AI-Driven Verification of Climate-Related Claims: Automated Evidence Retrieval from the IPCC Sixth Assessment Report

Iestyn Mullinor

4th Year Project Report
Artificial Intelligence and Computer Science
School of Informatics
University of Edinburgh

2024

Abstract

The pervasive spread of online climate change misinformation on social networks and in traditional media warrants the need for specialised claim verification tools.

Leveraging the comprehensive Sixth Assessment Report by the Intergovernmental Panel on Climate Change, we propose a novel approach to climate change fact-checking by retrieving authoritative evidence from this reputable corpus, which is subsequently analysed by a specially fine-tuned Large Language Model (LLM). Our work emphasises the need for solid and reputable evidence to establish a credible system, while considering real-world implications, and building upon the CLIMATE-FEVER claim-verification framework created by Diggelmann et al. (2020) [11].

Despite the relatively small scale of the knowledge base, we found that dense retrieval from this specialised evidence source is effective for verifying a wide array of climate-related claims, particularly those pertaining to climate science. After extensively testing each component and constructing this pipeline, the system demonstrated a 76% success rate in verifying real claims from social media, while providing supporting evidence from the Sixth Assessment Report.

The integration of LLMs has demonstrated that the performance of the system can be significantly enhanced through techniques such as In-Context Learning and Parameter-Efficient Fine-Tuning, opening up vast potential for future enhancements and extensions of the system.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Iestyn Mullinor*)

Acknowledgements

First and foremost, I'd like to thank my dissertation supervisors, Björn Ross and Sandrine Chausson, for the invaluable advice and helpful feedback throughout the year. I'd also like to thank Pasquale Minervini, for our insightful second marker meetings, and the interesting NLU+ lectures that have come in very handy when undertaking this project. Next, I'd like to thank Xue Li, for providing me access to the "Miscc" dataset to test the system. Thanks to all the other lecturers and staff in the School of Informatics as well, you've all been great.

Also thank you to my parents and friends for helping me proof read this very long dissertation!

Table of Contents

Chapter 1

Introduction

1.1 Motivation

As the sheer volume of misinformation on the internet rises, the need for robust and efficient fact-checking becomes more apparent. Misinformation has been ranked as one of the top 10 global trends threatening the world [8], and there is evidence that the volume of misinformation present online can cause people to disregard any information they read online whatsoever [33], due to a lack of trust. While many online media sites, such as Facebook and X, deploy human fact-checking strategies for specific posts, this process is extremely time-consuming and labour-intensive, and it is infeasible to keep up with the sheer volume of posts made online with this method [68].

Climate change is a field that has been particularly negatively impacted by misinformation, both online and in traditional media [65]. Climate change misinformation has likely contributed to confusing and polarizing the public on the issue, which can lead to political inaction [46]. For a problem as urgent as climate change [57], it can be argued that public confusion and political inaction warrant a robust and vigorous climate change fact-checking system to combat the spread of this misinformation. While automated fact-checking systems will likely not replace human fact-checkers altogether, they can potentially become valuable in tackling misinformation. There have been recently combined efforts between the fields of natural language processing and climate science to create tools and systems that can be used to fact-check claims about climate change [11], yielding promising results.

Emerging technologies in the field of natural language processing show the potential to be useful in creating such claim verification systems, especially with the recent advancements in Large Language Models. A robust fact-checking system for climate change-related claims on large social media sites could benefit society, as it could empower users to make informed decisions. This technology could work in conjunction with human fact-checkers to provide rapid verification for claims online. An issue with claim verification systems is ensuring that the evidence used does not contain misinformation, as they use extremely large and often un-fact-checked search spaces for finding evidence. Instead of using a large search space of potentially unreliable information, we want to assess evidence retrieval methods on a smaller but robustly

fact-checked evidence bank for complicated and scientific claims, as this may improve the reliability and trust of such a system.

In this dissertation, we construct a claim verification pipeline using the IPCC Sixth Assessment Report as our evidence bank. While developing this system, we aim to answer a set of research questions, as described below, to contribute to reducing the spread of climate-related misinformation.

1.2 Research Questions

Due to the scale and complexity of creating such a system, we break down our research goals into three distinct research questions. These questions aim to expand on existing works while also providing new approaches based on current state-of-the-art technology and analysing the results of these approaches by comparing them to existing methods.

Research Question 1: **What is the viability of using a single, comprehensive, domain-specific document, such as the IPCC Sixth Assessment Report, as the entire evidence corpus for various evidence retrieval methods in a claim verification system?**

The IPCC Sixth Assessment Report is an extremely comprehensive and rigorously fact-checked document which contains information on a vast range of climate science-related concepts. The completeness and reliability of the information present in this document makes it an interesting candidate to use as the evidence bank for a climate-related claim verification system. We aim to answer this question by constructing an evidence retrieval pipeline, using the IPCC AR6 as a knowledge base, shown in Chapter 3.

Research Question 2: **How well do Large Language Models (LLMs) perform for the task of classifying whether an evidence sentence supports or refutes a claim, and how can their performance be improved?**

With the recent advancements in generative LLMs, we want to assess their ability to classify whether given evidence can support or refute a given claim. This is an interesting task for LLMs, since it does not assess their ability to generate a large amount of text, but their ability to understand the meaning of sentences and the nuances of entailment classification. To answer this, we will incorporate LLMs into the evidence classification component of our evidence retrieval pipeline, and employ multiple techniques to improve their performance. We implement this in Chapter 4.

Research Question 3: **How well does the system we create perform on real claims? And how does its performance differ on different types of climate-related claims, such as scientific, political, and informal?**

Once we have created the claim-verification system, it will be valuable to assess how well it performs on different types of climate-related claims. Since the motivation for this work is tackling online misinformation related to climate change, understanding how well the system works on different types of claims will allow us to assess what changes or improvements could be made to the system to help achieve this goal. Also, this will allow us to understand how useful our system could be in a real-world setting. We analyse this in Chapter 5.

Chapter 2

Background

2.1 Claim Verification Systems

Automated claim verification is the task of finding supporting or refuting evidence for a given claim, to verify whether or not the claim is correct [60]. There have been multiple implementations of different solutions to the task, and a common and effective implementation consists of a pipeline of 3 stages [61]:

Document retrieval: The process of finding and ranking the documents containing potential supporting evidence for a given claim based on their relevance. Often, there will be a re-ranking step, where the top k retrieved documents are re-ranked using a more computationally expensive method.

Sentence Selection: The process of selecting which sentences from a document can be used as evidence for a given claim. There will often be a re-ranker in the sentence retrieval stage.

Sentence Classification: The process of classifying if an evidence sentence supports or refutes a claim, usually by classifying a claim-evidence pair as one of the following: $\{$SUPPORTS, REFUTES, NOT_ENOUGH_INFO$\}$ [61].

Other less common stages in the process of claim verification include:

Claim detection: A system which can take a passage of text and extract all "claims" from it [48], or a system which can take a complex sentence in a given context, and use this context to determine what the sentence is claiming, as defined by Levy et al. [32].

A visual representation of such a system can be seen in Figure 2.1

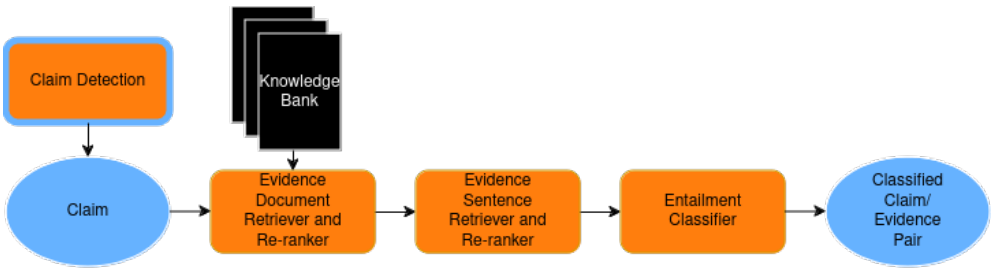

Figure 2.1: Common Claim Verification Pipeline. The blue border indicates the less common steps in the pipeline.

2.2 Datasets

FEVER [61] is a dataset consisting of 185,445 claims, each with corresponding pieces of evidence from Wikipedia labelled "SUPPORTS", "REFUTES", or "NOT_ENOUGH_INFO" based on whether the evidence supports or refutes the claim. For clarity, we will refer to "NOT_ENOUGH_INFO" as "NEI" throughout this paper. These claims were generated by annotators, and the paper introduces 2 main tasks. The first task involves retrieving evidence for each claim using the entirety of Wikipedia as a search space, and the second task consists of labelling each claim based on whether its evidence supports or refutes it, or if there is not enough information to decide.

Inspired by FEVER, CLIMATE-FEVER [11] is a dataset consisting of 1,535 climate-related claims, each with 5 related evidence sentences from Wikipedia. Each claim/evidence pair is labelled as "SUPPORTS", "REFUTES", or "NEI". Each claim also has an overall label based on the majority vote of its evidence sentences. Appendix A.2 shows an example of this. An important difference between FEVER and CLIMATE-FEVER is that all claims in CLIMATE-FEVER are real-world claims that have been extracted from the internet, whereas the claims in FEVER are all artificial. It has been shown [11] that the real-life nature of the claims in CLIMATE-FEVER can hugely increase the difficulty of classifying whether evidence supports or refutes a claim. FEVER and CLIMATE-FEVER can be used as training sets or test sets for various stages in the claim-verification pipeline, which will be described in the coming section.

The papers which introduce these datasets[61, 11] introduce baseline methods for carrying out the tasks, which we discuss in the coming sections. Furthermore, the paper "Evidence based Automatic Fact-Checking for Climate Change Misinformation" [69] shows the benefits that can be gained from using a large portion of CLIMATE-FEVER as a training set for various components in the claim verification pipeline, and using the rest of it as a test set, as it is specific to the domain of climate change. The authors also introduce methods for these components, which we will explore in the coming sections.

2.3 Document Retrieval Methods

Document retrieval involves finding the top k documents that may contain evidence for a given claim c. In the context of the FEVER shared task, the document retriever's

performance is measured by the number of claims that are supported or refuted by any of the k documents retrieved for the claim.

The original paper that defines the FEVER task provides a "baseline system description" that describes a simple method for tackling the problem. The authors opt for the approach of the "DrQA" system [5], which is a document retrieval method developed by Meta which returns the k nearest documents for a claim using cosine similarity between binned unigram and bigram Term Frequency-Inverse Document Frequency (TD-IDF) vectors of the document and the claim. TF-IDF vectors are mathematical representations that describe the importance of a word within a document, see Appendix A.1 for more details. The original CLIMATE-FEVER paper uses BM25 [53] for document retrieval, which is an approach that ranks how similar a document is to a query based on term frequency and document length, to find the top 10 documents. Alternatively, Wang et al. (2021) [69] employ Google search for this task, restricting it to a list of 55 "credible websites" in an attempt to ensure only reputable sources are used for evidence.

Other approaches to the task, such as "UKP-Athene" [20], deploy named entity recognition (NER) tools to identify the entities which the claim references, and then search Wikipedia for articles whose titles match the named entities. This is a popular approach, and is commonly used in many document retrieval systems [55, 62]. While many of these approaches exclusively use Wikipedia for the domain of their document search, some approaches have leveraged Google Search for document retrieval to achieve a wider search space of evidence [54].

Google search and Wikipedia provide a source of up-to-date information for document retrieval, which cannot be achieved with a classic pre-defined corpus of documents. However, despite efforts to only include credible sources in these searches, and the favourable reliability of Wikipedia [15], the unpredictability of Google search and Wikipedia documents could lead to some false information being retrieved in the document selection process. Furthermore, if such a system were to be deployed in a social media platform, the long-standing distrust many people have in Wikipedia [29] may cause users to disregard or distrust a fact-checking system which sources Wikipedia.

2.4 Sentence Selection Methods

Sentence selection is carried out once appropriate evidence documents have been found. In this stage, the goal is to find which sentences in the retrieved evidence documents are relevant to the claim being verified.

In the original FEVER shared task formulation paper [61], the authors modified the document retrieval component of DrQA [5] to pick sentences instead of documents, to extract the top l sentences from the top k documents using their bigram TF-IDF vector similarity to the claim. The baseline approach given in the original CLIMATE-FEVER paper [11] utilises more recent methods, and involves finding relevant sentences by comparing their sentence embeddings. The sentence embeddings are calculated using an ALBERT (large-v2) [31] model fine-tuned on FEVER.

Similarly, Wang et al. (2021) [69] approach the task of finding evidence sentences for climate-related claims by using a RoBERTa [36] model fine-tuned on both FEVER and CLIMATE-FEVER. The model encodes the claim concatenated with the potential evidence sentence and then feeds this encoding to a linear classifier that predicts whether it is an evidence sentence.

Many existing methods also employ a "sentence re-ranker" step, in which the top k retrieved evidence sentences are each given a new score, often using more computationally expensive methods, and the evidence sentences are re-ranked based on their new score. This allows the use of more advanced methods for allocating scores to evidence, as we only re-rank a small number of the top retrieved evidence sentences. A common sentence re-ranking strategy, which is employed by Diggelmann et al. (2020) in the original CLIMATE-FEVER paper [11], involves fine-tuning a BERT [10] based model as a binary classifier to distinguish between "evidence" and "non-evidence", and using the model's confidence in the "evidence" class as the new score.

2.5 Evidence Classification Methods

2.5.1 Traditional Methods

The goal of evidence classification is to take a claim and an evidence sentence and classify the pair as one of $\{$SUPPORTS, REFUTES, NOT_ENOUGH_INFO$\}$.

The baseline approach suggested in the original paper which proposes FEVER [61] involves a simple model which is outlined by Riedel et al. (2017) [52], and involves using a Multi-Layer Perceptron with one hidden layer which takes in the TF (Term Frequency) vector of a claim, the TF vector of an evidence sentence, and the cosine similarity of the TF-IDF vectors of the claim and evidence sentence as inputs. This is a simple approach, yet yields impressive results for a baseline model.

A more recent and common approach which yields impressive results is using a BERT-based model [10] with a classification head fine-tuned on the FEVER [61] dataset. This approach was adopted for the baseline method provided in the original CLIMATE-FEVER paper [11], where a fine-tuned ALBERT (large-v2) model [31] takes in the claim concatenated with an evidence sentence, and then enters the `[CLS]` token into a 3-way classifier, which predicts the class. Using this method, the authors achieved an unweighted F1-score of 32.85% on CLIMATE-FEVER.

This approach was also employed by Wang et al. (2021) [69], using RoBERTa fine-tuned on the CLIMATE-FEVER dataset. Since CLIMATE-FEVER is a relatively small dataset, Wang et al. (2021) [69] utilise unsupervised data augmentation to make use of unlabelled data in the training process. This was done by translating additional unlabelled claims into German, and then back-translating them into English. The model is then trained with the goal of reducing the loss on the supervised data, and ensuring that the additional unlabelled claims are labelled consistently with their back-translated counterparts. Using this method, the authors gain an impressive unweighted F1-score of 71.8% on their test set. See Appendix A.7 for an explanation of these metrics.

2.5.2 LLMs for Text Classification

The task of classifying a claim/evidence pair as "SUPPORTS", "REFUTES", or "NOT_ENOUGH_INFO" is a form of text classification. Many recent approaches to text classification involve using transformer-based [67] models, such as BERT, fine-tuned for specific text classification tasks [58]. These approaches often yield impressive results, especially compared to more traditional methods, for example, logistic regression using TF-IDF vectors [19].

More recently, since the widespread use and adoption of generative Large Language Models (LLMs) such as GPT-3, there have been many studies on using these models for text classification [59, 13, 71], which show promising results when compared to comparatively small transformer-based models, such as BERT. In the context of text classification, the LLM must generate a token of its predicted label, and this token can be extracted from the generated text to identify the model's class prediction.

Sun et al. (2023) [59] introduce a method of prompting LLMs for text classification called CARP (Clue And Reasoning Prompting). In this approach, the LLM is prompted to generate its Chain of Thought (CoT) [30], as well as what "clues" it finds in the original text to help with classification. This technique can enhance the performance of the model, as it gains additional context and reasoning for consideration before making a classification judgment. This method achieved state-of-the-art results for many text classification benchmarks. Wang et al. (2023) [71] discuss the performance improvement that can be gained by prompting LLMs with precisely crafted prompts which are designed to steer the model into providing a precise result in the correct format. These techniques are known as "zero-shot", as the models are provided with no training examples of correct classifications prior to inference.

The performance of LLMs can often be improved using "few-shot prompting" [4], also known as "in-context learning", where the prompt given to the model contains some examples of the task it is requested to carry out. In the context of text classification, this would entail providing examples of k pieces of text with their corresponding label in the prompt, for example, correctly classified claim/evidence pairs. LLMs can also be fine-tuned for specific tasks, such as text classification, which can greatly improve the performance of the model [35] when compared to in-context learning techniques.

