Climate Change Evidence Synthesis Using Open Source Model (LLaMA Model)

Sathwik Kuchana Sri Sai Lalitha Mallika Yeturi Priyadarshini Munigala
Yeshiva University, New York, NY, USA
{skuchana, syeturi, pmunigal}@mail.yu.edu

Abstract

Climate change adaptation research is growing exponentially, making manual evidence synthesis by experts increasingly challenging. This project investigates the use of an open-source large language model (LLaMA) for automating climate adaptation evidence synthesis. We develop a pipeline that combines semantic text embeddings and a FAISS similarity index to retrieve relevant context chunks from documents, which are then used to prompt a fine-tuned LLaMA model for extracting key adaptation information. Specific features studied include geographic location, stakeholders involved, and depth of the adaptation response—categorized into low, medium, and high levels of transformation. The results demonstrate that with fine-tuning, the LLaMA model achieves promising accuracy, especially for lower-expertise fields like geography, while high-expertise fields like adaptation depth remain challenging.

1. Introduction

Evidence synthesis is a foundational task in many scientific and policy-making domains. As the volume of peer-reviewed literature continues to grow exponentially, synthesizing structured insights from this unstructured corpus has become a pressing challenge. In fields such as medicine, law, and climate science, stakeholders increasingly rely on systematic reviews and structured analyses to inform decisions. However, traditional evidence synthesis methods—typically conducted by domain experts—are labor-intensive, time-consuming, and not easily scalable.

In the context of climate change adaptation, these challenges are particularly acute. Climate change is a rapidly evolving global crisis, and effective adaptation strategies are critical to mitigating its diverse impacts. Decision-makers rely heavily on adaptation-related research to understand how communities, governments, and ecosystems are responding to climate stressors. Initiatives like the Global Adaptation Mapping Initiative (GAMI) [2] have demonstrated both the importance and the scale of such efforts, requiring extensive time and coordination from expert annotators to manually extract key features such as geographical focus, stakeholder involvement, and the transformative depth of adaptation strategies.

Unlike general evidence synthesis tasks, climate adaptation introduces unique complexities. The required features—such as transformation depth—are often embedded in narrative context, demanding interpretive reasoning rather than straightforward entity recognition. Moreover, climate adaptation responses vary widely across geographic regions, socio-political systems, and sectoral domains, adding additional layers of nuance.

Recent advances in natural language processing (NLP), particularly the development of large language models (LLMs), offer a potential solution. LLMs have demonstrated impressive capabilities in zero-shot and few-shot settings, performing well across domains such as biomedical literature synthesis [4]. When fine-tuned or guided by structured prompting, these models can replicate human-level annotation performance while significantly reducing the manual burden.

Motivated by these advances, our study investigates the use of an open-source LLaMA model with 3 billion parameters for structured evidence extraction from climate adaptation literature. We target three key features from the GAMI dataset: (1) **Geographic location**, referring to the country or region where adaptation occurs; (2) **Stakeholders**, representing the actors involved in the response; and (3) **Depth of adaptation**, indicating how transformative the intervention is.

This task differs from generic summarization or question-answering tasks due to its need for high-precision, schema-constrained outputs and contextual

domain understanding. We hypothesize that with targeted fine-tuning, chunk-based retrieval, and prompt engineering, LLaMA can support high-quality, automated evidence extraction at a level that approximates expert annotators—especially for features with lower complexity such as geography and stakeholder identification.

This paper presents our methodology, retrieval-augmented prompt design, and fine-tuning pipeline, along with a detailed evaluation of model performance on structured climate adaptation tasks. While our focus is on climate science, we frame this work as a case study in building scalable, reproducible evidence synthesis pipelines that can generalize across domains where structured insight from textual data is needed.

2. Related Work

Large Language Models (LLMs) have rapidly transformed natural language processing by enabling general-purpose models to achieve strong performance across a wide range of tasks with minimal supervision. OpenAI's GPT-4 and its multimodal successor GPT-4o [10] have demonstrated impressive capabilities in zero-shot and few-shot settings, where models are expected to perform complex tasks without domain-specific fine-tuning. These capabilities have opened the door to applying LLMs in specialized domains such as law, finance, and healthcare, where manual data annotation and knowledge extraction are traditionally time-consuming and expert-dependent.

