Domain Adaptation to Climate Change with Improved BLEU Evaluation Method

Stanford CS224N {Custom} Project

Yunan Li
Energy Science Engineering
Stanford Doerr School of Sustainability
yunanli@stanford.edu

Abstract

This project aims to develop a climate-focused QA model that can generate reliable opinions (output) on scientific statements (input) to prevent misinformation in climate sciences. To achieve this goal, we begin our baseline model by fine tuning T5-small model on the climate dataset, and we realize a research problem that model responses are sensitive to input prompt formats. To address this, we conduct an analysis of different prompt formats on the FLAN T5-small model, and select the most effective one to use for fine-tuning with the climate dataset. Our improved FLAN T5-small model outperforms the baseline model by 32%. We continue to improve our model along two pathways: 1) using a larger pre-trained language model and 2) using a larger fine-tuning dataset. The results show T5-base model performs 29% better than FLAN T5-small model, and a larger dataset does not significantly improve performance. The T5-base model outputs are more elaborative, while the FLAN T5-small model responses are stable across different prompt formats. Additionally, we improve the BLEU score by counting synonyms occurrences to quantify model performance.

1 Key Information to include

- Mentor: Hong Liu
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Large language models have been very successful in the NLP community [1, 2] for topics such as machine translation [3], multi-task and few-shot learning [4], and question answer (QA) chatbot models [5]. Domain adaptation on pre-trained large language models has been widely used by fine tuning on dataset with specific topics or tasks [6]. This helps to improve the model capability on a particular topic or task that the original training process or dataset was not given enough emphasis [7]. One of the research problems is the model responses could be very sensitive to the input prompt formats [8]. It is motivated and important to investigate the performance of multiple prompt formats and select the best one for fine tuning. In addition, it is challenging to analyze the similarity between two sentences quantitatively. BLEU score is widely used for multiple NLP tasks [9] such as machine translation, QA models, etc. However, there are restrictions for sentiment level evaluation as well as the synonyms [10]. That means when two sentences show the same meaning for human beings but expressed in different sentence structures or synonym words, the BLEU score will underestimate the performance. Therefore, it is motivated and important to introduce a syn-based BLEU score that considers synonyms when counting the n-gram occurrences.

Climate change and climate sciences are trending topics in our society in recent years [11], and mis-information grows significantly [12]. Although researchers have been contributing efforts from multiple perspectives, it is important to have a QA language model that can interpret the scientific statements with correct opinions in order to avoid mis-information in this domain [13]. At present, we still highly rely on climate scientists to manually evaluate the opinions or claims on a statement, which is expensive regarding time and efficiency. This project aims to leverage the QA language model to fill the gap, avoid mis-interpretations on climate statements, and accelerate the labeling work in literature.

The goal of this project is to develop a QA model in climate domain by fine tuning the pre-trained language models on climate dataset. The prompt formats study paves the way to improve the model stability and accuracy for its performance, and an improved syn-based BLEU score quantify sentences similarities more accurately. We therefore have a better understanding of how far is the predicted sentence to the "golden sentence". Further, we aim to investigate the impacts of large pre-trained language model sizes and dataset sizes on the model performance.

3 Related Work

The state-of-the-art natural language models follow two general steps: pre-training a large language model on an auxiliary task, and then fine-tuning the model on a task-specific labeled dataset using cross-entropy loss [14]. The Text-to-Text Transfer Transformer (T5) model has an encoder-decoder structure [15] that is pre-trained on Colossal Clean Crawled Corpus [16]. We are motivated to start our baseline from the T5 pre-trained model because it works well on a variety of tasks out-of-the-box by pre-pending a different prefix to the input corresponding to each task. FLAN-T5 is an improved version that has been fine tuned in a mixture of tasks [17], and literature has pointed out that the prompt tuning could be comparable to fine tuning [18]. It instructs the work of this project by testing prompt formats analysis on FLAN T5, and explore the model performance on the best prompt format before and after fine tuning on FLAN T5 model using the climate dataset. Fine tuning the pre-trained language models are preferable because we do not need to train a model from scratch, and it also shows strong performance [19].