2.6 Knowledge Base - IPCC Sixth Assessment Report

While it is common to use Wikipedia or Google search to find potential evidence documents, in this research, we explore the viability of using one single comprehensive document, which provides comprehensive information in many areas of climate science. The IPCC Sixth Assessment Report (IPCC AR6) [23] was created by The Intergovernmental Panel on Climate Change, which is the United Nations body for assessing the science related to climate change. It is a comprehensive report of around 7500 pages, and provides a plethora of technical and scientific information on climate change, as well as in-depth analysis on its impacts and future risks. The report was finished in 2023 and is the most up-to-date comprehensive report on climate change at the time of

writing. The report is made up of four sections: a synthesis report, and 3 working group reports.

Synthesis Report: This is the shortest of the four reports making up the AR6 and a summarization of the three working group reports. The IPCC creates comprehensive reports every five years, each with a shorter synthesis report to be easily digestible.

Working Group 1 Report - The Physical Science Basis: This report includes the most recent physical understanding of climate change and details recent advances in the field of climate science. It consists of 12 chapters.

Working Group 2 Report - Impacts, Adaptation, and Vulnerability: This report details the impacts of climate change on many different areas, such as at the human and environmental levels, and goes in-depth into impacts within different ecosystems. It also assesses the ability of humans and nature to adapt to a changing climate environment. It consists of 25 chapters.

Working Group 3 Report - Mitigation of Climate Change: This report examines progress and pledges from governments and organizations to reduce climate change and investigates the largest contributors to global emissions. It consists of 17 chapters.

A major benefit of using the IPCC AR6 as our knowledge base is the reputation and status of the IPCC. The IPCC organisation has members from 195 countries [22], making it arguably the most trusted source of climate-related information in the world. The information in the IPCC reports is agreed upon by these member countries, which is likely to increase people's trust in the information in these documents.

2.7 Other Climate-Related NLP Research

ClimateBert: In the paper "ClimateBert: A pretrained model for climate-related text" [72], the authors propose "ClimateBert", a climate-specific transformer-based language model that was created by fine-tuning RoBERTa [36] on 1.6 million climate-related paragraphs. ClimateBert is publicly available on the HuggingFace hub, and it also has 5 other versions available, which are pre-trained for the a variety of downstream tasks, such as climate-related text detection.

ChatClimate: In "ChatClimate: Grounding Conversational AI in Climate Science" [66], the authors enhance OpenAI's GPT-4 by providing it access to the IPCC AR6, to create a ChatBot capable of answering climate-specific questions accurately with up-to-date answers. It was found that a hybrid model of GPT-4's prior knowledge combined with the IPCC AR6 information was able to provide accurate information during the task of climate-related Question Answering.

Chapter 3

Evidence Retrieval

3.1 Overall Evidence Retrieval System Architecture

The task of creating the evidence retrieval portion of our system consists of the following key components:

- Preparing the evidence bank for information retrieval
- Climate claim detection system
- Evidence sentence retrieval system
- Evidence sentence re-ranking model

Due to the smaller search space of evidence, we deemed it appropriate to include only the sentence retrieval step while not utilising the document retrieval step employed in many existing evidence retrieval methods. The previously described methods use search spaces of millions of documents, whereas our smaller search space of just the IPCC AR6 allows us to remove this step while still maintaining efficient evidence retrieval.

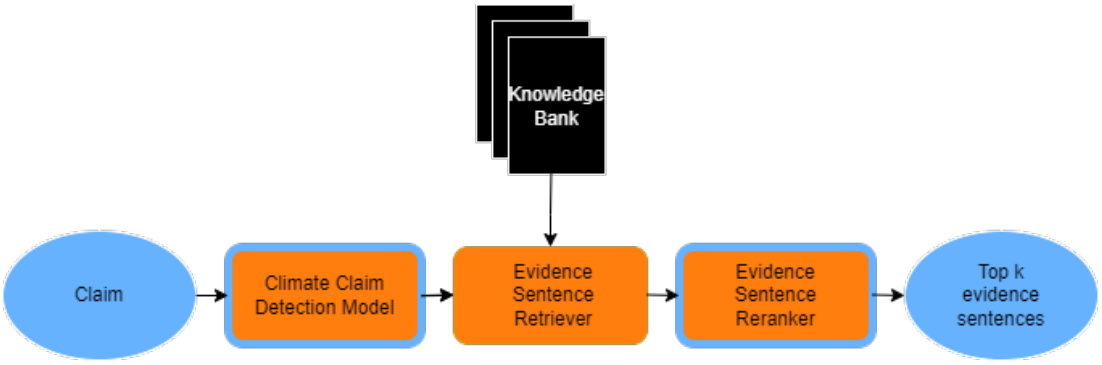

Figure 3.1: Overall design for the evidence retrieval system. Steps with a blue border represent the “extra” steps that are not in the “core” pipeline. This means we will test the overall system with and without these components.

Figure 3.1 shows the system of retrieving evidence for a given climate-related claim.

This consists of "core" components and "extra" components. The "extra" components are highlighted in Figure 3.1, and we will analyse the performance of the system with and without these steps. Each component will be tested individually before the overall system is evaluated, and the interactions between each component will be assessed.

3.2 Knowledge Base Extraction

The IPCC AR6 is available via the IPCC Website [24], where each chapter from each report is accessed via a separate URL. Each URL contains titled sections for each subsection of the chapter, as well as figures that are referenced in the text. Using a list of every URL for the Synthesis report [23], Working Group 1 report [25], Working Group 2 report [26], and Working Group 3 report [27], we implemented a program that extracts all text from the titled sections of the web page and stores them in separate files. This was implemented using the `bs4` (Beautiful Soup) Python library [51].

We only extract the raw text from the report and do not incorporate any text from figures or tables in our knowledge base. Many sentences in the text reference figures and tables, and these would not make sense in a text-based claim verification system, so the text is cleaned to remove these. Other features of sentences that were removed in the cleaning process were citations and references to other chapters. Also, certain abbreviations were expanded. For example, instances of "GHGs" were expanded to "greenhouse gasses". This cleaning process ensures that most evidence sentences are understandable outside of the context in which they originally appeared. Sentences under 7 words are also removed, as short sentences tend to lack relevance for claim verification, often serving as references to other sections or as headings.

Each section of text in the IPCC AR6 was cleaned, and then each sentence in the cleaned section was paired with its section title and added to the main knowledge base file. We pair each sentence with its numbered section title so the source of the evidence sentence in the IPCC report can be easily traced. An example of a section from the IPCC AR6 before and after the cleaning process can be seen in Appendix A.3.

3.3 Climate Claim Detection

3.3.1 Task Description

The climate claim detection system decides whether or not an input claim is climate-related before passing it on to the rest of the system. As seen in the pipeline Figure 3.1, this is not a "core" step in the system, and the system could still function without it. Despite this, we still include it for the following reasons:

- To demonstrate how a similar system may be deployed in a real setting - For example, on a social media platform, a post would be input into a climate-claim detector, and if it is decided to be climate-related, the post could then be verified by a claim verification system.
- To reduce error in later stages of the pipeline - If a non-climate related claim is

entered into the system, it is possible that the system may incorrectly support or refute the claim with IPCC AR6 evidence. We know for certain that the IPCC AR6 only contains evidence for climate-related claims, so this component prevents errors where we may incorrectly find evidence for a non-climate-related claim.

3.3.2 Methods

ClimateBert [72] is a fine-tuned RoBERTa [36] model that has been specifically adapted to the domain of climate-related text. The creators of this model also fine-tuned it for the downstream task of classifying whether or not a segment of text is climate-related. In the description of this model, it is warned that "*This model is trained on paragraphs. It may not perform well on sentences.*". This is a concern, as the majority of claims that our system will be handling are single sentences. Therefore, we must run additional tests to ensure this model is suitable for our task.

To determine whether or not this existing model is suitable for detecting climate-related claims, we created a test set consisting of 3070 claims, half of which are climate-related (extracted from CLIMATE-FEVER) and the other half are non-climate-related (extracted from FEVER). We input these claims into the existing model and found that it worked proficiently for our specific task of classifying climate-related claims.

From our tests, the model achieved an impressive F1-score of 0.96, despite being trained on paragraphs as opposed to claims. The classification report from this test can be found in Appendix A.4. Also, upon inspection, it was noticed that many sentences from CLIMATE-FEVER which were incorrectly classified as non-climate-related were not very obviously climate-related, such as "*Houlton has been exploring this possibility for years.*". Due to the high F1 score, we incorporate this existing model into our pipeline instead of fine-tuning a separate model for this task.

3.4 Evidence Sentence Retrieval

3.4.1 Task Description

In this section, we compare multiple methods for retrieving evidence sentences from the IPCC AR6. After cleaning, the evidence bank consists of 16,004 unique evidence sentences. This is small compared to many claim-verification methods, which often use Wikipedia or Google search as their evidence search space. In our case, it is sufficiently quick ($\sim$1 second) to search the entire evidence bank using all of our sentence retrieval methods, and we deem this efficient enough to warrant not using a document retrieval step. A potential benefit of using a single large document instead of many smaller documents is that the system will not miss any potentially important evidence sentences which are in documents which may not be ranked highly by a document retrieval step, and can instead search through every sentence in our document of evidence sentences.

3.4.2 Methods

3.4.2.1 Sparse Retrieval

Sparse retrieval involves searching for evidence sentences which are similar to the claim through comparative analysis of their words and phrases. BM25 is a well-known sparse retrieval method which is commonly used for document retrieval, and can also be used for sentence retrieval [2]. We implemented this as a potential evidence sentence retrieval component for our system, due to its proven high performance [53] and efficiency. This method was implemented using the `BM25Okapi` module from the `rank_BM25` library [3]. BM25 ranks potential evidence documents based on term frequency and document length. In our case, we use it to retrieve evidence based on term frequency and sentence length since we do not implement the document retrieval step. The implementation of this system is described below [7]:

$$\text{score}(D,Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{\text{TF}(q_i,D) \cdot (k_1+1)}{\text{TF}(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \tag{3.1}$$

$$\text{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right) \tag{3.2}$$

Where:

- $Q =$ Claim containing keywords $q_1, ..., q_n$
- $\text{TF}(q_i, D) =$ number of times that the keyword q_i appears in evidence sentence D
- $|D| =$ length of evidence sentence D
- avgdl = average evidence sentence length
- $\text{IDF}(q_i) =$ Inverse Document Frequency of a keyword q_i, which is calculated using Equation 3.2
- $N =$ total number of evidence sentences in the collection
- $n(q_i) =$ number of evidence sentences containing keyword q_i
- $k_1 =$ constant, described below
- $b =$ constant, described below

The k_1 parameter controls the impact of the Term Frequency (TF) value on the score, and is typically within the range of 0.5 to 2.0. b is a parameter, between 0 and 1, that affects the impact of sentence length normalisation on the scoring; for instance, a high value for b penalizes longer sentences. We set k_1 to 1.5 to ensure that the terms greatly impact the score. We set b to 0.75, as after informal testing, it was found that much longer sentences tend to contain less relevant evidence and often discuss a broad range of topics instead of specific evidence for a claim.

Using Equation 3.1, for a claim sentence Q, and each potential evidence sentence D, we assign each evidence sentence a $\text{score}(D,Q)$, which describes how similar the evidence

sentence is to the claim sentence. Using these scores, we can then rank every evidence sentence based on its relevance to the claim.

BM25 is a "bag-of-words" model, meaning it does not consider the meaning and order of words, only the words themselves. An advantage of BM25 is its simplicity and efficiency. If we scaled our evidence corpus to something much larger, BM25 would still quickly retrieve evidence compared to neural methods since it is purely a term-based retrieval method.

3.4.2.2 Dense Retrieval

Dense retrieval involves encoding evidence sentences into dense vectors which capture their semantic meaning, often using deep learning methods. A claim sentence can be encoded the same way, and we can then find the most semantically similar evidence sentences to the claim sentence by comparing these vector encodings. The evidence sentences only have to be encoded once, and these vector representations, also known as embeddings, are stored. Therefore, finding evidence sentences for a claim is efficient as only the claim sentence must be encoded during retrieval. This is visually represented in Figure 3.2.

For a given encoder model, we pre-calculate the embeddings for every evidence sentence in the knowledge base. This is a computationally expensive process, as we have 16,004 evidence sentences, but only has to be executed once. Once we have the sentence embeddings for each evidence sentence, we can take a claim, find its embedding using the model, and find the k nearest neighbours to the claim using cosine similarity. This gives us the k sentences that have the most similar vector representations to the claim. In order to find the most optimal model for our specific task, we conduct in-depth testing on various models.

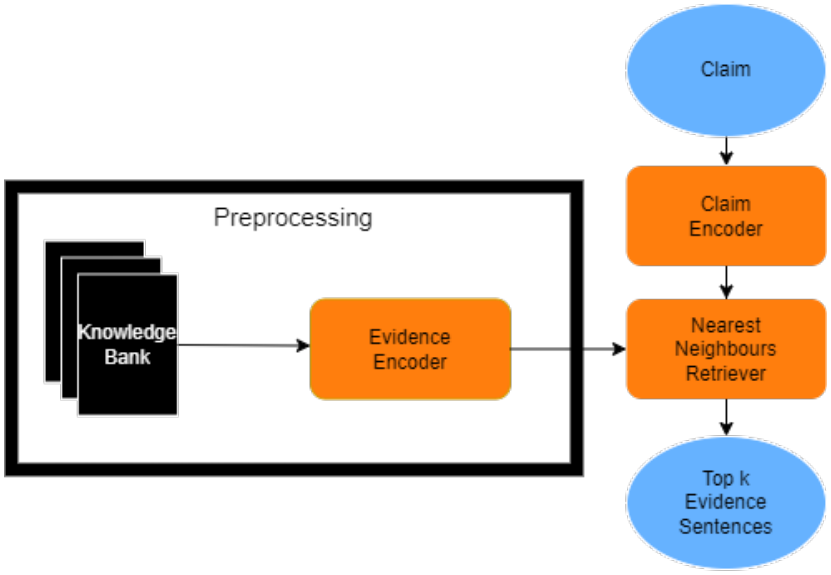

Figure 3.2: Dense Retrieval System

One method we use to generate these embeddings is the `sentence_transformers`

Python library [49]. This method allows us to generate a rich fixed-length vector representation that captures the sentence's overall meaning instead of the meanings of individual words. This is useful for our task, as we can recognise synonyms and other similarities in meanings in complex scientific literature that we cannot identify using a lexical search alone.

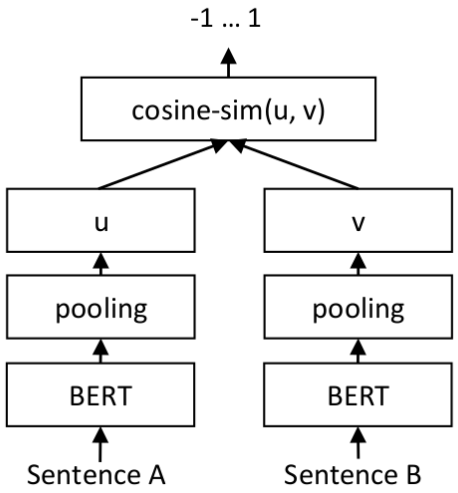

Figure 3.3: Diagram from "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks" [49] showing the architecture of using sentence transformers to find the similarity of two sentences.

Sentence transformer models work by adding a pooling layer to the output of a base transformer model, such as BERT [10], to create a fixed-length vector representation for a sentence. Sentence transformer models are then fine-tuned to ensure that the embeddings are meaningful and representative of the entire sentence, so they can be compared using cosine similarity. They are often fine-tuned using "Siamese networks" [42], which are two parallel neural networks that learn to assign similar input pairs with similar representations in the embedding space, while distancing pairs that are not similar.

Sentence transformers are proven to significantly improve the efficiency of searching for similar sentences in an evidence base [49] when compared to other BERT-based approaches, and they also exhibit increased performance on many benchmarks. Consequently, we employ them for our specific task, capitalising on their impressive capabilities. Figure 3.3 shows a visual representation of comparing two sentences using a sentence transformer, in this case using a (fine-tuned) BERT as the base model.

We analysed the performance for the following sentence transformer models, which consist of the most popular models on HuggingFace [14]:

- all-mpnet-base-v2 [49]
- all-roberta-large-v1 [50]
- all-MiniLM-L12-v2 [47]
- bge-base-en-v1.5 [79]

- all-MiniLM-L6-v2 [48]
- e5-large-v2 [72]
- gte-large [33]

As well as using these sentence transformer models, we also implemented dense retrieval methods via the "Contriever" [28] approach, aiming to further enhance our results. Contriever is a family of dense retrieval models developed by Meta that use contrastive learning in the training process. For a more in-depth description of this, see Appendix A.1. We analyse the performance of the base "Contriever" model and "Contriever-msmarco" for our evidence retrieval task. "Contriever-msmarco" is the base contriever model fine-tuned on the "msmarco" dataset [43], which is a well-known dataset consisting of many question and answer pairs.

