A CROSS-LINGUAL QUESTION ANSWERING ARCHITECTURE BASED ON PRE-TRAINED TOKEN-FREE MODEL WITH INSTRUCTION FINE-TUNING

Yi Dai
University of Michigan
yid@umich.edu

Franklin Kong
University of Michigan
ftk@umich.edu

Chenshun Ni
University of Michigan
nichensh@umich.edu

Xiyu Tian
University of Michigan
xiyutian@umich.edu

ABSTRACT

Most widely used cross-lingual question answering (QA) systems are operated with human/machine translation, which use large language models (LLM) based on sequences of tokens corresponding to words or subwords units. However, current cross-lingual QA systems have limitations, such as difficulty on answering questions independently on the source language. In this report, we propose a cross-lingual QA architecture that fine-tunes state-of-the-art token-free LLM ByT5 on the modified XQuAD to extract answers from the source in different language. We set ByT5-base without fined-tuned as baseline, and test several fine-tuned variants of ByT5. We evaluate the performance of not only all variants of ByT5 on the cross-lingual QA task, but also the architecture with machine translation with different variants of ByT5. Results demonstrate that mono-lingual instruction-tuned ByT5 model outperforms others on mono-lingual QA task, and cross-lingual instruction-tuned ByT5 model has the best performance on the cross-lingual QA task.

1 Introduction and Motivation

Question Answering (QA)[1, 2], which is concerned with building systems that automatically answer questions posed by humans in a natural language. Cross-lingual QA is a subfield of QA, and it focuses on creating systems that do not only answer questions but also allow the user to access information in other languages that used for asking questions.

Despite LLMs have reached huge development, cross-lingual QA is still one of the most challenging downstream tasks of natural language processing[3]. Lack of tools and resources and vocabulary gap between source and target languages, frustrate any effort to adapt existing approaches for cross-lingual QA task. It could be imagined that thousands of hundreds of people will benefit from a mature cross-lingual QA system in which people are allowed to ask questions and to acquire answers in mother language while source languages are independent on target languages. Thus, this system can effectively break down the language barriers between different cultures so as to improve the quality of tech communication between different countries and facilitate modernized education systems in some poor areas lacking access to high quality education resources.

In this work, we propose a cross-lingual QA architecture that utilizes the state-of-the-art LLM ByT5[4] on the modified XQuAD[5] and attempt to show that the cross-lingual question answering task can be achieved without machine translation module. Moreover, we implement instruction-tuning to fine-tune ByT5 in order to improve the performance of the proposed architecture. The machine translation module is also implemented into this architecture to compare with our results.

All codes and dataset have been submitted with this report.

2 Proposed Method

We propose a cross-lingual QA architecture that fine-tunes state-of-the-art token-free LLM ByT5 on the modified XQuAD or instruction tuning datasets to extract answers to a specific question about the context in a different language.

2.1 Dataset

2.1.1 Modified XQuAD

We use XQuAD[5] as a benchmark QA data set for our cross-lingual QA task, which consists of 240 context paragraphs and 1190 question-answer pairs related to the contexts. However, the context and QA pairs are completely parallel across all 11 kinds of languages, and thus XQuAD is more similar to a mono-lingual QA data set than a cross-lingual QA data set, whose QA pairs' language may be different from the corresponding context language. In view of this, we propose a modified XQuAD data set whose context and QA pairs are in different languages and fine-tune the language model on it attempting to make the model's ability to generate sensible answers less dependent on the context's language. Figure 1 shows an example of our modified XQuAD data set that has English context paragraph and Mandarin QA pairs.

2.1.2 Instruction Tuning Related Dataset

To further improve the LLM's learning capability, we implement instruction tuning which let our LM learn to perform many tasks via natural language instructions, such as chain-of-thoughts and natural language inference. Our instruction tuning related data sets consist of: (1) the FLAN 2022 Collection[6], (2) the Cross-lingual Mixed Instruction-Following Dataset (CMIFD) adapted from Alpaca[7].

