Auto-regressive Text Generation with Pre-Trained Language Models: An Empirical Study on Question-type Short Text Generation

Anonymous ACL submission

Abstract

This paper presents a multi-way parallel math word problem dataset, which covers English, Tamil and Sinhala. We employ this dataset in an empirical analysis of GPT-2, BART, and T5, as well as mT5 and mBART in auto-regressive text generation. Our findings show that BART and T5 perform noticeably better that GPT-2 for the considered task, and text generation with mBART50 and mT5 provides very promising results even for languages under-represented in these pre-trained models.

1 Introduction

Auto-regressive language models such as GPT-x (Radford et al., 2019) have been commonly used for Natural Language Generation (NLG) tasks such as patent claim generation (Lee and Hsiang, 2020), news generation (Mosallanezhad et al., 2020) and dialogue systems generation (Budzianowski and Vulić, 2019). Sequence-to-sequence (seq-seq) models such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2019) have also been used for NLG in an auto-regressive manner (Tan et al., 2020; Lewis et al., 2020). However, this option has been used to a lesser extent compared to GPT-x in similar text generation tasks. Consequently, no comparative study is available on the performance of these three pre-trained models. Comparative studies between mT5 and mBART for auto-regressive text generation have been limited to high-resource languages (Chen et al., 2021).

We present an empirical study on the effectiveness of GPT-x, BART, and T5 for question-type short text generation for English with respect to parameters such as the seed length and the fine-tuning dataset size. We also evaluate mBART50[1] and mT5 for text generation in the context of low-resource languages. The considered domain is math word problems (MWPs) used in elementary level.

An MWP is a narrative with a specific topic that provides clues to the correct equation with numerical quantities and variables therein (Zhou and Huang, 2019). MWPs can be in categories such as algebra, geometry and statics. Compared to text generation tasks such as story generation (Roemmele, 2016), lyrics generation (Potash et al., 2015) or news generation (Leppänen et al., 2017), MWP generation is challenging because MWPs have mathematical constraints, units and numerical values. Auto-regressive generation of MWPs has been tried out only with RNN models before (Liyanage and Ranathunga, 2020), and template-based MWP generation has been a common option until recently Wang and Su (2016).

We extended the dataset created by Liyanage and Ranathunga (2020) for MWP generation by adding questions with more diversity. Each English question was manually translated to Sinhala and English, creating a multi-way parallel dataset. Chen et al. (2021) also presented a multi-way parallel dataset for story generation. However, they focused only on 4 high-resource languages. Our dataset is released[2], and can be considered as a test set even for Machine Translation.

Our results reveal interesting observations. We show that sequence-to-sequence models significantly outperform auto-regressive GPT-2, for English question-type short text generation. mBART and mT5 also perform on par with their monolingual counterparts for English. Interestingly, performance of mBART and mT5 for the considered low-resource languages (which are underrepresented in mT5 and mBART) outperformed the GPT-2 results for English, highlighting the strong cross-lingual capabilities of the multilingual models. Thus this finding opens a new avenue for auto-regressive short-text generation for low-resource languages.

[1] referred to as mBART hereafter

[2] https://anonymous.4open.science/r/MWP-Dataset

2 Experiments

2.1 Dataset

Liyanage and Ranathunga (2020)'s dataset contains two types of MWPs: simple MWPs and algebraic MWPs. The simple MWP dataset contains 2000 questions and the Algebraic MWP dataset contains 2350 questions. This dataset contains questions in English, Tamil and Sinhala, but is not multi-way parallel. We extended this dataset using the Dolphin18K dataset (Huang et al., 2016) and al-lArith dataset (Roy and Roth, 2016) to add more diversity to the dataset. The extended dataset now contains 4210 Algebraic MWPs and 3160 simple MWPs. Mathematics tutors translated these questions to Sinhala and Tamil. All questions belong to the elementary level. Simple MWP dataset contains simple arithmetic questions. These questions contain constraints such as '*first number is always larger than the second one*'. Algebraic MWPs are more logical and require two or more equations to solve. Example questions corpus stats are given in the Table 14.