3.4.3 Evaluation Methods

To analyse how these methods perform for our specific task and to find the best method for our system, we set aside 50 claims from the CLIMATE-FEVER dataset. For each claim sentence, we employ our sparse and dense methods to find potential evidence sentences for the claim from the IPCC AR6 knowledge base. Each method returns the 3 highest ranked evidence sentences it retrieves for each claim, so we have 150 claim/evidence pairs for each method. We then manually annotate these as "ENOUGH_INFO" or "NOT_ENOUGH_INFO" based on whether the evidence sentence contains enough information to support/refute the claim or not. Details of this manual tagging process can be found in Appendix A.9. We can then compare the methods on two metrics:

- **Backed up claims** - The percentage of the 50 test claims with at least 1 piece of evidence supporting or refuting the claim.
- **Relevant evidence** - The percentage of the 150 retrieved evidence sentences that supported or refuted their corresponding claim.

It should be noted that the test size of 50 claims, and the sample size of 3 evidence sentences per claim, were selected due to the time-consuming process of manually annotating the retrieved sentences, and interpreting whether or not they can support/refute their corresponding claim.

3.4.4 Results

Table 3.1 shows the performance of each evidence retrieval method for this task after the retrieved evidence had been manually annotated. Our sparse method (BM25) has the lowest performance across both metrics, and our dense retrieval methods significantly outperform it. The large disparity between the performance of BM25 and most other models is possibly due to the scientific nature of the language in the evidence corpus. It is likely that many test claims use different phrasing to the related evidence sentences in the IPCC AR6, which BM25 is unable to consider since it focuses purely on the terms/words whereas the dense models generate a meaningful representation of the sentence.

Model	Method	Backed Up Claims	Relevant Evidence
BM25	Sparse	0.38	0.18
all-mpnet-base-v2	Dense	**0.78**	**0.47**
all-MiniLM-L12-v2	Dense	0.64	0.37
all-roberta-large-v1	Dense	0.68	0.4
bge-base-en-v1.5	Dense	0.58	0.33
all-MiniLM-L6-v2	Dense	0.62	0.35
e5-large-v2	Dense	0.44	0.2
gte-large	Dense	0.6	0.35
contriever	Dense	0.58	0.28
contreiver-msmarco	Dense	0.68	0.41

Table 3.1: Results for different evidence sentence retrieval models

The highest-performing model across both metrics, all-mpnet-base-v2, is a sentence transformer model that converts a given sentence to a 768-dimensional vector. This is a version of Microsoft's MPNet model [56], a transformer-based model that uses Permuted Language Modelling [77], to improve natural language understanding. See Appendix A.1 for more details. This has been fine-tuned on a set of over 1 billion pairs of sentences to create all-mpnet-base-v2, thereby enabling it to generate a rich and meaningful vector representation of a sentence.

To ensure significance in our tests for determining the best evidence retrieval model, we compared the results for our best model, all-mpnet-base-v2, to the results for every other model using a McNemer test [37] on the **backed up claims** metric. In doing so, we generate contingency tables based on which claims were backed up for each model. The construction of these contingency tables can be seen in Figure 3.4, where "correct" means valid evidence was found for the claim. We then compare the results of the highest-performing model with every other model by calculating a p-value using the McNemer test formula:

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{3.3}$$

Where χ^2 represents the p-value and b and c are quadrants from the contingency table shown in Figure 3.4.

Table 3.2 shows the p-values comparing the best-performing model to all other models on the **backed up claims** metric. The p-values represent the probability that the observed difference between the best-performing model and each other model is due to chance rather than a true difference in performance. If the p-value comparing 2 models is less than α, we can be confident that this is a statistically significant difference. We take $\alpha = 0.05$, so we can see there is no statistical significance between the performance of all-mpnet-base-v2 and 3 other models (all-MiniLM-L12-v2, all-roberta-large-v1, contriever-msmarco).

This means that any of these 3 models could perform just as well as all-mpnet-base-v2 for the task. For simplicity's sake, we will continue utilising all-mpnet-base-v2 for

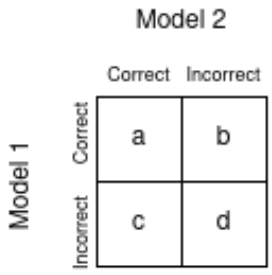

Figure 3.4: Contingency table for McNemer test for statistical significance. Model 1 is our best-performing model, and model 2 is the model we are comparing it to.

Model	P-value
BM25	8.8e-5
all-MiniLM-L12-v2	0.065
all-roberta-large-v1	0.267
bge-base-en-v1.5	0.006
all-MiniLM-L6-v2	0.039
e5-large-v2	0.0002
gte-large	0.022
contriever	0.021
contreiver-msmarco	0.226

Table 3.2: P-values comparing best model (all-mpnet-base-v2) to all other models on **backed up claims** metric.

dense retrieval within our system's framework, even though the other models would likely perform equally well.

In addition to sparse and dense methods, we also implemented and tested a hybrid approach, where we assign each evidence sentence a score based on a weighted sum of its sparse and dense scores. This approach has been found to perform better than dense retrieval in some cases [34], as it combines the ability of dense retrieval to encapsulate the semantic meaning of a sentence, along with the preciseness of term-based search methods. Unfortunately, we found these to be ineffective in our system. More details can be found in A.8.

3.5 Evidence Re-Ranker

3.5.1 Task Description

In this section, we explore methods to improve our evidence retrieval system by making use of a re-ranker model. Re-rankers have been shown to improve the performance of existing evidence retrieval systems [50], and we can train climate-specific models for this task, motivating the inclusion of this component. Our dense retrieval system can efficiently retrieve similar sentences using our pre-calculated embeddings, so we will explore more computationally expensive methods to re-rank the top k sentences from our dense retrieval system, in an attempt to find the best evidence sentences from the k most similar sentences. We take k as 30 in our case, as using the re-ranker for inference on a large set of inputs becomes slow and inefficient.

3.5.2 Architecture

For each of our top k evidence sentences from the sentence retrieval system, we concatenate the claim with the evidence sentence and then input this concatenated claim/evidence pair into a binary classifier. This classifier is trained to classify between "Relevant Evidence" and "Irrelevant Evidence". The model's confidence in the "Relevant Evi-

dence" class will be the new score for the claim/evidence pair. This process is illustrated in Figure 3.5. The sentences that have higher confidence for "Irrelevant Evidence" than "Relevant Evidence" will be removed and not considered.

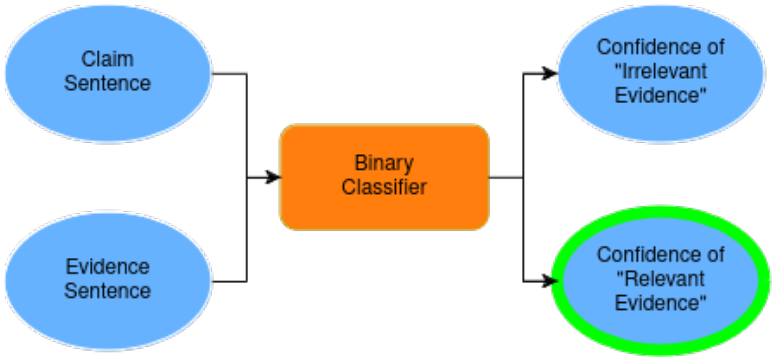

Figure 3.5: Architecture for calculating re-ranking scores of claim/evidence pairs. The green border indicates the new score used to re-rank the evidence sentences.

This is a computationally expensive process as we use transformer-based models to encode every claim/evidence pair, which is then passed on to a classification head which predicts the output class. For this reason, it would be infeasible to use this system to rank all evidence in the knowledge base, and we only use it for the top k sentences. This approach was taken because BERT-based models are shown to be highly effective in understanding and processing natural language [36], and we can fine-tune them for this climate-specific task. Also, as described in Figure 3.1, this is not a "core" component of the system, so we will assess the performance of the system with and without the re-ranker.

RoBERTa-base [36] is one of the models we fine-tune for the binary classification task due to its ability to take multiple input sentences for the claim and evidence and its proven track record in capturing complex language representations, which are key for understanding the nuances in claim/evidence pairs. Also, it can be easily fine-tuned for classification using `RobertaForSequenceClassification` from the HuggingFace `transformers` library [75]. We also fine-tuned ClimateBert [72], which in itself is a version of RoBERTa which has been fine-tuned on climate-related text. This choice aimed to evaluate if a model precisely fine-tuned for climate-related tasks outperforms a non-domain-specific counterpart.

3.5.3 Training

We fine-tune RoBERTa and ClimateBert using 3 different levels of fine-tuning:

- FEVER data alone
- FEVER and CLIMATE-FEVER data
- FEVER, CLIMATE-FEVER, and tagged claim/evidence pairs gained from evaluating the sentence retrieval models

The training hyperparameters used can be found in Appendix A.15.1. In total, this gives us 6 re-ranker models to evaluate. These training sets were constructed as follows:

FEVER: To generate training data for the re-ranker from FEVER, we extract all claim/evidence pairs which are labelled as "SUPPORTS" or "REFUTES" and label them as "Relevant Evidence". We disregard all pairs labelled "NEI", as these evidence sentences do not directly support or refute their respective claim. We then generate an equal number of claim/evidence pairs to be labelled "Irrelevant Evidence". These pairs consist of FEVER claims and a random evidence sentence for another claim. This provides the model with information on the difference between relevant and not-relevant evidence to a claim.

CLIMATE-FEVER: We generate training data from the CLIMATE-FEVER training set using the same method as described for FEVER. The majority of pairs in CLIMATE-FEVER are labelled as "NEI", and since we remove these from the re-ranker training set, the amount of training data gained from CLIMATE-FEVER is relatively limited.

Annotated Pairs From Sentence Retrieval: During the process of assessing the performance of the various sentence retrieval methods in Section 3.4.3, retrieved evidence sentences for the test claims were manually annotated as "ENOUGH_INFO" or "NEI". These annotated claim/evidence pairs were saved and used as training data for the re-ranker model. A considerable majority of pairs were labelled as "NEI", so these were downsampled to ensure an equal class split in this training set. Although this training set is smaller than the other two, the evidence sentences in this training set are directly from the IPCC AR6, so this may prove valuable for adapting the system to our domain.

3.5.4 Evaluation Methods

To assess the performance of the re-ranker models, we create a new test set of 50 claims from the CLIMATE-FEVER test set. A new test set is used here to ensure there is no crossover between training and test data, as we use the annotated pairs from the previously used test set as training data. For each claim, we use our sentence retrieval system (all-mpnet-base-v2) to retrieve the top 30 evidence sentences from the IPCC AR6, and then each model re-ranks these evidence sentences. Similar to how the retrieval models were evaluated, we manually annotate the top 3 re-ranked claims for each model for each claim sentence. The claim/evidence pairs are annotated as "ENOUGH_INFO" if the evidence sentence supports or refutes the claim, or "NEI" otherwise. Again, we assess each model on both the **backed-up claims** and **relevant evidence** metrics, as defined in Section 3.4.3.

3.5.5 Results

Table 3.3 shows the results on both metrics for the manual annotations for every re-ranker model, and compares these to dense retrieval without re-ranking. For our new set of 50 claims, all-mpnet-base-v2 performed slightly worse than the previous test set, and retrieved relevant evidence for only 64% of claims, as opposed to 78% of claims in the original test set.

Model	Backed Up Claims	Relevant Evidence
No Re-ranker	0.64	0.42
RoBERTa-F	**0.68**	0.41
RoBERTa-F-CF	0.64	0.39
RoBERTa-F-CF-IPCC	0.62	0.42
ClimateBert-F	0.52	0.28
ClimateBert-F-CF	0.62	0.42
ClimateBert-F-CF-IPCC	**0.68**	**0.47**

Table 3.3: Results for different evidence re-ranker models. An explanation of each model name can be found in Appendix A.6.

Furthermore, it can be seen that re-ranking using ClimateBert fine-tuned on FEVER, CLIMATE-FEVER, and the annotated IPPC report had a slightly better performance than all-mpnet-base-v2 alone across both metrics, however, this difference was relatively small. Interestingly, ClimateBert fine-tuned on FEVER performed the worst by a wide margin. This may be due to conflicting domains, as ClimateBert is primarily intended for use on climate-specific text, whereas FEVER is not climate-specific, so fine-tuning ClimateBert on just FEVER may have had adverse effects on the domain specificity of ClimateBert.

Similarly to assessing the performance of the evidence retrieval models, we used a McNemer test for statistical significance on the **backed-up claims** metric to check whether our re-ranker models were a significant improvement over the base evidence retrieval system. In doing so, we found that in comparing the result for our best re-ranker (ClimateBert-F-CF-IPCC) and our result for using no re-ranker (all-mpnet-base-v2), we gained a high p-value of 0.754, meaning we do not have a statistically significant improvement. Since we do not have a statistically significant improvement from using a re-ranker, we cannot conclusively say that the use of our re-ranker is beneficial to the overall system. Despite the lack of statistically significant improvement, the re-ranker's performance did not deteriorate compared to the dense retrieval system, indicating it does not negatively impact the system.

Our tests did not definitively confirm that any of our re-ranker models outperformed the dense retrieval system in retrieving relevant sentences. A potential reason for this is the nature of the knowledge base. For some test claims, there may be no valid supporting or refuting evidence in the IPCC AR6, especially if the claim is not related to climate science. Consequently, for these claims, it will be impossible for either the dense retrieval system alone or the re-ranker to retrieve valid evidence. The inverse of this is also true for some claims, where there may be an abundance of valid evidence sentences in the knowledge base, so both the re-ranker models and the dense retrieval model find multiple valid evidence sentences. Despite this, we will investigate the performance of our highest-performing re-ranker across both metrics, ClimateBert-F-CF-IPCC, as a component within the overall system.

Chapter 4

Evidence Classification

4.1 Methods

4.1.1 Task

The evidence classification task is as follows: Given a claim sentence and an evidence sentence, we must classify the claim/evidence pair into the categories "SUPPORTS", "REFUTES", or "NEI" ("NOT_ENOUGH_INFO"), based on whether the evidence supports or refutes the claim, or if the claim and evidence are unrelated. As done by Wang et al. (2021) [69], we extract a test of 95 random claims from CLIMATE-FEVER, each with 5 evidence sentences. This gives us a test set of 475 labelled claim/evidence pairs to evaluate our systems. We use the rest of CLIMATE-FEVER as a training set for our various models.

4.1.2 Models

This research explores the viability of using generative Large Language Models (LLMs) for evidence classification. Our approach involves prompting LLMs by describing the task of classifying a claim/evidence pair, and then providing the claim and evidence sentences to be classified. The LLM then generates the predicted class as part of its response. We also explore multiple techniques to improve the performance of the LLMs for the classification task.

For this task, we make use of the following Large Language Models:

- **Llama-2:** Llama-2 [64] is a family of open-source Large Language Models developed by Meta, which were pre-trained on 2 trillion tokens, and outperformed most other open-source LLMs on many benchmarks at the time of creation. Llama-2 comes in 3 size variants, 7 billion parameters (Llama-2 7b), 13 billion parameters (Llama-2 13b), and 70 billion parameters (Llama-2 70b). In this research, we make use of Llama-2 7b and Llama-2 13b, as we do not have access to the computational resources required for Llama-2 70b. These models were accessed via the HuggingFace hub [14]. A benefit of using Llama-2 is the carbon emissions saved by using smaller models, as explained in A.19.

- **GPT-4 Turbo:** Created by OpenAI, GPT-4 [44] is a multimodal Large Language Model which achieves state-of-the-art results for many benchmarks [1]. The exact size of this model is unknown; however, it is likely much larger than GPT-3, which is made up of 175 billion parameters [17]. This model was accessed via the OpenAI API [45].

4.1.3 Task Adaptation Techniques

We investigate the performance of various models for this classification task, as well as exploring multiple techniques for improving their techniques by adapting them for this specific task:

Base Models: We assess the performance of Llama-2 7b, Llama-2 13b, and GPT-4 Turbo out-of-the-box, without any additional context. We carry this out as a baseline, to assess how our methods can improve their performance.

In-Context Learning: We then attempt to improve the performance of these base models using "In-Context Few-Shot learning" [70], where we provide k examples of correctly classified claim/evidence pair in the prompt. This has been shown to improve the performance of LLMs for many classification tasks [16, 12]. For the Llama-2 models, we take $k = 9$, to provide 3 examples of each class. For GPT-4, we take $k = 3$, since the OpenAI API incurs cost per additional token in the prompt.

Paremeter-Efficient Fine-Tuning: We further attempt to improve the performance of Llama-2 7b and Llama-2 13b by fine-tuning them, using training examples constructed from the FEVER and CLIMATE-FEVER datasets. We employ an efficient method for fine-tuning these models with limited resources, as described below.

4.1.4 Parameter-Efficient Fine-Tuning

Due to the scale and computational cost of fine-tuning large language models, we make use of the HuggingFace `peft` (Parameter-Efficient Fine-Tuning) [39] library which provides methods to significantly decrease the computational cost and time of fine-tuning by only training a relatively small number of extra parameters. This allows us to fine-tune large-scale models with up to 13 billion parameters at a relatively low cost.