The FLAN 2022 Collection offers the most extensive publicly available set of tasks for instruction tuning, which have been compiled in one place. This dataset has been supplemented with hundreds more of Google's high-quality templates, richer formatting patterns, and data augmentations, which consists of 16 tasks and more than 70 text datasets, including both language understanding and language generation tasks, into a single mixture. The CMIFD (in Fig. 2) contains four versions of the same text dataset generated by GPT-4: the language combinations for writing instructions and answers are 4 different permutations$(2 * 2 = 4)$ of Chinese and English.

Both aforementioned datasets are our potential instruction tuning training datasets.

2.2 Fine-tune ByT5

ByT5 is a large-scale language model for token-free text-to-text transfer learning, a variant of mT5[8] that simplifies the NLP pipeline by doing away with vocabulary building, text preprocessing and tokenization. ByT5 outperforms mT5 in any of these five scenarios[4]:
(1) at model sizes under 1 billion parameters,
(2) on generative tasks,
(3) on multilingual tasks with in-language labels,
(4) on word-level tasks sensitive to spelling and/or pronunciation,
(5) in the presence of various types of noise.

We first use ByT5-base without fine-tuning to extract answers directly from the modified XQuad as the baseline. Note that ByT5-base is the largest ByT5 model we are able to train due to computational cost and RAM size limitation. Machine translation is not necessary, as the dark grey pathway shown in Fig.3, since the modified XQuAD contains two languages (English and Mandarin) and we think that the relationship between different languages can be learned by ByT5 via fine-tuning.

On the foundation of baseline, we fine-tune ByT5 mainly by two methods and compare its performances: (1) fine-tuning ByT5 on XQuAD with and without machine translation module[9], (2) conducting instruction fine-tuning[10] on FLAN 2022 Collection data set and CMIFD to get FLAN-ByT5-1 and FLAN-ByT5-2 respectively, then fine tuning on XQuAD data set with and without machine translation module. All different variants of ByT5 (fine (instruction)-tuned on different datasets) are shown in the green box of pipeline schematic (Fig. 3). To compare with the proposed cross-lingual QA architecture, an architecture with machine translation module via different variant of ByT5 is also implemented(yellow pathway in Fig.3). The novelty in this work has been pointed out in bold font, which are the modified XQuAD, and FLAN-ByT5.

2.3 Evaluation

Reliable evaluation metrics are needed to demonstrate performances of all variants of the proposed cross-lingual QA architecture.

2.3.1 BLEU

BLEU[11] is one of the most popular NLP evaluation metrics due to its intelligibility and language-independence. BLEU score is between 0 and 1. The higher BLEU Score is, the closer the predicted sentence is to the human-generated target sentence. BLEU Score combines N Clipped Precisions by using Geometric Average Precision Scores (GAPS) as the following formula below:

$$\text{Geometric Average Precision} = \prod_{n=1}^{N} p_n^{w_n}$$

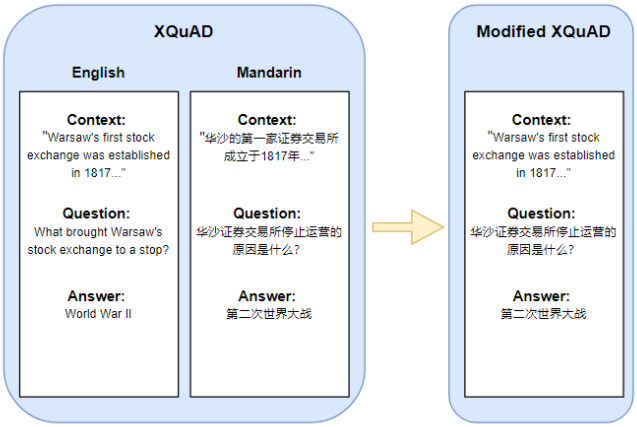

Figure 1: The modified XQuAD

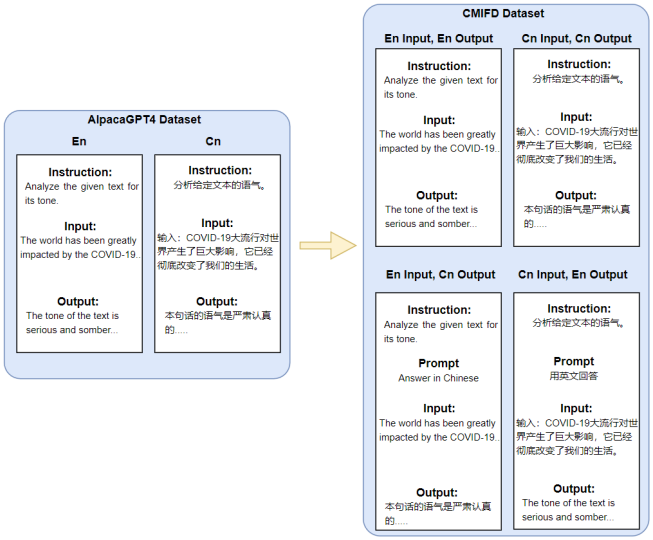

Figure 2: CMIFD DataSet

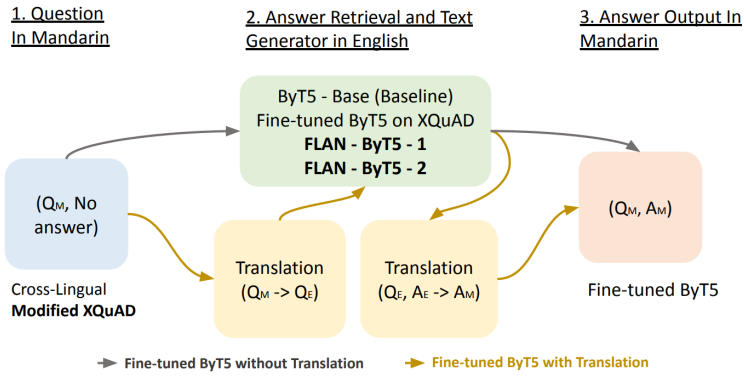

Figure 3: Schematic diagram of pipeline of cross-lingual QA

where typically $N = 4$, $w_n = 1/N$, and

$$p_n = \frac{\text{Number of correct predicted n-grams}}{\text{Number of total predicted n-grams}}$$

To encourage the model to generate a longer answer, the Brevity Penalty is used as the formula below:

$$\text{Brevity Penalty} = \begin{cases} 1, & \text{if } c > r \\ e^{(1-r/c)}, & \text{if } c \leq r \end{cases}$$

where

$$c = \text{number of words in the predicted sentence}$$

$$r = \text{number of words in the target sentence}$$

The BLEU score is calculated by multiplying the Brevity Penalty with the Geometric Average of the Precision Scores:

$$\text{BLEU} = \text{BP} \cdot \text{GAPS}$$

We can see that the Brevity Penalty(BP) is smaller for a shorter predicted sentence and vice versa. Note that BLEU score is proportional to BP, and thus a smaller BP value indicates a larger penalty.

2.3.2 ROUGE

ROUGE-n and ROUGE-L[12] are implemented to generate another two metrics. Both of them utilize Recall, Precision, and F1-Score as the following formulas below:

$$\text{Recall} = \frac{\text{count n-grams found in model and reference}}{\text{count n-grams in reference}}$$

$$\text{Precision} = \frac{\text{count n-grams found in model and reference}}{\text{count n-grams in model}}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score gives us reliable measures of our model performances that rely not only on the model capturing more (recall) but also doing so without outputting irrelevant words (precision).

ROUGE-L measures the longest common subsequence between the model output and reference. The recall, precision, and match (same as F1-score) are calculated in the same ways in ROUGE-n method.

