Examining Extractive and Abstractive Question Answering Methods for Climate Texts

Marisa Lenci and Grace Murnaghan

Spring 2025

Abstract

We seek to build on existing work using language models for climate change question answering and retrieval tasks using the dataset assembled by Rony, et al. for Climate Change Machine Reading Comprehension. Rony et al.'s work explored an application of ALBERT, pre-trained for SQuAD tasks, to extract answers using span detection from a question and context paragraph. We expand on this work by experimenting with various pre-trained models and architectures, comparing extractive and abstractive approaches to question answering tasks in the climate domain. We also incorporate an information retrieval component to our framework to perform a more complex and open-ended approach to question answering generation for climate related text. We find that abstractive question answering methods outperform extractive methods as measured by ROUGE and sentence embedding semantic similarity.

1 Introduction

In the United States where the science of climate change is heavily politicized, credible and authoritative information on climate science must be widely available to increase public understanding of climate change and to combat misinformation (van der Linden, et al., 2017). Organizations like the Intergovernmental Panel on Climate Change (IPCC) and NASA produce various reports on climate change. These reports provide detailed and accurate information but the length and technical language can hurt accessibility for the general public. Our work aims to make credible information about climate change more approachable by building a natural language question answering system specifically for climate-related communications. To do so, we compare performance of general pre-trained models to domain-specific models and task-specific models. Prior research in the climate change domain has shown the importance of pre-training for climate change domain terminology (Webersinke, et al., 2022; Rony, et al., 2022). We seek to evaluate the tradeoffs between pre-training for domain vs pre-training for NLP task (i.e. open-book question answering).

We also seek to expand upon methods for question answering in the climate domain by incorporating abstractive question answering, continuing from extractive question answering performed by Rony, et al. (2022). Given that answers generated using span detection are bound by the original wording of the context, we hypothesize that abstractive methods will produce better answers both by quantitative measures and by qualitative assessment.

Lastly, we incorporate a context retrieval component to give us a more realistic view of our final evaluation metrics. We imagine this work to be used in a setting such as a human-facing app where the system may be prompted with novel questions to answer using previously-compiled corpus of climate-related text.

2 Background

Rony, et al. published Climate Bot, in which their team builds the Climate Change Machine Reading Comprehension (CCMRC) dataset and climate question answering modeling that our research builds upon. In this paper, researchers fine tuned the language model ALBERT pre-trained on SQuAD to extract text span given a question and context paragraph. This component of the model achieved BLEU scores of 0.682 and 0.678 on validation and test data, respectively. The "reader" model was also evaluated using F1 score and METEOR. The primary focus of our work is to improve upon this component of the work done by Rony, et al. through experimentation with other pre-trained language model architectures as well as abstractive methods.

Prior research has shown that models pre-trained on domain-specific text exhibit enhanced performance compared to models pre-trained on only general text (Araci 2019, Lee et al., 2020). In particular, we consider prior approaches to related downstream tasks in the climate change domain to understand how domain adaptive pre-training impacted model performance. Webersinke, et al. introduce ClimateBERT, in which the authors adapt DistilRoBERTa for additional pre-training on over 2 million climate-related paragraphs using sources from news, corporate disclosures, and scientific articles for a variety of downstream tasks in the climate change domain (2022). The authors demonstrate that their domain-adaptive pre-training for ClimateBERT decreased loss by 48% compared to DistilROBERTA, experimenting with several downstream tasks for classification, sentiment analysis, and fact checking in climate change contexts. Previous work has also shown that LLM-based question answering systems can be effective in the climate domain. The work of Nguyen, et al. (2024) shows promising results for open-source LLMs like LLaMA as judged by human annotators.

3 Methods

3.1 Data

We use the Climate Change Machine Reading Comprehension (CCMRC) dataset, which consists of more than 7,400 context paragraphs and 21,000 question-answer pairs based on excerpts of text from climate-related research articles, IPCC and government reports, and news articles. The context paragraphs were assembled by the authors using a PDF-reader, and the question-answer pairs were manually written and created by humans from Amazon Mechanical Turk and the Smart Data Analytics group.

