "climate," "CO2," "atmosphere," and "gases" appear, they are surrounded by a chaotic jumble of irrelevant terms. The terminologies lack cohesion and appear randomly.

Relevance:
The answer does not provide any meaningful explanation about global warming. Key points like temperature rise, greenhouse gases, or human impact are fragmented and incoherent.

Grammar:
The grammar is completely broken, with misplaced punctuation, repeated words, and no syntactic structure. Words and symbols appear erratically.

Flow:
There is no flow or logical structure in the response. It reads as an incomprehensible series of words and characters.

— Below is a rating and a comment from ChatGPT—

Overall Rating: **1/10**
This output is unintelligible, with little relevance to the question. While a few terminologies are technically correct, they are buried in nonsensical text.

—

Below is a sample output of the GPT-2 model that was trained on both the Climate Fever dataset and the SQuAD 2.0 dataset.

```
Question: What is global warming?
Answer:, of,, the of of the the, and the and,.', ', to of and of.' the in
the.' of to the ' the to,., in of ' of in, a of a, that of. the that, (,
is, as, globallight,xaup,The to and in.' ' to to.' and to '.' to in and '
' and and.' in to that ' in in a to a the ( the a and that Lumin the as
the. and. of ( of that ducks, ", are opaque,0ordon the humandds,-,)
athletes thexaients the isplayer, greenhouse of0.'..'.' is the global of
surface of " of is of global, byplayer the) Vikings,\ the- the0 the " the
on toxa, TheUnit, temperature,° ofThe the are ofxa of are theThe, other,
on of as of greenhouse in ' that.') souven, iceicle, or, with,C to. that
to as.' that the or the greenhouse, climate, surface, have the byivers,
atmosphere workshops,- 223,', '. in. a.' global to is.'xa the°.' a in that
and ( to ( and a. ' global ' a a ' (.' are.' " that inxa ' is and0
manipulative the with. to are, human,
```

Terminologies:
This response also includes relevant terms such as "global," "greenhouse," "climate," "surface," and "atmosphere." However, like the first output, the terms appear sporadically without any meaningful connections.

Relevance:
While slightly improved, the response still fails to provide any coherent explanation of global warming. It contains fragments of potentially relevant ideas but lacks clarity.

Grammar:
The grammar remains broken, with repetitive words, misplaced punctuation, and nonsensical phrases. Some phrases like "global surface" and "greenhouse" are technically relevant but are obscured by surrounding gibberish.

Flow:
Although marginally better than the first response, the flow is still nonexistent. The text is disjointed, with no logical progression of ideas.

— Below is a rating and a comment from ChatGPT—

Overall Rating: **2/10**

This output is slightly better than the first, with a marginal increase in the presence of relevant words. However, it still lacks coherence, proper grammar, and logical structure.

—

Although the results are not what we were hoping for. It is not to surprise that the GPT-2 model was generating gibberish sentences. This is mainly because of the lack of data points within the context of our project. However, it is a good sign that combining the two datasets garnered a higher rating from ChatGPT. Assuming that the rating was valid, we can strongly infer that having more data with similar forms can improve the plausibility of the responses by the GPT-2 model.

What is more interesting is that when we tried increasing the complexity of the model by setting the embedding dimensions to 1024 and the number of layers to 12. The model's output not shown on this report had worse outcomes which makes it seem like the model is already overfitting. Thus, even though we used more complex models like GPT-3 or GPT-4, we will still not be able to generate plausible responses. This necessitates more data rather than using more complex models.