The technique provided in `peft` that we use for fine-tuning our Llama-2 models is LoRA (Low-Rank Adaption) [21]. LoRA works by freezing the pre-trained weights of the model, and adding rank decomposition matrices at each layer within the transformer. Rank decomposition matrices are a pair of low-rank matrices which, when multiplied together, can be used to adjust the original weight matrix. This can simulate the effects of fine-tuning the weights in the original weight matrix itself. These rank decomposition matrices are the only weights that are updated during the fine-tuning process, which greatly reduces the cost of training as the number of trainable parameters in these matrices can be up to 10,000 times smaller than the base model itself, greatly reducing the amount of GPU memory and processing power required. Figure 4.1 shows a visual representation of the weight matrix update in a transformer layer using LoRA, where A and B are the rank-decomposition matrices, and W is the original frozen weight matrix.

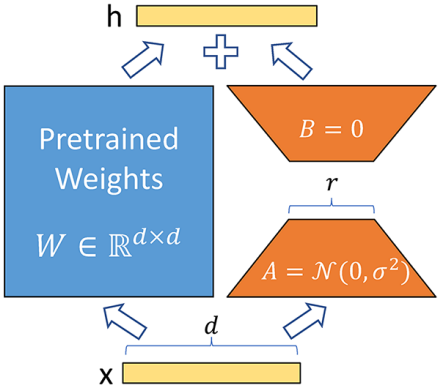

Figure 4.1: Visual Representation of a weight matrix in the transformer architecture during fine-tuning with LoRA [21]

To further reduce the amount of GPU memory required, we also employ "QLoRA" techniques (Quantized LoRA) [9], which introduces memory-saving techniques without sacrificing performance. This is achieved by quantizing the trainable parameters in the LoRA rank-decomposition matrices. This means that we map each value in the matrices to a set of discrete values, reducing their precision and memory footprint. To ensure we do not lose information during this quantization, we take advantage of the NF4 (4-bit NormalFloat) datatype, which reduces the memory requirement of each weight to 4 bits using quantization, and is "information theoretically optimal for normally distributed weights". This is ideal in our case, as the weights in pre-trained neural networks often follow a normal distribution [9]. Furthermore, we use "double quantization" which reduces memory usage by an average of 0.37 bits per weight by further quantizing the "quantization constants" to 8 bit-values. The quantization constants are a collection of 16 32-bit values mapped to each weight during quantization. These techniques are shown to reduce the computational resources required for fine-tuning greatly.

This is greatly beneficial as it allows us to fine-tune large and complicated models for this specific task, potentially leading to large performance increases, with limited resources. The hyperparameters used in this fine-tuning can be found in Appendix A.15.2.

4.1.5 Training and Evaluation Methods

4.1.5.1 Testing Methods

Using the test set of 475 claim/evidence pairs from CLIMATE-FEVER, we generate test prompts to assess the LLMs. These test prompts provide a task description along with the claim and evidence sentence, and the exact prompt can be seen in Appendix A.11.1. The prompts are formatted in a way where the last token generated by the model should be its class prediction.

It has been shown that the prompt format used can drastically impact the performance of an LLM [74], so we trialled multiple prompts on a small test set before deciding on the final prompt template. The alternate prompt templates which were not used in the final system, along with their performance, can be found in Appendix A.11.

We can then test the model by providing it with these test prompts, so it generates the predicted label as the next token. We can then extract this predicted label from the generated text, and compare it to the true label. Using these predicted labels for each prompt, we can measure the performance of each model. During inference, we employ "greedy decoding" [18], a method that guarantees the same prompt will always receive the same classification, since the models only generate the label. See Appendix A.1 for more details.

During the assessment of the "out-of-the-box" models, this test prompt is all the model is provided with. For the "In-Context Learning" setting, we modify this test prompt to contain k examples of correctly classified claim evidence pairs, using the same format. Creating the fine-tuned models is more complicated, as we first need to generate training prompts to fine-tune the models, before we can assess their performance using the test prompts.

As described the the original CLIMATE-FEVER paper [11], the real-life nature of the claims increases the difficulty of classifying whether evidence supports or refutes a claim. We conducted experiments to verify this, and more details can be seen in Appendix A.10.

4.1.5.2 Training Data

In creating our fine-tuned models, we make use of 2 datasets: FEVER and the training portion of CLIMATE-FEVER. We use the labelled claim/evidence pairs from these datasets to construct training prompts, the format of which can be seen in Appendix A.11.1. These training prompts contain the claim sentence, evidence sentence, and label, where the label is the final token. We can then fine-tune LLMs using the `peft` [39] library, and `SFTT` (Supervised Fine-Tuning Trainer) from the `TRL` [73] library. This allows the LoRA adapters to learn appropriate weights for our specific task. After each model is trained, we merge the adapter weights with the original LLM weights and upload the merged model to HuggingFace.

The proportions of each class in the CLIMATE-FEVER dataset are very unbalanced, with $\sim 64\%$ of the claim/evidence pairs being labelled as "NEI". To account for this, we explore models fine-tuned on the following sets of data:

- **CLIMATE-FEVER with Random Resampling:** We randomly resample the minority classes in CLIMATE-FEVER to ensure an equal class balance in the training data. This method intends to ensure we only use domain-specific examples as training data. Models fine-tuned with this data are indicated with "*FT:* CF".
- **FEVER with Equal Class Balance:** FEVER is an extremely large dataset, and it would require a huge amount of computational resources to fine-tune an LLM on the whole dataset. Instead, we sample 2000 training examples of each class from FEVER to use as a training set. Models fine-tuned with this data are indicated with "*FT:* F".
- **FEVER + CLIMATE-FEVER:** We employ an alternate technique to balancing

the classes in CLIMATE-FEVER, where we top-up the minority classes with samples of the same label from FEVER. This ensures that each training example is unique, while maintaining CLIMATE-FEVER's domain-specific properties. Models fine-tuned with this data are indicated with "*FT:* CF+F".

GPT-4 is not open-source, so it cannot easily be fine-tuned. Therefore, we only assess this model with few-shot in-context learning and without context.

4.2 Results

Base Model	Task-Adaption Method	Label Accuracy	Precision	Weighted F1-Score	Macro F1-Score
GPT-4 Turbo	N/A	0.73	0.73	0.73	0.64
GPT-4 Turbo	*ICL:* $k = 3$	0.72	0.73	0.73	**0.65**
Llama-2 7b	N/A	0.30	0.20	0.20	0.20
Llama-2 7b	*ICL*: $k = 9$	0.53	0.58	0.54	0.44
Llama-2 7b	*FT:* CF	0.65	**0.74**	0.66	0.60
Llama-2 7b	*FT:* F	0.42	0.63	0.48	0.38
Llama-2 7b	*FT:* CF+F	0.72	0.71	0.70	0.60
Llama-2-13b	N/A	0.63	0.60	0.61	0.44
Llama-2 13b	*ICL:* $k = 9$	0.57	0.61	0.59	0.50
Llama-2 13b	*FT:* CF	**0.76**	**0.74**	**0.75**	0.64
Llama-2 13b	*FT:* F	0.63	0.64	0.62	0.52
Llama-2 13b	*FT:* CF+F	0.74	**0.74**	0.72	0.61

Table 4.1: Performance of different classification models on the test set. "*ICL*" indicates k-shot In-Context Learning is used. "*FT*" indicates that Parameter Efficient Fine-Tuning is used.

Table 4.1 provides each model and method's label accuracy, precision, and weighted and unweighted (macro) F1-scores. The definitions for these metrics can be found in Appendix A.7.

We can see that "Llama-2 13b *FT:* CF" performs slightly better than GPT-4 turbo with and without few-shot learning for label accuracy, precision, and weighted F1-score. This is an impressive result, given that Llama-2 13b likely has a fraction of the parameters of GPT-4. It is also interesting to note that "Llama-2 13b *FT:* F" performs significantly worse across all metrics than the variant fine-tuned on CLIMATE-FEVER alone, and the variant fine-tuned on FEVER and CLIMATE-FEVER. This may be caused by the artificial nature of the claims in FEVER, which contrasts with the real-life and scientific claims in the CLIMATE-FEVER test set.

Both Llama-2 7b and Llama-2 13b show a dramatic performance improvement after fine-tuning. A particularly noteworthy example is the improvement gained when fine-tuning Llama-2 7b on CLIMATE-FEVER and FEVER, where we gain a 0.50 increase in weighted F1 and a 0.40 increase in unweighted F1 compared to the out-of-the-box model.

The improvement gained from employing in-context learning can be observed when we compare the performance of Llama-2 7b with and without this technique, where we gain a significant performance increase across all metrics. However, this method does not improve the performance to the same extent as fine-tuning. Interestingly, for Llama-2 13b, In-Context Learning provides little, if any improvement, suggesting that this technique may be more advantageous for smaller models.

4.3 Evaluation

4.3.1 Test Set Inspection

In the CLIMATE-FEVER test set, the proportion of claim/evidence pairs with the "NEI" label is much larger than "SUPPORTS" or "REFUTES". Some claims in the dataset are so specific or so vague that all 5 of their evidence sentences are labelled as "NEI". For example, all evidence sentences for the following claims are labelled "NEI":

- **Claim 1:** *Not only was 2016 the warmest year on record, but eight of the 12 months that make up the year - from January through September, with the exception of June - were the warmest on record for those respective months.*
- **Claim 2:** *A World Heritage site, it is currently under assault from unusually hot ocean temperatures*

Since Claim 1 contains many specific details about which individual months were warm for a specific year, most evidence sentences will not contain enough information to support or refute the claim explicitly, so they are all labelled "NEI". Claim 2 is very vague, as we are unsure of the context of the claim as to what specific "world heritage site" it is referring to, so every evidence sentence is labelled "NEI". This also provides an example of the nuances and difficulty of this classification task, as it is often up for interpretation whether an evidence sentence supports/refutes a claim or not.

The large proportion of "NEI" labels in the test set can make it difficult to assess the performance of a model, as a model which predicts "NEI" too often will achieve a high accuracy. We perform a more in-depth error analysis for several of our models to account for this.

4.3.2 Error Analysis

Figures 4.2, 4.3, 4.4 show confusion matrices for several of our models on the test set. The test set imbalance becomes more obvious from these confusion matrices, with the large number of samples labelled "NEI".

We can see in Figure 4.4 and 4.2 that GPT-4 performs similarly to "Llama-2 13b *FT:* CF", but correctly identifies refuting evidence more often. It misclassifies between "SUPPORTS" and "REFUTES" slightly more often, with 6 in total, and has an almost identical performance as "Llama-2 13b *FT:* CF" on the "SUPPORTS" class.

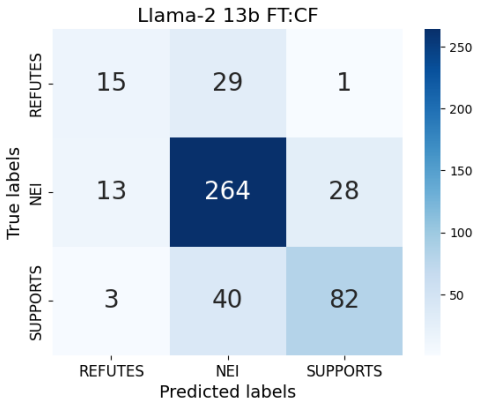

Figure 4.2: Confusion Matrix for Llama-2 13b fine-tuned on CLIMATE-FEVER

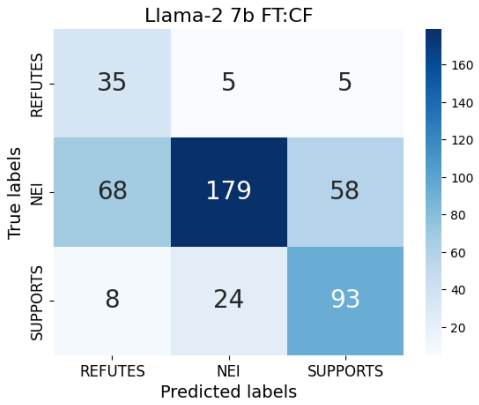

Figure 4.3: Confusion Matrix for Llama-2 7b fine-tuned on CLIMATE-FEVER

From Figure 4.2, we can see that "Llama-2 13b *FT:* CF" tends to predict "NEI" too often, even with the large proportion of this class in the test set. This model also rarely predicts the "REFUTES" class, and only correctly predicts it 15 times out of a possible 45. This model achieved the highest label accuracy and weighted F1-score, and upon inspection, this is likely due to the abundance of the "NEI" class in the test set. We can also see that this model rarely misclassified "SUPPORTS" as "REFUTES" and vice versa, as it only produces this error 4 times.

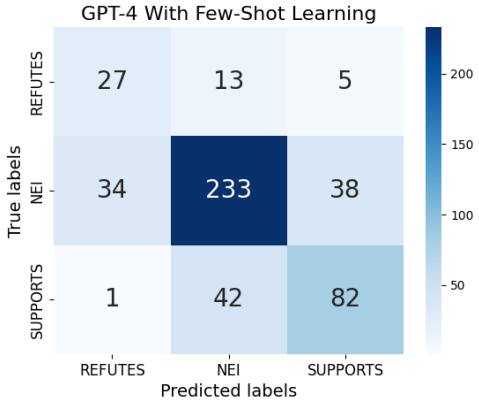

Figure 4.4: Confusion Matrix for GPT-4 with In-Context Learning, $k = 3$

In contrast to this, from Figure 4.3, "Llama-2 7b *FT:* CF" predicts the "SUPPORTS" and "REFUTES" classes more frequently. It often predicts "SUPPORTS" or "REFUTES" when the correct label is "NEI", however, it correctly predicts these classes more frequently than "Llama-2 13b *FT:* CF", despite predicting them too often. This model misclassifies between the "SUPPORTS" and "REFUTES" class frequently, with 13 times in total.

4.3.3 Implications for Overall System

From these tests, we gained several candidates for the evidence classification component of our system. The highest-performing model in Table 4.1 is Llama-2 13b *FT:* CF. However, after inspecting the confusion matrices, we can see some flaws in some of our higher-performing models, being that they predict "NEI" too often. With this insight, we will assess the system performance with our highest-performing models created in this section to assess what interacts more effectively with the rest of the system.

Chapter 5

Overall System

5.1 Overall System Description

5.1.1 Architecture

In this section, we describe the architecture of the overall claim verification system, comprised of the highest-performing components from our previous tests. We also test different combinations of these components to find the highest-performing system for our task.

Climate Claim Detection: As described in Section 3.3, we use ClimateBert fine-tuned for climate-related text detection. This is a binary classifier, which is used to check whether a given claim is climate-related or not.

Evidence Sentence Retrieval: As described in Section 3.4, after extensive testing of multiple evidence sentence retrieval methods, it was concluded that the most appropriate method for this specific task is dense retrieval, using the model "all-mpnet-base-v2". We do not incorporate any sparse or hybrid retrieval methods, as we found that these methods perform worse for this specific task.

Evidence Re-ranker: As described in Section 3.5, our tests were inconclusive in determining whether any of our proposed re-ranker models improved upon the performance of our evidence retrieval system. Due to this, we will assess the overall system's performance both with and without the re-ranker component. The re-ranker model used in the final architecture is "ClimateBert-F-CF-IPCC" since it achieved the highest score across both metrics. The exact details of this model can be seen in Appendix A.6.

Evidence Classification: As described in Chapter 4, we use fine-tuned variations of Meta's Llama-2 to classify whether an evidence sentence supports or refutes a claim sentence. We have multiple candidate models for this task, and we will measure the system's performance using several of these models.

Figure 5.1 shows a visual representation of this overall system.

