

Urdu-ClimateGPT: Adapting LLM for Climate Data in Urdu Language

Muhammad Farooq¹, Muhammad Asif Habib², Jabeen Sultana³, Muhammad Umar Aftab^{4*}

Department of Computer Science, National University of Computer and Emerging Sciences,
Chiniot-Faisalabad Campus, Pakistan^{1,4}

College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU),
Riyadh 11432, Saudi Arabia^{2,3}

Abstract—Climate change is increasingly recognized as a major global challenge that affects environmental systems, weather patterns, and human societies around the world. Rising global temperatures have been linked to more frequent extreme weather events and long-term shifts in climate patterns. Communicating climate information effectively, therefore, becomes essential, especially in a way that is accessible and inclusive. But languages like Urdu are under-represented in the sources of climate knowledge, thus leaving many communities with fewer reliable sources of climate knowledge. To address this issue, Urdu-ClimateGPT is introduced by this study, as a domain-adapted language model based on LLaMA 3.1, along with a retrieval-augmented conversational framework built around it. Domain specific fine-tuning is combined with retrieval-based grounding evidence by the system. This is an effort to make the hallucinatory responses less common and factual disalignments in the generated responses less frequent, in the context of conversations with climate-related topics. The model was evaluated on a held-out set of Urdu climate prompts, and compared to the baseline LLaMA 3.1 model. The findings reveal that Urdu-ClimateGPT outperforms in various automated evaluation metrics including: language fluency, domain-specific correctness, factual consistency, and response completeness. Overall, a normalized average score of 0.82 was achieved by the Urdu-ClimateGPT, whereas a score of 0.52 was scored by the baseline model. These results suggest that large language models for low-resource languages in specialized domains can be adapted, which is both feasible and beneficial. It is shown by the study that hallucination like behavior can be reduced by retrieval augmented architectures when evaluated using automated metrics. However, further evaluation by human experts will be necessary to determine the system's factual reliability and its potential real-world impact.

Keywords—Climate change communication; domain-specific AI; large language models; low-resource languages; natural language processing; retrieval augmented generation; Urdu language processing

I. INTRODUCTION

One of the most critical problems of the 21st century is climate change. As the world deals with the consequences of climate change: rising temperatures, extreme weather, events, and shifting climate patterns, the demand for a data-backed decision making system is paramount. Climate change is a global threat that affects communities with limited access to reliable information and resources more than others. This shows how important it is to make climate-related knowledge available to everyone [1], [2].

*Corresponding author

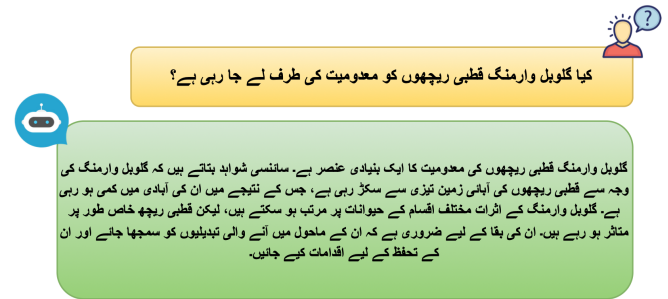


Fig. 1. Urdu-ClimateGPT conversation. User: Is global warming leading to the extinction of polar species? Response: Global warming is a key factor contributing to the decline of polar species' populations due to rapid habitat loss. Scientific evidence suggests that global warming accelerates land ice melting, leading to habitat shrinkage and decreasing population numbers.

While various species are affected, polar ecosystems are particularly vulnerable. Understanding these environmental changes is crucial for conservation efforts aimed at protecting these species.

Urdu is one of the most widely spoken languages in the world, with over 230 million speakers [3], [4]. However, there have been very few efforts to create centralized, climate-related LLM-based chatbots that are appropriate for their language and culture [5]. Fig. 1 presents a Urdu-ClimateGPT conversation.

This gap has to be closed at all costs in an attempt to create awareness and enable such communities to make data-driven decisions. Urdu also has been identified by UNESCO, as an institutional language that has been in use in both educational and administrative fields [6].

This study introduces Urdu-ClimateGPT, a fine-tuned language model, trained on an Urdu climate-specific dataset, and a retrieval augmented conversational system built on it. The system would be a useful source of climate-related information and fact-checking as it will allow Urdu speakers to check climate claims, learn the scientific concepts, and discover credible knowledge in their native language. It can support well-informed decisions of educators, journalists, policymakers, and the general population by offering precise and contextually sound information. Also, the system can be used to eliminate misinformation and myths concerning climate change and provide a more scientifically knowledgeable community more prepared to deal with environmental issues.

In recent years, several large language models (LLMs), including ChatGPT [7], Claude [8], and Google Gemini [9], have demonstrated exceptional performance as conversational

agents. However, since these models are not open-source, there remains, to the best of the authors' knowledge, no open-source AI system that matches their capabilities in the Urdu language. We fine-tuned LLaMa 3.1, an open-source model renowned for its multilingual capabilities.

A. Notable Contributions

Our work presents the following key contributions:

- We introduce Urdu-ClimateGPT, a domain-adapted language model based on LLaMa 3.1, fine-tuned on an Urdu-translated subset of the Clima500-Instruct dataset (10% of the english corpus) focused on climate change and sustainability.
- We design a retrieval augmented conversational system that integrates Urdu-ClimateGPT with a retrieval augmented Generation (RAG) framework, utilizing the Climate-FEVER dataset for evidence grounding and the translated Clima500-Instruct corpus for contextual retrieval, thereby reducing reliance on large-scale fine-tuning.

II. RELATED WORK

Recent advances in large language models (LLMs) have catalyzed growing interest in domain-specific conversational agents for climate science communication. Arabic Mini Climate-GPT [10] introduces a climate-focused conversational agent focused on Arabic language. Despite sharing a script derived from the Arabic alphabet, Urdu and Arabic are distinct languages with different phonetics, vocabulary, and grammatical structures. Similarly, ClimateQ&A [11] is a ChatGPT-based tool that distills climate science knowledge from IPCC reports and environmental articles into user-friendly responses. While available on HuggingFace, it is not open-sourced, as it relies on ChatGPT, and its responses are limited to English. ClimateBot [12] is another climate question-answering model based on a clustering approach. However, it is limited to localizing answers from a given context rather than generating responses like modern large language models (LLMs). Additionally, its responses and document embeddings are restricted to English.