2.2 Model Selection

According to Huggingface[3], GPT2-Medium, T5-base and BART-large variants have approximately 300M model parameters. Therefore these were used for further experiments. For multilingual MWP generation, we selected mT5-base and mBART50-large models, to correspond to their monolingual counterparts.

2.3 Experiment Setup

Fine-tuning for the selected models was set-up with 20 epochs, 16-batch size and 1e-4 learning rate. We tested with half of a question and a quarter of a question as the seed. For example, for the question: "*The sum of two numbers is 55. The smaller number is three less than the larger. What are the numbers?*" , the quarter seed is *"The sum of two numbers is 55"*, and the half seed is *"The sum of two numbers is 55. The smaller number is"*

2.4 Baseline

Since Liyanage and Ranathunga (2020) have provided the evaluation results for their dataset, we considered this as our baseline. They used 50-100 characters of a question as the input seed (i.e. more than half of a question). We followed the exact same experiment setup.

We divided our experiments into 4 steps.

1. Baseline experiments for English MWPs by fine-tuning GPT-2 medium, BART-large and T5-base as well as the baseline model.

2. Empirical study on English MWP generation by varying training set size (including zero-shot) and seed length.

3. Comparison of T5 vs mT5 models and BART vs mBART50 for English text generation.

4. Multilingual text generation experiments for Sinhala, Tamil and English by fine-tuning the mT5 and mBART models.

2.5 Evaluation Metrics

Test BLEU (Papineni et al., 2002) and ROUGE (ROUGE-1 and ROUGE-2) (Lin, 2004) scores were used as evaluation metrics.

Especially in the zero-shot generation, BLEU and ROUGE scores direct us to contradictions because they only consider the quality of the generated text. In such scenarios, we need lexical based quality metrics and semantic-based quality metrics (Tan et al., 2020). We used MS-Jaccard metric (Alihosseini et al., 2019) (higher the better), TF-IDF (lower the better) and Fréchet BERT Distance (FBD) (Alihosseini et al., 2019) (lower the better).

The generated MWPs should have correct spelling/grammar and satisfy different Mathematical constraints. A Maths tutor should be able to edit a generated MWP in less time compared to writing a question from scratch. We carried out a human evaluation to validate the quality of the generated questions and their practical usability.

3 Results and Evaluation

3.1 Model Performance for English NLG

We used the same training and testing sizes (train:validation: test 80:10:10) used in the baseline and obtained the English results for both half and quarter input seeds using our models. Results are shown in Table 1 . The results show that all three pre-trained models outperform the baseline and are able to generate quality MWPs. Also, we can conclude that the T5 model is generally better for this task. Table 6 in Appendix shows a sample of generated English MWPs from each model.

[3]https://huggingface.co/transformers/v3.3.1/pretrained_models.html

Table 1: BLEU and ROUGE scores (R1 and R2) for the baseline experiments of English MWPs.

Dataset Type	Model	Seed size	BLEU	R1	R2
Simple	baseline	>Half	22.97	-	-
	FT GPT-2	Quarter	67.00	0.785	0.671
		Half	81.28	0.863	0.798
	FT BART	Quarter	80.93	0.811	0.689
		Half	95.72	0.961	0.926
	FT T5	Quarter	**88.42**	**0.877**	**0.791**
		Half	**97.26**	**0.976**	**0.954**
Algebraic	baseline	>Half	33.53	-	-
	FT GPT-2	Quarter	48.93	0.659	0.489
		Half	59.86	0.799	0.678
	FT BART	Quarter	62.99	0.647	0.460
		Half	76.58	0.784	0.676
	FT T5	Quarter	**72.69**	**0.734**	**0.600**
		Half	**86.12**	**0.870**	**0.816**

3.2 Zero-shot generation for English

In zero-shot generation, we use the pre-trained models and just give the input seed to the model to get the generated output. Table 5 in Appendix shows sample MWPs generated in a zero-shot manner. We see that the generated sentences are not questions but more like stories. This is because these pre-trained models are not specifically trained on a question-type dataset.