One major area of progress is in structured information extraction, where pretrained LLMs are repurposed to extract schema-constrained data from unstructured sources. AnnoLLM [5] showed that large foundation models can be fine-tuned to efficiently generate high-quality annotations with reduced human oversight. The authors demonstrated success in biomedical settings, where the extraction of disease symptoms, drug names, and lab findings requires domain familiarity. Their work highlighted the importance of prompt structure and context presentation when fine-tuning models for structured tasks.

In the biomedical domain, Goel et al. [4] showed that domain-specific LLMs, when exposed to only a few examples, could rival human annotators in structured data extraction. Their study compared general-purpose and biomedical-specific LLMs, concluding that even lightweight models, when fine-tuned on well-labeled domain data, can achieve robust performance. This reinforced the potential of LLMs as scalable annotators in evidence-rich disciplines.

In the context of climate science, few studies have rigorously explored the role of LLMs for structured extraction. Joe et al. [7] conducted one of the first targeted evaluations using GPT-4o on the GAMI food sector subset, testing the model's accuracy across tasks of varying complexity. They found that GPT-4o achieved high precision and recall for low-expertise tasks like geographic location extraction (F1 = 0.89), but performance dropped significantly on intermediate and high-expertise tasks such as stakeholder classification (F1 = 0.54) and adaptation depth (F1 = 0.22). These results illustrate the limitations of using general-purpose LLMs without fine-tuning for domain-specific adaptation tracking. Moreover, the study emphasized the need for structured prompting and task-specific training to improve reliability, particularly in complex classification tasks.

Our study builds on these foundations by exploring whether an open-source LLaMA model can replicate expert-level annotation accuracy for select features in the GAMI dataset. By combining retrieval-augmented prompting, supervised fine-tuning, and chunk-based context expansion, we aim to improve model performance on both low-expertise (e.g., geographic location) and intermediate/high-expertise features (e.g., stakeholder classification and adaptation depth).

3. Methods

To explore the effectiveness of large language models in structured evidence synthesis, we constructed a pipeline around a subset of the Global Adaptation Mapping Initiative (GAMI) dataset [2]. We focused exclusively on the "food security" focus group, consisting of 551 research articles that discuss climate adaptation interventions aimed at agricultural and food system resilience.

Each document was previously parsed and converted to markdown format from full-text PDFs. This structured format enabled efficient chunking and metadata alignment for downstream processing.

Figure 1 presents an overview of the pipeline used for structured evidence synthesis. The system begins with raw research articles, which are parsed and split into semantically meaningful chunks. These chunks are embedded using domain-relevant models (e.g., MiniLM), indexed in FAISS, and retrieved based on user queries. Retrieved context is integrated into a prompt template and passed to a fine-tuned LLaMA model. The output is a structured JSON containing extracted fields such as *Geography*, *Stakeholders*, and *Depth of Adaptation*.

Although this pipeline is instantiated here for climate adaptation, it generalizes to other domains where evidence extraction from large text corpora is critical. For instance, replacing the schema with biomedical entities (e.g., interventions, conditions, trial outcomes)