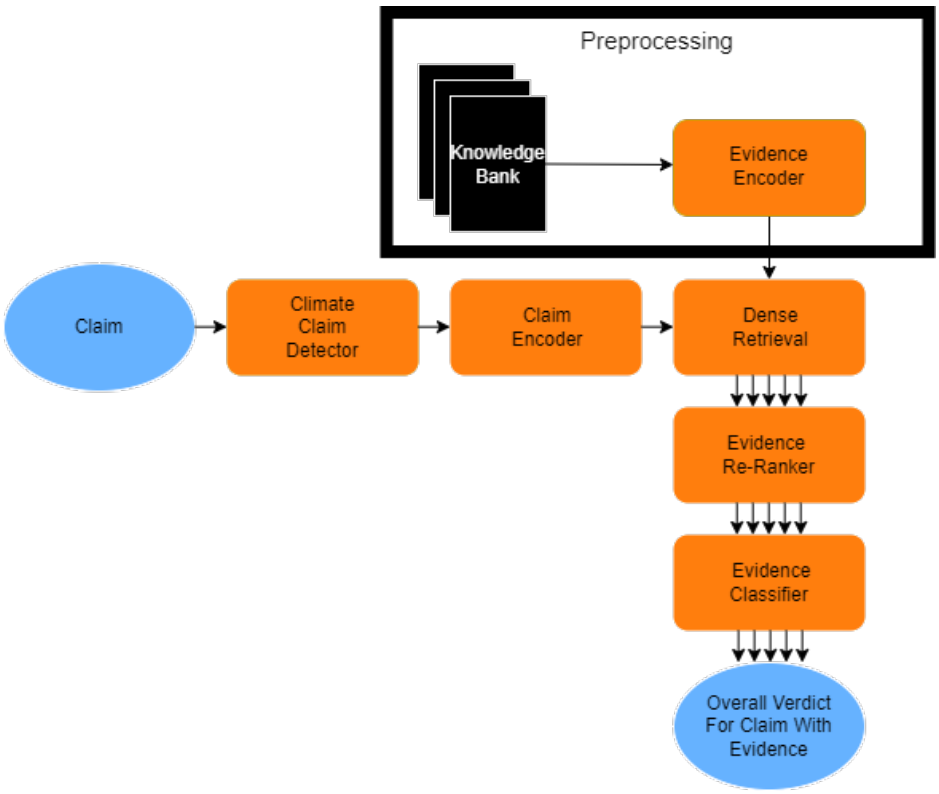

Figure 5.1: Overall Claim Verification System with All Components

5.1.2 Method of Verifying Claims

The goal of the overall system is to support or refute a claim while providing the evidence used to come to this decision. For each claim, the system retrieves the top 5 pieces of evidence, and decides if a claim is "SUPPORTED" or "REFUTED" by the IPCC AR6. We introduce the "Unable To Verify" ("UTV") label, which is assigned to claims that the system cannot support or refute. For a claim to be classified as "SUPPORTED" or "REFUTED", the following conditions must hold:

- The claim must be classified as a "Climate Claim" by the climate claim detector component
- The claim must have at least one piece of evidence classified as "SUPPORTS" or "REFUTES"
- The claim must not have any evidence that disputes each other, that is, if one evidence sentence supports the claim, then another evidence sentence cannot refute the claim, and vice versa.

If these conditions are not met, the claim is classified as "UTV". This differs from the "NEI" evidence sentence label, as a claim is tagged as "UTV" if **all** of its evidence sentences are labelled "NEI" **or** if it has both supporting and refuting evidence. If the conditions are met, the claim is classified as "SUPPORTED" or "REFUTED", depending on whether the evidence supports or refutes it. This is to prevent situations where a claim has both supporting and refuting evidence, as in these cases, we cannot

be confident that the model's evidence classifications are correct.

5.2 Testing of Overall System

5.2.1 Test Data

The motivation behind this research is to aid in reducing the spread of online misinformation relating to climate change. Because of this, we deemed it appropriate to evaluate the model's ability to verify claims on a test set of real climate-related claims on "X" (formerly Twitter). This test set consists of 50 climate-related claims found by searching X for climate-related keywords. This simulates a potential real-world usage for such a system. Examples from this test set can be seen in Appendix A.13.1.

Each of these claims was assigned a "veracity" label, which is either "True" (not misinformation) or "False" (misinformation). These are used to evaluate the system's performance, therefore, for many claims, it was necessary to find sources to support or refute them manually. For example:

- **"With melting glaciers and ice caps in Greenland and Antarctica sea levels will rise"** - This claim required in-depth verification to be assigned its veracity label of `True`. Supporting evidence for this claim was found in the IPCC AR6.
- **"Climate change is fake and stupid."** - This claim did not require in-depth verification, as it is subjective and non-factual, so it was assigned a veracity label of `False`.

5.2.2 Testing Approach

We assessed several system architectures using 4 different classifier models, and for each one, we assessed the system both with and without the re-ranker component. To measure the performance of a system architecture, we run every test claim through the system, to produce a label. We then assess the accuracy of each system-generated label by comparing it to the claim's veracity label, classifying it as correct if the labels align ("SUPPORTED" with "True" or "REFUTED" with "False"). The system architectures will then be assigned a score, calculated as follows:

$$\text{Score}(A) = \frac{1}{n} \sum_{c_i \in C} F(A, c_i)$$

$$\text{F(A, c}_i) = \begin{cases} +1 & \text{if } p(A, c_i) = g(c_i) \\ 0 & \text{if } p(A, c_i) = \text{"UTV"} \\ -1 & \text{if } p(A, c_i) \neq g(c_i) \text{ and } p(A, c_i) \neq \text{"UTV"} \end{cases} \tag{5.1}$$

where:

- A = system architecture
- C = set of claims $c_1, \ldots, c_n$
- n = number of claims in C
- $p(A, c_i)$ = prediction of the architecture A for claim c_i
- $g(c_i)$ = gold label of claim c_i
- $F(A, c_i)$ = Prediction accuracy for architecture A on claim c_i
- "UTV" = "Unable To Verify"

This metric ensures that the system is more heavily penalised when incorrectly supporting or refuting a claim, as in a real setting, this could be dangerous for the spread of misinformation. An "incorrect verification" occurs when the system supports a claim that contains misinformation or refutes a claim that does not. This gives us a score in the range of $[-1, 1]$ based on the ability of a system to predict the veracity of claims. This scoring metric is to allow us to compare our potential system architectures easily.

5.2.3 Results

Classifier Model	Re-ranker	Correct	UTV	Incorrect	Score	Label Accuracy	F1 Score
Llama-2 13b *FT:* CF	No	21	29	0	0.42	0.42	0.52
Llama-2 13b *FT:* CF	Yes	22	28	0	0.44	0.44	0.55
Llama-2 13b *FT:* CF+F	No	29	17	4	0.5	0.58	0.69
Llama-2 13b *FT:* CF+F	Yes	29	19	2	0.54	0.58	0.71
Llama-2 7b *FT:* CF	No	34	14	2	0.64	0.68	0.79
Llama-2 7b *FT:* CF	Yes	38	10	2	**0.72**	0.76	0.84
Llama-2 7b *FT:* CF+F	No	38	10	2	**0.72**	0.76	0.84
Llama-2 7b *FT:* CF+F	Yes	37	11	2	0.70	0.74	0.83

Table 5.1: Performance of each architecture on the test set of X claims. "UTV" stands for "Unable To Verify". An explanation of the Classifier Model names can be found in Section 4.1.5.2. The F1 score used is the weighted average of the "SUPPORTS" and "REFUTES" class, as no gold labels are "UTV".

Table 5.1 shows the score for each model, with and without the re-ranker component, for the test set of claims. It should be noted that every claim in the test set was classified as "Climate-Related" by the climate-claim detection system, so we do not assess the systems without this component. A further breakdown of the reasons each system returned "UTV" and how often they predict each class can be found in Appendix A.16. We include label accuracy and weighted F1-score to show how our new "score" metric more severely penalises systems that make incorrect verifications.

These tests yielded surprising results. There is no correlation between the performance of the classifier model on the CLIMATE-FEVER test set in Section 4.2, and the overall system's performance when that classifier model is used. For example, "Llama-2 13b *FT:* CF" had the highest performance on the CLIMATE-FEVER test set, but the system

architectures which use this model performed the worst on the X claims test set. This discrepancy is visualised and further explored in Appendix A.12.

One potential reason for this discrepancy is the difference in the test sets used. The CLIMATE-FEVER test set, which was used to assess these models in isolation, contains a large proportion of samples labelled "NEI", so models which predict this label more often were scored higher. It can be seen in the case of "Llama-2 13b *FT:* CF", while no claims were incorrectly verified, a large number of claims received the UTV tag. This indicates that this model predicts NEI too often, causing the overall system to frequently output UTV.

Perhaps more surprising is that our architectures that use smaller models, such as variants of Llama-2 7b, tend to outperform those that use Llama-2 13b. Again, this is likely due to the larger models' tendency to overpredict the "NEI" class, which causes architectures that use it to be more conservative about supporting or refuting claims.

Also, the use of a re-ranker seems to make little impact on the performance system architecture, as the score of a system with and without the re-ranker never differs by more than 0.08. This is possibly due to the abundance of evidence in the IPCC AR6 for many of these claims, where the top 5 evidence sentences may contain supporting or refuting evidence both with and without the re-ranker.

See Appendix A.17 for examples of system outputs from these tests, where it can be seen that the system returns the evidence sentences used to make its decision, along with the report and section title within the IPCC AR6 that this evidence was retrieved from.

5.3 Different Claim Domains - A Qualitative Experiment

5.3.1 Experiment Description

The IPCC AR6 predominantly consists of scientific text on climate change. It has a larger emphasis on pure climate science concepts and research, as opposed to more general discourse on climate change, such as politics and speculation. While it does include discussions on climate change mitigation strategies, these are based on scientific evidence rather than political discourse.

Since our evidence bank is focused on scientific climate-related concepts, it is valuable to gain insight into how well the system performs on unscientific climate-related claims as well. This will allow us to assess further the feasibility of using the IPCC AR6 as our evidence bank. To test this, we selected three "domains" of climate-related claims. We constructed a small test set of claims for each domain and tested the system on each test set using the same methods described in Section 5.2. The selected domains are as follows:

- **Scientific** - Specific claims relating to the science and fact behind climate change
- **Political** - Claims related to politics and government actions involving climate change

- **Informal** - Claims relating to climate change that use very informal or unprofessional language

These domains capture a wide range of claims found online and represent different challenges to the overall system. Political claims may prove challenging, as these claims are often specific or involve recent events, which may be outside of the scope of the IPCC AR6. Also, we categorise informal claims into their own separate category as the phrasing used in these claims is very different to that of the evidence sentences, potentially increasing the difficulty of evidence retrieval.

These test sets were constructed from a mixture of sources. Some claims were extracted from social media sites, particularly the informal set. Other claims were extracted from existing datasets, such as the dataset provided by Coan et al. (2021) in "Computer-assisted classification of contrarian claims about climate change" [6], and the "Miscc" dataset, constructed by Xue Li [76]. As demonstrated in Section 5.2.1, we assign each label a "veracity" claim based on its truthfulness. For many of these claims, it was challenging to determine the truthfulness of a claim, and some examples of these can be found in Appendix A.13.

5.3.2 Qualitative Analysis

We tested the system architecture using "Llama-2 7b *FT:* F+CF" with no re-ranker on the 3 test sets of different domains, since this was one of our highest-scoring architectures in Section 5.2.3. The results table from this experiment can be seen in Appendix A.14.

As expected, the system performs best on the set of scientific claims due to the scientific nature of the content within the IPCC AR6, achieving a score of **0.9**. Despite mainly containing scientific information, the IPCC AR6 contains a small amount of information about climate policy and political actions in the Working Group 3 report. This may explain why there is a reasonable amount of success on the political test set, where we get a score of **0.6**, albeit worse than the scientific test set. For example, the system refuted the political claim "*climate deniers are not associated with the right*" with the evidence sentence "*...left-wing parties tend to adopt more pro-climate policy positions*". This showcases the system's adaptability to verify some claims outside the domain of pure climate science.

The "informal" test set is designed to test the system's ability to retrieve evidence for and verify claims of a different writing style to the knowledge base and training data. For example, "*climate change is a huge effing HOAX!!!*" contains vastly different language from the IPCC AR6. In our tests, the system refuted this claim with "*Vested economic and political interests have organised and financed misinformation and 'contrarian' climate-change communication*". The system also performs reasonably well on this test set, achieving a score of **0.6**.

This small experiment provides examples of the system adapting to non-scientific domains and performing reasonably well. Since these tests are on a small scale, using only small sets of claims, they do not show conclusive results of the system's ability across all domains. However, they give an idea of how the system may perform in a real use case of verifying real climate-related claims.

Chapter 6

Discussion

6.1 Evidence Retrieval Discussion

6.1.1 Climate Claim Detection Component

From our experiments, we found that our climate-claim detection system was successful both in isolation and as a component in the end-to-end claim verification pipeline. Although this is not an essential part of the pipeline, it demonstrates how such a system may be deployed in a real setting. Furthermore, when used as a component in the overall system, this component correctly identified every test claim as "climate-related", highlighting its effectiveness.

As described in Section 3.3, we use a version of ClimateBert [72], which has been fine-tuned for the task of recognising climate-specific text. Using this model for this task comes with several limitations. Since the model is trained to recognise climate-specific **text** as opposed to climate-specific **claims**, it will classify climate-related questions and random strings of climate-related words as "climate claims". Examples of these misclassifications can be seen in Appendix A.5. Essentially, our climate claim detector is just a climate text detector, and we assume all inputs will be claims, as claim detection, in general, is a complicated task [32], and we consider it outside of the scope of this research.

6.1.2 Sentence Retrieval Component

6.1.2.1 Comparisons to Existing Work

From our tests, we found dense retrieval to be the most effective method of evidence retrieval from the IPCC AR6, using the "all-mpnet-base-v2" sentence transformers [49] model. The scientific nature of the corpus is likely a factor in this, as the dense vector representations of sentences will likely capture the complexities of these sentences more effectively than term-based methods.

Our approach to evidence retrieval differs from previous works using transformer-based evidence retrieval systems [55, 49, 69] by removing the necessity of the document

retrieval step. Many approaches still utilise term-based approaches for document retrieval, and by limiting our search space to a specialised climate science corpus, we gain the advantage of bypassing this step due to our smaller search space. This allows us to perform a dense search over every potential evidence sentence in our corpus, without missing vital evidence in documents that were not ranked highly.

In the context of CLIMATE-FEVER, Wang et al. (2021) [69] retrieved evidence for 98.1% of the claims in their test portion of CLIMATE-FEVER by employing a Google Search-based system. It is unclear what qualifies as "valid evidence" in their work, and we do not know the claims used as a test set, so we cannot draw direct comparisons with our system. However, our approach, which returned valid evidence for 78% of claims in our initial tests, is unlikely to achieve the same results. This is due to the smaller yet specialised evidence corpus we employ. Our unique approach prioritises searching for robustly fact-checked evidence, which is grounded in climate science, instead of online sources for all claims related to climate change.

6.1.2.2 Limitations

One of the limitations of our testing strategy on the different models in Section 3.4.3 for the climate-claim evidence retrieval task is the testing claims used. During the manual annotation process of claim/evidence pairs it was discovered that some claims used in evaluating the models were not suitable for the given task. For example, the claim "*The rate of warming according to the data is much slower than the models used by the IPCC*" cannot be supported or refuted by evidence in the IPCC report itself, so this claim, along with other similar ones, provided no benefit when comparing the evidence retrieved by the different models as they were always annotated as "NEI".

6.1.3 Re-ranker Component

Our re-ranker implementation follows a similar approach to the sentence retrieval component by Wang et al. (2021) [69] and the re-ranker component implemented by Diggelmann et al. (2020) [11]. We are unable to directly compare our re-ranker to these works, as they use Wikipedia and Google as search spaces. However, we build on top of their work by providing the additional training data of claims from CLIMATE-FEVER with evidence sentences from the IPCC AR6, which in future could be further extended to create a more robust re-ranker specifically for evidence retrieval from this document.

We developed multiple re-ranker models, in an attempt to improve the performance of our evidence retrieval component. One of the primary benefits we hoped to gain from the re-ranker was a system specifically engineered for ranking evidence for climate-related claims from the IPCC AR6. Unfortunately, we were unable to achieve significant improvement from using our re-ranker models. However, we still explored how effectively it integrates with the other components in the overall system, to which we saw promising results.

It would have been beneficial to carry out these tests over a larger set of claims, as this may have allowed us to achieve statistically significant results. This was not done, as the test set was based on claims from CLIMATE-FEVER, and using more of these claims

in the test set would reduce the amount of training data. Furthermore, manually tagging and evaluating the output of a model for 50 claims, each with 3 evidence sentences, is a time-consuming process.

6.2 Classification System Discussion

6.2.1 Comparison to Existing Work

Our best LLM-based classifier model achieved a label accuracy of 0.76, a weighted F1 of 0.75, and a macro F1 of 0.64 on our test set of 475 claim/evidence pairs.

We cannot directly compare our results to the original CLIMATE-FEVER paper [11], as they use the whole of CLIMATE-FEVER as a test set. Several of our models would likely outperform the approach given in this paper, as they achieve a weighted F1-score of 0.3285.

Similarly, we cannot directly compare our results with Wang et al. (2021) [69] since they use an unknown randomly selected test set of 95 claims. However, our result on our random test set is slightly worse than theirs, as we achieve an unweighted F1-score of 0.66 compared to 0.718 achieved by Wang et al. (2021). Given that these are evaluated using different test sets, these results do not conclusively show which method is superior for this task.

The indirect comparisons we make with these previous works raise an interesting question about the most insightful metric for comparing these models. In their original paper, Diggelmann et al. (2020) [11] measure performance using a weighted F1 score, to compensate for the unbalanced distribution of labels, since $\sim$65% of claim/evidence pairs are labelled as "NEI". However, this can lead to the issue where models that predict "NEI" too often may not be correctly penalised by getting the other labels incorrect, since there is a smaller number of them so they have a smaller impact on the weighted F1 score. Wang et al. (2021) [69] report their performance using an unweighted (macro) F1 score. This metric ensures that the model is appropriately penalised for misclassifying the minority classes, as the unweighted F1 score does not consider the proportion of each class.

