Abstractive Summarization of Scholarly Articles in the Environmental Science Domain: Fine-Tuning State-of-the-Art Language Models

Woojae Chung
School of Information
University of California, Berkeley
woojae.chung@berkeley.edu

Brendan Lundquist
School of Information
University of California, Berkeley
brendan.lundquist@berkeley.edu

Summer McGrogan
School of Information
University of California, Berkeley
summermcgrogan@berkeley.edu

Abstract

Abstractive text summarization, a crucial task in natural language processing, aims to condense complex articles into concise summaries, making that knowledge more accessible to the general public. This paper focuses on creating a high-performance abstractive summarization model for environmental science articles using three state-of-the-art language models, PEGASUS, FLAN-T5, and BART, fine-tuning them on a curated subset of environmental science papers from the Semantic Scholar Open Research Corpus (S2ORC). We evaluate the models using both automated metrics (ROUGE, BERTScore) and human assessments, noting that, while traditional metrics may not fully capture summary quality, human evaluations reveal the strengths of our fine-tuned models. Results indicate that the fine-tuned PEGASUS model provides the most accurate and coherent summaries for this task. These results highlight the potential for these models to enhance the accessibility of scientific literature on critical environmental issues.

1 Introduction

Abstractive text summarization is one of the most important and challenging natural language processing tasks in artificial intelligence. A perfect model would be capable of taking any input, and producing a reliable output that provides the reader with an accurate overview of the article. This task can spread important information, such as news or research studies, to individuals that may have previously forsaken them due to time constraints. Effectively summarizing complex papers for the layperson to consume is an important task, as often it is the average person that is affected by esoteric scientific knowledge like climate change.

As, most notably, climate change ravages the landscape of the world today, the importance of environmental science cannot be understated. While many are aware of environmental issues, only few have an academic understanding of the complexities of the problem, and misinformation on environmental science has become too easy to spread. As such, understanding scientific research around this topic is imperative even for laypersons. However, academic articles are long, filled with jargon and scientific concepts that the average reader might find difficult to digest. Additionally, abstracts are not required for all media content, providing another barrier for everyday reading. Our goal is to build off of existing breakthrough models and make an abstractive text summarization model that can intake an article and produce an accurate output summary, with a focus on environmental science journals.

Text summarization aims to generate an abridged version of an input document(s) based on some relevant criteria such as word count. There are two types of summarization, extractive and abstractive. Extractive summarization extracts subsets of relevant sentences based on a scored relevance, and appends those passages to the summary. Abstractive summarization generates a new summary of the input by identifying salient information, paraphrasing key sentences, and generating new words. Abstractive summarization is typically more challenging than extractive summarization because it requires novel language generation, rather than pulling directly from the passage.

Conceptually, the abstractive text summarization problem has been broached with a sequence-to-sequence encoder-decoder style architecture. The encoder-decoder model consists of two distinct components with two distinct roles. In the context of text summarization, the encoder interprets an input and assigns that input a vector representation to derive its underlying meaning. The input from the encoder is subsequently passed to the decoder as its initial input, where its task is to generate an output. When a token is generated, an additional context token is gathered via the self-attention mechanism

from previously generated tokens and the output from the encoder itself. Despite the success that these large, pre-trained models may have in the task of text summarization, there are challenges that have yet to be solved.

Summarization for our task comes with two primary challenges stemming from common abstractive summarization issues: incoherent output sentences and hallucinations. Incoherent output sentences are gibberish and uninterpretable by a reader. Hallucinations are interpretable by a reader, but can create false information that lead to an inaccurate representation of the whole article. Both of these common issues make for summaries that readers cannot understand or trust. Abstract summarization issues are challenging to solve because we must teach a model to identify and understand important information, and then repeat it, all without having a robust scoring system.

Domain-specific text summarization is a sparse field of research. Most text-summarization research focuses on larger issues surrounding summarization, such as scoring or model architecture. Works that have been conducted on this topic focus on financial or medical document summarization, and often utilize extractive text summarization (Lee et al., 2021, Li & Xu, 2023). Current research has leveraged transformer-based sequence-to-sequence models to improve rouge scores through pre-training and fine-tuning (Lee et al., 2021). Novel approaches tend to also prefer extractive methods (Li & Xu, 2023).