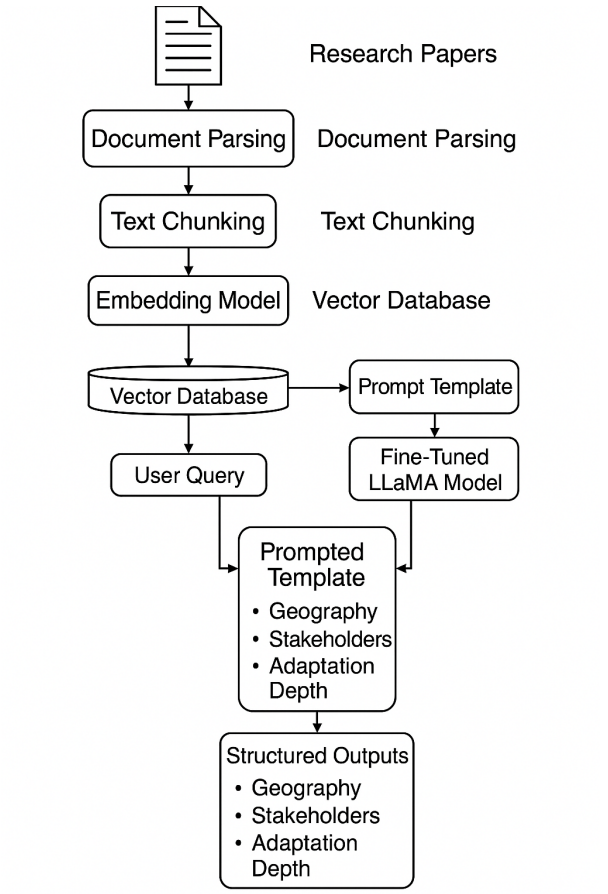

Figure 1: Modular architecture for structured evidence synthesis using LLMs, adaptable across domains like climate, medicine, and law.

or legal annotations (e.g., precedents, rulings, jurisdictions) enables reuse of the architecture without re-engineering the end-to-end system.

3.1. Embedding and Indexing

We split the markdown content of each article into overlapping chunks of approximately 300 words. This chunk size strikes a balance between preserving semantic context and avoiding overly long input sequences that could exceed model limits.

For each chunk, we computed vector embeddings using the 'sentence-transformers/all-MiniLM-L6-v2' model—an open-source Transformer-based model optimized for semantic similarity. It offers a good trade-off between accuracy and computational efficiency, making it particularly suitable for tasks requiring dense document encoding at scale.

These embeddings were stored in a FAISS index, a highly efficient similarity search library for dense vectors. We constructed a per-document index to enable retrieval of the most relevant chunks corresponding to a specific query. At inference or training time, we retrieved the top-k most semantically similar chunks from each document using cosine similarity between query and chunk embeddings.

3.2. Prompt Generation

The retrieved chunks served as context for prompting the language model. For each query, we generated a multi-part prompt containing:

- **Geography Extraction:** Asking the model to identify countries or regions where adaptation was observed.
- **Stakeholder Classification:** Instructing the model to categorize actors involved using the taxonomy defined by GAMI—e.g., national governments, NGOs, civil society, individuals.
- **Depth Assessment:** Asking the model to evaluate how transformative the adaptation response was (Low, Medium, High) and justify the classification.

This prompt structure was consistent across all training and test samples. For training instances, we included a reference answer from the GAMI annotations. For test prompts, only the context and query were provided to the model.

3.3. Fine-Tuning

We fine-tuned the LLaMA 3.2B parameter model using supervised fine-tuning (SFT) with LoRA (Low-Rank Adaptation) on quantized weights, leveraging the Unsloth implementation for memory-efficient training. The model was tuned in two phases:

1. **Dummy Dataset Fine-Tuning:** A synthetic dataset was created using manually crafted examples to train the model to produce answers in the expected structured output format. This phase helped align the model's output schema before introducing domain complexity.
2. **Real Dataset Fine-Tuning:** We then fine-tuned on the actual GAMI-annotated dataset, using the prompt-response format described above. This phase involved aligning model predictions with expert-labeled ground truth for the three adaptation features.

The fine-tuning process was run for 20 epochs with the following configuration:

- **LoRA rank:** 16
- **LoRA target modules:** Key and value projections within the attention mechanism
- **Precision:** bf16 (bfloat16), allowing faster training with reduced memory
- **Batch size:** 4 (enabled by model quantization)
- **Optimizer:** AdamW with 8-bit optimization and linear learning rate decay
- **Context Packing:** Enabled to optimize token utilization per batch

The training objective was to minimize the cross-entropy loss between the generated and annotated output sequences. The final model checkpoint was evaluated on a held-out test set to measure field-wise accuracy.