It may be more appropriate to evaluate these models on a separate balanced test set, with equal representation for each class. In this case, the weighted and unweighted F1 scores would align, allowing for a more direct comparison between implementations while still appropriately penalising models for misclassifications.

6.2.2 Potential Improvements

One major limitation of our LLM-based evidence classifier is that we only allow the LLM to generate a single word, which is its class prediction based on the input. It has been shown that prompting an LLM to generate its "Chain of Thought" (CoT) and reasoning before it makes a classification decision [30] can greatly improve its performance for classification tasks. We did not employ this technique in our research, mainly due to resource constraints, as generating a CoT requires more GPU time than

just the classification label. In future, it would be valuable to explore this method to see how it compares with the techniques we used.

Alternatively, a recent paper by Thulke et al. (2024) discusses alternative approaches to the CLIMATE-FEVER task, relying more heavily on LLMs. More details can be found in Appendix A.18.

6.3 Overall System Discussion

After integrating the individual components to form the overall system, we observed that numerous components that performed well independently exhibited diminished effectiveness within the context of the integrated system. For example, system architectures that utilised our highest-performing evidence classifiers performed considerably worse than those that used our other models. This discrepancy is likely due to our higher performing models' tendency to over-predict the "NEI" class, so when presented with a list of 5 evidence sentences, they were often all labelled as "NEI". Also, the test set used to measure the system's performance was constructed from real claims found on X, which may differ in terms of wording and content from CLIMATE-FEVER. This shift in domain is likely a factor in why our lower-scoring models outperformed our highest-scoring models in this task.

Several of our system architectures performed strongly for this task, and could correctly predict the veracity of 76% of our test claims, while providing evidence, and only incorrectly verifying 4% of the test claims. The "UTV" (Unable To Verify) label allows the system to abstain from supporting or refuting a claim when it retrieves conflicting evidence or only irrelevant evidence. This likely reduces the number of incorrectly verified claims, as in these cases, we cannot be confident in the classifier's decision.

One benefit of our system, as opposed to leveraging an LLM to carry out the entire claim verification process, is the ability of our system to provide a source of the evidence used to support or refute a claim. LLMs are often regarded as "black-box" models, as it is often unclear why they make their predictions. Furthermore, they are liable to "hallucinate" [40], and generate false information. Providing the evidence used to support or refute a claim is invaluable, as it clearly shows what information was considered when coming to a decision. To further improve the transparency of the system, Chain of Thought prompting could be utilised to explain how the model concluded that evidence supports or refutes a claim.

After constructing this system, and running various tests on real claims, we can see that the system can effectively retrieve evidence for and validate a variety of claims. However, there are still some obvious flaws. For example, the classifier component could be improved, as the errors it makes are often detrimental to the performance of the system. In its current state, the system is likely more useful as a tool used to assist human fact-checkers in more efficiently verifying claims and reducing misinformation, as opposed to being used or deployed in a real setting for verifying claims.

Chapter 7

Conclusions

7.1 Research Findings

7.1.1 Viability of IPCC AR6 for Evidence Retrieval

Research Question 1: **What is the viability of using a single, comprehensive, domain-specific document, such as the IPCC Sixth Assessment Report, as the entire evidence corpus for various evidence retrieval methods in a claim verification system?**

After constructing a claim verification pipeline using the IPCC AR6 knowledge base, and conducting extensive tests, we have found that evidence retrieval from the IPCC AR6 is an effective method of claim verification for certain types of climate-related claims.

We experimented with multiple methods to find the most effective evidence retrieval method from this document. After extensive testing and a large manual annotation process, it was found that dense retrieval significantly outperforms both sparse and hybrid methods when retrieving evidence at the sentence level from this corpus. We can likely attribute this to the scientific nuances and complexity of the content in this document. We also created several climate-specific sentence re-rankers in an attempt to improve the performance of the evidence retrieval system, which showed promising results. However, we could not conclusively say that they are an improvement over the base dense retrieval system.

Using the dense retrieval component constructed in 3.4.3, we were able to retrieve valid evidence for 78% of claims from our CLIMATE-FEVER test set and valid evidence for 76% claims in our X claims test set. We also found that the system was often more effective for scientific claims than political or informal claims. This is caused by the limited scope of the IPCC AR6, being a primarily scientific document.

In future, it may be worth expanding the knowledge base to contain other documents, to increase the variety of claims that can be verified. However, a key factor in the choice of using the IPCC AR6 for this task was the reputability it has, so expanding to less reputable documents may harm the trust people may have in such a system.

7.1.2 Large Language Models for Classifying Evidence

Research Question 2: **How well do Large Language models perform for the task of classifying whether an evidence sentence supports or refutes a claim, and how can their performance be improved?**

After evaluating multiple Large Language Models, we found them to be effective tools for classifying whether evidence supports or refutes a climate-related claim. In particular, we found that the efficacy of smaller LLMs, such as Llama-2 13b, can be greatly improved to perform just as well as much larger models, such as GPT-4, after they are fine-tuned for this specific task with training examples, achieving a label accuracy of 76% on our test set.

We also found that it can be challenging to evaluate and compare the performance of classifiers using the CLIMATE-FEVER dataset. The lack of a pre-defined training/test split of this dataset leads to inconsistencies between how different papers record their results, with some, including ourselves, creating a random training/test split of the dataset. Also, the abundance of evidence sentences labelled as "NEI" in this dataset can skew the results, as models that tend to overpredict this class often achieve higher scores. Perhaps, in future, it may be worth extending this dataset to include a training/test split, as well as finding more evidence sentences which support/refute their respective claim, to balance the class labels. However, we cannot understate how valuable the CLIMATE-FEVER dataset was in creating and testing our system.

While we cannot directly compare our results to existing work, through indirect comparative analysis we found that LLMs have the potential to produce impressive results for this task. There are also many potential methods to improve this classifier as future work that we did not explore, such as using CoT (Chain of Thought) prompting, leading to exciting potential.

7.1.3 System Performance on Real Claims

Research Question 3: **How well does the system we create perform on real claims? And how does its performance differ on different types of climate-related claims, such as scientific, political, and informal?**

After assembling an overall claim-verification pipeline from the various components created to answer the previous research questions, we ran initial tests on a set of real climate-related claims from X. We found the system could correctly verify 76% of these claims, which is a promising result given the small scope of the knowledge base compared to existing methods.

We found that the system is unsurprisingly most effective on climate-science-related claims, due to the scientific nature of the knowledge base. However, we also found that the system is surprisingly versatile, and seems to perform reasonably well with non-scientific claims, such as political and informal claims. The versatility and adaptability of the system to other domains are promising, as it increases the scope of use cases where such a system may be useful.

7.2 Key Contributions

Our developed system builds upon the CLIMATE-FEVER framework created by Diggelmann et al. (2020) [11], by integrating the IPCC AR6 into our evidence knowledge base, offering a streamlined and transparent method for climate change fact-checking. This approach prioritizes the use of reputable sources and solid facts for verifying claims by avoiding potentially unreliable web sources, hopefully preventing distrust in such a system, while still containing enough information to verify a large proportion of claims. By focusing exclusively on evidence grounded in climate science, we eliminate ambiguity around what qualifies as "evidence" for supporting or refuting claims. Moreover, by incorporating LLMs as the classification mechanism within our system, we demonstrate how these models can be effectively improved and utilised in fact-checking systems. Additionally, in testing our system, we've developed a novel metric for evaluating the performance of claim-verification systems to compare our potential system architectures. This metric more harshly penalizes systems that mistakenly endorse misinformation or reject accurate information, providing a sense of how useful they may be in practice.

7.3 Wider Implications

Our approach leverages recent NLP technologies, by integrating evidence-retrieval mechanisms with LLMs to provide reliable claim verification. This approach addresses a major limitation of LLMs, their tendency to hallucinate, by grounding their outputs with the reliable information provided by the IPCC.

We have shown that such a system is a feasible tool for verifying claims by correctly verifying 76% of claims from a test set of real claims from X. This may potentially aid in reducing the spread of climate-change-related misinformation online. We have found that the IPCC AR6 and likely future IPCC reports provide an effective knowledge base for verifying a plethora of climate claims, especially scientific ones.

Furthermore, incorporating LLMs into the system opens up many possibilities for extending and improving such a system. Using basic techniques such as prompt engineering and parameter-efficient fine-tuning, we significantly improved their performance in classifying whether an evidence sentence supports or refutes a claim. In the future, the adaptability and scalability of LLMs can be leveraged in exciting ways to further improve their ability to interpret the nuances of evidence classification. This framework of combining traditional evidence retrieval techniques with LLMs provides opportunities to greatly increase the reliability of such a system.

While climate change is a contentious issue, this work encourages climate-related discourse based on fact and science by providing grounded evidence. With the rate at which NLP technologies are evolving, it is likely that in the near future, such a system may be used in real applications, not only as a tool for fact-checkers but also as a tool employed on social media sites to encourage fact-based discussions.

In a time when the stakes of climate change are at an all-time high, with rampant misinformation spreading throughout the internet, collaboration between the fields of Natural Language Processing and Climate Science is essential going forward.

Bibliography

[1] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

[2] Ivan Boban, Alen Doko, and Sven Gotovac. “Improving sentence retrieval using sequence similarity”. In: *Applied Sciences* 10.12 (2020), p. 4316.

[3] D. Brown. *rank-bm25*. https://pypi.org/project/rank-bm25/. Accessed: December 2023. 2022.

[4] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[5] Danqi Chen et al. “Reading wikipedia to answer open-domain questions”. In: *arXiv preprint arXiv:1704.00051* (2017).

[6] Travis G Coan et al. “Computer-assisted classification of contrarian claims about climate change”. In: *Scientific reports* 11.1 (2021), p. 22320.

[7] Wikipedia contributors. *Okapi BM25 — Wikipedia, The Free Encyclopedia*. [Online; accessed 27-March-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Okapi_BM25&oldid=1194828429.

[8] John Cook, Peter Ellerton, and David Kinkead. “Deconstructing climate misinformation to identify reasoning errors”. In: *Environmental Research Letters* 13.2 (Feb. 2018), p. 024018. DOI: 10.1088/1748-9326/aaa49f. URL: https://dx.doi.org/10.1088/1748-9326/aaa49f.

[9] Tim Dettmers et al. “Qlora: Efficient finetuning of quantized llms”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[10] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[11] Thomas Diggelmann et al. “Climate-fever: A dataset for verification of real-world climate claims”. In: *arXiv preprint arXiv:2012.00614* (2020).

[12] Qingxiu Dong et al. “A survey on in-context learning”. In: *arXiv preprint arXiv:2301.00234* (2022).

[13] Lautaro Estienne. “Unsupervised calibration through prior adaptation for text classification using large language models”. In: *arXiv preprint arXiv:2307.06713* (2023).

[14] Hugging Face. *Hugging Face Hub*. https://huggingface.co. Accessed: December 2023. 2023.

[15] Don Fallis. “Toward an epistemology of Wikipedia”. In: *Journal of the American Society for Information science and Technology* 59.10 (2008), pp. 1662–1674.

[16] Christopher Fifty, Jure Leskovec, and Sebastian Thrun. “In-Context Learning for Few-Shot Molecular Property Prediction”. In: *arXiv preprint arXiv:2310.08863* (2023).

[17] Luciano Floridi and Massimo Chiriatti. “GPT-3: Its nature, scope, limits, and consequences”. In: *Minds and Machines* 30 (2020), pp. 681–694.

[18] Ulrich Germann. “Greedy decoding for statistical machine translation in almost linear time”. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 2003, pp. 72–79.

[19] Santiago González-Carvajal and Eduardo C Garrido-Merchán. “Comparing BERT against traditional machine learning text classification”. In: *arXiv preprint arXiv:2005.13012* (2020).

[20] Andreas Hanselowski et al. “Ukp-athene: Multi-sentence textual entailment for claim verification”. In: *arXiv preprint arXiv:1809.01479* (2018).

[21] Edward J Hu et al. “Lora: Low-rank adaptation of large language models”. In: *arXiv preprint arXiv:2106.09685* (2021).

[22] Intergovernmental Panel on Climate Change. *About the IPCC*. `https://www.ipcc.ch/about/`. Accessed: 2024-03-23. 2024.

[23] Intergovernmental Panel on Climate Change. *Climate Change 2022: Synthesis Report.* `https://www.ipcc.ch/report/ar6/syr/`. Accessed: [September 2023]. 2022.

[24] Intergovernmental Panel on Climate Change. *IPCC - Intergovernmental Panel on Climate Change*. `https://www.ipcc.ch/`. Accessed: [September 2023]. 2024.

[25] Intergovernmental Panel on Climate Change, Working Group 1. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report*. `https://www.ipcc.ch/report/ar6/wg1/`. Accessed: [September 2023. 2021.

[26] Intergovernmental Panel on Climate Change, Working Group 2. *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report*. `https://www.ipcc.ch/report/ar6/wg2/`. Accessed: [September 2023]. 2022.

[27] Intergovernmental Panel on Climate Change, Working Group 3. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report*. `https://www.ipcc.ch/report/ar6/wg3/`. Accessed: [September 2023]. 2022.

[28] Gautier Izacard et al. “Unsupervised dense information retrieval with contrastive learning”. In: *arXiv preprint arXiv:2112.09118* (2021).

[29] Aniket Kittur, Bongwon Suh, and Ed H Chi. “Can you ever trust a Wiki? Impacting perceived trustworthiness in Wikipedia”. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 2008, pp. 477–480.

[30] Takeshi Kojima et al. “Large language models are zero-shot reasoners”. In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

[31] Zhenzhong Lan et al. “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942* (2019).

[32] Ran Levy et al. “Context dependent claim detection”. In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2014, pp. 1489–1500.

[33] Stephan Lewandowsky, Gilles E Gignac, and Samuel Vaughan. “The pivotal role of perceived scientific consensus in acceptance of science”. In: *Nature climate change* 3.4 (2013), pp. 399–404.

[34] Jimmy Lin et al. “Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2356–2362.

[35] Haokun Liu et al. “Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 1950–1965.

[36] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[37] Yunqing Lu. “A revised version of McNemar’s test for paired binary data”. In: *Communications in Statistics—Theory and Methods* 39.19 (2010), pp. 3525–3539.

[38] Priyanka Mandikal and Raymond Mooney. “Sparse Meets Dense: A Hybrid Approach to Enhance Scientific Document Retrieval”. In: *arXiv preprint arXiv:2401.04055* (2024).

[39] Sourab Mangrulkar et al. *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods*. `https://github.com/huggingface/peft`. 2022.

[40] Ariana Martino, Michael Iannelli, and Coleen Truong. “Knowledge injection to counter large language model (LLM) hallucination”. In: *European Semantic Web Conference*. Springer. 2023, pp. 182–185.

[41] Luca Massaron. *Fine-tune Llama 2 for sentiment analysis*. `https://www.kaggle.com/code/lucamassaron/fine-tune-llama-2-for-sentiment-analysis/notebook`. Accessed: December 2023. 2023.

[42] Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. “Learning text similarity with siamese recurrent networks”. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. 2016, pp. 148–157.

[43] Tri Nguyen et al. “Ms marco: A human-generated machine reading comprehension dataset”. In: (2016).

[44] OpenAI. *GPT-4: OpenAI’s Generative Pre-trained Transformer 4*. `https://openai.com/`. Accessed: December 2023. 2023.

[45] OpenAI. *OpenAI API*. Accessed: 2024-01. 2024. URL: `https://openai.com/api/`.

[46] Warren Pearce et al. “Beyond counting climate consensus”. In: *Environmental Communication* 11.6 (2017), pp. 723–730.

[47] F. Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[48] Revanth Gangi Reddy et al. “Newsclaims: A new benchmark for claim detection from news with background knowledge”. In: *arXiv preprint arXiv:2112.08544* (2021).

[49] Nils Reimers and Iryna Gurevych. “Sentence-bert: Sentence embeddings using siamese bert-networks”. In: *arXiv preprint arXiv:1908.10084* (2019).

[50] Ruiyang Ren et al. “RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking”. In: *arXiv preprint arXiv:2110.07367* (2021).

[51] Leonard Richardson. *Beautiful soup documentation*. 2007.

[52] Benjamin Riedel et al. “A simple but tough-to-beat baseline for the Fake News Challenge stance detection task”. In: *arXiv preprint arXiv:1707.03264* (2017).

[53] Stephen Robertson, Hugo Zaragoza, et al. “The probabilistic relevance framework: BM25 and beyond”. In: *Foundations and Trends® in Information Retrieval* 3.4 (2009), pp. 333–389.

[54] Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. “COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic”. In: *arXiv preprint arXiv:2106.03794* (2021).

[55] Amir Soleimani, Christof Monz, and Marcel Worring. “Bert for evidence retrieval and claim verification”. In: *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*. Springer. 2020, pp. 359–366.

[56] Kaitao Song et al. “Mpnet: Masked and permuted pre-training for language understanding”. In: *Advances in neural information processing systems* 33 (2020), pp. 16857–16867.

[57] Nicholas Stern. *Why are we waiting?: The logic, urgency, and promise of tackling climate change*. Mit Press, 2015.