Abstractive approaches to scientific text summarizations have been explored before. These studies utilize models including sequence-to-sequence RNNs (Nallapati et al., 2016), pointer generator networks (See et al., 2017, Rehman et al., 2023), and the same pointer generator networks with coverage mechanisms (Tu et al., 2016). However, these models have mostly been dethroned with current state-of-the-art summarization models. To expand on this limited field of research, we will be using current abstractive summarization models on an unexplored domain-specific dataset.

2 Background

Recent literature has made major milestones in abstractive text summarization. Specifically, we approach our abstract summarization task by applying three state-of-the-art pre-trained language models, T5-Flan, BART, and PEGASUS, to our dataset and further fine-tuning them on a subset from our dataset.

2.1 Overview of Models

FLAN-T5 (Chung, et al., 2022): Scaling Instruction-Finetuned Language Models utilizes a transformer architecture, but is instruction-fine tuned (Flan) using 1,836 tasks such that it can better generalize to unforeseen ones. These tasks are broken down into four different types of "task mixtures" phrased as instructions for the model to learn how to answer questions.

BART (Lewis et al., 2019): Bidirectional and Auto-Regressive Transformers (BART), is a transformer-based sequence to sequence autoencoder that combines a bidirectional encoder with an autoregressive decoder. It is trained by corrupting text with an arbitrary noising function and optimizing a reconstruction loss between the decoder's output and the model's input.

PEGASUS (Zhang et al., 2020): Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) is a transformer-based sequence-to-sequence encoder-decoder model that has produced state of the art results for abstractive text summarization. It is trained using Masked Language Modeling (MLM) and Gap Sentence Generation (GSG) where entire sentences are masked, and the model is trained to predict the masked sentences. PEGASUS has been shown to apply well to unseen summarization datasets.

3 Methods

3.1 Data

The Semantic Scholar Open Research Corpus (S2ORC) consists of 81.1 million English-language academic papers spanning many disciplines. By accessing the API directly provided we identified a subset of 4,160 appropriate articles in the environmental science field. We extract 3,328 papers and their abstracts for training, including 666 for validation, and 832 test papers and their abstracts.

Our dataset consists of the full-length text of the paper (the input) and the corresponding paper abstracts (the label). For each summary that we generate, the full text of the paper is passed into the model, and a generated summary is returned. The generated summary is then compared to the paper's abstract for evaluation.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTPrecision	BERTRecall
Baseline					
T5Level 1	0.198	0.047	0.132	0.845	0.803
T5Level 2	0.200	0.049	0.132	0.845	0.804
T5Level 3	0.207	0.049	0.135	0.847	0.804
BART	0.173	0.046	0.114	0.860	0.807
PEGASUS	0.204	0.042	0.128	0.832	0.804
Fine Tuned					
T5Level 1	0.278	0.073	0.167	0.850	0.820
T5Level 2	0.280	0.074	0.168	0.851	0.820
T5Level 3	0.278	0.072	0.165	0.849	0.819
BART	0.283	0.065	0.162	0.850	0.819
PEGASUS	0.284	0.070	0.165	0.848	0.819

Table 1: Model Scores

3.2 Models

Since our goal is to select and optimize the model that has the highest propensity for abstract summarization on a domain-specific dataset, we applied each model to our test dataset to get a baseline success rate of each model without fine-tuning to gauge the ability of each model to increase performance from fine-tuning. The logic for each model, and why it was explored in the context of our environmental article summarization task, is as follows:

FLAN-T5: The FLAN-T5 model is trained on a multitude of instructions and tasks to better generalize to unforeseen ones. Providing various "levels" of tasks of increasing complexity allows us to determine whether more detailed instructions passed to the model would result in better performance. We align with a task type of "Chain-of-Thought" annotation types, without examples, as outlined in the paper. Chain-of-Thought reasoning provides additional information on how a task should be accomplished and is included as part of the input of what is passed to the model. We establish three levels for the T5 modeling, with each model's input having the defined phrase prefixed to it:

- Level 1: "Summarize: "
- Level 2: "Summarize this scholarly article: "
- Level 3: "Summarize this scholarly article for someone that wants to learn about environmental science: "

BART: The BART model has the potential to be very successful for this summarization task, as it has achieved state-of-the-art results for a variety generation tasks and remains one of the most successful abstractive summarization models to date. A particularly applicable model for our task is the BART large size model, trained on the CNN Daily Mail dataset, a common dataset used for summarization.