The BLEU score is initially developed to evaluate machine translation problems automatically [9]. With more applications in NLP research, we find the benefits and limitations of BLEU [10, 20, 21]. A variety of improved versions of BLEU scores emerge in the literature and show better performance on specific tasks, such as for morphologically rich languages [22], sentence or sub-sentence level evaluation [23], and so on. The typical way to test the BLEU score performance is to compare with human evaluation, and it indicates positive signals when a strong correlation appears [24]. Considering this specific work, we aim to develop a syn-based BLEU score due to the large portion of synonyms in the climate dataset. So the syn-based BLEU score is a nice next step for implementation and analysis.

It is more and more important and urgent to address the growth of climate change misinformation [12]. Researchers have attempted to analyze climate disclosures [25], develop pre-trained models for climate-related text [26], and analyze climate policy and actions [27] using NLP. However, those explorations are still quite early in climate science. Literature also points out the need to have a QA based model that can interpret the current opinions based on scientific statements [13]. It motivates this project to address the challenges to develop a climate QA chatbot and contribute to the literature.

4 Approach

In this project, the fine tuning task aims to train a QA model that specializes in climate-related topics. The model takes a climate-related scientific statement as input and generates an opinion or claim as output, both in the form of text sentences. We fine tune the model using the climate dataset as the baseline, and as a stretch goal, we also experiment with the larger wikiQA augmented climate dataset to observe the effect of dataset size on model performance. To evaluate the model's performance, we use the averaged BLEU score on the testing set as the baseline evaluation approach. We also introduce an improved version of BLEU, called Syn-based BLEU, which counts synonyms to improve sentence evaluation. We define the loss function in Eqn. 1 using cross-entropy loss [2], which we minimize to ensure that the model-generated output responses are as close to the target "golden sentences" as possible.

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} y_{i,j}\log(\hat{y}_{i,j}) \tag{1}$$

where θ denotes the model parameters; N is the number of training examples; J is the length of the answer span (or the number of tokens in the answer span); $y_{i,j}$ is the true probability distribution of the answer span for example i at token position j; $\hat{y}_{i,j}$ is the predicted probability distribution of the answer span for example i at token position j.

Baseline: fine tuning on T5-small pre-trained model using climate dataset

The T5-small pre-trained model [16] is selected as baseline because 1) the size of the model is approachable given the computational resources; 2) the encoder-decoder architecture (Fig. 6) shows great potential for the tasks in this project. The model performance is evaluated on the testing set by computing the average syn-based BLEU score of predicted responses compared to the "golden sentences".

Syn-based BLEU score

Syn-based BLEU score is an improved version of BLEU score by considering the synonyms in a sentence. In BLEU score computation, we count the occurrences of n-grams, and we possibly ignore synonym words although they represent the correct meaning. An example in appendix (Fig. 7) shows more details. The syn-based BLEU score counts synonyms by taking advantage of the WordNet, which is a lexical database designed to capture the relationships between words [28]. The implementation is shown in Fig. 1 with the following steps: 1) identify the unique synonyms for each word in the reference sentence; 2) aggregate the reference sentences by combining all identified synonyms; 3) apply the BLEU score metric to find the highest BLEU score by comparing the predicted sentence to all aggregated reference sentences. In step 1), the WordNet could possibly provide more than 10 items, and we first get the unique synonyms and select the top 10 if beyond. A lot of words (such as "this") do not have a synonym output from WordNet, so we take the word itself for aggregation considerations. We evaluate the syn-based BLEU approach in Fig. 1 by comparing with BLEU score on 100 examples randomly selected in the climate dataset during the stability test. We compute both syn-based BLEU and BLEU scores for the same pairs of (prediction, "golden sentence") data. It shows syn-based BLEU score performs better because it is always greater or equal to the BLEU score. For the dots staying on the diagonal line, it shows the syn-based BLEU ends up to be the same with BLEU, that means synonyms do not challenge BLEU score. In general, we observe quite a lot of cases that syn-based BLEU is higher than BLEU, proving it is important to consider synonyms when evaluating model performance. Therefore, we use syn-based BLEU as the approach to quantify model performance.