Table 2: BLEU (BL), ROUGE(R1 and R2), MS Jaccard(MSJ), TF-IDF distance(TID) and Fréchet BERT Distance (FBD) for zero-shot generation (with quarter seed) of simple and algebraic English MWPs. G- GPT-2, B-BART, T-T5. Seed size: Quarter of the question

Type	M	BL	R1	R2	MSJ	TID	FBD D
ES	G	13.24	0.201	0.132	0.063	160.65	98.07
	B	**49.22**	**0.606**	**0.511**	0.038	118.77	95.55
	T	24.12	0.363	0.149	**0.075**	**90.92**	**78.71**
EA	G	16.44	0.225	0.131	0.060	267.35	107.29
	B	**43.75**	**0.532**	**0.398**	0.046	201.65	93.75
	T	25.00	0.428	0.317	**0.065**	**181.58**	**67.05**

Results for zero-shot generation are shown in Table 2. BLUE results of the BART model are pretty good even if the generated questions are not related to the math domain. This is because, (1)The same words generate repeatedly without any meaning, and (2) Most of the time only a few words were generated. This in fact is a commonly reported problem (Martin et al., 2020). Thus the generated text is evaluated using MS-Jaccard (MSJ), TF-IDF Distance (TID) and Fréchet BERT Distance (FBD).

Evaluation scores suggest that the generated MWPs have low lexical and semantic quality and have low diversity. However, T5 model is a step ahead of the other two models for zero-shot MWP generation for both simple and algebraic cases. Across all three matrices, simple MWP generation achieves better performance gains than algebraic MWP generation because the latter contains more domain-specific words included in the appendix.

3.3 MWP Generation with Different Fine-tuning Dataset Sizes

We conducted comprehensive experiments on our models to analyze how the quality of the results varies with different fine-tuning dataset sizes. We split the dataset within the train:validate:test with a ratio of 80:10:10, 40:10:50 and 20:10:70.

Figure:2 BLEU score variation for Algebraic MWPs (English)

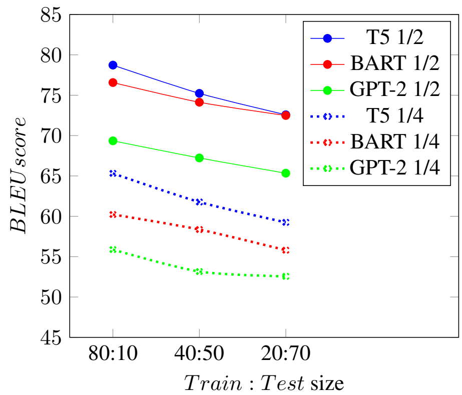
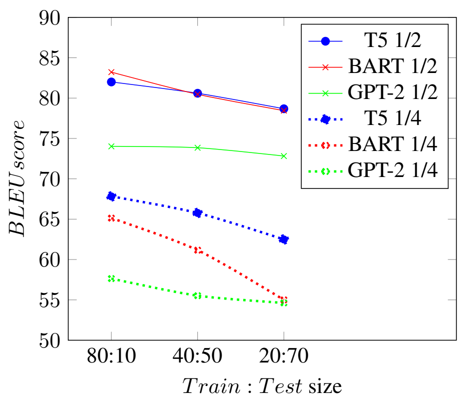


Figure:3 BLEU score variation for Simple MWPs (English)



Figures 2 and 3 show comparative results. Corresponding numerical results are reported in Tables 7-12 in Appendix. The size of the fine-tuning dataset and the seed size affect the output, which of course is not surprising. The former has been a common observation for similar seq-seq tasks (Rothe et al., 2021), and even for other types of pre-trained models (Wu and Dredze, 2020). However, even a small amount (around 600 data points) of fine-tuning dataset is enough for obtaining a sufficient result with the pre-trained models (GPT2: 54.63, BART: 55.00, T5: 62.49). For all but one cases, fine-tuned T5 model has the best result. However, when the amount of fine-tuning data reduces (below 800), the gap between T5 and BART (for ½ seed), and BART and GPT-2 (for 1/4 seed) becomes negligible.