ClimateGPT introduces a family of climate-oriented LLMs trained on interdisciplinary climate data and optimized with retrieval-based methods to improve grounded responses, addressing both scientific breadth and multilingual access challenges [11]. Similarly, specialized models such as Arabic Mini-ClimateGPT leverage open-source instruction tuning on curated climate change corpora to produce conversational responses in Arabic, including retrieval mechanisms for improved fidelity [10]. The potential of tailoring LLMs for domain-specific tasks is demonstrated by these efforts, but they are still limited to higher-resource languages or single language families, leaving low-resource linguistic regions.

In related climate NLP research, model understanding of climate content has been assessed through the development of benchmark datasets and evaluation frameworks. Laud et al. introduce Climabench, a comprehensive dataset for climate change text understanding, facilitating more rigorous evaluation of models on domain-specific tasks [13]. Climate-Fever provides labeled examples for factual verification of

climate claims, enabling supervised learning and evaluation of fact-checking performance in climate discourse [14]. Recent work at the ClimateNLP workshop is explored by retrieval augmented generation (RAG) approaches for climate change information communication as a complement to these benchmarks. For example, Biswas et al. test RAG systems with the UK climate projections data and show that retrieval-based agents are much more effective in making the right answers in the context of the situation compared to generic agents [15]. Collectively, these datasets point to the growing interest in benchmarking, as well as improving model reliability in climate NLP.

To minimize the hallucinations and enhance the factual consistency, retrieval augmented generation (RAG) has become an important architecture to facilitate the verification of climate claims tasks [16]. RAG models incorporate dense retrieval and large language model generation, allowing models to condition external evidence at inference [17]–[19]. Ram et al. extend RAG to in-context retrieval paradigms as retrieval processes may be combined during the generation phase to enhance completeness and accuracy of answers [20]. Surveys conducted recently place RAG as the key to sound domain-specific systems, and it may be assumed that the more generative responses are based on structured knowledge bases, the more dependable they are and the lower the level of unjustified inference [21], [22]. This is the direct motivation behind our application of the retrieval pipelines in Urdu-ClimateGPT wherein we assure the factuality of climate claim responses.

In modern NLP research, the issue of adapting LLMs to low-resource languages like Urdu has been identified. Arif et al. compare a variety of general and specialized models to a set of Urdu NLP tasks and find a significant performance difference between generalist and finetuned models in low-resource conditions [23]. Also, benchmarks like UrBLiMP provide a systematic assessment of Urdu linguistic competence across syntactic phenomena, highlighting the shortcomings of current multilingual models in capturing fine-grained language properties [24]. The datasets of bias and contextual evaluation such as PakBBQ also highlight the need for culturally and regionally sensitive benchmarks to measure fairness and contextual soundness of question answering tasks with Urdu [25]. These results support the need for targeted adaptation and assessment when extending LLM to under-served languages and groups.

One of the main problems present throughout LLM applications is hallucination and transmission of fake information, particularly within high stakes fields, such as climate change. Mitigating hallucinations often involves combining retrieval grounding, domain-specific training, and rigorous evaluation. ClimateGPT itself incorporates hierarchical retrieval strategies to reduce unsupported outputs. In addition, domain benchmarking efforts stress the need for multi-facet evaluation of faithfulness and relevance when deploying AI for scientific communication [15]. By integrating a retrieval augmented architecture with a secondary verification mechanism, Urdu-ClimateGPT aims to address both the multilingual adaptation gap and the need for trustworthy, climate-grounded responses in low-resource contexts.

In contrast to prior work, Urdu-ClimateGPT uniquely combines (1) domain-specific fine-tuning on a translated Clima500-

Instruct dataset, and (2) retrieval grounding using ClimateFEVER evidence. To the best of our knowledge, this is the first open-source system to integrate fine-tuning, retrieval-augmented generation for climate discourse in Urdu.

III. DATA SOURCES

A. CLIMATE-FEVER Dataset

This dataset is a publicly available resource designed for the verification of climate change related claims. It comprises 1,535 real-world claims collected from the internet, each annotated with evidence from reputable information sources and labeled as “Supports”, “Refutes”, “Disputed”, or “NotEnoughInfo”. Fig. 2 shows the distribution of data across different categories. Each claim in the CLIMATE-FEVER dataset is paired with five manually annotated evidence sentences sourced from English Wikipedia, totaling 7,675 claim evidence pairs. These evidence sentences are categorized as supporting, refuting, or providing insufficient information regarding the claim. The dataset presents complex claims that encompass multiple aspects and includes contentious cases where evidence both supports and refutes the claim.

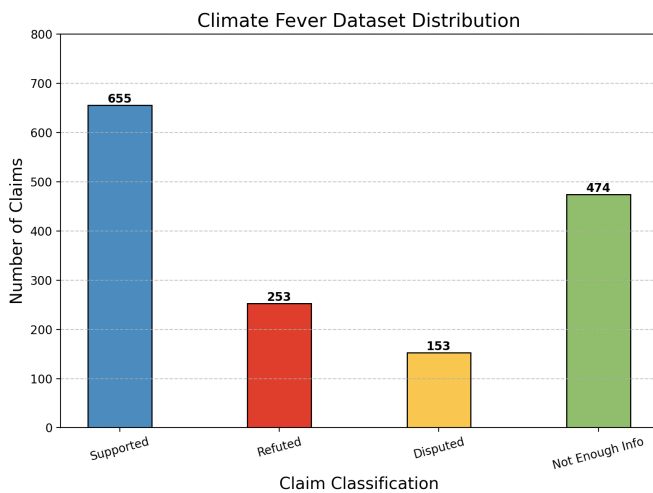


Fig. 2. Climate-FEVER dataset distribution statistics.