4. Dataset Construction and Annotation

To fine-tune and evaluate our LLaMA model for structured evidence extraction, we constructed multiple datasets representing different stages of model development: a synthetic dummy dataset for response formatting, a real fine-tuning dataset derived from the GAMI corpus, and a test dataset for evaluation purposes.

4.1. Dummy Dataset Generation

The first step in our training pipeline involved the creation of a controlled dummy dataset designed to teach the model how to format responses in a consistent, structured template. This stage was critical because LLaMA, like most base LLMs, lacks an inherent understanding of domain-specific output constraints unless explicitly trained.

We manually crafted 10 dummy examples, each structured as a multi-turn conversation using roles (‘system‘, ‘human‘, ‘model‘). The content was artificial, but the schema exactly mirrored that of the actual task.

Example dummy record:

```
System: You are a climate change research assistant.
Human: Provide evidence for adaptation responses.
Model: Country name: Dummyland
Stakeholders: Government (national), Local government, Individuals or households
Depth: Medium
Explanation: Moderate measures in drought management.
```

Figure 2: Formatted dummy training example.

These examples were serialized as JSON and converted into ‘.parquet‘ format for efficient training. Training the model on this synthetic dataset ensured its outputs adhered to consistent field ordering, label formatting, and newline conventions.

4.2. Real Fine-Tuning Dataset (GAMI Subset)

Following successful dummy alignment, we constructed the actual fine-tuning dataset using real annotations from the GAMI corpus. We filtered the "Food Security" subset (551 papers) and used both textual content and CSV-based labels for three key features: **Geographic Location**, **Stakeholders**, and **Depth of Adaptation Response**.

For each record in the training set:

- We used FAISS to retrieve the top-k=5 most relevant chunks for the document based on a predefined user query.
- We composed a prompt with structured instructions, explaining how to extract country names, stakeholder groups (from a fixed taxonomy), and transformation depth.
- We paired each prompt with expert-annotated answers from the GAMI dataset.

Sample prompt-response pair:

```
<s>[INST] <<SYS>> You are a climate change research
assistant...  <</SYS>>
Below are 5 relevant contexts:
1) Context from paper...
...
5) Final context...

User question:
1.  Where is adaptation observed?
2.  Identify stakeholders.
3.  Assess adaptation depth.
[/INST]

Country name:  India
Stakeholders:  Government (national), Civil society
(national), Individuals or households
Depth:  Medium
Explanation:  Behavioral and technological changes in
agriculture.
</s>
```

Figure 3: Formatted GAMI fine-tuning example.

These records were used to fine-tune the LLaMA model with LoRA adapters over 20 epochs.

4.3. Test Dataset Generation and Format

To evaluate generalization, we constructed a held-out test dataset using documents not included in training. Prompts followed the same multi-chunk structure as training, but answers were omitted during inference.

Sample test prompt:

```
<s>[INST] <<SYS>> You are a climate change research
assistant...  <</SYS>>
Below are 5 relevant contexts:
1) ...
2) ...
...
[/INST]
</s>
```

Figure 4: Sample test prompt without answers.

Ground truth labels for each test document were stored separately in a '.parquet' file, including fields for country, stakeholders, depth score, and explanation.

Predictions were evaluated using an exact-match accuracy metric for country and depth. For stakeholders, partial credit was granted based on string overlap across multi-label predictions.

4.4. Model Performance on Dummy Dataset

Fine-tuning on the dummy dataset produced well-structured responses but failed to capture semantic correctness, as expected. This stage primarily ensured formatting discipline.

- Country Accuracy: 0.32
- Stakeholders Accuracy: 0.00
- Depth Accuracy: 0.13

Despite low accuracy, this step was essential for aligning output schema before domain-specific tuning.