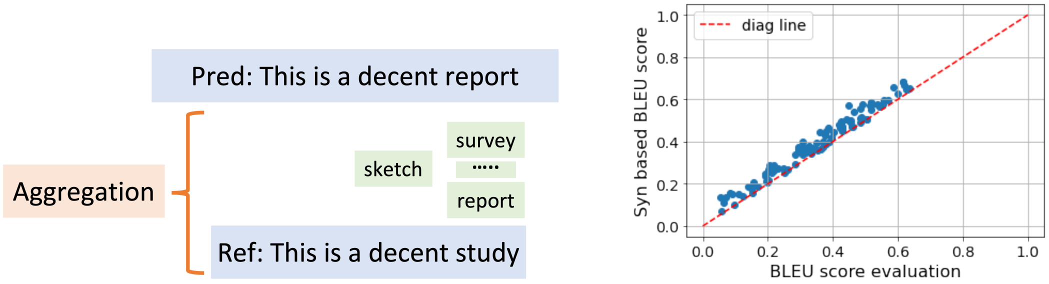


Figure 1: Syn-based BLEU score implementation and performance. Left: illustration of syn-based BLEU score, and the synonyms are based on WordNet [28]. Right: Syn-based BLEU score compared with the BLEU score based on randomly selected 100 examples in the climate dataset during the stability test.

FLAN T5-small for prompt format analysis

The Few-shot Learning with Attentive Networks (FLAN) T5 is an extension of T5 pre-trained model with improvements of enhanced attention mechanism, few-shot and multi-task learning, better scalability, and so on [17]. To address the research problem of model responses stability, it is important to study the input sentence prompt formats to generate output responses in a stable and reliable way [29]. There are 5 steps for prompt format analysis workflow: 1) propose 10 different candidates of prompt formats; 2) apply the prompt formats to the climate dataset; 3) evaluate the performance of all prompt formats using FLAN T5-small pre-trained model directly; 4) select the best prompt format and fine tune on FLAN T5-small using the climate dataset; 5) evaluate the model performance and compare with baseline.

Table 1: Candidates of prompt formats

Case ID	**Prompts formats**
Case 1	What is the opinion on the following scientific statement: **[input]**
Case 2	Do you agree or disagree on **[input]**
Case 3	What do you think of **[input]**
Case 4	What does the following **[input]** mean?
Case 5	What does this climate related scientific statement **[input]** tell us?
Case 6	The scientific statement is **[input]**, so the opinion or claim is
Case 7	The claim based on the following climate statement **[input]** is
Case 8	The correct understanding of **[input]** is
Case 9	Question: **[input]** answer:
Case 10	Scientific statement: **[input]** claim or opinion:

Table 1 shows all candidates of prompt formats we proposed in this work. The **[input]** represents the text sentences (scientific statements) in the climate dataset, and we add prefix or suffix for each case to feed into FLAN T5-small model directly for responses. We generate 5 candidates in the question tone and another 5 candidates in the declarative tone. The evaluation of prompt formats is shown in Eq. 2 that is composed of two aspects: stability and accuracy. For stability test, we have an additional version of **[input]** that adds an extra prefix of "**Recent literature points out that [input]**" before feeding into different prompt formats to compare with the **[input]**.

$$\begin{aligned} Stable\ ratio &= \sum_{i}^{N} \frac{c}{N} \\ Stability &= \sum_{i}^{N} \frac{1}{N} Syn-based\ BLEU(\hat{y}_{1,i}, \hat{y}_{2,i}) \\ Accuracy &= \sum_{i}^{N} \frac{1}{2N} \{Syn-based\ BLEU(\hat{y}_{1,i}, y_i) + Syn-based\ BLEU(\hat{y}_{2,i}, y_i)\} \\ Overall\ score &= 0.25 * Stable\ ratio + 0.25 * Stability + 0.5 * Accuracy \end{aligned} \tag{2}$$

where c is the number of examples that outputs from FLAN T5-small are exactly the same using two different versions of input (**[input]** and "**Recent literature points out that [input]**") under each prompt format case; N is the total number of examples; $\hat{y}_{1,i}$ is the prediction of FLAN T5-small on input prompts based on "**[input]**"; $\hat{y}_{2,i}$ is the prediction of FLAN T5-small on input prompts based on "**Recent literature points out that [input]**"; y_i is the "golden sentence" in the climate dataset.

Stretch goals

The stretch goals of this work are 1) testing a larger dataset and 2) a larger pre-trained language model to observe the model performance. The performance is computed as the averaged syn-based BLEU score on the testing set. We select T5-base for the larger pre-trained language model in this study. We augment the climate dataset by adding 2,325 examples selected from the wikiQA dataset that are related to climate sciences. The climate dataset includes 7,675 examples and the augmented dataset (named as wikiQA+climate) has 10,000 examples. The training, validation, testing split is 80%, 16%, and 4% respectively.