The data are structured as sets of context paragraphs with associated questions and answers. Answers are structured as excerpts of text from the corresponding context paragraph with an answer span start index position also provided. Labeled answers range in length from a single word up to a span of 282 words; example questions, contexts, and answers are provided in Appendix A. For our experiments with retrieval, we construct a corpus of text composed of the shuffled context paragraphs, maintaining the provided train, validation, and testing splits.

3.2 Experimental Design

Our experimental design includes experiments across extractive and abstractive methods for question answering, retrieval, and combining both tasks in an end-to-end implementation. Building on Rony et al.'s work with ALBERT on the CCMRC dataset, we first implemented the simplest baseline, using out-of-the-box pre-trained models. We selected BERT for question answering on SQuAD 2.0 (bert-base-cased-squad2) for our extractive baseline for its performance and task-specific architecture for question answering (Deepset, 2024; Zhang & Xu, 2019), and T5 (t5-base) for our abstractive baseline, for its text-to-text framework and performance on downstream tasks including question answering (Raffel, et al., 2020). We then experimented with fine-tuned models in the climate change domain and compared them to our baseline models. We selected the best performing models to experiment with our own fine-tuning for our question answering task. Finally, building upon our fine-tuned model, we experiment with hyperparameters to optimize performance.

We build upon our question answering task to include retrieval for our context data. We experiment with two methods for retrieval, tf-idf and semantic search. We select tf-idf as our baseline model for retrieval, for its simplicity and performance in search applications. We compare the tf-idf baseline to a vector-matching method for semantic similarity. We select the sentence transformer pre-trained for question answering (sentence-transformers/msmarco-distilbert-cos-v5) for its semantic search ability and performance on question answering in the MS Marco dataset of question-answer pairs (Reimers & Gurevych 2019). Finally, we combine the best-performing retrieval model with our best-performing extractive and abstractive question answering models to evaluate performance as an end-to-end system.

3.3 Model Selection for Experimentation

For extractive question answering, we experimented with BERT- and BART-based models. We selected BERT models for experimentation for their strength in building contextual

understanding. Since BERT's pre-training applies masked language model techniques and next sentence prediction, it has shown strong performance in building contextual understanding of words and sentences, which can be easily applied for downstream tasks like question answering (Chang et al., 2018). We experimented with BERT-base for question answering on SQuAD. Given Webersinke, et al.'s application of their model ClimateBERT, which is fine tuned for climate change domain on top of DistilRoBERTa, we also experimented with the smaller DistilRoBERTa-base (Sanh, et al., 2019) and ClimateBERT on our dataset, to compare how our results respond to both efficiency tradeoffs and domain-specific models. We also experimented with BART models, using BART-large fine-tuned for extractive question answering on SQuADv1 (Patil, 2020) as well as for summarizing climate-related texts (Dickson, 2023). The pre-training for BART extends the masked language model task of BERT to a variety of de-noising tasks, building contextual language understanding in its encoder. It is a sequence to sequence model but also performs well on token classification tasks like SQuAD (Lewis, et al., 2019). To fine-tune BART models for extractive question answering, we put a span classification head on top of the final hidden state of the decoder, leveraging information learned in seq2seq pre-training for extractive question answering. We compare the performance of the BART model fine tuned on SQuAD to the model fine tuned on climate texts to better understand performance given task- versus domain-specific pretraining.

We experimented with additional seq2seq models to perform abstractive question answering, starting with T5. T5 is trained to perform SQuAD-like question answering using the input format "Question: … Context: …" so we expect returned answers to be similar to those returned by extractive methods, though it is being emitted as a sequence by the model rather than classifying tokens as in extractive methods. In addition to our T5- baseline, we experimented with Flan-T5 (Chung, et al., 2022) which is fine tuned on instruction with the goal of better generalizing to new tasks than T5 which is trained on a limited set of tasks. We examined a Flan-T5 model fine tuned on climate domain texts (Khanal, 2024) as well to test if training with domain-specific language improves answer quality. We also experimented with LLaMA, a large language model, because of its strong performance given instructions and in few-shot scenarios (Touvron, 2023).