3 Related Work

3.1 Question Answering

Question answering is a growing research field because its powerful ability to deliver short, precise and question-specific answering[13]. Question answering systems enable understanding questions and retrieving answers by using natural language queries[2]. Current NLP pipelines often use transfer learning which fine-tunes a pre-trained model on some specific downstream tasks of interest[14], for instance, QA. Cross-lingual QA, as one of the most challenging subfields of QA, has some state-of-the-art frameworks. XOR is a new task framework that involves cross-lingual document retrieval from multilingual and English resources by implementing translation[15], while CORA, the first unified many-to-many QA model[16], answers directly in the target language without any translation or in-language retrieval modules.

3.2 Multilingual LMs

Some of state-of-the-art pre-trained language models are BERT[17, 18], T5[19], GPT3[20], PaLM[21, 22, 23]. BERT (Bidirectional Encoder Representations from Transformers) is one of the first developed Transformer-based self-supervised language models and an encoder-only bidirectional transformer with 430 million parameters, while GPT3 (Generative Pre-Training model), powering the famous ChatGPT, is a decoder only unidirectional autoregressive model with 175 billion parameters. PaML (Pathways Language Model) is the most recent breakthrough from Google which has 540 billion parameters and is efficiently trained with the Pathways system[24]. T5, Text-To-Text Transfer Transformer, is an encoder-decoder, multi-task learning model that proposes reframing all NLP tasks into a unified text-to-text-format where the input and output are always text strings. This formatting makes T5 model fit for multiple tasks.

Recent multilingual models are pre-trained on data set covering not only English, but also other frequently used language, such as Germany, French, Mandarin, etc. The state-of-the-art, such as, mBERT[17, 25], mGPT[26] and mT5[8] are multilingual variant of the corresponding pre-trained language models. Among above, mT5 outperforms many cross-lingual NLP tasks. ByT5[4], is a general-purpose pre-trained text-to-text (token-free) model covering over 100 languages based on mT5 model. As state-of-the-art LM, ByT5 outperforms mT5 on in some scenarios, such as general tasks, which

3.3 Instruction Tuning

In order to give NLP models a wider range of applications, the concept of instruction tuning is introduced to help the model perform tasks it was not trained on, as shown in Fig. 4. Instruction tuning, fine-tuning a language model on a collection of NLP tasks described via instructions, can improve the zero-shot performance of language models on unseen tasks. Researchers instruction-tuned a 137B pretrained LM by dividing 62 NLP tasks into 12 categories[10], finetuning the model on 11 categories and testing the zero-shot effect on the left one category to see

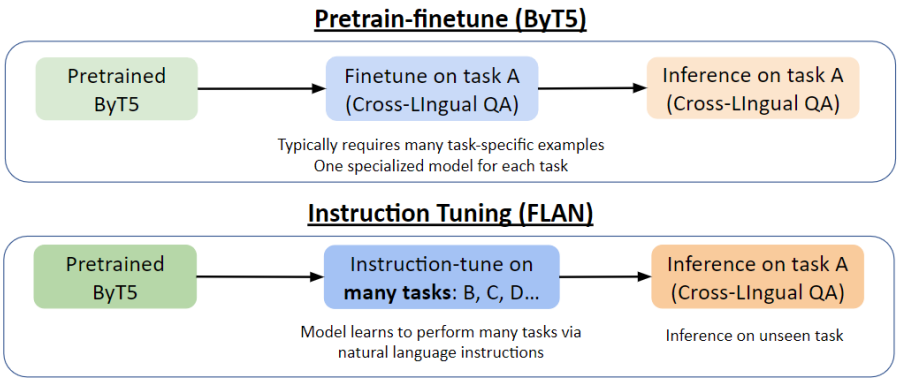

Figure 4: Instruction Tuning Schematic

Configuration	BLEU Score	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ByT5 (Base)	0.50	2.15	0.58	0.30	0.16
ByT5 + Finetune	4.15	20.00	6.45	4.00	2.82
ByT5 + Finetune + MT	27.50	61.47	51.46	42.86	41.46
FLAN-ByT5-1 + Finetune	2.42	14.29	4.09	2.48	2.47
FLAN-ByT5-1 + Finetune + MT	65.05	75.71	70.91	67.84	65.41
FLAN-ByT5-2 + Finetune	4.57	32.40	15.13	8.33	6.98
FLAN-ByT5-2 + Finetune + MT	37.76	37.75	46.77	40.40	36.21