[58] Chi Sun et al. “How to fine-tune bert for text classification?” In: *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*. Springer. 2019, pp. 194–206.

[59] Xiaofei Sun et al. “Text Classification via Large Language Models”. In: *arXiv preprint arXiv:2305.08377* (2023).

[60] James Thorne and Andreas Vlachos. “Automated fact checking: Task formulations, methods and future directions”. In: *arXiv preprint arXiv:1806.07687* (2018).

[61] James Thorne et al. “FEVER: a large-scale dataset for fact extraction and VERification”. In: *arXiv preprint arXiv:1803.05355* (2018).

[62] James Thorne et al. “The fact extraction and VERification (FEVER) shared task”. In: *arXiv preprint arXiv:1811.10971* (2018).

[63] David Thulke et al. “ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change”. In: *arXiv preprint arXiv:2401.09646* (2024).

[64] Hugo Touvron et al. “Llama 2: Open foundation and fine-tuned chat models”. In: *arXiv preprint arXiv:2307.09288* (2023).

[65] Kathie M d’I Treen, Hywel TP Williams, and Saffron J O’Neill. “Online misinformation about climate change”. In: *Wiley Interdisciplinary Reviews: Climate Change* 11.5 (2020), e665.

[66] Saeid Ashraf Vaghefi et al. “ChatClimate: Grounding conversational AI in climate science”. In: *Communications Earth & Environment* 4.1 (2023), p. 480.

[67] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[68] Hong Tien Vu, Annalise Baines, and Nhung Nguyen. "Fact-checking climate change: An analysis of claims and verification practices by fact-checkers in four countries". In: *Journalism & Mass Communication Quarterly* 100.2 (2023), pp. 286–307.

[69] Gengyu Wang, Lawrence Chillrud, and Kathleen McKeown. "Evidence based automatic fact-checking for climate change misinformation". In: *International Workshop on Social Sensing on The International AAAI Conference on Web and Social Media*. 2021.

[70] Yaqing Wang et al. "Generalizing from a few examples: A survey on few-shot learning". In: *ACM computing surveys (csur)* 53.3 (2020), pp. 1–34.

[71] Zhiqiang Wang, Yiran Pang, and Yanbin Lin. "Large Language Models Are Zero-Shot Text Classifiers". In: *arXiv preprint arXiv:2312.01044* (2023).

[72] Nicolas Webersinke et al. "Climatebert: A pretrained language model for climate-related text". In: *arXiv preprint arXiv:2110.12010* (2021).

[73] Leandro von Werra et al. *TRL: Transformer Reinforcement Learning*. `https://github.com/huggingface/trl`. 2020.

[74] Jules White et al. "A prompt pattern catalog to enhance prompt engineering with chatgpt". In: *arXiv preprint arXiv:2302.11382* (2023).

[75] Thomas Wolf et al. "Huggingface's transformers: State-of-the-art natural language processing". In: *arXiv preprint arXiv:1910.03771* (2019).

[76] Vaishak Belle Xue Li Bjorn Ross. *Miscc dataset*. Funded by Edinburgh Laboratory for Integrated Artificial Intelligence. Not yet published.

[77] Zhilin Yang et al. "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32 (2019).

Appendix A

Supplementary materials

A.1 Definitions

Here, we define several terms that are used throughout the dissertation and are not comprehensively explained:

- **TF-IDF Vectors:** Term Frequency-Inverse Document Frequency (TF-IDF) vectors are measurements of a word's importance relative to a document within a corpus. The importance increases with the number of times a word occurs in the document but is then decreased by the frequency of the word in the corpus.
- **Greedy Decoding:** Greedy decoding [18] is a method for generating text from a model, where the token with the highest probability is selected at each step in the sequence. This method is efficient but does not guarantee an optimal sequence of words for most sequences. However, in our case, where the only token the LLM generates is its prediction, it is appropriate.
- **BERT-based Models:** BERT (Bidirectional Encoder Representations from Transformers) [10] based models are a family of transformer-based models designed to learn deep bidirectional representations from unlabeled text by conditioning on both left and right context.
- **Permuted Language Modelling:** Permuted Language Modelling is a training method used in Microsoft's MPNet [56] model, where the order of the words in the input is rearranged in a way that allows the model to predict missing tokens in the context of all possible permutations.
- **Contrastive Learning:** Contrastive learning is the process of learning how to discriminate between textually similar and different data. In the context of Contriever [28], it is achieved by using data augmentation for the model to learn from a large corpus, as opposed to the labelled sentence pairs used to fine-tune a sentence transformer model. To achieve this, the InfoNCE loss function is used, which gives positive scores to text from the same documents, and negative scores to text from different documents.

A.2 Example of CLIMATE-FEVER claim with evidence

Votes : 4
Entropy: 0.23
Claim: Extreme weather isn't caused by global warming
Evidence:
Refutes: Extreme Weather Prompts Unprecedented Global Warming Alert. [wiki/Extreme_weather]
Refutes: Scientists attribute extreme weather to man-made climate change. [wiki/Extreme_weather]
Refutes: Researchers have for the first time attributed recent floods, droughts and heat waves, to human-induced climate change. [wiki/Extreme_weather]
Refutes: Climate change is more accurate scientifically to describe the various effects of greenhouse gases on the world because it includes extreme weather, storms and changes in rainfall patterns, ocean acidification and sea level."". [wiki/Global_warming]
Refutes: The effects of global warming include rising sea levels, regional changes in precipitation, more frequent extreme weather events such as heat waves, and expansion of deserts. [wiki/Global_warming]
Verdict: Refutes

Figure A.1: Example of claim in CLIMATE-FEVER database, with labelled evidence sentences [11]

A.3 Example of IPCC data before and after data cleaning process

Here, we show an example of a section of the IPCC before and after the cleaning process. We have truncated the section for this example, as the original text segment is very long. This example is taken from Working Group 2, Chapter 5, Section 5.6.1 of the IPCC AR6. It should also be noted that in the actual system, each cleaned evidence sentence is stored as a tuple along with their section title, so the origin of the sentence can easily be traced.

Before Cleaning

The IPCC AR5 stated that there is high confidence that numerous plants and animal species have already migrated, changed their abundance, and shifted their seasonal activities as a result of climate change (Settele et al., 2014). The report highlighted the widespread deaths of trees in many forested areas of the world. Forest die back could significantly affect wood production among other impacts.The SRCCL (Barbosa et al., 2019) concluded that climate change will have positive and negative effects on forests, with varying regional and temporal patterns. For example, the SRCCL noted the increasing productivity in high-latitude forests such as those in Siberia. In contrast, negative impacts are already being observed in other regions such as increasing tree mortality due to [wildfires.In](http://wildfires.in/) the past years, tree mortality continued to increase in many parts of the world. Large pulses of tree mortality were consistently linked to warmer and drier than average conditions for forests throughout the temperate and boreal biomes (high confidence) (Sommerfeld et al., 2018; Seidl et al., 2020). Long-term monitoring of tropical forests indicates that climate change has begun to increase tree mortality and alter regeneration (Hubau et al., 2020; Sullivan et al., 2020).

After Cleaning

- The IPCC AR5 stated that there is high confidence that numerous plants and animal species have already migrated, changed their abundance, and shifted their seasonal activities as a result of climate change.
- The report highlighted the widespread deaths of trees in many forested areas of the world.
- Forest dieback could significantly affect wood production among other impacts.
- The SRCCL concluded that climate change will have positive and negative effects on forests, with varying regional and temporal patterns.
- For example, the SRCCL noted the increasing productivity in high-latitude forests such as those in Siberia.
- In contrast, negative impacts are already being observed in other regions, such as increasing tree mortality due to wildfires.
- In the past years, tree mortality continued to increase in many parts of the world.
- Large pulses of tree mortality were consistently linked to warmer and drier than average conditions for forests throughout the temperate and boreal biomes.
- Long-term monitoring of tropical forests indicates that climate change has begun to increase tree mortality and alter regeneration.

A.4 Classification Report for Climate Claim Detection System

Class	Precision	Recall	F1-Score	Support
`no`	0.95	0.97	0.96	1535
`yes`	0.97	0.95	0.96	1535
Accuracy			0.96	3070
Macro Avg	0.96	0.96	0.96	3070
Weighted Avg	0.96	0.96	0.96	3070

Table A.1: `scikit-learn` [47] Classification Report for climate claim detection

A.5 Examples of Misclassification from Climate Claim Detector

Input: *sea level animals warming forest fire*
Classification: `climate-related claim`

Input: *Is this a climate-related claim or a question?*
Classification: `climate-related claim`

Input: *climate*
Classification: `climate-related claim`

Input: *Jordan Knight is a vegan.*
Classification: `climate-related claim`

A.6 Explanation of Re-ranker model names:

Name of Re-ranker Model	Retrieval Method Before Re-ranking	Base Model For Re-ranker	Training Data For Fine-Tuning
No Re-Ranker	all-mpnet-base-v2	N/A	N/A
RoBERTa-F	all-mpnet-base-v2	RoBERTa-Base	FEVER
RoBERTa-F-CF	all-mpnet-base-v2	RoBERTa-Base	FEVER CLIMATE-FEVER
RoBERTa-F-CF-IPCC	all-mpnet-base-v2	RoBERTa-Base	FEVER CLIMATE-FEVER Annotated Pairs
ClimateBert-F	all-mpnet-base-v2	ClimateBert	FEVER
ClimateBert-F-CF	all-mpnet-base-v2	ClimateBert	FEVER CLIMATE-FEVER
ClimateBert-F-CF-IPCC	all-mpnet-base-v2	ClimateBert	FEVER CLIMATE-FEVER Annotated Pairs

Table A.2: Explanation of how each re-ranker model is created. “Annotated Pairs” refers to the manually annotated pairs from testing the evidence retrieval systems in Section 3.4.3. These consist of a claim sentence from CLIMATE-FEVER, and an evidence sentence from the IPCC AR6. It was also ensured that no claims used in the training data were also used in the test set.

A.7 Explanation of Metrics

Throughout this dissertation, we make use of several evaluation metrics for assessing models. These metrics are calculated in terms of True Positives TP, False Positives FP, False Negatives FN, and True Negatives TN:

Precision:

$$Pr = \frac{TP}{TP + FP} \tag{A.1}$$

Recall:

$$Re = \frac{TP}{TP + FN} \tag{A.2}$$

F1-Score:

$$\mathrm{F1} = \frac{2 \cdot Pr \cdot Re}{Pr + Re} \tag{A.3}$$

Using these formulae, we calculate an F1-score for each class. In our case, these classes are "SUPPORTS", "REFUTES", "NOT_ENOUGH_INFO". For an overall model, we can then find a weighted F1-Score, and an unweighted (macro) F1-Score.

Unweighted (macro) F1-Score:

$$\mathrm{F1}_{\text{unweighted}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{F1}_i \tag{A.4}$$

Weighted F1-score:

$$\mathrm{F1}_{\text{weighted}} = \sum_{i=1}^{N} \left(\frac{n_i}{N_{\text{total}}} \right) \times \mathrm{F1}_i \tag{A.5}$$

Where:

- N = Number of classes
- $\mathrm{F1}_i$ = F1 score for class i
- n_i = Number of samples of class i
- N_{total} = Number of samples across all classes

Unweighted (macro) F1-score provides the average F1-score for each class, regardless of the number of samples in each class. Weighted F1-score provides a weighted average, where classes with a larger number of samples have a larger weight. Both metrics have their uses: unweighted (macro) F1-score treats all classes equally, suitable for datasets with balanced importance across classes, while weighted F1-score accounts for class imbalance, ideal for datasets with imbalanced classes.

A.8 Hybrid Evidence Retrieval

A.8.1 Hybrid Methods - Combining Sparse and Dense Scores

Inspired by the work of Mandikal et al. [38], and Lin et al. [34], we considered a hybrid approach of ranking potential evidence sentences based on both their sparse and dense retrieval scores. This allows us to encapsulate the advantages of both methods: Dense retrieval's ability to capture the sentence's overall semantics, and the preciseness of sparse retrieval in searching for exact terms. We explore multiple hybrid models, which are created from our BM25 sparse retrieval model, and the best-performing dense retrieval model.

Given a claim sentence, we assign each potential evidence sentence in the evidence bank a "dense score" s_{dense} and a "sparse score" s_{sparse}. These scores are calculated based on the methods previously described on sparse and dense retrieval. Since BM25 has no upper limit for its score values, and the cosine similarity for the dense score is always in the range $[-1, 1]$, we normalise both the sparse and dense scores using L1 normalisation. This is to ensure sparse scores do not cause too much impact, as they can be much larger than dense scores. We then combine the dense and sparse score for each potential evidence sentence using interpolation as follows:

$$s_{\text{hybrid}} = \lambda s_{\text{dense}} + (1 - \lambda) s_{\text{sparse}} \tag{A.6}$$

Where λ is a pre-selected constant between 0 and 1.

Each evidence sentence is assigned a hybrid score, and the sentences with the top k hybrid scores are retrieved. We experiment with different values of λ, to determine whether we can improve the performance of the sentence retriever using this hybrid approach.

A.8.2 Hybrid Evidence Retrieval Results

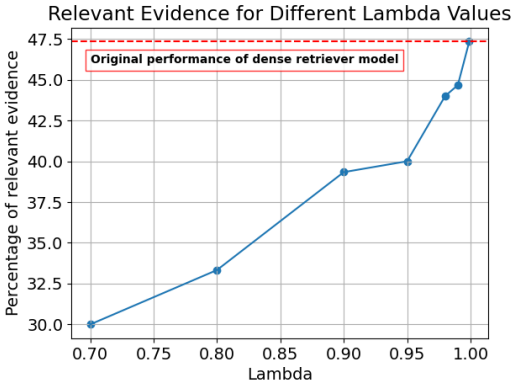

Figure A.2: Performance of Hybrid Retrieval systems for an increasing value of λ for the **relevant evidence** metric. The red dotted line is the original performance of all-mpnet-base-v2.

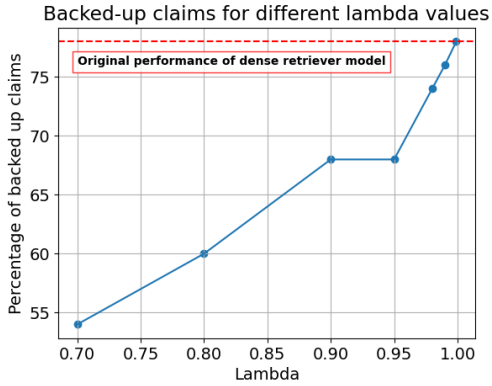

Figure A.3: Performance of Hybrid Retrieval systems for an increasing value of λ for the **backed-up claims** metric. The red dotted line is the original performance of all-mpnet-base-v2.

The hybrid methods were implemented using the highest performing model, all-mpnet-base-v2, for the dense retrieval component, and BM25 for the sparse retrieval component. We experiment with different values of λ: $[0.7, 0.8, 0.9, 0.95, 0.98, 0.99, 0.999]$. Relatively high values were selected due to the massive difference in performance between sparse and dense retrieval in our previous tests, so it was important to ensure the sparse component was not overly influential.

We were unable to improve the performance of our evidence retrieval system using this hybrid approach, as for all values of λ, the hybrid approach performed worse

than the dense approach. Figures A.2 and A.3 show the performance of the hybrid retrieval on the **relevant evidence** metric and the **backed-up claims** metric, for an increasing value of λ. These hybrid models were manually assessed using the same evidence-tagging process as the other methods, using the same set of test claims. It can be seen that as the value for λ increases, the performance of the hybrid retrieval model also increases, until it reaches the same performance as the dense retrieval model for a value of $\lambda = 0.999$. When the value if λ reaches this point, the sparse retrieval system has very little influence over the scoring of evidence sentences, and it can be seen that the model has the same performance as pure dense retrieval.

These results indicate that incorporating input from the sparse retrieval system hinders the performance of the evidence retrieval system, as with larger λ values, there is less input from the sparse retrieval system, and larger λ values have a higher performance. Hybrid approaches often improve the performance of evidence retrieval systems [34], however, for our system, this is not the case. This is likely due to the complex and scientific nature of the text in the IPCC AR6, where there may be more value gained from a semantic understanding of the sentence as opposed to finding sentences with the same terms. As well as this, the IPCC AR6 contains many keywords repeated consistently throughout, such as "climate", "carbon", etc, which may throw off a purely term-based retrieval approach as these words appear in many potential evidence sentences.

A.9 Manual Annotation Process

In order to assess the evidence retrieval and evidence re-ranker systems, we had to manually inspect and annotate the retrieved evidence for each claim. Each claim/evidence pair annotation was saved, so if another model retrieved the same evidence sentence for the same claim, then the annotations are consistent between models. From this process, we gather 2 metrics to evaluate the performance of a model, being "backed-up claims" and "Relevant Evidence". The manual annotation process for each method was carried out as follows:

- For each claim, we retrieve the top 3 evidence sentences using the model we are evaluating.
- Each of these evidence sentences is checked against our claim/evidence pairs that have been previously annotated. If the claim/evidence pair has been annotated before, it receives the same label.
- If the claim/evidence pair is not found in our set of annotated pairs, it is displayed to the annotator. The annotator then labels the claim/evidence pair as "ENOUGH_INFO", if the evidence sentence contains enough information to support or refute the claim, otherwise, it is labelled as "NOT_ENOUGH_INFO".
- Newly annotated claim/evidence pairs are then added to the set of saved annoyed claim/evidence pairs, so the annotations can be reused.