B. Clima500-Instruct Dataset

This dataset is a conversational instruction-tuning corpus that is specifically designed to address the topics of climate change and sustainability. It has been created with the aim of improving conversational agents by converting the existing simple question answering datasets such as CCMRC and Clima-QA into an interactional dialogue. The dataset was created by following several steps: first, extracting the question-answer pairs, then generating the conversation answers with the help of ChatGPT-4, translating the answers into Arabic while preserving the contextual integrity, and finally manually post-processing by native Arabian speakers. The dataset covers various climate-related topics. Fig. 3 shows the distribution of data across different climate-related categories. This dataset provides a rich and structured resource to improve AI driven climate discussions in Arabic, making it a valuable contribution to sustainability-focused NLP research. From the full

Clima500-Instruct corpus (512,081 English), we selected a stratified 10% subset (50,000 samples) for Urdu translation and model fine-tuning. This subset size was chosen to balance domain coverage, computational feasibility, and diminishing performance returns observed beyond this scale. Our approach ensures that the category distribution in the training data closely mirrors the original dataset, with major categories such as Greenhouse Emissions and Climate Policy/Laws being proportionally represented. Due to the inherent class imbalance in the source dataset, no additional re-balancing techniques were applied. We believe this preserves the natural distribution of climate topics. While we did not perform a fine-grained analysis of how class imbalance affects hallucination rates during fine-tuning, the combination of domain-specific fine-tuning and RAG consistently improved factual consistency across categories, as shown in our ablation study, Table IV.

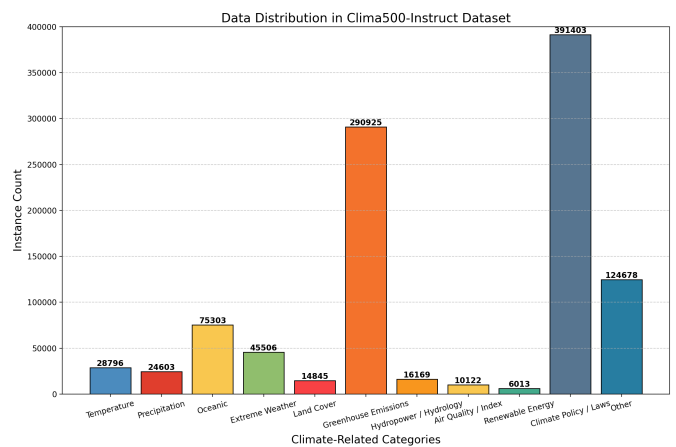


Fig. 3. Clima500-instruct dataset distribution statistics.

C. Urdu Translation of Clima500-Instruct

To construct an Urdu climate instruction dataset, a subset of Clima500-Instruct corpus was translated to Urdu using ChatGPT-4. This approach was selected due to its strong multilingual capabilities and it has been proven to yield grammatically correct, context sensitive translations in low resource languages [26], [27]. To ensure the quality and usability of the ChatGPT-translated Urdu dataset, we applied a systematic post-processing pipeline including removal of corrupted samples, manual correction of English fragments, and quality inspection by two native Urdu speakers on randomly selected samples. The native speakers reviewed a randomly selected subset of 500 samples (1% of the translated dataset), assessing fluency, semantic fidelity, and correctness of climate-specific terminology. Minor edits were made based on their feedback. The same translation pipeline was used consistently for both training and test sets.

1) *Translation method:* An entirely automated LLM-based pipeline was used for the translation. Every English instruction-response pair of Clima500-Instruct was entered into ChatGPT-4 with explicit prompts to preserve semantic meaning, technical terminology, and conversational tone. Special consideration was given to maintenance of domain specific vocabulary relevant to climate science (greenhouse gases,

deforestation, sea-level rise) and the prevention of literal, word-for-word translations that could distort meaning in Urdu.

2) *Quality assurance and validation*: To ensure the quality and usability of the ChatGPT-4 translated Urdu dataset, a systematic post-processing and manual filtering procedure was used. First, all the translated entries were automatically scanned in order to identify and remove malformed tokens, undefined symbols, and formatting artifacts introduced during generation. Second, remaining English words or phrases in Urdu questions or answers were manually detected and translated to Urdu to maintain linguistic consistency.

Third, the corrupted samples, where the response or the instruction was partially or fully in English, were eliminated. Lastly, to validate overall translation quality, two native Urdu speakers inspected, randomly selected subset of 500 samples (1% of the translated dataset), assessing fluency, semantic fidelity, and the correctness of climate-specific terminology. Their suggestions were then followed up with minor revisions for polishing and correcting the technical mistranslations in the downstream part of the phrasing. This post-processing pipeline guaranteed the final Urdu Clima500-Instruct subset was coherent, domain-faithful, and suitable for fine-tuning and retrieval augmented generation.

IV. METHODOLOGY

Fine-tuning is an important machine learning technique, especially in the field of natural language processing (NLP) [28]. This method involves a pre-trained model being further trained on a specialized dataset to adapt it for a specific task. Initially, large language models (LLMs) are trained on large amounts of general-purpose text data allowing them to develop a broad understanding of language. However, to obtain a higher degree of accuracy and contextual relevance in domain-specific tasks, fine-tuning is used.

In our work, we leverage fine-tuning to develop UrduClimateGPT, a model explicitly designed to produce retrieval-grounded and climate-related responses in the Urdu language. This was achieved by fine-tuning LLaMA 3.1 exclusively on the Urdu-translated 10% subset of the Clima500-Instruct dataset. This translated dataset plays a crucial role in fine tuning the model, allowing it to comprehend and verify claims in Urdu effectively. The fine-tuned Urdu-ClimateGPT model is a retrieval-grounded text generator, able to generate climate-focused responses based on external evidence under an automated evaluation protocol.

On top of fine-tuning, our approach incorporates retrieval augmented generation (RAG), a powerful technique that enhances contextual understanding and reliability of the generated responses. RAG combines the advantages of retrieval-based and generative models to reduce hallucination by grounding the generated text in verifiable and external sources of knowledge. In this method, relevant information is retrieved from a vector database, containing structured embeddings from authoritative sources such as Climate-FEVER. The retrieved context is then inputted into the language model, increasing the accuracy of its response and factual alignment.