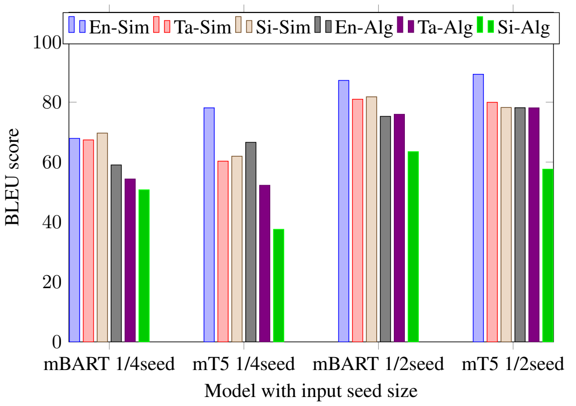
3.4 Mono vs Multilingual Text Generation

The purpose of this experiment is to see whether it is better to have monolingual or multilingual models for English text generation. For this experiment, we fine-tuned the T5 and mT5 models, BART and mBART models with 40:10:50, train:validation: test sets using ½ seed[4]. Results in Table 5 suggest that these multilingual models are capable of providing almost the same results as T5 and BART with a reasonable amount of fine-tuning data.

3.5 Multilingual Text Generation

In this experiment, we fine-tuned mT5 and the mBART models for Sinhala and Tamil with train:validation:test with a ratio of 40:10:50. Also, we fine-tuned the English language for better comparison. Results are in Figure 4, with the numerical results in Table 13 in the Appendix.

Figure:4 Multilingual Simple and Algebraic MWP generation results



On average mBART model shows better results(for Sinhala +7.58 and for Tamil +4.32 BLEU score on average) than the mT5 model for both Sinhala and Tamil languages. However for English, the mT5 model shows better results than mBART. The amount of data in the pre-trained model has shown to have an impact on performance of models s.a. mBERT and XLM-R (Hu et al., 2020). However, we get mixed results wrt this. In mBART, Sinhala is the most under-represented, followed by Tamil (refer to Table 4 in appendix for stats). Although Tamil Algebraic result is better than Sinhala (in both mT5 and mBART), for simple questions both models perform better for Sinhala except in one case. Sinhala and Tamil results slightly outperform English results except for Simple quarter seed of mBART, Algebraic half seed for both mBART and mT5 in 3 of the experiments. This indicates model performance depends on other factors such as the domain of pre-trained and fine-tuned data.

Results of the manual analysis are reported in Tables 16 and 15 in Appendix. For English MWPs, mT5 model takes the smallest time to correct and for Sinhala MWPs, mBART model takes lesser time to correct. Note that all these times are less than what Liyanage and Ranathunga (2020) has reported, who in turn have shown that writing questions from scratch takes considerably more time than text generation from their technique. We identified, subject/object, unit, spelling and grammar as the main possible errors in the generated text (Table 17). However, these errors are usually less than 20% even in the worst performing model.

Table 3: BLEU scores for MWP generation with T5 vs mT5 and BART vs mBART with train/test set sizes

Data	Tr	Te	T5	mT5	BART	mBART
Sim	788	986	**90.54**	89.31	**88.26**	87.25
Alg	939	1175	**80.85**	78.15	**76.32**	75.26

4 Conclusion

This paper made 3 contributions: 1. A multi-way parallel MWP dataset including 2 low-resource languages, 2) a comprehensive analysis of GPT-2, BART and T5 for auto-regressive question-type short text generation and 3) analysis on the performance of mT5 and mBART for text generation with respect to the language representation in the pre-trained model. Our experiments reveal that 1) the multilingual and monolingual seq-seq models are equally capable of short text generation for English, while T5/mT5 is generally better, 2) Even for languages under-represented in the models, results show gains over GPT-2 results reported for English, 3) Model performance generally depends on pre-trained data amounts, but other factors s.a. data domain can have an influence. In future we plan to improve these models in few and zero-shot scenarios.

[4]This dataset contains only 2350 Algebraic MWPs and 1972 Simple MWPs samples as the final set of multilingual dataset was finalized at the last minute. We will update the result upon paper acceptance.