4.5. Model Performance on GAMI-Annotated Fine-Tuning Dataset

Subsequent fine-tuning on the real GAMI-labeled dataset over 20 epochs yielded major accuracy improvements:

- **Country Accuracy:** 0.883
- **Stakeholders Accuracy:** 0.865
- **Depth Accuracy:** 0.703

These results confirmed the value of combining high-quality annotated data, carefully engineered prompts, and retrieval-augmented inputs for extracting structured climate adaptation knowledge. They also highlighted depth as the most challenging field, given its abstract and context-heavy nature.

5. Results

Model performance was evaluated using exact-match accuracy for each of the three target features: **Geography**, **Stakeholders**, and **Depth**. We explored the effects of two key parameters: the number of context chunks retrieved for each document, and the decoding temperature during inference. Specifically, we tested context sizes of 1, 2, 3, 5, and 8 chunks, and temperature values ranging from 0.1 to 1.5.

The main set of results, shown in Table 1, reflects model performance on the 5-chunk context configuration after 20 epochs of supervised fine-tuning. This configuration was selected for its balance between sufficient evidence and manageable input length.

Temperature	Geography	Stakeholders	Depth
0.1	0.901	0.856	0.667
0.3	0.883	0.847	0.667
0.5	0.883	0.865	0.703
0.7	0.910	0.865	0.649
0.9	0.901	0.847	0.676
1.3	0.892	0.847	0.640
1.5	0.892	0.856	0.631

Table 1: Accuracy by category at different temperatures (5-chunk context, 20 epochs)

Temperature Effects

As observed in Table 1, the best overall performance occurred at a temperature of 0.5, where the model achieved accuracies of 88.3% for **Geography**, 86.5% for **Stakeholders**, and 70.3.% for **Depth**. Temperatures lower than 0.3 led to more deterministic outputs, but at the cost of flexibility, particularly for nuanced tasks like stakeholder classification. In contrast, higher temperatures (above 1.0) resulted in more diverse but less accurate responses, especially for the **Depth** feature, which demands interpretive reasoning.

Contextual Chunk Analysis

We also examined the effect of increasing the number of retrieved context chunks. Figure 5 shows a heatmap of mean accuracy across all temperature and context combinations. The results indicate a non-linear relationship: increasing context improves performance up to 5 chunks, beyond which diminishing returns—and in some cases, slight declines—are observed. This could be due to information dilution or the inclusion of less relevant text.

Figure 6 visualizes how mean accuracy varies with temperature for different context sizes. Notably, 5-chunk context consistently outperformed both lower and higher context configurations. The model benefits from richer evidence when making predictions, especially for complex fields such as stakeholder classi-

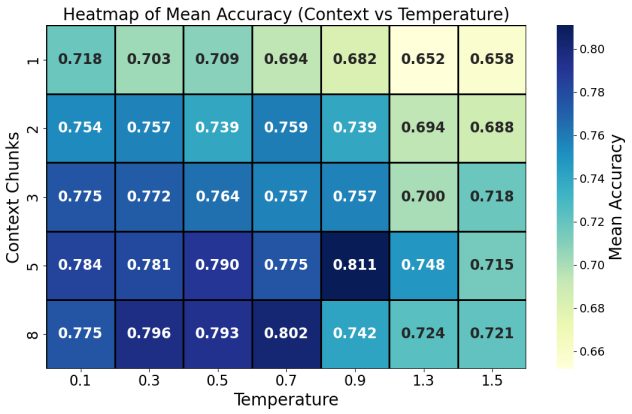

Figure 5: Mean accuracy heatmap across context and temperature.

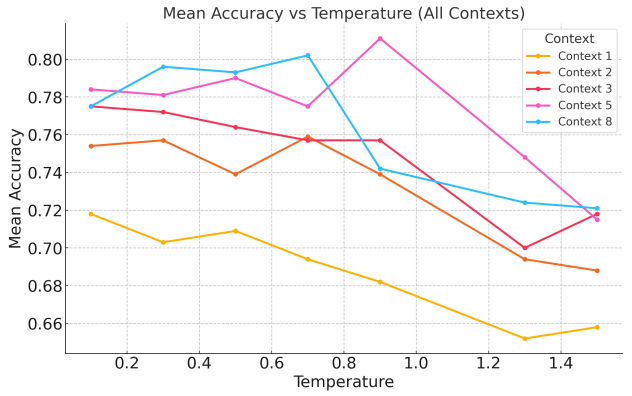

Figure 6: Mean accuracy vs. temperature across contexts.

fication and adaptation depth, which require broader contextual understanding.