Originality

1) Implement multiple pre-trained language models in workflow and then fine tune with climate data for domain adaptation; 2) Clean and reformat the climate dataset to select appropriate pairs of text sentences from the source; 3) Develop the Syn-based BLEU function to improve the baseline sentence evaluation method (BLEU); 4) Prompt format analysis workflow on FLAN T5-small model to select the best candidate for fine tuning; 5) Dataset augmentation for a larger dataset by selecting examples from wikiQA dataset as stretch goals.

5 Experiments

5.1 Data

The climate dataset is from the CLIMATE-FEVER data [13], including a collection of 19,341 pairs of text in format of (statements, claim/thoughts) pairs. An example is shown in Fig. 2. The preprocessing includes identifying examples that show relations (opinions are supported or refused by statements). It ends up to have 7,675 pairs of sentences for this project. We split the dataset for training (80%), validation (16%), and testing (4%). The tasks of this dataset include fine tuning T5-small for a QA model focusing on climate change, prompt format analysis, and language model size study.

The wikiQA dataset [30] includes 29,258 pairs of question and answer sentences for open-domain QA. There are 2,325 examples selected from wikiQA that are related to climate sciences. The task of wikiQA is to augment the climate dataset to 10,000 examples to analyze the fine tuning dataset size impacts on model performance.

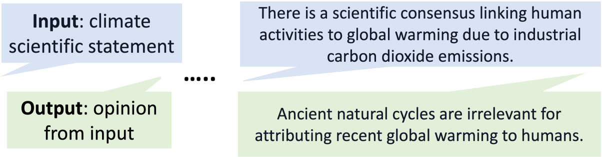

Input: climate scientific statement
There is a scientific consensus linking human activities to global warming due to industrial carbon dioxide emissions.
.....
Output: opinion from input
Ancient natural cycles are irrelevant for attributing recent global warming to humans.

Figure 2: An example of climate dataset with a pair of text sentences (statement, opinion).

5.2 Evaluation method

To measure model performance, we use BLEU score and syn-based BLEU as our evaluation metrics to measure the testing set predictions compared to the "golden sentences". The baseline evaluation metric is BLEU score [9] because it has been widely used to evaluate NLP tasks (especially machine translation and natural language generation like QA) [10]. We also develop our own Syn-based BLEU to improve model performance evaluation based on BLEU.

For prompt format analysis, we use the stable ratio and stability (Eq. 2) to measure model outputs variations given inputs in different formats. The overall score is calculated considering both stability and accuracy for 10 prompt format candidates.

5.3 Experimental details

The experiment details for pre-trained models and datasets are documented in Table 3. They are all conducted on the Tesla T4 GPU with 40 cores and 16 GB memory.

For T5-small fine tuning experiments, we used its tokenizer to process the text sentences, and then define the sequence to sequence training with learning rate of 0.0004, batch size of 8, drop rate of 0.1, maximum input length of 512 tokens, using syn-based BLEU as the evaluation method. We conduct evaluation every 100 steps with appropriate information logged. The training time is around 65 minutes for 2 epochs, and we save the model with the best performance on testing set. When using the augmented dataset, it takes 80 minutes for training. For fine tuning on T5-base using the climate dataset, the training process takes around 133 minutes.

For FLAN T5-small fine tuning experiments, we follow the same set of hyper-parameters but finish training much faster (26 min for the climate dataset and 31 min for the wikiQA+climate dataset) in 2 epochs.

5.4 Results

Figure 3 shows T5-small fine tuning model responses to an example of climate statement. The opinion responses are very different for two types of input prompt formats, and this is not what we expect. A language model should generate stable outputs given similar inputs. In order to improve the baseline results, we further do experiments to fine tune: 1) on FLAN T5 pre-trained model with the best prompt format selected, 2) on a larger pre-trained language model (T5-base), and 3) with a larger dataset (wikiQA+climate).