5 Ethical Considerations

We have obtained the permission to republish the baseline (Liyanage and Ranathunga, 2020) datasets. In Dolphin18K dataset (Huang et al., 2016) and allArith dataset (Roy and Roth, 2016), they have not mentioned any restrictions on using the data. We cited their papers as requested in their repos. We paid tutors and other parties for multilingual dataset creation and manual evaluation according to the rates in the country. We verbally explained the purpose of the dataset and the process they have to follow. Annotator information was not collected nor included in the dataset, as this is not relevant ot the task, In the fine-tuning process, we only focused on elementary-level MWPs therefore the fine-tuned language models won't introduce any offensive language.

References

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. 2021. Mtg: A benchmarking suite for multilingual text generation. *arXiv preprint arXiv:2108.07140*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Danqing Huang, Shuming Shi, Chin-Yew Lin, Jian Yin, and Wei-Ying Ma. 2016. How well do computers solve math word problems? large-scale dataset construction and evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 887–896.

Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning openai gpt-2. *World Patent Information*, 62:101983.

Leo Leppänen, Myriam Munezero, Mark Granroth-Wilding, and Hannu Toivonen. 2017. Data-driven news generation for automated journalism. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 188–197.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Vijini Liyanage and Surangika Ranathunga. 2020. Multi-lingual mathematical word problem generation using long short term memory networks with enhanced input features. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4709–4716.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Ahmadreza Mosallanezhad, Kai Shu, and Huan Liu. 2020. Topic-preserving synthetic news generation: An adversarial deep reinforcement learning approach. *arXiv preprint arXiv:2010.16324*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Melissa Roemmele. 2016. Writing stories with help from recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 140–145.

Subhro Roy and Dan Roth. 2016. Unit dependency graph and its application to arithmetic word problem solving. *arXiv preprint arXiv:1612.00969*.

Bowen Tan, Zichao Yang, Maruan AI-Shedivat, Eric P Xing, and Zhiting Hu. 2020. Progressive generation of long text with pretrained language models. *arXiv preprint arXiv:2006.15720*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Ke Wang and Zhendong Su. 2016. Dimensionally guided synthesis of mathematical word problems. In *IJCAI*, pages 2661–2668.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Qingyu Zhou and Danqing Huang. 2019. Towards generating math word problems from equations and topics. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 494–503.

A Appendix

A.1 Example MWPs

Simple MWP: *Pala used 90 kilograms of cement and 125 kilograms of sand for the house. How much more sand did Pala use than the cement?*

Here relevant units (i.e kilograms for cement and sand) and appropriate combinations (i.e cement and sand for building a house) should be matched.
Algebraic MWP: *Find two numbers whose sum is 53 and whose difference is 27, what is the larger number, What is the smaller number.*

A.2 Language Data Statistics of the Pre-trained Models

Language data statistics reported in Table 4 are from (Xue et al., 2020) (Tang et al., 2020)

Table 4: Language Data Statistics of the Pre-trained Models

Model		Pre-trained Dataset		
		English	Sinhala	Tamil
mT5	Token(B)	2,733	0.8	3.4
	Pages(M)	3,067	0.5	3.5
mBART	Token(B)	55.61	0.243	0.595
	GiB	300.8	3.6	12.2

Table 5: Sample Zero shot Generation results

Model	Generated MWPs
GPT2	The difference between a "first," and an ordinary, job is that the former often requires significant skills.What's next? Well. . . not much really right now though!
BART	The... The difference between the two
T5	The difference between the two is that the difference between the two is the difference between the

Table 6: Sample English MWPs generated using the baseline and the fine-tuned models. Seed size: Quarter of the question

Model	Generated MWPs
Baseline	the sum of two numbers is 12. their differenct are the two consecutive integers if the sum of the second integers is 10.
Fine-tuned GPT2	The sum of two numbers is 76, the second is 8 more than 3 times first, what are these 2 numbers?
Fine-tuned BART	The sum of two numbers is 60. three times the smaller number minus twice the larger number is 56. Find the larger number.
Fine-tuned T5	The sum of two numbers is 91. the larger number is 1 more than 4 times the smaller number. Find the numbers.

.