A. System Architecture Overview

The overall architecture of Urdu-ClimateGPT follows a hybrid retrieval-generation design, integrating fine-tuning and retrieval augmented generation (RAG) to improve factual reliability and domain specificity. The system consists of four main components: 1) a preprocessing and embedding module, 2) a vector database for document storage, 3) a retrieval engine, and 4) a fine-tuned Urdu LLaMa model.

Climate-FEVER evidence sentences and translated Clima500-Instruct entries are embedded using a sentence transformer model and stored in the vector database. When a user query is received, semantically relevant documents are retrieved using similarity search and are appended as context to the fine-tuned Urdu-ClimateGPT model.

B. Vector Database Construction

To enable efficient semantic retrieval, we built a vector database containing dense embeddings based on two main sources, the Climate-FEVER dataset, and the translated Clima500-Instruct corpus. All pairs of evidence sentences and instructions-responses were embedded using a pre-trained sentence-transformer model.

The resulting vectors were indexed using a similarity search backend to handle quick approximate nearest-neighbor queries. Each of the vectors was accompanied by metadata such as source document ID, claim label (Support, Refute, Disputed), and topical category to facilitate filtered retrieval and structured grounding during inference.

C. Retrieval Augmented Generation Pipeline

Urdu-ClimateGPT uses a retrieval augmented generation (RAG) to ground responses in authoritative climate knowledge. After a user query, it first computes its embedding and retrieves the semantically-closest top- k entries of the vector database.

The documents returned are formatted into a structured context block appended with the original query using a pre-defined prompt template. The resulting enriched prompt is then passed to the fine-tuned Urdu-ClimateGPT model, which allows generating contextually aware and fact-grounded responses.

In our implementation, we set $k = 5$ to balance retrieval diversity and prompt length constraints. This design reduces hallucinations and improves factual consistency, particularly for climate claim verification tasks.

D. Claim Detection and Routing Logic

To differentiate factual claims from general informational queries, we implemented a lightweight claim detection mechanism using a general-purpose LLM (Base LLaMa 3.1). The claim detection model was not fine-tuned on the Urdu climate dataset and operates independently of the response generation pipeline to avoid training leakage or bias. Each user query is first classified into one of two categories: 1) factual claim or 2) informational inquiry.

If the query is classified as a factual claim, the system triggers the retrieval and verification pipeline, retrieving

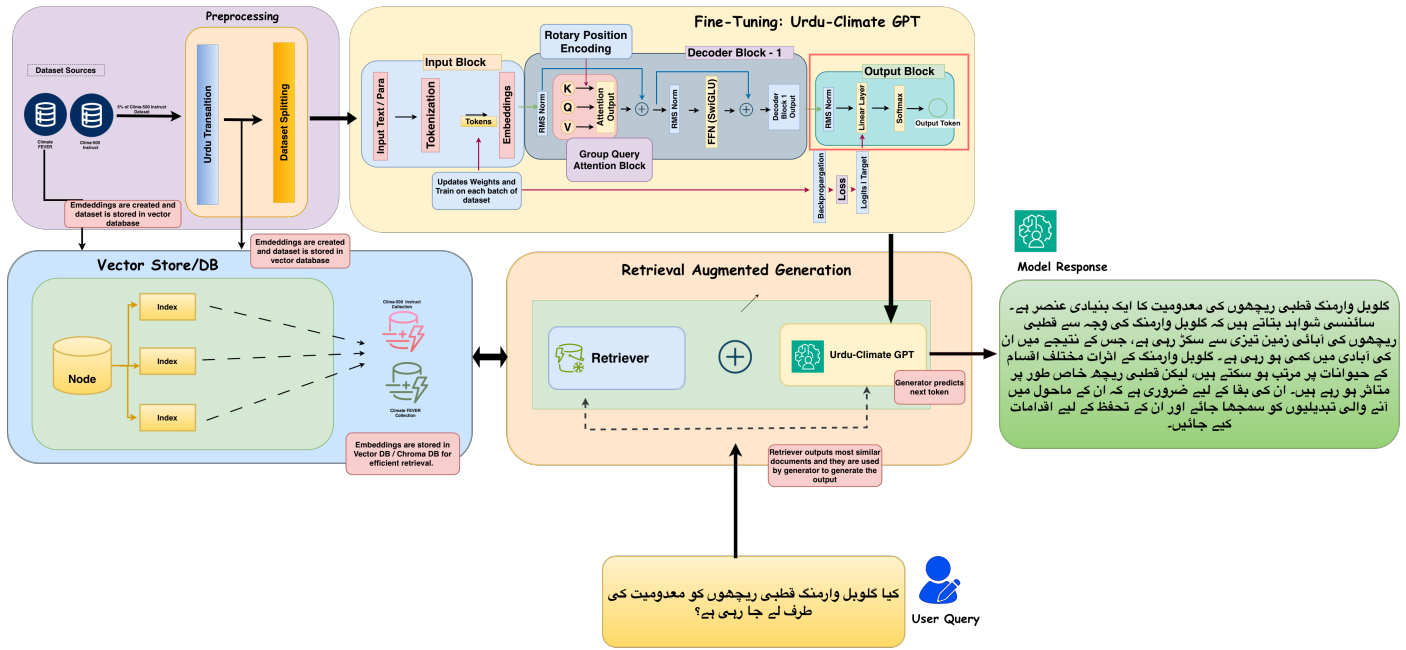


Fig. 4. Overview of the Urdu-ClimateGPT architecture integrating fine-tuning and retrieval augmented generation (RAG). Climate-FEVER embeddings are stored in a vector database for factual claim verification, while translated Clima500-Instruct embeddings are used for contextual grounding.

supporting or refuting evidence from the Climate-FEVER vector database. If the query is categorized as an informational inquiry, it is directly forwarded to the fine-tuned Urdu-ClimateGPT model.

This routing facilitates the optimal use of computational resources while maintaining high factual accuracy of claim-based queries.