3.4 Evaluation

To evaluate the responses returned by the question answering system, we use ROUGE-1, ROUGE-2, and ROUGE-L scores (Lin, 2004). Since ROUGE is recall-based, this metric allows us to see how many of the words or sequences in the reference answer are present in the candidate answer which one may think of as the "completeness" of the candidate answer. In this dataset, reference sentences are spans of the context that best answer a given question, therefore it's appropriate to use a word overlap metric like ROUGE to assess performance of extractive question answering methods.

For abstractive question answering using sequence to sequence models or LLMs, ROUGE is still applicable as we want to make sure that key ideas, terms, or phrases from the references are mentioned in the candidate answers. Still, we recognize that abstractive candidate answers may be appropriate but score poorly in ROUGE metrics if the candidate answer is semantically similar to the reference answer but different words are used. To address this, we use cosine similarity of the candidate and reference answers encoded using all-MiniLM-L12-v2 embeddings, available in the sentence-transformers package. The all-MiniLM-L12-v2 model was selected for its strong performance on sentence embeddings and relatively fast encoding speed. Given that we intend this to be a public-facing tool, we want to highly penalize a system that returns misinformation, so evaluation using word overlap measured with ROUGE is a more conservative choice than using other methods, but we still consider semantic similarity for abstractive methods. This is less of a concern for extractive question answering where answers come directly from the text. Additionally, using a common evaluation method allows us to directly compare extractive and abstractive results.

We treat the context retrieval component of our work as a binary classification task. Given the text of a question, the model retrieves a context paragraph from our corpus of context passages. Since our original data is structured as having a single context paragraph

corresponding to each question, we can classify the retrieved paragraph as the correct or incorrect context for the given question. We measure performance using accuracy. Since each question-answer pair is only associated with one context paragraph, we have elected to return a single context paragraph, k=1. Thus we consider context retrieval only as a measure of accuracy, as measuring precision and F1 are not appropriate for our application where we return the top 1 context paragraph.

4 Results and Discussion

Table 1: Extractive Methods

Model	ROUGE-1	ROUGE-2	ROUGE-L
Baseline: bert-base-cased-squad2	0.27399	0.23738	0.27115
distilroberta-base	0.20000	0.13269	0.16757
distilroberta-base (train QA head for 3 epochs)	0.34177	0.22910	0.28975
distilroberta-base-climate-f	0.18772	0.13806	0.16642
distilroberta-base-climate-f (train QA head for 3 epochs)	0.12461	0.07747	0.10801
bart-large-finetuned-squadv1	0.43411	0.37962	**0.43218**
bart-large-finetuned-squadv1 (fine tuned with Lora for 2 epochs, r=8, alpha=32)	0.44285	0.36825	0.40077
bart-large-finetuned-squadv1 (fine tuned with Lora for 2 epochs, r=8, alpha=16)	**0.46334**	**0.39130**	0.42004
bart-large-finetuned-squadv1 fine tuned and run with tf-idf context retrieval	*0.35688*	*0.27123*	*0.31737*
bart-large-cnn-climate-change-summarization (train QA head for 2 epochs)	0.32481	0.23233	0.28355
bart-large-cnn-climate-change-summarization (train QA head for 5 epochs)	0.32636	0.23451	0.28362
bart-large-cnn-climate-change-summarization (train using LORA for 2 epochs)	0.32957	0.22386	0.28006

Table 2: Abstractive Methods

Model	ROUGE-1	ROUGE-2	ROUGE-L	Semantic Similarity
Baseline: T5-base	0.39642	0.33858	0.39496	0.5784
T5-small (fine tuned for 1 epoch)	0.57041	0.52688	0.56100	0.7301
T5-small fine tuned and run with tf-idf context retrieval	*0.41986*	*0.35099*	*0.40546*	*0.5813*
T5-small (fine tuned for 3 epochs)	**0.57813**	**0.53708**	**0.57066**	0.7341
T5-small fine tuned and run with tf-idf context retrieval	*0.42608*	*0.35809*	*0.41258*	*0.5852*
flan-t5-base	0.34952	0.28607	0.34610	0.5487
flan-t5-climate-qlora	0.34902	0.28667	0.34593	0.5487
Meta-Llama-3.1-8B-Instruct (temperature=1, top_p=0.3)	0.53869	0.44541	0.49941	**0.7679**
Llama-3.1-8B-Instruct run with tf-idf context retrieval	*0.42384*	*0.32620*	*0.38713*	*0.6696*

Note on Reading Tables 1 and 2: Experiments are organized by model architecture with fine tuned iterations listed in indented blocks underneath the pre-trained model they iterate upon. For models where we tested the end-to-end process of tf-idf context retrieval and question answering, results are listed in italics underneath the model used.