Table 1: BLEU metric

whether the model can truly understand instructions. The performance of FLAN model fine-tuned by the above multitasking instructions can exceed that of zero-shot or even few-shot of GPT-4[7] in many task categories(especially in tasks of natural language inference and reading comprehension). However, they mainly focus on instruction-tuning monolingual LLMs, and in view of this, we can attempt to instruction-tune some multilingual LLMs.

In this report, we introduce a cross-lingual variant of ByT5[4] which preserves the token-free text-to-text model and the feature of robustness to noise. Our model avoids suffering from error propagation of the machine translation component into the downstream QA because of no existence of translation on question/answer processing. This enables our model to answer questions whose answers can be found in resources written in languages other than English or the target languages.

4 Results and Discussion

We conduct our experiments with seven configurations. The configurations differ by whether fine-tuning, instruction-tuning, or machine translation (MT) took place. For instruction-tuning we use two different data sets: The Chain of Thought suite of FLAN 2022 Collection (FLAN-ByT5-1) and CMIFD instruction set (FLAN-ByT5-2).

Note that when fine-tuning and machine translation occur together, it implies that the XQuAD data set used to fine-tune the model is monolingual (pure English). If the fine-tuning is by itself, it utilizes the mixed-lingual XQuAD data set (Chinese and English).

For each configuration, we use ByT5-base (581 million parameters) and fine-tune it with 10 epochs, with a learning-rate of 0.0001. The XQuAD data set consists of 1130 data points in the training set. FLAN-Byt5-1 is instruction-tuned with 70000+ data points, and FLAN-Byt5-2 is instruction-tuned with 180000+ data points.

4.1 Observation

Table. 1 shows the BLEU scores as well as the n-gram precision of the configurations. As expected, ByT5 without any fine-tuning or instruction tuning yields the worst performance. The ByT5 model with zero-shot generation does not seem to understand the task of cross-lingual QA. Instead of answering questions, ByT5 without fine-tuning is attempting to generate new question-context pairs in Chinese and English, with mostly gibberish information that is usually only partially readable, while being riddled with spelling errors and non-sequitur. Table. 2 reflects this issue by having 0 values for all of its ROUGE scores. Almost all of the generated responses are irrelevant and oblivious to our cross-lingual QA task.

Configuration	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1-Score	Recall	Precision	F1-Score	Recall	Precision	F1-Score
ByT5 (Base)	0	0	0	0	0	0	0	0	0
ByT5 + Finetune	0.133	0.164	0.144	0.011	0.025	0.015	0.132	0.164	0.143
ByT5 + Finetune + MT	0.605	0.631	0.593	0.265	0.286	0.269	0.606	0.631	0.594
FLAN-ByT5-1 + Finetune	0.117	0.135	0.123	0.021	0.022	0.021	0.114	0.135	0.121
FLAN-ByT5-1 + Finetune + MT	0.792	0.771	0.776	0.525	0.506	0.511	0.792	0.771	0.777
FLAN-ByT5-2 + Finetune	0.265	0.333	0.284	0.056	0.067	0.059	0.261	0.329	0.280
FLAN-ByT5-2 + Finetune + MT	0.569	0.563	0.559	0.332	0.344	0.334	0.565	0.559	0.555

Table 2: ROUGE metric

We can also observe that fine-tuning on the mixed-lingual modified XQuAD has some effect on performance. ByT5 with fine-tuning is able to produce readable and most relevant answer text in Chinese, a result of fine-tuning making the model learn our task. But the improvement is limited, since most predictions are incorrect. An observation could be made that the model recognizes what type of predictions to produce. For example, when the expected answer is numeric, the fine-tuned model would often output a numeric prediction. Otherwise, the model would produce an answer text in Chinese that is at least somewhat related to the English context and the Chinese question, albeit occasionally it utilizes and exploits spurious patterns, such as certain keywords in the question, likely ignoring the English context altogether. There are cases where the model naively repeats a key-phrase from the question and does not answer the questions in an intuitive fashion. Examples of these cases can be found in Appendix A.