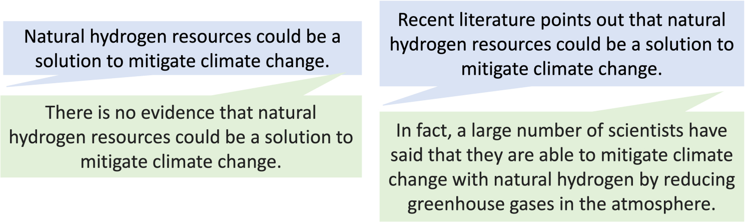

Natural hydrogen resources could be a solution to mitigate climate change.
There is no evidence that natural hydrogen resources could be a solution to mitigate climate change.
Recent literature points out that natural hydrogen resources could be a solution to mitigate climate change.
In fact, a large number of scientists have said that they are able to mitigate climate change with natural hydrogen by reducing greenhouse gases in the atmosphere.

Figure 3: T5-small fine tuning model responses to a statement example in different prompt formats.

The prompt format analysis results in Table 2 show case 7 is the best among all. For prompt cases 6 to 10, they are in the declarative tone, and they tend to have higher overall score performance compared to cases 1 to 5 in the question tone. It is expected to see the best candidate is of the declarative tone. The syn-based BLEU score is always greater than the BLEU score due to its designed algorithm and implementation. Case 7 prompt format is therefore used for fine tuning on FLAN T5-small model.

Table 2: Performance of prompt format candidates on FLAN T5-small model before fine tuning

Case ID	**Stability**			**FLAN T5-small accuracy**		**Performance**
	Stable ratio	BLEU	Syn-based BLEU	BLEU	Syn-based BLEU	Overall score
Case 1	0.4391	0.2117	0.2368	0.1679	0.2027	0.2703
Case 2	0.5505	0.2857	0.2973	0.1048	0.1538	0.2889
Case 3	0.2280	0.0893	0.0942	0.0793	0.0811	0.1211
Case 4	0.4397	0.1903	0.2074	0.1383	0.1700	0.2468
Case 5	0.4169	0.1627	0.1819	0.0950	0.1254	0.2124
Case 6	0.4332	0.1684	0.1826	0.0873	0.1250	0.2165
Case 7	**0.3941**	**0.5255**	**0.5304**	**0.1933**	**0.2149**	**0.3386**
Case 8	0.3694	0.4081	0.4279	0.1661	0.2067	0.3027
Case 9	0.3550	0.1541	0.1616	0.0541	0.0868	0.1726
Case 10	0.3453	0.4555	0.4750	0.1778	0.2079	0.3090

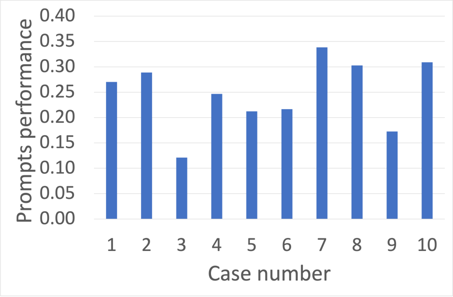


Figure 4: Overall performance of prompt format candidates.

More experiments are conducted to analyze size impacts of pre-trained language model and fine tuning dataset. Table 3 shows the performance for all experiments in this work. First, it is expected to see the fine tuning on FLAN T5-small using the climate dataset improves from 0.2149 (before fine tuning) to 0.2835 (after fine tuning). Second, it shows the larger size of the language model tends to improve the performance. Third, the larger dataset improves very slightly on model performance.

The observations are expected and it tells the approach to fine tune on larger language model with the best prompt format is preferable compared to dataset augmentation in this work.

Table 3: Results of 6 designed experiments with details of pre-trained models and fine tuning datasets

Exp ID	**Pre-trained model**		**Data set**		**Performance**
	Name	**No. params**	**Name**	**No. samples**	**Syn-based BLEU**
Exp 1	T5-small	60M	climate	7,675	0.2137
Exp 2	FLAN T5-small	80M	climate (train)	6,140	Table2
Exp 3	FLAN T5-small	80M	climate	7,675	0.2835
Exp 4	T5-small	60M	wikiQA+climate	10,000	0.2198
Exp 5	FLAN T5-small	80M	wikiQA+climate	10,000	0.2906
Exp 6	T5-base	220M	climate	7,675	0.3654

Exp ID	**Notes**
Exp 1	**Baseline case**
Exp 2	FLAN T5-small on climate (train) dataset with 10 prompt format candidates
Exp 3	Fine tune FLAN T5-small + best prompt format on climate dataset
Exp 4	wikiQA+climate is the augmented dataset with 2,325 examples from wikiQA
Exp 5	Fine tune FLAN T5-small + best prompt format on augmented dataset
Exp 6	**Stretch goal**: investigate the performance of a larger language model