PEGASUS: The PEGASUS model uses MLM and GSG in the pre-training to effectively learn and generate key information. A model learning to effectively predict masked sentences will make summaries more coherent, which we believe is especially effective with academic writing such as in our case. We use the XSum downstream dataset, which is trained on a vast corpus of news articles, but ones that are fine-tuned on scientific writing such as arXiv or PubMed would also have been appropriate.

All model outputs are subject to the same hyperparameters so they can be compared across models. To improve efficiency of summary generation, we choose a maximum summary length of 150 words and a minimum summary length of 75 words. While a maximum length of 150 is considerably shorter than most abstract labels, and provides lower scores, a longer maximum length requires a prohibitive amount of compute power that we simply did not have for our study. We also specify a no-repeating n-grams of size 4, which we find to be an appropriate number to remove replicating phrases, to applicable models, in order to avoid repetitions of generated text.

3.3 Evaluation Methodology

Abstract summarization is challenging to evaluate. Words that are spelled the same can take on multiple meanings (e.g. flies the insect and flies the verb) and will not be penalized using traditional

Model	Overall Quality	Fluidity	Coherence	Relevance	Total
Baseline					
T5Level 1	1.90	3.00	2.12	2.38	9.40
T5Level 2	2.48	3.45	2.58	2.78	11.30
T5Level 3	2.57	3.43	2.63	2.77	11.40
BART	3.47	4.45	4.12	3.27	15.30
PEGASUS	3.12	4.07	3.67	2.85	13.70
Fine Tuned					
T5Level 1	3.72	3.92	3.77	3.65	15.05
T5Level 2	3.95	4.12	4.00	3.87	15.93
T5Level 3	3.98	4.13	4.03	3.85	16.00
BART	3.83	4.07	3.93	4.10	15.93
PEGASUS	4.07	4.38	4.17	3.93	16.55

Table 2: Human Evaluations

scoring methods like ROUGE, but words that fundamentally mean the same thing (e.g. feeling good, feeling well) will be penalized. To evaluate the generated summaries, an evaluation that takes into account both the authors' words (e.g. matching n-grams) and whether a summary has a similar meaning is needed. To evaluate these two criteria, we evaluate data using ROUGE and BERTScore metrics.

ROUGE (Lin, Chin-Yew, 2004): These metrics can be used to calculate the level of n-gram overlap the generated summary has with its reference texts. With varying levels of ROUGE scores (e.g. ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-LSUM), we evaluate the extent of varying lengths of n-grams between reference and generated texts to generate similarities.

BERTScore (Zhang et al., 2020): The BERTScore (in the form of BERT-Precision, BERT-Recall, and BERT-F1) is a way for us to evaluate our summaries by taking advantage of the pre-trained embeddings from BERT models, and calculating the cosine similarity between tokens within reference and predicted texts. Where there are semantically correct sentences between reference and generated text, BERTScore provides a way for us to score them highly. For example, the sentence "our planet is growing warmer" and "Earth's temperature is rising" are fundamentally the same thing. BERTScore would calculate the precision and recall for these particular phrases by first matching the most similar tokens between the two phrases via cosine similarity. Precision is then calculated by taking this cosine similarity by the tokens in the generated sentence, whereas recall is calculated by dividing instead by the tokens in the reference sentence. In this sense, BERT-Precision evaluates the "false positive" or irrelevant content in a reference, whereas BERT-Recall evaluates the "false negative" or missing meaning in a reference.

Human Evaluation: We also incorporate a human evaluation to supplement the above scores. We implement a grading criteria, stemming from existing methods of human evaluation of model generated summaries (Iskender et al., 2021; Barbella et al., 2022). Our criteria aligns with our goals presented in this paper and addresses issues that automated scoring methods can struggle on. Our criteria is as follows:

1. Overall quality of the summary
2. Fluidity, which refers to the correct use of words and grammar within a sentence
3. Coherence, which refers to the proper structure of the spans of multiple sentences
4. Relevance, which refers to how close the generated summary is to the given abstract and accurate information

The human evaluations were evaluated on twenty randomly selected article results from each model, holding articles constant across all models for fair evaluation. All three of our group members graded all 20 generated summaries from each model in comparison to the article abstracts on a scale of 1 (very bad) to 5 (very good) for all aforementioned criteria. Our individual scores were then averaged between the three of us to remove bias.