If ByT5 is fine-tuned on the mono-lingual English-only XQuAD data set, after translating the inputs from Chinese to English and the predictions from English into Chinese, the performance of these predictions, by either human judgment or metrics, improved drastically. This result is expected as mono-lingual tasks should experience fewer noises during fine-tuning. Most predictions from this setup are either correct or are near correct, with some outputs being only a few characters away from the label, but the information contained in the predictions are consistent with that of the expected answers.

Instruction tuning with fine-tuning on mixed-lingual XQuAD also has a limited effect on the performance. It should be noted that FLAN-ByT5-1 is instruction-tuned on an English-only instruction set, whereas FLAN-ByT5-2 is instruction-tuned on a mixed-lingual Chinese-English instruction set. Although FLAN-ByT5-1 yields a relatively unremarkable performance, FLAN-ByT5-2 is able to perform somewhat better than ByT5 without instruction tuning. The output behaviors of these two setups(FLAN-ByT5-1,FLAN-ByT5-2) are similar to those of ByT5 with mixed-lingual fine-tuning(without instruction tuning), but with comparatively better or worse predictions.

However, if we are to fine-tune FLAN-ByT5 setups on the English-only XQuAD and conduct a similar translation process to that of ByT5 + Finetune + MT, we are able to generate exceptional predictions that almost completely match the label answer texts (via human judgment). This is especially the case for FLAN-ByT5-1, which is already instruction-tuned on a mono-lingual English-only instruction set. For both BLEU and ROUGE metrics, we see an extreme boost in performance.

4.2 Discussion

From the data we have collected, we can make some key speculations, and also reflect upon the limitations of our conjecture and methodology.

An important observation that can be made is that under the same circumstances and with the same resources, mono-lingual tasks would outperform cross-lingual tasks. Directly fine-tuning a multi-lingual language model such as ByT5 on a cross-lingual QA task would not provide a highly desirable output. Training a model to conduct cross-lingual comprehension and carry out question-answering, without any external module's assistance with translation, is a burdensome task that requires a much larger data-set and significantly lengthier training cycles. Our mixed-lingual XQuAD data-set is of the same size as the original XQuAD data-set of a specific language. With small epoch number to 10, it seems unrealistic that the model could handle cross-lingual QA tasks reasonably.

Even with the addition of instruction-tuning, the mixed-lingual nature of the data sets we used are too noisy and small to be effective with a short training period. However, if we switch to the mono-lingual QA task after instruction-tuning, the precision and recall performances of BLEU and ROUGE metrics received an explosive gain, demonstrating the effectiveness of instruction-tuning at least for mono-lingual tasks.

A portion of our unremarkable results could be attributed to the limited resources at hand. ByT5 as a LLM uses a more lengthy tokenization strategy in comparison to other

LLMs. Although its aim is to provide character-specific tokenization on a more granular level to produce better results, this strategy inadvertently put more strain on the training process. Longer tokenization implies the need for larger VRAM during batch training, yet we only had access to NVIDIA A40, which has up to about 40 GB. Without truncating our inputs it is impossible to fine-tune without running out of memory. This creates great difficulty especially for the instruction sets, for its inputs can be extremely stretched-out. Setting a size limit for our tokenized inputs means that a large part of the context is cut off. Therefore, we were unable to use our data sets to their fullest in the learning process, most likely causing us to waste too many resources for subpar results. Should we have used a multi-lingual LLM that uses a shorter tokenization strategy, such as mT5, we may be able to alleviate the strain on memory space to some extent.

We were also unable to use ByT5 models larger than ByT5-base, given the scale of our task and the limited resource we have access to. With a larger model and more parameters, we could produce better results for QA tasks, but it is unrealistic for us to utilize it without sacrificing even more aspects of our data sets.