GPT-2 Model on the comprehensive environment_data dataset

Here are the results for Experiment 1 using a 1-layer GPT-2 model on 200k data points:

```
Number of beams: 1
Generated Text:
Flooding precautions are not only used to produce the water for water. .S.
Environmental Protection Agency (EPA) and the U.S., the total water
consumption. . .3)4.1.5% of water withdrawal.3.4%5 percent of our water
usage.2% water use.g.6% in water and water used in our manufacturing
processes. and manufacturing facilities. 3.0% and waste management. 4.7%
reduction in the production of waste.8% by 2025.9% from the previous year.
— the equivalent of the main water withdrawals.com/water and wastewater
treatment.html.pdf - in a year, and recycling of wastewater consumption of
a water footprint.e. - and reuse and disposal of hazardous waste is not
yet the waste generated. 5. 1. 6. in 2020.D.10% is a waste that
result.11.A. (2. 2.P.ON.V.a. 8.00. 7.E. for the use of all of total of
production.17% decrease.'s total. 3% (3%, and in% for our waste,000.
b.B.15.16 — in 2018.12. n. 10. 12. ·. 13.org/en. �.13.aspx. 9.18. - 5% 1,
water-3 percent reduction of which
--------------------------------------------------
Number of beams: 5
```

Generated Text:
Flooding precautions to ensure that the Company’s natural gas pipelines, and pipelines are subject to a high degree of subjectivity in the oil and gas and NGLs and natural resources and the availability of oil, gas production and production of crude oil production, which could be adversely affected by our operations and adversely affect our results of operations, financial condition and cash flows of drilling, development and development activities and capital expenditures, including but are not limited to: drilling and drilling activities; the timing of exploration and evaluation, production activities, such as well as the exploration activities that are required in accordance with the requirements; and (ii) the extent of the U.S. Environmental Protection Agency (“”); and operations; (i.P.e.g., and in which we operate in connection with a variety of future operations or in order to comply with our properties, or (v) or production processes that may not be required for a manner in a third-party properties that could not in place to be incurred in relation to operations of production facilities; or otherwise not have on which may affect the amount of a result of such activities on a material effect on the environment, may have a number of wells and may be recoverable quantities of an estimate of our ability to remediation of properties in such such oil or to the future exploration, operations on our production.A.5% of proved reserves and equipment, exploration or at the cost of capital flows and otherwise be subject

----

Number of beams: 10

Generated Text:
Flooding precautions are used to reduce the amount of water used in order to minimize the impacts of our operations on the environment, as well as water, water use and water resources, and the availability of fresh water for our water usage, including water consumption and use of drinking water in the communities where we serve and our customers’ needs to be able to meet water needs for customers, customers and communities, communities and other stakeholders in our supply chain, we are committed to reducing our impact on our planet and we operate and manage our environmental impact, social and environmental risks and opportunities to improve the quality of life and reduce our overall environmental footprint, but we have on-related risks that could have a negative impacts on people and how we do not have the greatest impact of products and to our products that we use, in place where our business needs of their own operations, manage the Company” and resources and services we believe we need to manage their health, while also have not only we can be required to ensure that they are required for the environmental impacts to protect our employees and products we work with our communities in a material impact our stakeholders and supply chains, the health and provide our people who are not operate; and develop our own resources that are also also not to comply with the planet, our clients and in turn, which we‟water resources

we live in an impact from our ability to minimising our manufacturing processes and protect the company're proud to help to use.