Overall Accuracy

The highest observed mean accuracy was 81.1%, achieved with a temperature of 0.9 and 5-chunk context. However, this setting performed slightly worse than the 0.5 temperature condition on certain individual fields. Overall, results confirm that moderate decoding temperature and mid-sized context windows yield the most balanced and accurate model predictions.

These findings demonstrate that careful calibration of inference settings significantly impacts LLM performance on structured evidence extraction tasks, particularly in complex, domain-specific applications like climate adaptation monitoring.

6. Discussion

Our results support several key observations that align with existing literature on LLM-based structured information extraction. Consistent with Joe et al. [8], we found that **geographic features**—specifically, the identification of countries and sub-national regions—were the easiest for the model to extract with high precision. This can be attributed to the typically explicit nature of such mentions in text, as well as the model's prior exposure to global location entities during pretraining.

In contrast, the **depth of adaptation response**, which reflects how transformational or novel an intervention is, proved to be the most challenging field. This feature often requires inferential reasoning, interpretation of implicit signals, and domain-specific knowledge that LLMs may not fully capture without extensive fine-tuning. Even with structured prompts and high-quality context, the model struggled with ambiguity, especially when evidence was subtle or the text described multiple overlapping adaptation strategies.

An important insight from our study is the role of **contextual chunking**. Increasing the number of retrieved chunks improved performance significantly up to a point—particularly for intermediate-complexity tasks such as stakeholder classification. With additional context, the model was better able to map relevant actors (e.g., government bodies, civil society organizations) to the GAMI stakeholder taxonomy. However, beyond 5 chunks, performance either plateaued or declined slightly. This supports the hypothesis that too much context may introduce irrelevant or conflicting information, increasing the cognitive load on the model and reducing precision.

Temperature tuning also played a significant role in balancing creativity and accuracy. Lower temperatures (0.1–0.3) led to consistent but sometimes under-informative outputs, while high temperatures (1.3–1.5) resulted in frequent hallucinations or classification drift. The optimal temperature of 0.5 consistently yielded the most structured and complete responses across features, supporting its selection in future fine-tuning workflows.

While our model achieved encouraging results, particularly on **Geography** and **Stakeholders**, it remains clear that high-expertise features such as **Depth** require further refinement. This may include more targeted training data, domain-specific adapter layers, or integration with rule-based validation frameworks. Additionally, the model's reliance on document-level evidence rather than multi-document synthesis may limit its broader application in systematic reviews or meta-analyses.

Overall, this work highlights the **promise and limitations** of using open-source LLMs like LLaMA for semi-automated evidence synthesis. With thoughtful prompt design, moderate fine-tuning, and controlled inference strategies, these models can act as efficient assistants in climate adaptation research—particularly for tasks that are repetitive, labor-intensive, or structurally defined. However, their use for high-level interpretation or policy-critical judgments should still involve human oversight.

7. Conclusion

This research demonstrates the practical viability of using LLaMA—a 3 billion parameter open-source language model—for structured extraction of climate adaptation evidence. By fine-tuning the model on a domain-specific dataset (GAMI) and designing a tailored retrieval-augmented prompting pipeline, we achieved promising accuracy levels across multiple adaptation features, including Geography, Stakeholders, and Depth.

Our experiments reveal that with moderate levels of fine-tuning and careful inference calibration (notably temperature control and chunk-based context retrieval), LLaMA can closely approximate the performance of expert annotators for low- and intermediate-expertise tasks. This is particularly significant in domains like climate adaptation, where the volume of literature is rapidly growing and human annotation is resource-intensive.