4.1 Extractive model results

In our extractive experimental results, we consider both how climate change domain-specific models perform compared to their general pre-trained model alternatives, and how fine tuning impacts performance for both domain-specific and general models. We evaluate DistilRoBERTa (distilroberta-base: Sanh, et al., 2019) with ClimateBERT (distilroberta-base-climate-f) and BART for SQuAD (bart-large-finetuned-squadv1) with BART for Climate QA (bart-large-cnn-climate-change-summarization tuned for QA).

Our ClimateBERT results indicate reductions in performance relative to DistilRoBERTa, the general pre-trained model alternative, for all metrics except ROUGE-2 where we see very slight improvement. These performance reductions in ClimateBERT are exacerbated further with fine tuning for our task and dataset, where we see even stronger reductions relative to ClimateBERT, DistilRoBERTa, and to the overall baseline. For the general pre-trained DistilRoBERTa, fine tuning over the same conditions yields a substantial increase in performance, relative to initial DistilRoBERTa results and to the overall baseline. To understand why we observe this reduction in performance in the climate change domain model, we considered both the pre-training corpora and task-specific training. We expect this decline in performance is due to the climate change domain corpora being less relatable for our dataset. We expect that the additional climate vocabulary is not needed for our dataset, and rather most of the words in our dataset are well represented in general domains. An example of a common representation style in our dataset is shown in Appendix A.a, in which the question refers to common, factual-based information from the context paragraph for its reference answer.

We next experiment with BART for SQuAD and BART for Climate QA. Our BART for Climate QA results indicate reductions in performance compared to BART for SQuAD, the general pre-trained model alternative. Even with further fine tuning on BART for Climate QA for our dataset and question answering task, we do not observe any further improvement. For the general pre-trained BART for SQuAD, fine tuning for our dataset and task yields some improvement, yet still moderate. Notably, when fine tuning BART for SQuAD we observe that turning down the LoRA alpha parameter yields further improvement, suggesting some noise in our dataset having larger influence on models trained with higher alpha. These results indicate that fine tuning with our data improves to some degree, but then begins to fall because of existing noise.

Across our extractive models, we see some increased performance when fine tuning for our question answering task in the general pre-trained domain. However, fine tuning in the climate-change domain models yields decreased or static performance. Considering these results of domain-specific and general models with results of fine-tuning over these different models, we conclude that once the model can handle the question answering task well, then layering in our climate change domain works. However, trying to solve the problem the other way around by adapting a domain-specific model to our question answering task does not work well. We believe that some of this is due to limited time and resources to train, where models not trained for question answering needed much more time to adapt to our task than we could manage given available computing resources. We also believe that some of this is due to the format and style of our dataset, where question-answer pairs are not facts and figures about climate change concepts from context paragraphs, but more closely resemble reading comprehension style questions from context paragraphs.

4.2 Abstractive model results

We find that the T5 model applied to our data before any fine tuning to be a strong baseline. Though it does not outscore the best performing extractive model in terms of ROUGE, the baseline T5 model performs better than most of the extractive methods we experimented with. We found that the Flan-T5 architecture performed worse in both ROUGE and semantic similarity versus the baseline model. We verify that the baseline model performs well on SQuAD-like tasks and find that it does well on the CCMRC dataset, so we expect there is not any benefit to the additional instruction architecture that Flan-T5 introduces, and perhaps the additional instruction architecture actually degrades performance for this task.

Testing the performance of Flan-T5 allows us to better contextualize the performance of flan-t5-climate-qlora, a fine tuned version of Flan-T5 built for NLP tasks related to climate change and environmental policy. We find that this model does not show improvement over Flan-T5 in ROUGE-1, ROUGE-L, or semantic similarity; improvement to ROUGE-2 is very small (less than 0.001 points). For similar reasons as we noted in our analysis of extractive methods results with ClimateBERT, we expect that the additional climate corpora used in training Flan-T5 Climate does not directly benefit our dataset, due to CCMRC's reliance on specific, fact-based questions about associated context paragraphs.