The hybrid methodology provides factual credibility to the climate-related responses in the Urdu language, diminishes the hallucination-like effect, improves factual consistency, and ensures contextual accuracy in verification tasks. The overall methodology is illustrated in Fig. 4.

E. Embedding Model and Similarity Search

To generate dense vector representations for semantic retrieval, the pretrained sentence-transformer model all-MiniLM-L6-v2 was used to encode Climate-FEVER evidence sentences and translated Clim500-Instruct entries. This model was chosen because of high quality semantic representation and computational efficiency [29], therefore making it suitable for real-time retrieval in a low resource deployment environment.

All text inputs were normalised before being embedded to ensure consistent vector representations across datasets. The resulting embeddings were saved in a vector database for near instant ANNS (approximate nearest-neighbor search).

The cosine similarity between the user query embedding and stored document embeddings was used for semantic retrieval. The top-k most similar entries were retrieved with a k value of 5 as it was found to be empirically validated for a balance between this retrieval diversity and the constraint of prompt length.

F. Vector Database Backend

To build the retrieval backend, ChromaDB, a high-dimensional similarity search vector store was used. All embedded documents were stored, each having metadata fields [source dataset (Climate-FEVER), claim label, topic category].

This metadata-based filtering allowed for the specific retrieval of claims verification queries and thus factual grounding to be based on Climate-FEVER evidence. The vector store has cosine similarity search and approximate nearest neighbor indexing to enable low-latency retrieval in real-time interactions.

G. Prompt Engineering for Retrieval Augmented Generation

Evidence passages were retrieved and inserted into the generation prompt through a template that consists of three parts: 1) system instructions to enforce factual grounding, 2) a contextual evidence block with the top-k retrieved passages and 3) the original user query.

The specific instructions to the model prompted it to stick to the evidence and refrain from speculative or unsubstantiated claims. The factual consistency of the generated statements and the avoidance of hallucination were improved with this controlled prompting strategy for climate claim verification.

H. Computational Efficiency and Latency

The retrieval and verification components introduce modest computational overhead compared to direct generation. On average, the full pipeline required approximately $1.4\times$ the inference time of the base LLaMa 3.1 model due to embedding lookup and vector search. Despite this overhead, the system remained suitable for near real-time interaction in practical deployment scenarios.

V. TRAINING CONFIGURATION

Key dataset, training, and hyperparameter configurations are summarized in Table I and Table II. We utilized a stratified 10% subset (50,000 samples) of the Clima500-Instruct (English) dataset, split into 49,800 training and 200 test prompts for fine-tuning Urdu-ClimateGPT. Fine-tuning was performed over 3 epochs on a Google Colab Pro GPU, requiring approximately 5 hours. Low-Rank Adaptation (LoRA) was applied with rank 16, scaling factor 32, and dropout 0.1, alongside AdamW optimization at a learning rate of 2×10^{-5} and a batch size of 8.

TABLE I. LoRA AND OPTIMIZER HYPERPARAMETERS

Parameter	Value
LoRA rank (r)	16
LoRA scaling (α)	32
LoRA dropout	0.1
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	8
Quantization	4-bit (bnb)

TABLE II. DATASET AND TRAINING CONFIGURATION

Parameter	Value
Dataset	Urdu-translated Clima500-Instruct
Total entries used	50,000
Sampling method	Stratified random sampling
Training set	49,800 entries
Test set	200 entries
Epochs	3
Training time	≈ 5 hours
Compute environment	Google Colab Pro GPU

VI. EXPERIMENTS AND RESULTS

To evaluate the effectiveness of our fine-tuned LLaMa model on Urdu climate change data, we conducted a pairwise comparison experiment against the baseline LLaMa model using ChatGPT-4 as a blinded LLM-based preference assessor. We curated a test set of climate change-related prompts in Urdu that were not seen during training. For each prompt, we generated responses from both the baseline and our fine-tuned model.

To assess qualitative performance, we employed a blind evaluation approach using a separate, unbiased chatbot-based preference assessor [30]. The two anonymized outputs (in randomized order) for each test prompt were presented to the assessor, who was asked to choose the better response or indicate if neither was preferable. This process is mirrored by the evaluation protocol shown in similar comparative studies, though it is adapted for Urdu and the climate change domain.

The preference percentages are presented in Table III. Our fine-tuned model was significantly preferred over the baseline, indicating strong performance improvements in generating relevant and fluent Urdu responses.

These results highlight the effectiveness of domain- and language-specific fine-tuning in improving response quality for specialized, underrepresented languages like Urdu.

TABLE III. PREFERENCE RESULTS COMPARING OUR FINE-TUNED LLaMA-URDU MODEL WITH THE BASELINE LLaMA ON URDU CLIMATE CHANGE PROMPTS.

Model	Ours	Competitor	Neither
Baseline Lama3.1	86.50%	12.00%	3%

A. Component-Wise Performance Comparison

The incremental impact of fine-tuning and retrieval augmented generation (RAG) on response quality was analyzed through a component-wise performance evaluation conducted across three system variants: 1) the baseline LLaMa 3.1 model, 2) the fine-tuned Urdu LLaMa model without RAG, and 3) the full Urdu-ClimateGPT system with both fine-tuning and RAG enabled.

The same held-out test set of 200 Urdu climate-related prompts was used to evaluate each model. The 200 held-out test prompts were obtained via stratified random sampling from the Clima500-Instruct dataset to preserve the original topic distribution. The resulting topic-wise distribution in the test set is as follows:

- Greenhouse Emissions: 57 prompts
- Climate Policy / Laws: 38 prompts
- Oceanic: 15 prompts
- Extreme Weather: 9 prompts
- Temperature: 6 prompts
- Precipitation: 5 prompts
- Land Cover, Hydrology, Air Quality, Renewable Energy: 70 prompts combined

This provides reasonable coverage across major climate topics aligned with the source dataset 3. For every prompt, scores were assigned along four qualitative dimensions: linguistic fluency, domain accuracy, factual consistency, and response completeness. The scores were normalized to a 0–1 scale and averaged across all test samples.