4 Results and Discussion

In Table 1, we compare the results for each models' baseline and fine-tuned summaries using the evaluation methods discussed in the previous section. Overall, the ROUGE scores appear to be low. We attribute this partially to the fact that we limited the output length of each generated summary to about half of the length of the average ground truth summary, and as we will discuss in this section, ROUGE is sensitive to the difference in lengths of the generated text and the reference text. These scores show an average of 10% jump from baseline to fine-tuned ROUGE-1 scores, which is on par with the 12% jump extractive summarization scores in other domain-specific summarization research. The difference between ROUGE scores from the baseline models to the fine-tuned models, is nearly equal to the difference seen in domain-specific, extractive summarization scores (Lee et al., 2021).

For example, the model summary shown in Appendix Figure 1 scored low from a ROUGE perspective. Its ROUGE-1, ROUGE-2, and ROUGE-L metrics for this individual summary, for the T5 Level 3 Fine Tuned model, are 0.380, 0.146, and 0.216. In the highlighted portions of the text, we can see that the low-scoring matches, based off of n-gram overlaps, share very similar meanings. In the yellow text of both summaries, we see semantically similar explanations for the problem presented in the paper regarding the Circumpolar North. Similarly, in the green and purple sections of the summaries, we see semantically similar phrases that are used (such as "This review" vs. "This paper" in green, "changes in climate and environmental conditions that affect population health incomes" vs. "distribution of human health outcomes associated with climate change and related environmental shifts.") While these phrases are similar, the ROUGE scores will not reflect this fact properly.

Despite the imperfections of automatic scoring methodologies, we did see improvement for almost every score when comparing corresponding model baselines to their fine-tuned counterparts. Despite clear improvements, all of the models that we test show similar scores throughout. There is no clear indicator that would lead us to trust one model over the other.

Importantly, we address limitations in summarization evaluation metrics. While ROUGE family of scores and related metrics are the standard metrics in summarization tasks, studies have shown that they are not always reliable outside of its original intent, and this has specifically been demonstrated in the context of scientific article summarization (Fabbri et al., 2021; Cohan & Goharian, 2016). ROUGE measures the lexical overlap between generated and human-written text and is weak to nuances in terminology and extensive compression of detailed text, which by nature, are common in academic abstracts. ROUGE will also score incorrect summarizations highly, as long as similar wording is used. For example, "The weather outside was beautiful" and "The weather outside was terrible," are two sentences with similar wording, but do not convey the same idea. The significant n-gram overlap of the two sentences that would cause them to have a similar score, regardless of meaning. BERTScore has inherent limitations as well as an evaluation metric. Being that the core of its calculation is based off of the embeddings of the BERT LLM, it is subject to the same biases and limitations for which the embeddings were generated. Additionally, BERTScore has demonstrated a tendency to assign high performing scores to a poor translation candidate (that has clear grammatical, negating, or phrase-altering errors) if that candidate has a high semantic overlap (Hanna & Bojar, 2021). While studies have called for an improved summarization evaluation metric, a perfect scoring system has yet to emerge.

As a result, we employ a human evaluation on a sample of the model generated summaries. Human evaluation reveals that the model-generated summaries are in fact quite strong, despite the low scores previously seen from automatic scoring methods. In our judgment, many of the fine-tuned model summaries adequately summarized the reference text and aligned with the ground-truth summaries. An example BART summary can be seen in Appendix A Figure 2, which is highlighted to represent the effectiveness of the summarization, despite misleadingly low scores from ROUGE. In the yellow text of the BART example summary, we can see an introduction to the study and a review of the data collection methods. The blue reference text explains background information on the topic, while the blue BART text creates a condensed sentence to introduce the background for the study topic. The purple text summarizes the number of measures to be taken to limit earthquake effects in both the reference text and BART output.

Table 2 shows the results of the human-evaluated scores of all models. Overall, BART shows the highest total score across the baseline models. The fine-tuned BART model, despite having the least improvement over the baseline, produces summaries directly relevant to the articles it is summarizing and hallucinated the least in the sample. The most common issue with BART, baseline and fine-tuned, is the repetition of information across multiple sentences.

The FLAN-T5 baseline models score considerably lower than other models, producing repetitive sentence structures and inadequate relevance to the articles. However, it achieves on-par results as the BART in the fine-tuned model. In the sample text, you can see that the Flan-T5 Level 3 model does a sensible job at creating a summary for the article, that includes the same or similar information as in the reference text.