5 Conclusion

We build a cross-lingual QA architecture by fine-tuning the state-of-the-art language model ByT5. We test seven configurations that contain ByT5 without fine-tuned, with fine-tuned by the modified XQuAD, and FLAN-ByT5s fine-tuned by two instruction tuning data sets. Results demonstrate that (1) ByT5 without any fine-tuning gives the worst performance, (2) for instruction tuning, cross-lingual data set (CMIFD) improves performance better than mono-lingual data set (FLAN 2022 Collection), (3) mono-lingual tasks (with machine translation) outperform cross-lingual tasks by a large margin with limited model size.

For future work, we would try more data sets from FLAN 2022 Collection to improve instruction tuning and use larger ByT5 LMs, for instance, ByT5 xl if the computational resource is available. We would also test our architecture on other multilingual language models that don't require long tokenizer embeddings, such as mT5 and GPT-4.

6 Author Contribution

All co-authors equally contribute to the project. The table format of author contribution is shown in Table. 3. The detailed description of contribution is:

Yi Dai was involved in writing Abstract, Introduction and Motivation, Proposed Method, Related Work, and Conclusion, helped with sanity check of data pre-processing and experiments to get preliminary results of fine-tuned ByT5.

Franklin Kong researched and wrote the code framework for finetuning Byt5, finetuned and tested the ByT5 LLM with the data sets, collected metric results, wrote section 4 the report.

Chenshun Ni provided ideas of implementing the machine translation module and instruction tuning together to enhance the performance of the model dramatically which is a major discovery of our group. He also participated in the writing of most parts of the report, correcting many description inaccuracies and providing some useful suggestions in the format.

Xiyu Tian created the modified cross-lingual XQuAD and CMIFD dataset, worked on finetuning byT5 and wrote section 2

7 Acknowledgement

We are thankful to Professor Lee and all GSIs' help, especially Anthony Liu who met with us several times to provide guides and insights for this project.

References

[1] S. K. Dwivedi and V. Singh, "Research and reviews in question answering system," *Procedia Technology*, vol. 10, pp. 417–424, 2013.

[2] M. A. C. Soares and F. S. Parreiras, "A literature review on question answering techniques, paradigms and systems," *Journal of King Saud University-Computer and Information Sciences*, vol. 32, no. 6, pp. 635–646, 2020.

[3] B. Ojokoh and E. Adebisi, "A review of question answering systems," *Journal of Web Engineering*, vol. 17, no. 8, pp. 717–758, 2018.

[4] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, "Byt5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022.

[5] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," *arXiv preprint arXiv:1910.11856*, 2019.

[6] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei *et al.*, "The flan collection: Designing data and methods for effective instruction tuning," *arXiv preprint arXiv:2301.13688*, 2023.

[7] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," *arXiv preprint arXiv:2304.03277*, 2023.

[8] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[9] "Googletrans," https://github.com/ssut/py-googletrans.

[10] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021.

Types of Contribution	Yi	Franklin	Chenshun	Xiyu
Report Writing	✓	✓	✓	
Data Pre-processing				✓
Experiments	✓	✓		✓

Table 3: Table of author contribution

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[12] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[13] S. Pudaruth, K. Boodhoo, and L. Goolbudun, "An intelligent question answering system for ict," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*. IEEE, 2016, pp. 2895–2899.

[14] S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf, "Transfer learning in natural language processing," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, 2019, pp. 15–18.

[15] A. Asai, J. Kasai, J. H. Clark, K. Lee, E. Choi, and H. Hajishirzi, "Xor qa: Cross-lingual open-retrieval question answering," *arXiv preprint arXiv:2010.11856*, 2020.

[16] A. Asai, X. Yu, J. Kasai, and H. Hajishirzi, "One question answering model for many languages with cross-lingual dense passage retrieval," *Advances in Neural Information Processing Systems*, vol. 34, pp. 7547–7560, 2021.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[19] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[21] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *arXiv preprint arXiv:2204.02311*, 2022.