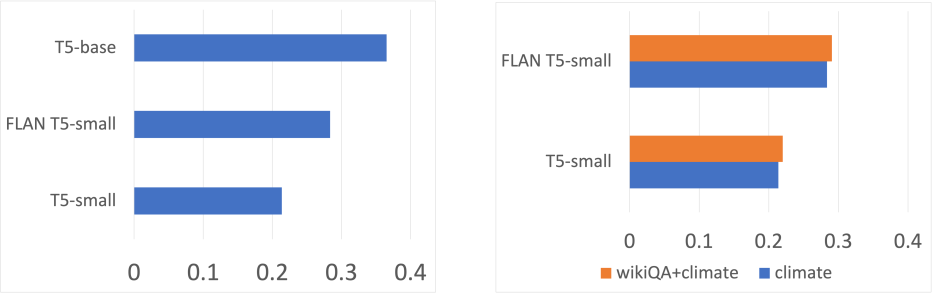


Figure 5: Model performance compared with multiple sizes of pre-trained language models (left) and multiple sizes of dataset (right). For both figures, horizontal axis represents the model performance evaluated using Syn-based BLEU.

6 Analysis

Baseline observations

The baseline model generates outputs in Fig. 3 that is sensitive to the input prompt formats. For the first example (Fig. 3 left), it is possibly that the climate dataset does not specifically include concepts related to natural hydrogen. In fact, natural hydrogen is a very new concept in the climate and energy research domain. However, if we tell the model that recent literature points out that statement, the baseline model changes its mind immediately and then get to agree with the statement. It tells the baseline model can understand the meaning of the prefix we added, and then interpret it correctly.

Prompt formats analysis

The input prompt sentence tones, keywords, structures are important to the overall performance considering stability and accuracy. The prompt analysis in Table 2 shows the best candidate from 10 different cases (Table 1). In general, a declarative tone of input shows better overall performance, considering both stability and accuracy. The question tone cases tend to have higher stable ratio, meaning more examples in question tone format will get the exact same outputs. However, its accuracy is in general less than the declarative tone cases. By comparing different cases in Table 1, case 7 prompt format specifically mentions keywords such as "climate statement" and "claim". This gives more clear indications to the FLAN T5-small model before fine tuning. Case 1 also mentions the keywords ("opinion", "scientific statement") specifically, and its overall performance score is relatively high among other cases in a question tone. Also, case 7 does not break a whole sentence by a comma, which might also be an advantage.

Model improvements

A larger fine tuning dataset does not improve a lot (shown in Fig. 5) because the wikiQA examples are pairs of (question, answer) instead of (statement, opinion). A larger pre-trained language model is expected to show better performance because it is capable to handle more complicated patterns. T5-base model performance is the best among all, and FLAN T5-small with the best prompt format shows improvements compared to baseline. We also test an example in Table 4 for the improved models.

Table 4: Improved models responses to a statement example in different prompt formats.

Input 1	Natural hydrogen is a carbon-free energy.	
Input 2	Recent literature points out that natural hydrogen is a carbon-free energy	
Model	**Input**	**Output**
FLAN T5-small	Input 1	Natural hydrogen is a renewable source of energy.
FLAN T5-small	Input 2	Natural hydrogen is a carbon-free energy.
T5-base	Input 1	The fact that natural hydrogen is a carbon-free energy is proving to be a good thing.
T5-base	Input 2	In fact, there is a lot of evidence that natural hydrogen is a carbon-free energy.

For the FLAN T5-small model, the responses to both input1 and input2 are very close to each other. *The renewable source of energy* has the equivalent meaning compared to *a carbon-free energy* in this context. This example shows great stability performance due to the way we design the fine tuning workflow.

For the T5-base model, the responses are more elaborative compared to the FLAN T5-small because T5-base is huge, having more than twice of parameters to FLAN T5-small. The T5-base model generates slightly different output responses when adding a prefix of *"Recent literature points out that"*. The words in output ("a lot of evidence") are correlated to this input prefix.