Here are the results for Experiment 2 using a 2-layer GPT-2 model on 100k data points:

```
Number of beams: 1
Generated Text:
Flooding precautions to reduce the environmental impact of our products
and the environment. . and we operate. .
.com/en/or/environmental.com.pdf/p.aspx
https://www.org/doc/t/diversity/responsibility/files/pdf
https/Pages/sustainability/and/about-environment.html https.5%20%2%% https
http://d.net/us/a.3%/social/media/c.2/v.7/corporate/h/cscc/2021/index.g.6%
pdf – and/www, and environmental impacts.youtube/re.v/ httpswww., https
https-t) and social and other issues.cd/climate/watch/2020.4%
httpwww-responsage/com% .pdf to the Group.gov/r.S.c% reduction in our
operations.pustain.au/default.ca/investment of the company.d/1%
1.0%D.1/content/b.8% of climate change.ON%respons.9%1.e.t.A.edu/2m%5.bc.,
and our business.00%7 httpsporate and safety, we have of
stakeholders.co.r., we are a positive impact.04% we can have a result to
our stakeholders .org% to
--------------------------------------------------
Number of beams: 5
Generated Text:
Flooding precautions to the environment, and the communities where we
operate and we work to ensure that our employees operate in a safe and
healthy environment is responsible for our communities and communities are
in which they operate, we have a positive impact on our business and our
operations, as well as our ability to operate our customers, our
stakeholders and their communities, which we do our planet, in our supply
chain, results of operations and results, including our products and
suppliers, communities. to our people and business, partners, customers
and society.com and other communities that we serve, suppliers and
customers are committed to meet our clients and partners and employees,
the planet.'s business partners to make our suppliers to manage our value
chain and resources and contractors, while our industry and to create a
right to develop our environment. . and people, where they have the right.
– our global communities in the world, employees and in their
stakeholders, through our shareholders, to serve as we believe we can have
our own our businesses, contractors and planet and shareholders and how we
continue to work, by our efforts to support our community, from our
activities.pdf and services, with our facilities, their business are not
able to engage, society and stakeholders" and local communities; our
workforce and consumers, investors, all communities through the company,
especially our reputation and supply and those communities – we are more
communities as the business in and environment and a business to do
business – – the industry.
--------------------------------------------------
Number of beams: 10
Generated Text:
```

```
Flooding precautions to minimize the environmental impact of our products
and services to our customers, and the environment in which we operate, as
well as reducing our impact on our own operations, while reducing the
amount of water and water use, in our supply chain, we use and use of
energy, water, waste and energy and waste disposal of materials and raw
materials in the production of hazardous materials, which are used to
landfills, including our manufacturing processes and our operations and
other processes, to ensure that we have a result of products are not
available for our suppliers, our facilities, from our business, such as
part of raw material use in line with the products, or services and
resources and packaging and products we work with suppliers and materials
we can be used in order to operate and we do not limited to reduce our
packaging, are managed in a products that can not managed by our ability
to meet our employees and to the use we are required to comply with our
sites and in compliance with all products in accordance with other
products.g., and/or to manage our stakeholders, suppliers's
operations.Sustainability, so as a material resources, the ability of the
company, where we need to make products" and their own products to use
chain and processes. products; and systems and suppliers.5% materials that
ensure we serve the supply of goods and practices, for the communities and
product processes that are responsible sourcing, processes; - and
technologies, by suppliers; the packaging.
--------------------------------------------------
```

—Below is the analysis of ChatGPT on the responses—

Summary

1. **Experiment 1 (1-Layer, 200k Data Points)**:
   - **Beams 1**: **2/10** (Nonsensical and repetitive)
   - **Beams 5**: **3/10** (Slightly better, but still incoherent)
   - **Beams 10**: **5/10** (Improved coherence and relevance)
2. **Experiment 2 (2-Layer, 100k Data Points)**:
   - **Beams 1**: **1/10** (Unusable output with URLs and gibberish)
   - **Beams 5**: **4/10** (Slight improvements with relevant terms but verbose)
   - **Beams 10**: **6.5/10** (Most plausible, relevant, and structured)

Conclusion: The outputs from **Experiment 2 (2-layer GPT-2, 10 beams)** are the most plausible and relevant, demonstrating improved structure and alignment with environmental topics. However, the model still struggles with verbosity and focus, indicating room for further fine-tuning and dataset refinement.

—

Using this comprehensive dataset is indeed an improvement from only using the available QnA datasets. From only scoring a maximum of 2 out of 10 from ChatGPT, the responses from the GPT-2 model trained on the comprehensive dataset were rated 5 and 6.5 out of 10 for the respective experiments. This goes to show that the model only needs more data ten-fold.