Table 7: BLEU and ROUGE scores for the train:test 80:10 of simple English MWPs.

Model	Seed size - Half of the question			Seed size - Quarter of the question		
	BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Fine-Tuned GPT2	74.02	0.802	0.684	57.64	0.648	0.462
Fine-Tuned BART	83.22	0.852	0.782	65.13	0.668	0.514
Fine-Tuned T5	82.00	0.872	0.808	67.82	0.721	0.588

Table 8: BLEU and ROUGE scores for the train:test 80:10 of algebraic English MWPs.

Model	Seed size - Half of the question			Seed size - Quarter of the question		
	BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Fine-Tuned GPT2	69.35	0.735	0.658	55.87	0.610	0.446
Fine-Tuned BART	76.57	0.777	0.667	60.22	0.618	0.424
Fine-Tuned T5	78.73	0.825	0.741	65.32	0.669	0.507

Table 9: BLEU and ROUGE scores for the train:test 40:50 of simple English MWPs.

Model	Seed size - Half of the question			Seed size - Quarter of the question		
	BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Fine-Tuned GPT2	73.85	0.784	0.612	55.52	0.621	0.447
Fine-Tuned BART	80.47	0.829	0.740	61.16	0.638	0.475
Fine-Tuned T5	80.60	0.858	0.786	65.77	0.698	0.557

Table 10: BLEU and ROUGE scores for the train:test 40:50 of algebraic English MWPs.

Model	Seed size - Half of the question			Seed size - Quarter of the question		
	BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Fine-Tuned GPT2	67.23	0.718	0.597	53.15	0.616	0.431
Fine-Tuned BART	74.13	0.757	0.635	58.36	0.601	0.408
Fine-Tuned T5	75.23	0.799	0.699	61.77	0.650	0.479

Table 11: BLEU and ROUGE scores for the train:test 20:70 of algebraic English MWPs.

Model	Seed size - Half of the question			Seed size - Quarter of the question		
	BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Fine-Tuned GPT2	65.34	0.667	0.547	52.56	0.596	0.503
Fine-Tuned BART	72.49	0.743	0.616	55.80	0.583	0.390
Fine-Tuned T5	72.58	0.774	0.664	59.25	0.635	0.634

Table 12: BLEU and ROUGE scores for the train:test 20:70 of simple English MWPs.

Model	Seed size - Half of the question			Seed size - Quarter of the question		
	BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Fine-Tuned GPT2	72.82	0.767	0.591	54.63	0.605	0.436
Fine-Tuned BART	78.45	0.819	0.725	55.00	0.592	0.418
Fine-Tuned T5	78.68	0.845	0.766	62.49	0.674	0.526

Table 13: BLEU and ROUGE scores for multilingual text generation experiments of simple and algebraic MWPs.

Dataset	Seed size	mT5-base			mBART-large		
		BLEU	ROUGE-1	ROUGE-2	BLEU	ROUGE-1	ROUGE-2
Sinhala Simple	seed ¼	61.96	0.632	0.467	**69.68**	0.684	0.529
	seed ½	78.22	0.798	0.687	**81.79**	0.819	0.719
Sinhala Algebra	seed ¼	37.56	0.469	0.278	**50.75**	0.547	0.365
	seed ½	57.64	0.639	0.480	**63.48**	0.688	0.529
Tamil Simple	seed ¼	60.32	0.632	0.446	**67.39**	0.684	0.505
	seed ½	79.95	0.808	0.681	**80.98**	0.829	0.714
Tamil Algebraic	seed ¼	52.26	0.556	0.380	**54.36**	0.572	0.417
	seed ½	68.85	0.725	0.620	**75.95**	0.782	0.698
English Simple	seed ¼	**78.10**	0.782	0.657	67.90	0.693	0.531
	seed ½	**89.31**	0.903	0.847	87.25	0.888	0.825
English Algebraic	seed ¼	**66.59**	0.677	0.515	59.04	0.608	0.420
	seed ½	**78.15**	0.815	0.727	75.26	0.763	0.645

Figure 1: Sample Generated Math Word Problems

T5 English Generated Examples -

- Quarter seed - Twice the larger number is 3 more than
- Generated questions- Twice the larger number is 3 more than five times the smaller number. The sum of the 2 numbers is 27. What are the two numbers?