The most successful model across all human evaluations is the PEGASUS model. The most common issue we saw from the PEGASUS model was incorrectly identifying the topic of the article. From the sample text shown in Appendix A Figure 2, it is clear that the model has identified the incorrect topic and re-stated details that have motivated the article, but are not the primary focus. While the baseline hallucinates on a few summaries in the sample, the model still produces accurate and well-written summaries of most articles. The fine-tuned PEGASUS model was the most successful of all the models, as it created nearly human-like summaries for most samples. The success of the PEGASUS model was expected, as it uses GSG for pre-training, which we hypothesized would make it optimal for abstractive summarization.

5 Conclusion

In this paper, we explored the impact of pre-trained and fine-tuned models on a domain-specific dataset. The highest scoring model across ROUGE and BERT metrics is the fine-tuned FLAN-T5 Level 3, and the highest scoring model across human evaluation metrics is the fine-tuned PEGASUS model. Our main contributions are:

- Established a new human-evaluation method for domain-specific abstractive summarization.
- Improved ROUGE scores with fine-tuning, nearly equivalent to those seen in previous domain-specific, extractive summarization scores.

6 References

Barbella, Marcello, et al. "Different Metrics Results in Text Summarization Approaches." *Proceedings of the 11th International Conference on Data Science, Technology and Applications*, 2022, pages 31-39.

Chung, Hyung Won, et al. "Scaling Instruction-Finetuned Language Models." 2022.

Cohan, Arman and Nazli Goharian. "Revisiting Summarization Evaluation for Scientific Articles." *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pages 806-813.

Fabbri, Alexander R, et al. "SummEval: Re-evaluating Summarization Evaluation." *Transactions of the Association for Computational Linguistics*, Vol. 9, 2021, pages 391–409.

Hanna, Michael and Ondrej Bojar. "A Fine-Grained Analysis of BERTScore." *Proceedings of the Sixth Conference on Machine Translation (WMT)*, 2021, pages 507–517.

Iskender, Neslihan, et al. "Reliability of human evaluation for text summarization: Lessons learned and challenges ahead." *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, 2021, pages 86–96.

Lee, Huije, et al. "Optimizing Domain Specificity of Transformer-based Language Models for Extractive Summarization of Financial News Articles in Korean". *Korea Advanced Institute of Science and Technology*, 2021.

Lewis, Mike, et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." *Facebook AI*. 2019.

Li, Shuaimin and Jungang Xu. "MRC-Sum: An MRC framework for extractive summarization of academic articles in natural sciences and medicine." *Information Processing & Management*, Volume 60, Issue 5, 2023.

Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries". 2024.

Lo, Kyle, et al. "S2ORC: The Semantic Scholar Open Research Corpus." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Nallapati, Ramesh, et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs

and Beyond". *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, 2016.

Rehman, Tohida, et al. "Generation of Highlights From Research Papers Using Pointer-Generator Networks and SciBERT Embeddings". *IEEE*, Vol. 11, 2023.

See, Abigail, et al. "Get To The Point: Summarization with Pointer-Generator Networks". *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pages 1073-1083.

Tu, Zhaopeng, et al. "Modeling Coverage for Neural Machine Translation". 2016.

Zhang, Jingqing, et al. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization." 2020.

Zhang, Tianyi, et al. "BERTScore: Evaluating Text Generation with BERT." 2020.