In contrasting the two experiments, we can see that configuring the GPT-2 model to have two layers generates better responses despite it running on fewer data points. This phenomenon was also seen when we were training the model on the TinyStories dataset. From the TinyStories dataset, it was found that increasing the number of layers improves the contextual understanding of the GPT-2 model. This very much explains this improvement from using 1 to 2 layers. However, it is important to note that an increase in layers means an increase of parameters to train and therefore increase in the energy assumption. This is where the tradeoff between model complexity and model sustainability comes in. It would be interesting to see where the optimal complexity of the model is with its energy consumption also being considered.

More ideally, if we have this amount of dataset with only plain text, surely, there will be a way to extrapolate possible questions and answers from them. If we acquire this ability, we can then sift through millions of data samples and easily convert them into the target format, context-question-answer, with the context being the entire text, where the question and answer are extracted.

<u>RAG: Combining the BERT model and GPT-2 model</u>

Feeding the top chunk generated by the BERT model, the following is the response GPT-2 model generated.

Context: `Water, food and agriculture are the sectors most frequently cited by developing countries as key  sectors for investment in both capacity-building and technology transfer for addressing climate change adaptation,`

Query: `"How can we help developing countries adapt to climate?"`

Answer

```
…and climate-related risks, as well as in the supply chain, such as
energy, water and water use, waste management, energy and energy
efficiency, climate change and greenhouse gas (GHG) emissions,
emissions and other risks associated with the impact of our
operations, in our products and the world's ability to our business,
we operate, including our customers, which we have a result in order
to reduce our own operations and environmental impact on our ability
```

```
of the environment, while we continue to meet our suppliers, our
stakeholders") and our global operations are more sustainable
development and we use of new products, the communities and
resources, by our clients and society and emissions reduction of
water-saving products we are a global supply chains, to operate in
line with our manufacturing processes and to make our facilities and
business and waste reduction efforts to ensure we believe we can have
on the Group'and our employees, through our value chain - and/or the
Company"Sustainability, for our activities, are committed to improve
our sustainability and to-based business - as a key sustainability
goals we do not only communities that are responsible and in a
responsible business partners, from our communities, where we need to
create a business practices, with climate and communities we work and
how we serve as we expect our…
```

—Below is the rating from ChatGPT—

Relevance Rating: 2/10

The response fails to answer the query in a meaningful way or utilize the provided context effectively. While it includes loosely related terminology, the lack of coherence, structure, and focus makes it nearly unusable.

—

Despite using some somewhat related language, the RAG system's response is mainly nonsensical and does not meaningfully answer the question. Words like "climate change," "sustainability," "emissions reduction," and "waste management" are used, but they are haphazardly repeated and don't make sense concerning the context that is given. The response veers into a stream of business jargon rather than expanding on the main notion of the context, which is that food, agriculture, and water are crucial areas for adaptation in developing nations. The output is badly organized grammatically, with superfluous words and run-on sentences that obscure its meaning.

Conclusion and Recommendations

To sum up, this study demonstrates the possibilities and difficulties of using a Retrieval-Augmented Generation (RAG) system to address the dissemination of knowledge about climate change/disasters. The system tries to recover pertinent text chunks from a database and produce educational responses to user queries by fusing the advantages of a GPT-2-based generator with a BERT-based retriever. The generated responses frequently suffer from a lack of coherence, excessive verbosity, and repetition, even if the retrieval process shows the ability to recognize contextually relevant information, such as important sectors like food, agriculture, and water for climate adaptation. This emphasizes how crucial it is to optimize both the generation and retrieval processes to guarantee that the system generates outputs that are precise, targeted, and useful in relation to the given environment.

<u>Manual Preparation of a Question-and-Answer Dataset</u>
Future efforts could focus on manually curating a high-quality QnA dataset by leveraging outsourcing platforms like Amazon Mechanical Turk or similar services. This approach would ensure the creation of reliable and contextually accurate data tailored to climate-related topics.