7 Conclusion

In this project, we aim to develop a a climate-focused QA language model that generates accurate opinions on scientific statements. We start with fine-tuning T5-small on the climate dataset, but find that the model's outputs are sensitive to input prompt formats. To address this problem, we conduct a prompt format analysis based on FLAN T5-small model, and develop a workflow, as well as performance evaluation metrics (stability and accuracy) to select the best prompt format before fine-tuning. We find that different prompt formats show very different performance, and prompt format selection is a necessary step. A declarative tone of prompt format with essential keywords tends to be the best out of all candidates.

After fine-tuning with the best prompt format, the FLAN T5-small model shows a 32% improvement compared to the baseline, and a 31% improvement compared to the model before fine-tuning. The T5-base fine-tuned model shows the best performance among all experiments, with an improvement of 29% compared to the FLAN T5-small fine-tuned model. This is because T5-base is a much larger language model, and the size of the pre-trained model is very important for model performance. Comparing the model outputs from FLAN T5-small and T5-base, the responses from FLAN T5-small have excellent stability performance, while T5-base outputs are more elaborate due to its larger model size. Both models are capable of generating reasonable opinion responses to climate-related scientific statements. We also develop a syn-based BLEU to improve sentences evaluation counting synonyms. Our results show that the syn-based BLEU is always greater or equal than the BLEU score, by the definition and implementation of syn-based BLEU.

For future work, we would like to continue our approach and fine-tune larger pre-trained language models to further improve model performance. Human evaluation from multiple researchers in this domain could also provide valuable references and avoid evaluation biases from a single person.

References

[1] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. 2007.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[6] Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. Contrastive domain adaptation for question answering using limited text corpora. *arXiv preprint arXiv:2108.13854*, 2021.

[7] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogerio Feris. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4805–4814, 2019.

[8] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. *arXiv preprint arXiv:2006.04884*, 2020.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[10] Ehud Reiter. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401, 2018.

[11] Valérie Masson-Delmotte, Panmao Zhai, Anna Pirani, Sarah L Connors, Clotilde Péan, Sophie Berger, Nada Caud, Y Chen, L Goldfarb, MI Gomis, et al. Climate change 2021: the physical science basis. *Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change*, 2, 2021.

[12] Justin Farrell. The growth of climate change misinformation in us philanthropy: evidence from natural language processing. *Environmental Research Letters*, 14(3):034013, 2019.

[13] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

[14] Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*, 2020.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[16] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[17] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[18] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[19] Christoph Alt, Marc Hübner, and Leonhard Hennig. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. *arXiv preprint arXiv:1906.08646*, 2019.

[20] Ngoc Tran, Hieu Tran, Son Nguyen, Hoan Nguyen, and Tien Nguyen. Does bleu score work for code migration? In *2019 IEEE/ACM 27th International Conference on Program Comprehension (ICPC)*, pages 165–176. IEEE, 2019.

[21] Bogdan Babych. Automated mt evaluation metrics and their limitations. *Revista Tradumàtica: tecnologies de la traducció*, (12):464–470, 2014.

[22] Shweta Chauhan, Philemon Daniel, Archita Mishra, and Abhay Kumar. Adableu: A modified bleu score for morphologically rich languages. *IETE Journal of Research*, pages 1–12, 2021.

[23] Xingyi Song, Trevor Cohn, and Lucia Specia. Bleu deconstructed: Designing a better mt evaluation metric. *Int. J. Comput. Linguistics Appl.*, 4(2):29–44, 2013.

[24] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of bleu in machine translation research. In *11th conference of the european chapter of the association for computational linguistics*, pages 249–256, 2006.

[25] Alexandra Luccioni and Hector Palacios. Using natural language processing to analyze financial climate disclosures. In *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*, 2019.

[26] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

[27] Pradip Swarnakar and Ashutosh Modi. Nlp for climate policy: Creating a knowledge platform for holistic and effective climate action. *arXiv preprint arXiv:2105.05621*, 2021.

[28] Christiane Fellbaum. Wordnet and wordnets. 2005.