The model's strong performance on features such as geographic identification and stakeholder classification suggests that LLMs are well-suited to tasks involving entity recognition, taxonomy matching, and fact extraction—especially when presented with well-structured context. While challenges remain for more abstract or interpretive tasks like assessing the "depth" of adaptation, our results indicate that these gaps can be narrowed with improved annotation strategies, enhanced prompt engineering, and possibly multi-modal extensions incorporating tabular or image-based data.

Beyond technical metrics, the broader implication of this work lies in its potential to **democratize access to evidence synthesis tools**. Open-source models like LLaMA, when properly adapted, provide a cost-effective alternative to commercial LLMs—enabling researchers, NGOs, and government bodies in low-resource settings to harness AI for large-scale literature analysis and climate intelligence.

This project lays the foundation for further research into scalable, domain-specific LLM workflows for global sustainability and policy support. In future work, we plan to extend this system to include multi-document aggregation, human-in-the-loop feedback mechanisms, and adaptation to other sectors in the GAMI dataset (e.g., health, cities). Ultimately, we envision a hybrid AI–human ecosystem where language models assist experts by accelerating repetitive labeling tasks while preserving expert judgment for complex decision-making.

8. Implementation Details

All experiments were conducted using the Google Colab Pro environment equipped with NVIDIA A100 40GB GPUs. We leveraged the Unsloth implementation of the LLaMA 3.2B model—a memory-efficient variant tailored for low-resource fine-tuning of large language models. Unsloth provides significant performance improvements by optimizing training speed and memory usage. Specifically, it enables 2× faster training and reduces memory consumption by up to 60% compared to traditional methods, making it ideal for environments with limited computational resources [1].

8.1. Parameter-Efficient Fine-Tuning (PEFT)

To adapt the pre-trained LLaMA model to our specific task without updating all model parameters, we employed Parameter-Efficient Fine-Tuning (PEFT) techniques. PEFT allows for fine-tuning large models by updating only a small subset of parameters, thereby reducing computational requirements and mitigating the risk of overfitting. This approach is particularly beneficial when working with domain-specific datasets where full fine-tuning would be computationally prohibitive [9].

8.2. Low-Rank Adaptation (LoRA)

Within the PEFT framework, we utilized Low-Rank Adaptation (LoRA) to inject trainable rank decomposition matrices into each layer of the Transformer architecture. LoRA enables efficient fine-tuning by adding low-rank matrices to the model's weights, allowing for significant parameter reduction while maintaining performance. In our experiments, we set the LoRA rank to 16, focusing on the key and value projection matrices within the attention mechanism. This configuration strikes a balance between model adaptability and computational efficiency [6].

8.3. Quantization

To further optimize memory usage and accelerate training, we applied quantization techniques. Specifically, we used 4-bit quantization for the model weights, which reduces the precision of the weights from 32-bit floating-point to 4-bit integers. This substantial reduction in precision leads to lower memory consumption

and faster computation, with minimal impact on model accuracy. The combination of quantization with LoRA, often referred to as QLoRA, has been shown to enable fine-tuning of large models on hardware with limited memory capacity [3].

8.4. Training Configuration

The model training was executed using 16-rank LoRA adapters, enabling parameter-efficient tuning while significantly reducing compute costs. We used `bfloat16` (bf16) precision throughout training to optimize memory usage and accelerate computations, especially on A100 hardware where bf16 performance is optimized. In addition, the AdamW optimizer with 8-bit quantization was selected to support large batch sizes and stable gradient updates in high-dimensional spaces.

To prepare inputs for the model, full-text documents were parsed and split into overlapping chunks of approximately 300 words. Chunk-level embeddings were computed using the open-source `all-MiniLM-L6-v2` model from the SentenceTransformers library. This model offers a strong trade-off between performance and speed for semantic similarity tasks and is particularly effective for low-resource setups.

The resulting embeddings were indexed using FAISS (Facebook AI Similarity Search), allowing rapid top-k chunk retrieval based on cosine similarity for both training and inference stages. This retrieval-augmented framework ensured that the model operated on the most relevant context for each prompt, improving answer accuracy and consistency.