Fluency measures grammatical correctness and naturalness of Urdu output. Factual correctness and alignment with established climate knowledge are reflected by accuracy. Whether the response fully addresses the user query without missing key points is captured by completeness.

Fig. 5 illustrates the comparative performance of the three model variants. Limited accuracy and incomplete responses are shown by the baseline model. Fluency and domain accuracy are substantially improved by fine-tuning, while factual grounding and completeness are further enhanced by the addition of RAG. The highest overall performance across all four metrics is achieved by the full Urdu-ClimateGPT system, demonstrating that complementary benefits are contributed by fine-tuning and RAG.

VII. ABLATION STUDY

To quantify the individual contributions of fine-tuning and retrieval augmented generation (RAG) in Urdu-ClimateGPT, an ablation study was conducted by systematically disabling

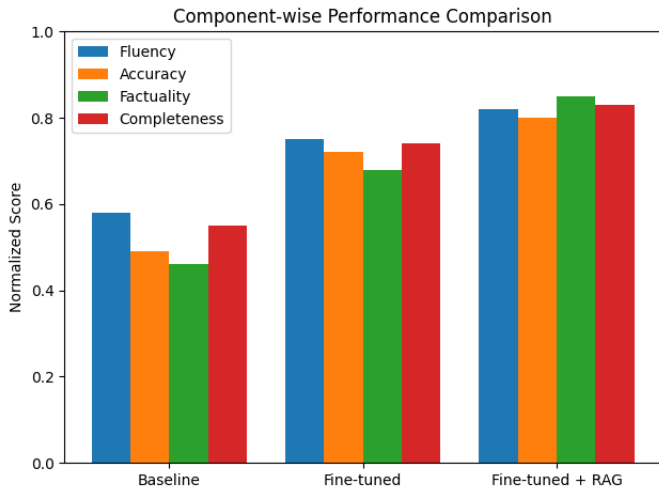


Fig. 5. Component-wise performance comparison of baseline LLaMa 3.1, fine-tuned Urdu LLaMa, and full Urdu-ClimateGPT across fluency, accuracy, factuality and completeness.

key components of the proposed system. Four configurations were evaluated: 1) the baseline LLaMa 3.1 model without any adaptation, 2) LLaMa 3.1 that was fine-tuned on the Urdu Clima500-Instruct dataset without RAG, 3) the baseline LLaMa 3.1 with RAG enabled but without fine-tuning, and 4) the full Urdu-ClimateGPT system with both fine-tuning and RAG enabled. The same held-out Urdu climate prompt set described in Section VI was used to test all configurations. Performance was measured by four criteria: linguistic fluency, domain accuracy, factual consistency, and response completeness. Each metric was normalized to a 0–1 scale and averaged across 200 evaluation prompts.

Table IV reports the comparative results. The baseline model performs weakest across all metrics, indicating limited capability in handling Urdu climate-specific queries. Fine-tuning alone yields substantial improvements in fluency and domain accuracy, while RAG alone enhances factual consistency by grounding responses in external climate evidence, while yielding moderate gains in domain accuracy and response completeness relative to the baseline. The full Urdu-ClimateGPT system achieves the highest overall score, demonstrating that fine-tuning and RAG provide complementary benefits and jointly reduce hallucinations while improving response quality.

TABLE IV. ABLATION STUDY ACROSS FLUENCY, ACCURACY, FACTUALITY, AND COMPLETENESS.

Model Variant	Fluency	Accuracy	Factuality	Completeness	Overall
Baseline LLaMa 3.1	0.58	0.49	0.46	0.55	0.52
+ Fine-tuning only	0.75	0.72	0.68	0.74	0.72
+ RAG only	0.64	0.63	0.76	0.65	0.67
Urdu-ClimateGPT (Full)	0.82	0.80	0.85	0.83	0.82

A. Urdu-ClimateGPT response vs Baseline LLaMa

Example 1: Fig. 6

Prompt: The impact of rising temperatures on the Earth’s surface due to climate change.



Fig. 6. Urdu question on impact of rising temperature on sea level with model response.

Response 1 (Urdu-ClimateGPT): When temperatures rise on Earth due to climate change, many environmental and natural systems are affected. Climate change is caused by an increase in greenhouse gases in the atmosphere, which trap more heat. As a result, more heat is added to the Earth’s surface, leading to several consequences. This not only increases the average temperature but also intensifies heatwaves. Due to rising temperatures, glaciers and polar ice are melting, which causes sea levels to rise. This increases the risk of flooding in coastal areas. In addition, weather patterns are changing, leading to extreme events such as droughts, floods, and storms. Ecosystems are also affected by rising temperatures. Difficulties are faced by many plants and animals in adapting to new conditions, which threatens biodiversity. The productivity of agriculture is affected, as crops could fail due to high temperatures and lack of water. Further, increases in temperature, which will reduce the availability of clean drinking water, can contribute to water scarcity. In conclusion, the rising temperature of the earth’s surface as a result of climate change is a serious issue that is impacting human life, ecosystems and natural resources. Global actions like greenhouse gas reduction and sustainable practice are needed to mitigate these impacts.

Response 2 (LLaMa 3.1): When temperatures rise on Earth due to climate change, it leads to many major environmental changes. One major effect is that ice is melting in the poles causing sea levels to rise. This poses a risk of flooding for coastal cities and island countries. Furthermore, elevated temperatures makes extreme weather events, like heatwaves, droughts, hurricanes and heavy rainfall more likely and intense. They inflict damage to infrastructure, losses to agriculture and endanger human lives. Ecosystems are also impacted by the increase of temperature. Many species are struggled with to survive as their natural habitats are changed or disappeared. Coral reefs suffer from bleaching caused by warmer ocean temperatures, and forest are increased in their susceptibility to fire. Another major impact that is recognized as serious is the scarcity of water. The evaporation rates are increased with rising temperature, which decreases the availability of freshwater. This poses challenges for drinking water supply, irrigation and energy production. Overall, widespread and long-term impacts on the Earth’s surface, human societies, and natural systems are had by climate change–driven temper-

ature increases. This issue demands international collaboration, carbon emissions reduction, and a transition to renewable energy.