[22] Y. Tay, J. Wei, H. W. Chung, V. Q. Tran, D. R. So, S. Shakeri, X. Garcia, H. S. Zheng, J. Rao, A. Chowdhery *et al.*, "Transcending scaling laws with 0.1% extra compute," *arXiv preprint arXiv:2210.11399*, 2022.

[23] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[24] P. Barham, A. Chowdhery, J. Dean, S. Ghemawat, S. Hand, D. Hurt, M. Isard, H. Lim, R. Pang, S. Roy *et al.*, "Pathways: Asynchronous distributed dataflow for ml," *Proceedings of Machine Learning and Systems*, vol. 4, pp. 430–449, 2022.

[25] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[26] O. Shliazhko, A. Fenogenova, M. Tikhonova, V. Mikhailov, A. Kozlova, and T. Shavrina, "mgpt: Few-shot learners go multilingual," *arXiv preprint arXiv:2204.07580*, 2022.

Appendix A Prediction Examples

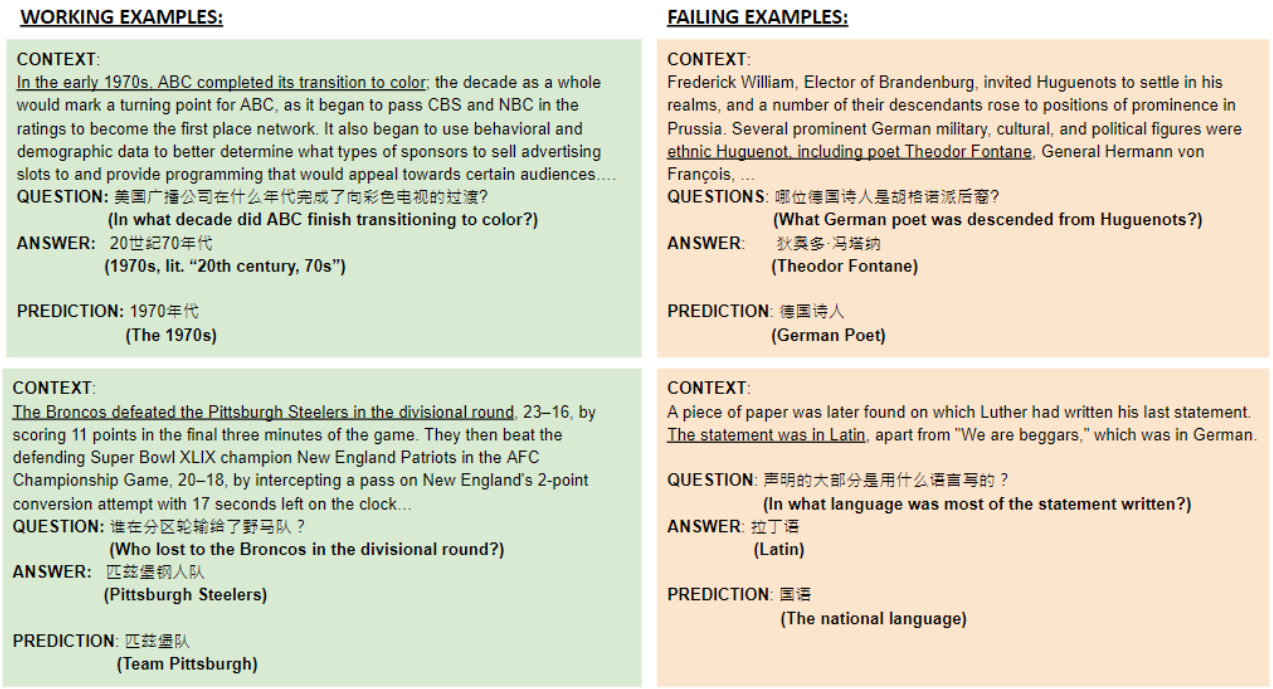

Figure 5: Answer prediction examples from fine-tuned ByT5 on the modified XQuAD.