We fine tune T5, our best performing model, with the CCMRC dataset and see strong improvement even after 1 epoch, despite having to use T5-small rather than T5-base due to compute restrictions. After fine tuning for 3 epochs, we report an improvement over the abstractive baseline of 0.18, 0.20, and 0.18 points of ROUGE-1, -2, and -L respectively. Average semantic similarity also improved by 0.16 points. We expect that this increase in performance is due in part to the model learning the structure of our data and style of question-answer pairs rather than solely building contextual understanding in the climate domain. The model is learning to exploit the format and types of questions relative to their context. In the following example, the provided answer in our dataset is a full sentence, and the fine tuned T5 model is rewarded in both ROUGE and semantic similarity scores for being more verbose than T5-base although the generated answer is not of higher quality, ending mid-clause.

Question	Label Answer	T5-base Answer	Fine Tuned T5 Answer
"How much must carbon emissions be cut in order to diminish the risk of further substantial change in climate?"	"the ipcc has estimated that at least a 60 percent reduction in carbon emissions from 1990 levels is necessary to reduce the risk substantially of further climate change substantially"	"60 percent"	"the ipcc has estimated that at least a 60 percent reduction in carbon emissions from"

The structure of our training data and our chosen evaluation metric reward wordier responses which ultimately do not hurt human comprehension but may produce answers less useful for our intended end user. In many cases though, the lengthier responses are helpful for human understanding (appendix section A.d).

Unlike T5 which is emulating extractive methods when prompted using "Question:... Context:...", the LLaMA-produced answers more often rephrase the relevant portion of the provided context. For the example above, the LLaMA response is "According to the IPCC, at least a 60 percent reduction in carbon emissions from 1990 levels is necessary to reduce the risk substantially of further climate change," a reorganized and more coherent expression than what is produced by T5-base as well as the fine tuned T5 model. This rephrasing of the answer is reflected in lower ROUGE scores because of changes to word order and word choice, such as "according to the IPCC" rather than "the ipcc has estimated". However, we see an improvement in semantic similarity scores between LLaMA answers and the label answer versus the T5 models, implying that the LLM answers are more similar in meaning than those produced by T5. Given that, LLaMA still shows strong results as measured by ROUGE.

4.3 Comparing Extractive and Abstractive Methods

Despite the CCMRC dataset being oriented towards extractive question answering, we find abstractive methods to perform better in generating complete and coherent answers. ROUGE metrics are also more geared toward extractive models in the way that they measure word overlap. We find that our best abstractive question answering methods perform well in both word overlap and meaning, the combination of which helps show that these models are producing useful answers, not just regurgitating the question and context nor being overly verbose. Qualitatively, we find the answers provided by abstractive methods to be both useful and readable to a human which are key qualities given our intended use case.

We test the question answering system end-to-end with our best performing retrieval model. We find that tf-idf performs better than embedding-based semantic search, with 66%

accuracy in context retrieval on the test dataset. Tf-idf search likely outperforms semantic search because of the presence of proper nouns and uncommon words in many of the CCMRC questions (ex: “What was Jim Haywood thanked for?”) that may favor a token-matching system over one that is based on meaning. We see an expected decrease in final ROUGE and semantic similarity scores when combining the question answering task with context retrieval, but the reduction in score was less than expected. We find that some answers generated using an incorrect context passage have non-zero ROUGE scores, indicating that there may be some relevance to the top retrieved passage even if it is not the one that matches the question. An example of an answer we find reasonable despite being based on incorrect context can be found in appendix section A.c.

5 Conclusion

Providing credible and authoritative information on climate science that is widely available is an important effort in order to increase public understanding of climate change and to combat misinformation. Our work contributes to making credible information about climate change more accessible to the public by building a natural language question answering system specifically for climate-related communications. We build on prior research with extractive methods to experiment with alternative methods and improve performance and accessibility by end-users. We find that abstractive methods performed better than extractive, demonstrating strong performance in capturing both word overlap and overall meaning. This is an interesting finding given that the CCMRC dataset is oriented towards extractive question answering, and ROUGE metrics favor extractive methods. Future research may consider further experimentation with fine tuning and hyperparameter tuning with more training time and computational resources, as well as truncating vs. chunking texts to further improve performance.