Example 2: Fig. 7

The screenshot shows a question in Urdu asking about the causes of global warming and the benefits of planting trees. It includes two model responses. The first response explains that trees absorb CO2 and release oxygen, helping to reduce greenhouse gas concentrations. The second response states that planting trees can help reduce carbon dioxide by absorbing it from the atmosphere through photosynthesis, but it is not a complete solution as tree growth is limited and land-use conflicts exist.

Fig. 7. Urdu question on carbon dioxide and tree planting with model responses.

Prompt: Reducing carbon dioxide in the atmosphere by planting trees.

Response 1 (Urdu-ClimateGPT): Trees play a vital role in reducing carbon dioxide because they absorb CO₂ during photosynthesis and release oxygen. When more trees are planted, more carbon dioxide is removed from the air, helping to reduce greenhouse gas concentrations. This process slows global warming and improves air quality. Trees also help regulate temperatures by providing shade and releasing moisture into the atmosphere, which cools the surrounding environment. In addition, forests support biodiversity by providing habitats for animals and plants. Planting trees also improves soil quality, prevents erosion, and enhances water retention. Overall, increasing tree cover is a natural and effective way to mitigate climate change, protect ecosystems, and promote environmental sustainability.

Response 2 (LLaMa 3.1): Planting trees can help reduce carbon dioxide by absorbing CO₂ from the atmosphere through photosynthesis, but it is not a complete solution to climate change. Time is taken by trees to grow, and the carbon absorption capacity is limited. In some regions, large-scale tree planting may not be deemed suitable due to water scarcity or land-use conflicts. Forests can also be lost to wildfires, pests, or deforestation, with stored carbon being released back into the atmosphere. Tree planting is seen as helpful but it should be coupled with the phasing out of fossil fuel utilization, the safeguarding of existing forests and switching to renewable energy.

Example 3: Fig. 8

Prompt: Discuss the causes of air pollution in cities and suggest solutions.

Response 1 (Urdu-ClimateGPT): Urban air pollution is a serious problem affecting both the environment and human health. It is caused by different factors such as vehicle emissions, industrial activities, construction dust, fossil fuel combustion and inadequate waste management. The deterioration

The screenshot shows a question in Urdu asking about the causes of air pollution in cities and the health risks it poses. It includes two model responses. The first response lists factors like traffic congestion, industrial emissions, and fossil fuel use as causes, and mentions respiratory diseases, heart disease, and premature death as health risks. The second response adds that poor urban planning and high population density worsen the situation, and notes that air pollution also impacts human health through asthma and lung infections.

Fig. 8. Urdu question on urban air pollution and model responses.

of air quality is due to the emission of polluting gases like carbon monoxide, nitrogen oxides, sulfur dioxide and particulate matter (PM2.5). Exposure to polluted air over a long period of time increases the risk of respiratory diseases, heart disease, and premature death. Air pollution makes it difficult to see, damages buildings and affects urban ecosystems. In response to this, it is necessary to encourage public transport, the use of clean energy, environmental laws, the creation of green areas and sensitizing the public. Good policies and community engagement play a vital role in the process of improving air quality and thereby creating better living conditions in urban areas.

Response 2 (LLaMa 3.1): In cities, traffic congestion, industrial emissions, and the use of fossil fuels are the three main factors contributing to air pollution. Bad urban planning and high population density due to the growing number of vehicles and energy needs make the situation worse. Air pollution also has a negative impact on human health, contributing to asthma, lung infections and other respiratory diseases. Climate change and environmental degradation are also contributed to by it. This problem needs to be addressed by coordinated efforts of governments, industries and citizens. Improving public transport systems, promoting the use of electric cars, controlling industrial pollution, and planting more trees in urban areas are some of the measures that can help to reduce air pollution. In the end, mitigating urban air pollution is a public health issue as well as the environment.

VIII. LIMITATIONS

Urdu-ClimateGPT is a fine-tuned language model based on LLaMa architecture. Various data, such as large-scale web and publicly available text data, were used for pretraining the base model. Although a wide language functionality is given as a benefit by this, the risk of inaccuracy with the training data is present.

We employed a vector-based retrieval augmented generation (RAG) database to check the factuality of model-generated claims to enhance factual correctness and context relevance. These parts help considerably to improve the credibility of outputs, especially with regard to the use of climate discussion. There are, however, a few issues with these safeguards. The bias behaviour, however, can be continued in the underlying language model and/or the model can give over-confident and

speculative and subtly wrong answers [31]. These problems are particularly probable in case of insufficient context, ambiguous prompts or where the retrieval and judgment systems incorrectly categorize the query.

Therefore, users and researchers are advised to take Urdu-ClimateGPT's answers with a pinch of salt and think critically. The results of the models are not considered to be definitive until they are externally verified. It is seen as important to acknowledge these limitations for AI systems to be deployed more responsibly and to continue to be developed in the most sensitive and high-impact areas, including climate communication. Further, it is noted that automated evaluation may not be sufficient to cover the factual reliability in the real world, which is a major focus of our human expert validation.

IX. ETHICS STATEMENT

The research is entirely based on publicly accessible and open source data which does not contain any personal or identifiable data hence there is no privacy or ethical risk. All contributions and ideas are original and reference to all the datasets and models used has been made. It is important to acknowledge that, similar to any generative language model, the system can generate dangerous or offensive results when used irresponsibly and therefore, responsible use and careful monitoring during deployment are crucial.