- Half seed - Twice the larger number is 3 more than five times the smaller number, the sum
- Generated questions- Twice the larger number is 3 more than five times the smaller number, the sum of the 2 numbers is 7 less than 3 times the larger number, Find the smaller number.

mBART Sinhala Generated Examples -

- Quarter seed - 150 ක
- Generated questions- 150 ක එකතුවක් ඇති අඛණ්ඩ සංඛ්‍යා තුනක් සොයා ගන්න.

- Half seed - 150 ක එකතුවක් ඇති අඛණ්ඩ
- Generated questions - 150 ක එකතුවක් ඇති අඛණ්ඩ ඉරට්ටේ සංඛ්‍යා දෙක කුමක්ද?

mBART Tamil Generated Examples -

- Quarter seed - சேஹனும் சரித்தும் ஒரு தொழிற்சாலை
- Generated questions - சேஹனும் சரித்தும் ஒரு தொழிற்சாலை வைத்திருக்கிறார்கள். செஹான் 600 பங்குகளையும், நிமல் 1600 பங்குகளையும் வாங்கினர். நிமலை விட கமல் எத்தனை குறைவான பங்குகளை வாங்கினார்?

- Half seed - சேஹனும் சரித்தும் ஒரு தொழிற்சாலை வைத்திருக்கிறார்கள். செஹான் 800, நிமல் 2000
- Generated question - சேஹனும் சரித்தும் ஒரு தொழிற்சாலை வைத்திருக்கிறார்கள். செஹான் 60, நிமல் 20 பங்குகளை வாங்கினார். நிமலை விட கமல் எத்தனை குறைவான பங்குகளை வாங்கினார்?

Table 14: Dataset Statistics

Dataset type	Number of Questions	Minimum Number of sentences per question	Average Number of words per question	Average Number of characters per question
English Simple	3160	2	15	54
English Algebraic	4210	2	14	62
Sinhala Simple	3160	2	19	61
Sinhala Algebraic	4210	2	17	59
Tamil Simple	3160	2	13	49
Tamil Algebraic	4210	2	16	57

Table 15: Human evaluation results for Simple MWPs in minutes. SE: SimpleEnglish, SS: Simple Sinhala. TTE: Time to Edit 10 generated MWPs, TTG: Time To Generate 10 MWPs

	Baseline				mBART		mT5	
	TTG		TTE		TTE		TTE	
	SE	SS	SE	SS	SE	SS	SE	SS
Tutor 1	18	15	2	2.5	0.5	0.38	0.66	0.66
Tutor 2	20	25	2.2	3	0.75	0.45	0.48	0.58
Tutor 3	15	17.5	1	1.5	0.55	0.38	0.71	0.51
Tutor 4	15	28	2.5	1	0.6	0.83	0.6	0.75
Tutor 5	21	26.5	3	2	0.63	0.91	0.45	0.6
Average	**17.8**	**22.4**	**2.14**	**2**	**0.60**	**0.59**	**0.58**	**0.62**

Table 16: Human evaluation results for Algebraic MWPs in minutes AE: Algebraic English, AS: Algebraic Sinhala, (Time taken to Edit 10 generated MWPs)

	mBART		mT5	
	AE	AS	AE	AS
Tutor 1	2	0.66	1.16	2
Tutor 2	0.73	0.65	0.58	0.73
Tutor 3	0.42	0.75	0.83	0.78
Tutor 4	0.9	0.88	1.26	1.41
Tutor 5	1.25	1.08	0.91	0.95
Average	**1.06**	**0.80**	**0.95**	**1.17**

Table 17: Different types of error percentages found in simple MWPs

Errors%	mBART				mT5			
	SE	AE	SS	AS	SE	AE	SS	AS
Subject %	4	4	6	4	8	2	6	2
Unit %	4	1	1	1	2	1	1	1
Spelling %	0	0	4	2	2	0	0	2
Grammar %	16	12	16	10	8	10	14	10