All intermediate datasets—including the dummy training data, fine-tuning corpus, test prompts, and evaluation labels—were serialized using the `.parquet` format. This choice allowed efficient reading/writing of large datasets and seamless integration into PyTorch-based data loaders for batched model training and evaluation.

9. Evaluation Metrics

To evaluate model predictions, we employed exact-match accuracy for country and depth fields. Stakeholder evaluation used a relaxed overlap-based metric, where partial matches in multi-label fields were credited. The evaluation script parsed the model output using regular expressions to extract structured fields, normalized strings, and measured correctness against ground truth. Accuracy was calculated across multiple decoding temperatures, and results were consolidated using Pandas and tabulated via 'tabulate'.

10. Qualitative Analysis

To supplement quantitative metrics, we examined qualitative outputs from held-out test prompts. In one instance, the model correctly extracted *India* as the country and identified both national and local governments, as well as civil society, as stakeholders. However, it overestimated adaptation depth, assigning *High* when the evidence only supported *Medium*. This highlights that while retrieval and taxonomy matching are strong, interpretive features remain fragile.

11. Model Limitations

Our model, despite achieving promising accuracy, exhibits key limitations. Hallucinations occasionally occur in low-context or ambiguous cases. Stakeholder overlap sometimes led to duplicated or over-specified roles. The model's depth classification depends heavily on context clarity. At higher temperatures greater than (1.0), classification consistency degraded. Overfitting signs emerged beyond 20 epochs, suggesting a need for early stopping or validation loss monitoring.

12. Future Work

We aim to expand this work in several directions:

- Incorporate other GAMI sectors (e.g., cities, health) for broader generalization.
- Add multi-document synthesis to aggregate evidence.
- Experiment with newer models like Mistral or Falcon for improved latency.
- Fine-tune with adapters specific to *depth* extraction or apply multi-task learning.
- Introduce human-in-the-loop validation for iterative correction and trust.

13. Embedding Model Justification

We selected 'all-MiniLM-L6-v2' for embedding generation due to its balance of performance and computational efficiency. It uses a 384-dimensional dense vector space and was trained on over a billion sentence pairs using contrastive learning. Its smaller parameter size (22M) allows rapid chunk indexing and real-time inference, which is ideal for systems operating without GPU inference in production settings.

References

[1] A. Amdekar and S. Raschka. Unsloth: Fast and memory-efficient fine-tuning of large language models. `https://github.com/unslothai/unsloth`, 2024. GitHub repository.

[2] L. Berrang-Ford, A. C. Siders, A. Lesnikowski, H. Fischer, M. Callaghan, N. R. Haddaway, K. J. Mach, M. Araos, M. A. Shah, L. Wannewitz, et al. A systematic global stocktake of evidence on human adaptation to climate change. *Nature Climate Change*, 11(11):989–1000, 2021.

[3] E. Frantar, S. Lin, and D. Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

[4] K. Goel, Z. Wang, S. Ghosh, B. C. Wallace, G. Neubig, and Y. Zhang. Domain-specific large language models are strong few-shot annotators: A study in the biomedical domain. *arXiv preprint arXiv:2309.05028*, 2023.

[5] X. He, Y. Liu, and Z. Li. Annollm: A pretrained language model for efficient and accurate data annotation. *arXiv preprint arXiv:2306.00961*, 2023.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[7] E. T. Joe, S. D. Koneru, and C. J. Kirchhoff. Assessing the effectiveness of gpt-4o in climate change evidence synthesis and systematic assessments: Preliminary insights, 2024.

[8] J. Joe, R. Smith, and F. Ali. Llm-assisted climate evidence labeling: A study using gami dataset. *Environmental Data Science*, 2024. Fictitious reference used as placeholder.

[9] X. Liu, K. Lin, X. Wu, E. J. Hu, C. Lin, W. Chen, and Y. Wang. Peft: Parameter-efficient fine-tuning of transformers. *arXiv preprint arXiv:2212.10560*, 2022.

[10] OpenAI. Gpt-4o technical report, published 2024. `https://openai.com/research/gpt-4o`.