After annotation, we then calculate the "Backed-Up Claims" and "Relevant Evidence" metrics for a model, where:

- **Backed-Up Claims** is the percentage of claims that have at least 1 corresponding evidence sentence labelled as "ENOUGH_INFO"
- **Relevant Evidence** is the percentage of evidence sentences a model retrieves that are labelled as "ENOUGH_INFO"

NOTE: When we carry out this manual evaluation for the re-ranker models, we ensure that none of the claims in the test set are used as part of its training data.

A.10 Comparison of FEVER and CLIMATE-FEVER

As described in the original CLIMATE-FEVER paper, the real-world claims in the CLIMATE-FEVER dataset prove to cause a significant challenge in this classification task when compared to the FEVER dataset, which uses artificial claims which are created just for the dataset, making them less realistic. Diggelmann et al. (2020) [11] demonstrate this by providing an example of a model which achieves a low label accuracy of 38.78% on CLIMATE-FEVER, but achieves a significantly higher label accuracy on FEVER at 77.69%. To further explore the increased challenge of claim/evidence classification on CLIMATE-FEVER compared to FEVER, we select a test set of 475 randomly sampled claim/evidence pairs from FEVER and use GPT-4 to classify them. We can then compare these results to GPT-4's performance on CLIMATE-FEVER.

From this test, we observed the following results:

Dataset	Accuracy	Weighted F1-score	Unweighted F1-score
FEVER	0.87	0.86	0.82
CLIMATE-FEVER	0.73	0.73	0.64

This supports the claim that the claim/evidence pair classification task is more challenging for CLIMATE-FEVER than for FEVER.

A.11 LLM Prompts

A.11.1 Final Prompt Used

Final Prompt Used

```
Given a "Claim" sentence and an "Evidence" sentence, determine
if the evidence supports or refutes the claim, or if there is
not enough information to decide.
Return the answer as the corresponding label "SUPPORTS" or
"REFUTES" or "NOT_ENOUGH_INFO"

[Claim: '<Claim>', Evidence: '<Evidence>'] = <Label>
```

This prompt template was adapted from Luca Massaron [41], and was used for our

experiments and within our final system. When used as a training prompt, we replace "<Label>" with the correct label. When used as a test prompt, we leave it empty.

A.11.2 Alternate Prompt Template

Alternate Unused Prompt

```
Given a "Claim" sentence and an "Evidence" sentence, determine
if the evidence supports or refutes the claim, or if there is
not enough information to decide.
Return the answer as the corresponding label "SUPPORTS" or
"REFUTES" or "NOT_ENOUGH_INFO"
Claim: '<Claim>'
Evidence '<Evidence>'
Label: <Label>
```

This prompt template was not ultimately used, as after fine-tuning Llama-2 13b on CLIMATE-FEVER using this prompt, we achieved inferior metrics of:

- Label Accuracy: **0.72**
- Unweighted F1: **0.57**
- Weighted F1: **0.71**

A.12 Comparison of Models Between Tasks

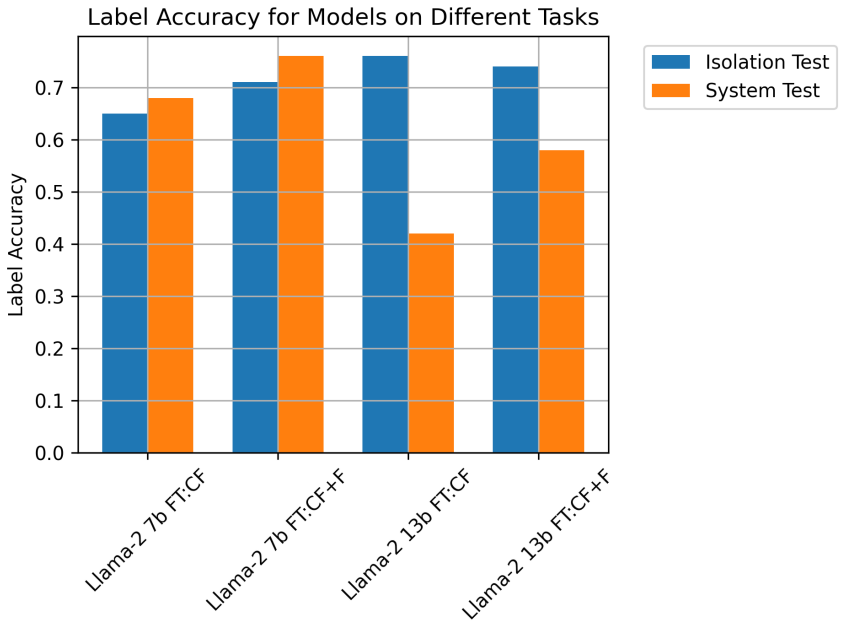

Figure A.4: Comparison of how model's performed when they were tested in isolation on CLIMATE-FEVER (Blue), compared to how the overall system architectures that incorporated these models performed on the X claims test set (Orange). Note: For simplicity, this graph only considers the overall system architectures with no re-ranker component.

Figure A.4 compares how classifier models performed when they are tested in isolation on CLIMATE-FEVER, compared to how architectures that incorporate these models perform in the tests in Section 5.2.3. This is not a direct comparison, as the results are on different test sets and for different tasks, and is intended as a visual representation of how our models perform in isolation compared to when they are integrated as a component in the overall system.

We can see that models that performed the best when they are tested in isolation tended to produce architectures which performed worse overall after they are combined with the other components. The "Label Accuracy" metric used in both tasks uses the same scale, so we can compare them next to each other to represent this discrepancy.

A.13 Examples from Gathered Datasets

A.13.1 Examples from X Claims Test Set

Claim: *There is no climate crisis that kids need to be worried about the climate will change like it always has because of the sun.*
Veracity: `False`

Claim: *Climate change is not only driving global heatwaves, droughts and glacier melting, it has also altered the flow of rivers across the world.*
Veracity: `True`

Claim: *Forest fires have been going on since the earth was formed. Some year are always worse than others. They are not due to climate change*
Veracity: `False`

A.13.2 Examples from Scientific Claims Test Set

Claim: *The stormier waters are caused by climate change, primarily driven by greenhouse-gas emissions from burnt fossil fuels like coal, oil, and gas, which continue to warm the planet and create conditions like rising sea levels and increased drought contribute to more severe storms.*
Veracity: `True`

Claim: *Global warming is not caused by the increase in atmospheric carbon dioxide concentration*
Veracity: `False`

A.13.3 Examples from Political Claims Test Set

Claim: *Corporations are largely responsible for pollution that contributes to climate change.*
Veracity: `True`

Claim: *to take on climate change in a serious way would involve making some tough political choices*
Veracity: `True`

A.13.4 Examples from Informal Claims Test Set

Claim: *climate change is a huge effing HOAX!!! ...that they are all benefiting from!!*
Veracity: `False`

Claim: *There is NO climate crisis only an ignorant controlled agenda crisis led by fools like you.*
Veracity: `False`

A.14 Domains Qualitative Experiment Results

Domain	Correct	UTV	Incorrect	Score
Scientific	9	1	0	0.9
Political	6	4	0	0.6
Informal	7	2	1	0.6

Table A.3: Performance of system on different types of climate-claim.

A.15 Fine-Tuning Hyperparameters

A.15.1 Re-Ranker Fine-Tuning

The following hyperparameters were used when fine-tuning RoBERTa and ClimateBert for the Re-ranker:

- Learning rate: 2e-5
- Batch size: 8
- Number of epochs: 5
- Weight decay: 0.01
- Warmup steps: 500

These hyperparameters were initially adapted from Liu et al. (2019) [36], and then adjusted accordingly for improvement.

A.15.2 Evidence Classifier Fine-Tuning

The following hyperparameters were used when fine-tuning Llama-2 7b and Llama-2 13b for the evidence classifier:

LoRA Config:

- Rank: 64
- LoRA α: 16
- LoRA Dropout: 0.1

Training Hyperparameters:

- Learning Rate: 2e-4
- Optimizer: AdamW 32 bit
- Weight Decay: 0.001
- Warmup Ratio: 0.03
- Max Gradient Norm: 0.3
- Number of Epochs: 1

These hyperparameters were adapted from Dettmers et al. (2024) [9]. We train for 1 epoch due to the computationally expensive nature of fine-tuning LLMs.

A.16 In-Depth Breakdown of System Test Results

Classifier Model	Re-ranker	SUPPORTS	REFUTES	Total
Llama-2 13b *FT:* CF	No	18	3	21
Llama-2 13b *FT:* CF	Yes	18	4	22
Llama-2 13b *FT:* CF+F	No	11	18	29
Llama-2 13b *FT:* CF+F	Yes	12	17	29
Llama-2 7b *FT:* CF	No	18	16	34
Llama-2 7b *FT:* CF	Yes	20	18	38
Llama-2 7b *FT:* CF+F	No	17	21	38
Llama-2 7b *FT:* CF+F	Yes	18	19	37

Table A.4: Number of claims correctly classified as “SUPPORTS” or “REFUTES” for each system architecture on the X claims test set

Classifier Model	Re-ranker	All NEI	Disputed	Total
Llama-2 13b *FT:* CF	No	29	0	29
Llama-2 13b *FT:* CF	Yes	28	0	28
Llama-2 13b *FT:* CF+F	No	11	6	17
Llama-2 13b *FT:* CF+F	Yes	16	3	19
Llama-2 7b *FT:* CF	No	6	8	14
Llama-2 7b *FT:* CF	Yes	6	4	10
Llama-2 7b *FT:* CF+F	No	6	4	10
Llama-2 7b *FT:* CF+F	Yes	10	1	11

Table A.5: Reasons for each system returning "UTV" on the X claims test set. "All NEI" means every evidence sentence was classified as "NOT_ENOUGH_INFO". "Disputed" means the system found both supporting and refuting evidence.

A.17 End-to-end Pipeline Examples

A.17.1 Successul Examples

Example 1:

This example shows the system correctly supporting a claim, by providing evidence and a source relating to methods for reducing carbon pollutions.

Input Claim:

Walk, bike, carpool, take public transportation or use electric vehicles to reduce air pollution and carbon emissions for a cleaner, greener future.

System Output:

This claim is **SUPPORTED** by the IPCC Sixth Assessment Report by the following evidence:

- **Evidence Sentence:** Compact cities with shortened distances between housing and jobs, and interventions that support a modal shift away from private motor vehicles towards walking, cycling, and low-emissions shared and public transportation, passive energy comfort in buildings, and urban green infrastructure can deliver significant public health benefits and have lower GHG emissions.
- **Report:** Working Group 3 - Climate Change 2022: Mitigation of Climate Change
- **Section Title:** Executive Summary

Example 2:

Here, the system correctly supports a claim about the impact of climate change on rivers

around the world, by providing 3 supporting sentences.

Input Claim:

Climate change is not only driving global heatwaves, droughts and glacier melting, it has also altered the flow of rivers across the world.

System Output:

This claim is **SUPPORTED** by the IPCC Sixth Assessment Report by the following evidence:

- **Evidence Sentence:** Climate change is increasing the frequency and intensity of these sudden and slow-onset disasters, among them, hydrological changes in major river basins where 1.5 billion people live .
- **Report:** Working Group 2 - Climate Change 2022: Impacts, Adaptation and Vulnerability
- **Section Title:** 10.5.6.2 Findings

- **Evidence Sentence:** Patterns in flow regimes can be directly linked to a variety of processes shaping freshwater biodiversity, so any climate change-induced changes in flow regimes and river connectivity are expected to alter species composition as well as having societal impacts) .
- **Report:** Working Group 2 - Climate Change 2022: Impacts, Adaptation and Vulnerability
- **Section Title:** 2.3.3.3 Observed Changes in Discharge

- **Evidence Sentence:** Already, one-third of wetlands have been lost and 9000 freshwater species are threatened with extinction without considering the effects of climate change , and only 13% of world rivers were recently classified as least impacted Globally, observed climate-driven changes in biodiversity are typically of species distributions shifting to higher latitudes .
- **Report:** Working Group 2 - Climate Change 2022: Impacts, Adaptation and Vulnerability
- **Section Title:** CCP1.2.1.1 Observed Impacts

Example 3:

This example shows the climate claim detection component correctly identifying that the claim is not climate-related, so evidence is not retrieved.

Input Claim:

Edinburgh is the capital city of Scotland and one of its 32 council areas

System Output:

Climate Claim not detected. Evidence will not be retrieved.

Example 4:

This example shows the system correctly refuting an incorrect claim, by providing three refuting statements.

Input Claim:

Climate Change is a natural cycle.

System Output:

This claim is **REFUTED** by the IPCC Sixth Assessment Report by the following evidence:

- **Evidence Sentence:** Anthropogenic climate change is unequivocal and ongoing.
- **Report:** Working Group 2 - Climate Change 2022: Impacts, Adaptation and Vulnerability
- **Section Title:** 1.3.2.1 Detection and Attribution of Climate Change and its Impacts

- **Evidence Sentence:** The science on climate change, its impacts and the opportunities to mitigate is continuously being updated.
- **Report:** Working Group 3 - Climate Change 2022: Mitigation of Climate Change
- **Section Title:** 4.4.3.3 New Information Becomes Available

- **Evidence Sentence:** Observations of past compound events may not provide reliable guides as to how future events may evolve, because human activity and recent climate change continue to interact to influence both system functioning and a climate state not previously experienced
- **Report:** Working Group 2 - Climate Change 2022: Impacts, Adaptation and Vulnerability
- **Section Title:** 2.3.2 Projected Impacts of Increases in Extreme Events

A.17.2 Unsuccessful Examples

Here is an example of the system returning “UTV”, as all retrieved evidence sentences received the “NOT_ENOUGH_INFO” tag.

Example 1:

Input Claim:

Nuclear has very low emissions and can give us large amounts of power

System Output:

Unable to verify claim.
Reason: Not Enough Information

Example 2:

This example shows an error in which the system finds both supporting and refuting evidence for a claim, so its output is "Unable To Verify". This is a political claim, which likely contributes to the system's inability to support or refute it properly.

Input Claim:

many Conservatives don't want to be part of leading the world in climate action.

System Output:

Unable to verify claim.
Reason: Supporting and Refuting evidence found.

A.18 ClimateGPT

In "ClimateGPT: Towards AI Synthesizing Interdisciplinary Research on Climate Change" [63], the authors discuss an alternate take on the CLIMATE-FEVER claim/evidence pair classification task for LLMs. The authors propose a task called "Fever-Boolean", where instead of presenting the model with a claim and evidence, we simply provide it with the claim and prompt the model to classify whether the claim is correct or misinformation. This task relies on a model's prior knowledge of a topic and would allow us to remove the evidence retrieval step completely. They achieve impressive results for this task by training an LLM from scratch on climate-specific text. An inevitable limitation of this approach is that LLMs are likely to hallucinate [40] and provide false information, unlike using an evidence bank of correct information. Also, such a system may reduce the "trust" from users in a claim verification system, as being presented with AI-generated text as evidence pales in comparison to rigorously peer-reviewed text from the IPCC AR6.

In the same paper, they also propose another approach to the task, which involves providing an LLM with a claim along with all 5 evidence sentences for the claim in CLIMATE-FEVER and prompting the model to classify if the evidence supports or

refutes a claim. This interesting approach would allow the model to take in more evidence information to classify a claim. In the future, we could fine-tune models to work with 5 evidence sentences to see if this improves the overall system.

A.19 Energy Consumption

Given that the motivation behind our system is to tackle online climate-change misinformation, we must be conscious of the energy consumption and carbon footprint of our work. While our fine-tuned Llama-2 13b model achieved the same performance on the test set as GPT-4, this is still an achievement due to our model size being much smaller than GPT-4. GPT-4 is likely much larger than GPT-3, which was made up of 175 billion parameters. This suggests that inference using GPT-4 will be much more computationally expensive, and therefore consume more power, than Llama-2 13b. If such a system were to be used on a large scale, it would be important to consider the energy consumption of the models used to ensure that it does not cause unnecessary harmful emissions.

Another benefit of using models created by Meta as opposed to models created by OpenAI is Meta's openness regarding the power consumption and carbon emitted when training their models [64]. For example, Meta has disclosed that the total power consumption when training Llama-2 13b was equivalent to 62.44 tonnes of CO_2 emissions. This allows us to consciously select models that emit less harmful gasses during their training, aligning with our goal of reducing the impact of climate change.

Furthermore, taking advantage of memory-saving techniques such as Quantized Low-Rank adaption greatly reduces our energy consumption in fine-tuning these models.