[29] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[30] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018, 2015.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A Appendix

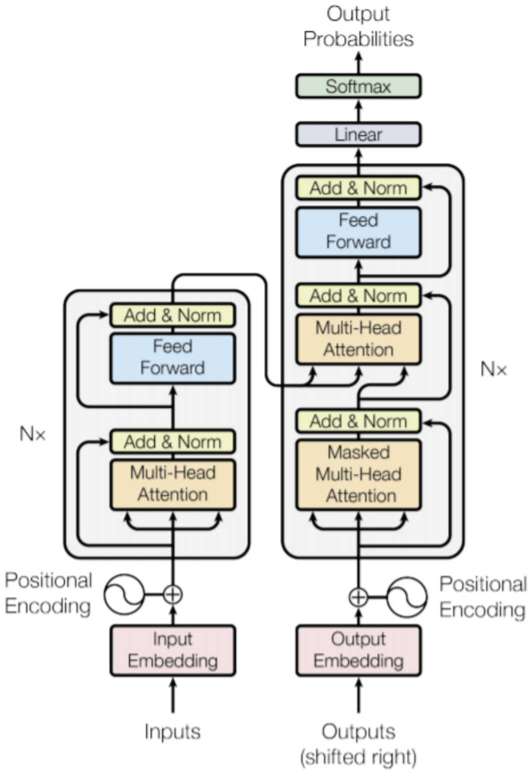
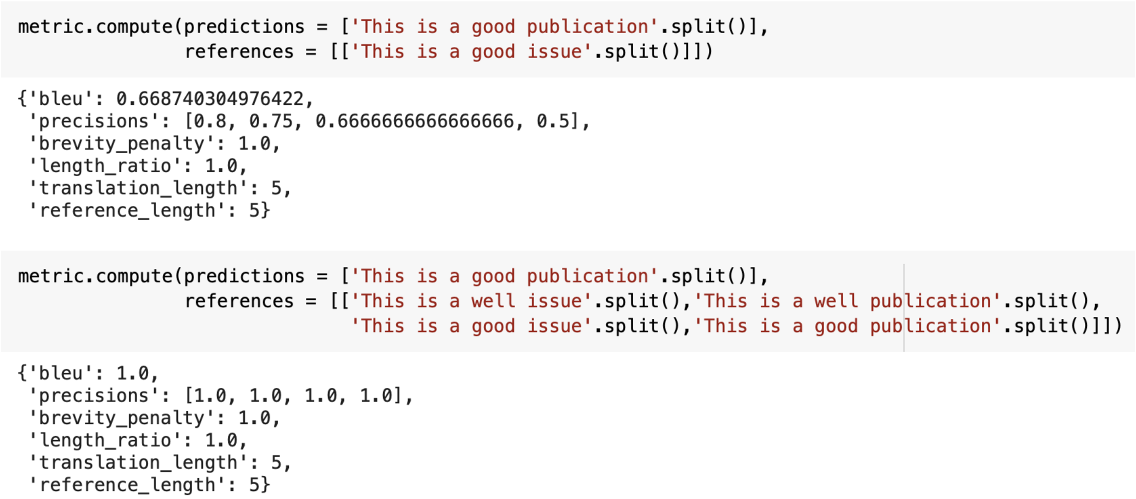


Figure 6: An example architecture of the transformer model [31]. The encoder is on the left and the decoder is on the right. The architecture of T5 pre-trained language model is very similar to this figure.



```
metric.compute(predictions = ['This is a good publication'.split()],
               references = [['This is a good issue'.split()]])
```

```
{'bleu': 0.668740304976422,
 'precisions': [0.8, 0.75, 0.6666666666666666, 0.5],
 'brevity_penalty': 1.0,
 'length_ratio': 1.0,
 'translation_length': 5,
 'reference_length': 5}
```

```
metric.compute(predictions = ['This is a good publication'.split()],
               references = [['This is a well issue'.split(),'This is a well publication'.split(),
                              'This is a good issue'.split(),'This is a good publication'.split()]])
```

```
{'bleu': 1.0,
 'precisions': [1.0, 1.0, 1.0, 1.0],
 'brevity_penalty': 1.0,
 'length_ratio': 1.0,
 'translation_length': 5,
 'reference_length': 5}
```

Figure 7: An example of Syn-based BLEU evaluation method showing improved score compared to the baseline of BLEU evaluation. The predicted sentence is *This is a good publication.* The reference sentence is *This is a good issue.* According to the Syn-based BLEU method, we check the synonyms and it shows the word *good* has a synonym *well*, and the word *publication* has a synonym *issue*. The aggregated reference sentences will be: *This is a good issue.*, *This is a well issue.*, *This is a good publication.* and *This is a well publication.* The Syn-based BLEU improves from BLEU score of 0.669 to 1.