X. FUTURE WORK

While this research uses LLM-based automated evaluation for scalability and consistency, the critical next step is the introduction of the structured evaluation by human experts to ensure factual correctness and real-world utility. In future work, we plan to use a multi-rater human assessment including climate science graduate students, environmental researchers, and Urdu-language educators. The human evaluation protocol will be used to evaluate model outputs in four dimensions, including: 1) linguistic fluency and grammatical correctness in Urdu, 2) factual accuracy in relation to established climate science, 3) evidence consistency with retrieved sources, and 4) response completeness and usefulness to end users. Each response will be independently rated by at least three annotators using a five-point Likert scale, with inter-annotator agreement measured using Krippendorff's alpha. The evaluators will not be told for any given response, which system variant produced it, to enable methodological rigor. The disagreements will be solved by means of majority voting and expert adjudication. In addition, a set of questions will be developed that will test extreme edge cases of ambiguous assertions, misinformation elicitation and culturally-relevant climate scenarios relevant to Urdu-speaking communities. This human-centered assessment will offer authoritative confirmation of factual accuracy, complement automated metrics, and establish a strong empirical foundation of Urdu-ClimateGPT implementation in educational, journalistic, and policy-support practice.

REFERENCES

- [1] S. C. Moser, "Communicating climate change: History, challenges, process and future directions," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 1, no. 1, pp. 31–53, 2010.
- [2] K. Nissen and H. Wong, "Democratizing climate change knowledge: The role of digital platforms," *Environmental Communication*, vol. 13, no. 3, pp. 345–360, 2019.

- [3] Ethnologue, "What are the top 200 most spoken languages?" 2025, accessed: May 2025. [Online]. Available: <https://www.ethnologue.com/insights/ethnologue200/>
- [4] "25 most spoken languages in the world in 2025," <https://www.berlitz.com/blog/most-spoken-languages-world>, 2024, accessed: 2025.
- [5] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the nlp world," *Proceedings of ACL*, 2020.
- [6] UNESCO, "Urdo — silk roads programme — unesco language profile," <https://en.unesco.org/silkroad/silk-road-themes/languages-and-endanger-languages/urdo>, 2025, accessed: 2025.
- [7] OpenAI, "Chatgpt," 2024, accessed: 2025-02-18. [Online]. Available: <https://openai.com/index/chatgpt/>
- [8] Anthropic, "Claude ai," 2024, accessed: 2025-02-18. [Online]. Available: <https://claude.ai>
- [9] Google, "Gemini ai," 2024, accessed: 2025-02-18. [Online]. Available: <https://gemini.google.com/>
- [10] S. S. Mullappilly, A. Shaker, O. Thawakar *et al.*, "Arabic mini-climategpt: A climate change and sustainability tailored arabic llm," *arXiv preprint*, no. 2312.09366, 2023. [Online]. Available: <https://arxiv.org/abs/2312.09366>
- [11] D. Thulke, Y. Gao, P. Pelsler, R. Brune, R. Jalota, F. Fok, M. Ramos, I. van Wyk, A. Nasir, H. Goldstein, T. Tragemann, K. Nguyen, A. Fowler, A. Stanco, J. Gabriel, J. Taylor, D. Moro, E. Tsybalov, J. de Waal, E. Matusov, M. Yaghi, M. Shihadah, H. Ney, C. Dugast, J. Dotan, and D. Erasmus, "Climategpt: Towards ai synthesizing interdisciplinary research on climate change," *arXiv preprint*, no. 2401.09646, 2024. [Online]. Available: <https://arxiv.org/abs/2401.09646>
- [12] Y. Z. Md Rashad and Others, "Climatebot: A machine reading comprehension bot for climate change documents," in *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 5243–5250. [Online]. Available: <https://www.ijcai.org/proceedings/2022/729>
- [13] T. Laud, D. Spokoyny, T. Corringham, and T. Berg-Kirkpatrick, "Climabench: A benchmark dataset for climate change text understanding in english," *arXiv preprint*, 2023. [Online]. Available: <https://arxiv.org/abs/2301.04253>
- [14] T. Diggelmann *et al.*, "Climate-fever: A dataset for verification of real-world climate claims," *arXiv preprint*, no. 2012.00614, 2020.
- [15] A. Biswas, H. Chahout, T. Pigram, H. Dong, H. T. Williams, F. Fung, and H. Xie, "Evaluating retrieval augmented generation to communicate UK climate change information," in *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*. Association for Computational Linguistics, 2025, pp. 126–141. [Online]. Available: <https://aclanthology.org/2025.climate-nlp-1.9/>
- [16] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, "Retrieval augmentation reduces hallucination in conversation," *Findings of ACL*, 2021.
- [17] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
- [18] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," *arXiv preprint*, no. 2007.01282, 2020.
- [19] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. tau Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, 2020, pp. 6769–6781.
- [20] O. Ram *et al.*, "In-context retrieval-augmented language models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1316–1331, 2023.
- [21] L. Huayang *et al.*, "A survey on retrieval-augmented text generation," *arXiv preprint*, no. 2202.01110, 2022.
- [22] Y. Gao *et al.*, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint*, no. 2312.10997, 2023.
- [23] S. Arif *et al.*, "Evaluating large language models for urdu," *arXiv preprint*, 2024. [Online]. Available: <https://arxiv.org/abs/2407.04459>

- [24] F. Adeeba, B. Dillon, H. Sajjad, and R. Bhatt, "Urblimp: A benchmark for evaluating the linguistic competence of large language models in urdu," *arXiv preprint*, 2025. [Online]. Available: <https://arxiv.org/abs/2508.01006>
- [25] A. Hashmat, M. A. Mirza, and A. A. Raza, "Pakbbq: A culturally adapted bias benchmark for qa," *arXiv preprint*, 2025. [Online]. Available: <https://arxiv.org/abs/2508.10186>
- [26] W. Jiao, X. Wang, S. Zhang, Y. Li, Y. Zhang, and Z. Tu, "Is chatgpt a good translator? a preliminary study," *arXiv preprint arXiv:2301.08745*, 2023.
- [27] L. Wang, Z. Tu, A. Way, and Q. Liu, "Large language models are state-of-the-art translators," *Transactions of the ACL*, 2023.
- [28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *JMLR*, 2020.
- [29] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *EMNLP*, 2019.
- [30] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Y. Wu, H. Zhuang, Z. Lin, Z. Zhang, J. Gonzalez, and I. Stoica, "Judging llm-as-a-judge: Bias and limitations of llm-based evaluation," *NeurIPS*, 2023.
- [31] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots," *Proceedings of FAccT*, 2021.