Authors' Contributions

Marisa tested and fine tuned the BERT baseline, DistilRoBERTa, ClimateBERT, and Flan-T5-climate-qlora models. Grace experimented with BART SQuAD, BART Climate, Flan-T5, and retrieval systems. Grace and Marisa both worked on T5 and LLaMA. Marisa and Grace also shared the work on formulating a project concept and writing the report.

References

Araci, D. 2019. Finbert: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.

Deepset. (2024, January 4). bert-base-cased-squad2. Huggingface.co. https://huggingface.co/deepset/bert-base-cased-squad2

Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., … Wei, J. (2022). Scaling Instruction-Finetuned Language Models. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/2210.11416

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR, abs/1810.04805. Retrieved from http://arxiv.org/abs/1810.04805

Dickson, Z. P. (2023). Bart-large CNN Climate Change Summarization. Retrieved from https://huggingface.co/z-dickson/bart-large-cnn-climate-change-summarization

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. CoRR, abs/2106.09685. Retrieved from https://arxiv.org/abs/2106.09685

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics, Volume 36, Issue 4, February 2020, Pages 1234–1240, https://doi.org/10.1093/bioinformatics/btz682.

Khanal, K. (2024). FLAN-T5 Climate Action QLoRA. Retrieved from https://huggingface.co/kshitizkhanal7/flan-t5-climate-qlora

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., … Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. CoRR, abs/1910.13461. Retrieved from http://arxiv.org/abs/1910.13461

Nguyen, V., Karimi, S., Hallgren, W., Harkin, A., & Prakash, M. (2024, August). My Climate Advisor: An Application of NLP in Climate Adaptation for Agriculture. In D. Stammbach, J. Ni, T. Schimanski, K. Dutia, A. Singh, J. Bingler, … M. Leippold (Eds.), Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024) (pp. 27–45). doi:10.18653/v1/2024.climatenlp-1.3

Patil, S. (2020). BART-LARGE finetuned on SQuADv1. Retrieved from https://huggingface.co/valhalla/bart-large-finetuned-squadv1

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. CoRR, abs/1910.10683. Retrieved from http://arxiv.org/abs/1910.10683

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Retrieved from http://arxiv.org/abs/1908.10084

Reimers, N., & Gurevych, I. (2025). Pretrained models - Original Models. Pretrained Models - Sentence Transformers documentation. https://www.sbert.net/docs/sentence_transformer/pretrained_models.html#original-models

Rony, M. R. A. H., Zuo, Y., Kovriguina, L., Teucher, R., & Lehmann, J. (7 2022). Climate Bot: A Machine Reading Comprehension System for Climate Change Question Answering. In

L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22 (pp. 5249–5252). doi:10.24963/ijcai.2022/729

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. Retrieved from https://huggingface.co/distilbert/distilroberta-base

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., … Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv [Cs.CL]. Retrieved from http://arxiv.org/abs/2302.13971

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. Global Challenges, 1(2) Retrieved from https://www.proquest.com/scholarly-journals/inoculating-public-against-misinformation-about/docview/2289669567/se-2

Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). ClimateBert: A Pretrained Language Model for Climate-Related Text. CoRR, abs/2110.12010. Retrieved from https://arxiv.org/abs/2110.12010

Zhang, Y., Xu, Z. (2019). BERT for Question Answering on SQuAD 2.0. Retrieved from https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15848021.pdf