<u>Extensive Training Using Large GPUs and TPUs</u>
For approximately a week, researchers with access to large-scale hardware architectures such as GPUs and TPUs could further train and fine-tune models. An interesting direction would involve experimenting with a deeper GPT-2 model (adding more layers) and training it on the entire environment_data dataset to explore potential performance improvements.

<u>Incorporating More Advanced Models like GPT-3 or GPT-4</u>
Given the limited capabilities of GPT-2, future work could explore using more powerful models like GPT-3 or GPT-4 from OpenAI to train on additional hyperparameters and datasets. However, this approach raises concerns about accessibility and equity, as these models require significant computational resources that are often prohibitively expensive. Since the project aims to help inform people in developing regions about climate adaptation strategies, deploying power-hungry models would not align with this goal. Such advanced architectures may further exacerbate the digital divide between the Global North and the Global South, where access to high-end computing resources remains limited. Additionally, ample research highlights the environmental impact of training such large-scale models, underscoring the need for more sustainable and accessible solutions.

Broader Impacts

Climate disasters are a pressing topic that requires greater representation within emerging technologies, particularly transformer-based architectures. However, this progress hinges on the availability of adequate and relevant data. This challenge is not unique to climate-related contexts but reflects a broader issue: the powerful models being released today demand significantly more data than we currently have. To address this, there must be a shift in focus among researchers toward generating more data, ideally through automated and scalable methods.

At a high level, this task compels us to consider not just the superior performance of these models but also the contexts in which they are deployed. For instance, it is crucial to think about the energy requirements of running such models, especially in resource-constrained areas, and to ask: What kinds of data are truly needed? If general public datasets are insufficient, why not prioritize generating localized and domain-specific data? These considerations highlight the importance of balancing model performance with accessibility and relevance. Moving forward, the ultimate question is how we can incentivize the labor-intensive task of data preparation to encourage the development of massive, high-quality datasets necessary for transformative progress.

References

[1] N. Koldunov and T. Jung, “Local climate services for all, courtesy of large language models,” *Commun. Earth Environ.*, vol. 5, no. 1, pp. 1–4, Jan. 2024, doi: 10.1038/s43247-023-01199-1.

[2] J. Bulian *et al.*, “Assessing Large Language Models on Climate Information,” May 28, 2024, *arXiv*: arXiv:2310.02932. doi: 10.48550/arXiv.2310.02932.

[3] D. Thulke *et al.*, “ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change,” Jan. 17, 2024, *arXiv*: arXiv:2401.09646. doi: 10.48550/arXiv.2401.09646.

[4] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Apr. 12, 2021, *arXiv*: arXiv:2005.11401. doi: 10.48550/arXiv.2005.11401.

[5] J. Wang *et al.*, “Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges,” *ACM Comput Surv*, vol. 56, no. 7, p. 185:1-185:33, Apr. 2024, doi: 10.1145/3648471.

[6] S. McLean, “The Environmental Impact of ChatGPT,” Earth.Org. Accessed: Dec. 17, 2024. [Online]. Available: https://earth.org/environmental-impact-chatgpt/

[7] O. F. Rokon, “Transformers in NLP: BERT and Sentence Transformers,” Medium. Accessed: Dec. 17, 2024. [Online]. Available: https://medium.com/@mroko001/transformers-in-nlp-bert-and-sentence-transformers-3faab61918ea

[8] “SentenceTransformers Documentation — Sentence Transformers documentation.” Accessed: Dec. 17, 2024. [Online]. Available: https://sbert.net/

[9] T. Diggelmann, J. Boyd-Graber, J. Bulian, M. Ciaramita, and M. Leippold, “CLIMATE-FEVER: A Dataset for Verification of Real-World Climate Claims,” Jan. 02, 2021, *arXiv*: arXiv:2012.00614. doi: 10.48550/arXiv.2012.00614.

[10] “SQuAD - the Stanford Question Answering Dataset.” Accessed: Dec. 17, 2024. [Online]. Available: https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/