A Appendix

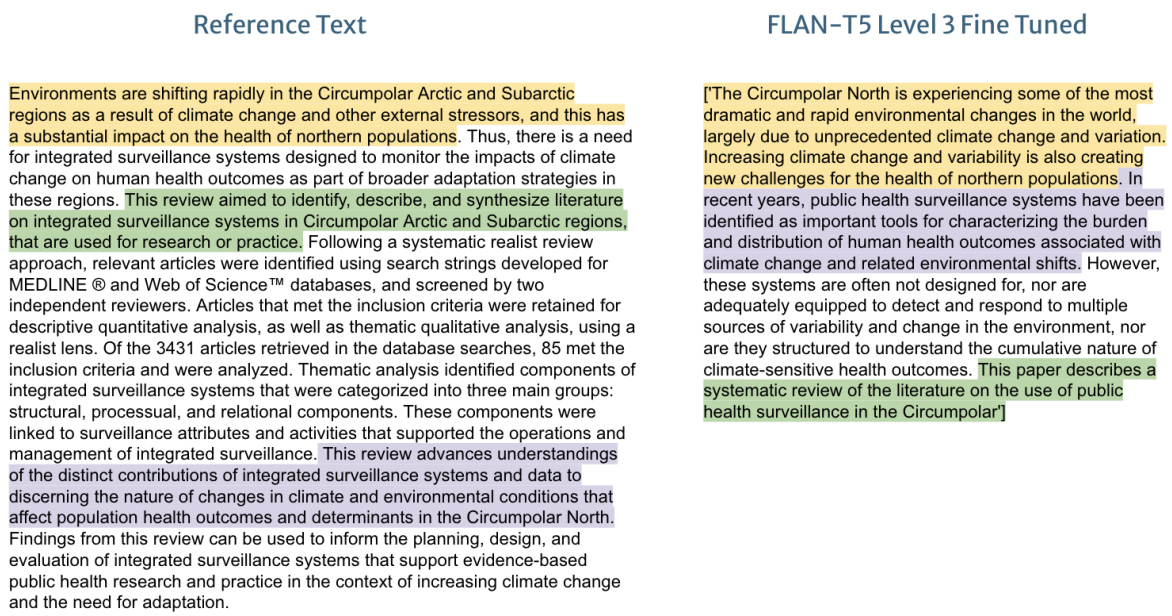

Figure 1: FLAN-T5 Fine Tuned Example Summary

Reference Text

Strong earthquakes can cause buildings to collapse, fires, and jammed doors and elevators. In this study we used the Internet of Things to develop a smart earthquake disaster prevention system. The system continuously monitors the real-time earthquake warning data of the Central Weather Bureau (CWB). When real-time seismic data are updated, the system immediately obtains the intensity and location of the earthquake and calculates the intensity and time of arrival at different locations. When the earthquake is judged to be dangerous, the system immediately shuts off the gas supply. When residents are determined to be at home by a human body detector, a sound and light earthquake alarm is immediately activated, and real-time earthquake data are displayed on LINE instant messaging platform, short message service (SMS), and electronic displays. Doors are also automatically opened to prevent them from being stuck. Earthquake information is immediately sent to regional elevator management centers to instruct elevators to stop at the next floor, open the doors, and broadcast information on the upcoming strong earthquake. By developing this system, the latest earthquake warning information from the Central Meteorological Administration can be effectively captured, and necessary evacuation and disaster mitigation measures can minimize deaths and property losses from earthquakes.

BART Fine Tuned

Geological disasters affect the development of society, economy, and politics across the globe. Major catastrophic earthquakes in history were mostly caused by abrupt crustal movements. However, the times and locations of earthquakes cannot be accurately predicted, so information on epicenter, depth, and intensity of earthquakes can be determined from seismic waves. In this study, a web crawler system was used to instantly access earthquake data published by the Central Weather Bureau (CWB) and calculate the epicenter distance, arrival intensity, and arrival time. Accordingly, effective quake prevention measures can be implemented to reduce the impact of strong earthquakes. In this study, we used the earthquake data published by the CWB to calculate the epicenter distance and other information. If the earthquake

Pegasus Fine Tuned

The purpose of this study is to investigate the impact of seismic waves on earthquake-related deaths in Taiwan. The study is based on the data collected from the Taiwan Geological Survey (TGS) and the National Science Foundation of Taiwan (NSFT). The data were analyzed using the Web of Science (WoS). The results show that the earthquake-related death rate in Taiwan is higher than the global average. The earthquake-related mortality rate in Taiwan was higher than that in the United States (US), China (US), Japan (Japan), and South Korea (South Korea). The earthquake-induced death rate was higher in Taiwan than in the US, China, Japan, South Korea, and South Korea. In addition, the earthquake-induced mortality

Flan-T5 Level 3 Fine Tuned

The Central Weather Bureau (CWB) publishes earthquake data online. The data are used to calculate the epicenter distance, arrival intensity, and arrival time of earthquakes. The data can be used to predict the magnitude of the earthquake, the epicenter location, and the epicenter depth. The data also can be used for the prediction of the earthquake intensity. The results show that the earthquake intensity can be predicted by using the CWB's earthquake data.

Figure 2: Model Fine-Tuned Example Summaries