Appendix

A. Question, Context, and Answer Examples
   a. Standard example of data structure:
      i. **Question:** When did Lost Creek wildfire occur?
      ii. **Context:** here, impairment of water quality by wildfires in forested source water regions was examined as a critical vulnerability of downstream water treatment processes. in 2003, one of the most severe recorded fires (lost creek wildfire) occurred in the eastern slopes of the rocky mountains of southern alberta, canada and impacted several aspects of water quality and streamflow in the upper oldman river basin (orb). data from source watersheds with varying degrees of wildfire associated land disturbance (reference [unburned], burned, and post-fire salvage-logged) were collected and evaluated during the four years post-fire. some of the water quality impacts during these recovery years have been reported elsewhere, while others are reported herein. rather than attempt to predict or demonstrate the impacts of wildfire and salvage-logging on a specific downstream drinking water treatment plant, all of the studied water quality impacts of wildfire in the orb are synthesized and analyzed to provide a holistic discussion of downstream threats to drinking water "treatability" that can be associated with upstream wildfire and post-fire intervention (salvage-logging). accordingly, this analysis of water quality impairment resulting from wildfire is used as a case study to demonstrate 1) the impacts of wildfire and post-fire salvage-logging on drinking water "treatability", 2) a general approach for assessing potential drinking water "treatability" implications of land disturbance, and 3) the need for developing strategies for effectively and sustainably managing water resources in anticipation of local climate change and other natural or anthropogenic land disturbances.
      iii. **Label Answer:** in 2003
   b. Example of vague question that is challenging for retrieval system:
      i. **Question:** What it would take?
      ii. **Context:** there is a need for careful case studies of what adaptation would involve in particular locations, and what component of this is from infrastructure. these case studies would need to consider the infrastructure deficit, and the institutional/governance underpinning

necessary for addressing the deficit and climate-proofing all new and existing infrastructure. this could lead to a better idea of the kind of funding needed for adapting infrastructure to climate-change risks, and then to some thoughtful discussion of what this implies for adaptation costs and adaptation funding in general. it would take only a few such studies of major cities particularly at risk from climate change and with large infrastructure deficits to show that the unfccc estimates for africa and for 'developing asia' are far too low. it is also likely that studies of major cities in latin america and the middle east at high risk from climate change would show the unfccc estimates for these regions to be far too low. even with a growing number of careful location-based estimates for costs, however, it will be difficult to extrapolate these to figures for whole regions. this is because: * there are very large differences in contexts (risks and vulnerability), including the scale of infrastructure deficits and the extent of local governance failures. in most of the locations with the largest infrastructure deficits and governance failures, much of the data needed to assess such costs are not there.

iii. **Label Answer:** it would take only a few such studies of major cities particularly at risk from climate change and with large infrastructure deficits to show that the unfccc estimates for africa and for 'developing asia' are far too low

c. Example where tf-idf retrieval is incorrect
   i. **Question**: What is resilience to climate change and Promoting resilience?
   ii. **Correct Context:** adaptation examples include constructing fuel breaks around a vulnerable population of a valued plant species to prevent extinction from climate-aggravated wildfire; rescuing a highly valued and climate-vulnerable animal species by captive propagation (e.g., california condor gymnogyps californianus shaw]); prescribing methods otherwise socially undesired (e.g., insecticides) to aggressively combat insect mortality that threatens high-value resources (e.g., insectand pathogeninfected young bristlecone pine pinus longaeva d.k. bailey] forests in the white mountains); requesting more than otherwise allotted water rights to maintain a unique and ecologically critical aquatic ecosystem (e.g., mono lake, california, relative to water delivery to los angeles for human use); and aggressively remov ing invasive species (box 19). actions that attempt to resist climate change are usually successful only in the short term, and become less effective over time as effects of climate change accumulate or management priorities change. as climate pressure increases, not only will it become more difficult to resist change, but when change occurs, it may exceed physical and biological thresholds and result in undesirable outcomes (e.g., severe wildfire, forest mortality, species extinction). some resistance approaches, such as managing high-value species in designated refugial networks, involve relatively low risk or investment. an example is refugial networks proposed for american pika ochotona princeps richardson 1828), a small mammal that lives primarily in high-elevation habitats and is considered at risk from a warmer climate (box 20) (millar and westfall 2010). another interpretation of the resistance strategy is to defer proposed (or approved) projects that are unlikely to succeed because of increasing climate pressure and future conditions. examples include removing lodgepole pine pinus contorta subsp. murrayana [balfour] engelmann) seedlings that invade alpine meadows such as tuolumne meadows, yosemite national park; reintroducing salmon in streams where future water temperatures will be too high to support them; and chaining (removing) junipers

juniperus spp.) that are becoming estab lished in great basin sagebrush artemisia spp.) steppe communities. develop resilience to climate change --promoting resilience is the strategy most often recommended for adaptation (folke et al. 2004, hansen et al. 2003). resilience, which has both ecological and socioeconomic implications, can range from short-term response to disturbance to long-term tolerance of prolonged droughts. as mentioned above, agreement on definition is less important than how the range of meanings informs development of effective adaptation plans. in an en gineering context, resilience refers to the capacity of a system or condition to return

iii. **Retrieved context**: two other adaptation concepts are resilience and vulnerability.85-87the apa report refers to resilience as the 'inner strengths and coping resources for necessary adaptation to situational demands'6(p. 117). at the community level, resilience refers to the resources the group can draw upon, including knowledge, support systems, and social capital.88-93vulnerability is the extent to which individuals and communities are at risk and are unable to cope with the adverse impacts of climate change. the differential resilience and vulnerability of individuals and groups influence their adaptive capacity. various models and frameworks in the existing literature that can help to explain individual and group responses to current and future climate change impacts. although mitigation is an important way to reduce climate change, many individuals and communities will be forced to adapt to changes in climate. psychologists and allied professionals must continue to investigate, understand, and reduce the impact of these adaptations that will inevitably occur.

iv. **Correct answer:** develop resilience to climate change --promoting resilience is the strategy most often recommended for adaptation (folke et al. 2004, hansen et al. 2003). resilience, which has both ecological and socioeconomic implications, can range from short-term response to disturbance to long-term tolerance of prolonged droughts. as mentioned above, agreement on definition is less important than how the range of meanings informs development of effective adaptation plans

v. **T5-small 3 epoch answer:** at the community level, resilience refers to the resources the group can draw upon, including knowledge,

d. Example where fine tuned T5 being wordier is helpful
   i. **Question**: What are the impact of the climate change in the disturbances?
   ii. **Context**: no model, statistical or otherwise, can yet include all the life history traits, that is, the biological characteristics of the species or their responses to various disturbances that may influence a species' response to changes in climate. we focus here on some of these types of uncertainty as related to nine biological and twelve disturbance modification factors (modfacs) that influence species' distribution, as determined from literature surveys (figure 4 matthews and others, in press). the biological factors attempt to assess the species capacity to adapt to changing conditions, especially those expected in the future following current trends. for example, higher capacities to regenerate after fire, regenerate vegetatively, or disperse are all positively associated with adaptability to expected climate changes. similarly, the disturbance factors assess the resilience of the species to twelve l. r. iverson and others disturbance types, or the species' capacity to withstand these disturbances, as best as we can determine from the literature (for example, burns and honkala 1990a b). many of the disturbances we

evaluated are expected to increase with climate change or other human-influenced stresses and

iii. **Label answer:** many of the disturbances we evaluated are expected to increase with climate change or other human-influenced stresses

iv. **T5-base answer:** increase
    1. ROUGE-1: 0.10526, ROUGE-2: 0.0, ROUGE-L: 0.10526, semantic similarity: 0.2911

v. **T5-small fine tuned 3 epoch answer:** “many of the disturbances we evaluated are expected to increase with climate change or other human-influenced stresses”
    1. ROUGE-1: 1.0, ROUGE-2: 1.0, ROUGE-L: 1.0, semantic similarity: 1.0

B. Test Accuracy of Various Retrieval Systems

a.

Model	**Test Accuracy**
tf-idf	66.08%
Sentence encoder cosine similarity	
sentence-transformers/msmarco-distilbert-cos-v5	47.66%
sentence-transformers/msmarco-distilbert-base-tas-b	47.33%
sentence-transformers/msmarco-MiniLM-L6-cos-v5	44.13%

C. LLaMA Hyperparmeter Tuning Results - based on a sample of 50 training examples

a.

max_new_tokens	do_sample	temperature	top_p	rouge-1	rouge-2	rouge-L
512	TRUE	0.3	0.95	0.47826	0.44444	0.47826
512	TRUE	1	0.95	0.24167	0.13445	0.20000
512	TRUE	1	0.3	0.50000	0.42857	0.44000
512	TRUE	0.5	0.3	0.46809	0.43478	0.46809
512	TRUE	0.5	0.4	0.50000	0.42857	0.44000