Question Generation With Different Datasets, Training Strategies, and Evaluation Metrics

Bronte Baer **Jean-Luc Jackson** **Richard Robbins**
UC Berkeley School of Information
MIDS 266: Natural Language Processing with Deep Learning
{bronte.baer, jlj, richard.robbins}@berkeley.edu

December 3, 2022

Abstract

We explore the ability of transformer-based encoder-decoder models to produce a question when given context and an answer, and the impact of using different datasets and training strategies. We built models from four source datasets and evaluated each with validation data from each source, without regard to whether the model being tested was trained on that source. We also built models by blending the datasets in different ways. We used traditional lexical similarity-focused evaluation metrics and more current semantic similarity-focused metrics to assess which produce scores more reflective of successful question generation.

Our models can produce fluent and high-quality questions. We discuss how we measure success and the degree to which our models succeed. Our best models were trained on a randomly shuffled blended dataset. Metrics that account for semantic similarity produce scores more reflective of successful question generation than those based on lexical similarity. Lexical metrics can be used together with semantic metrics for a more comprehensive evaluation that is especially well-suited to identifying higher-quality questions.

1 Introduction

Machine reading comprehension, question answering, and question generation are vibrant research and commercial domains that are well-studied yet still evolving. Machine reading comprehension requires a machine to take input text (context) and derive meaning. The question answering task requires a machine to produce an answer (or sometimes recognize that it cannot answer) when given context and a question based on that context (Liu et al., 2019). Question generation, on the other hand, requires a machine to produce a question based on context and an answer (Pan et al., 2019). We chose question generation because it is a comparatively novel task that allowed us to explore text generation and evaluation.

We had three primary goals:

1. Build successful question generation models.
2. Identify evaluation metrics that are well suited to question generation.
3. Explore a model's performance on datasets it had not been trained on and the performance of models built on blended datasets.

2 Background

Several papers have addressed question generation with attention-based or transformer architectures, often working from the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) and using lexical similarity metrics such BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) for evaluation.

Du and colleagues introduced a trainable sequence-to-sequence learning model with an RNN encoder-decoder architecture and an attention mechanism. Their approach to question generation stands in contrast to previously used rule-based systems and over-generate and rank approaches (Du et al., 2017). Chan and Fan investigated the use of BERT (Devlin et al., 2018) for question generation (Chan and Fan, 2019). More recent work in the question generation space has focused on using T5 (Rajapakse, 2020) and other transformer-based architectures, such as BART (Cao and Wang, 2021).

3 Methods

We began by fine-tuning T5 (Raffel et al., 2019) for question generation using SQuAD.[1] We built three more models by training on the Natural Questions (NQ) (Kwiatkowski et al., 2019), Question Answering in Context (QuAC) (Choi et al., 2018), and TriviaQA (Joshi et al., 2017) datasets.[2] We built a fifth model by blending the four source datasets and randomly shuffling the data. We repeated the exercise replacing T5 with BART (Lewis et al., 2019). We tested each of the ten models against the validation data from the four source datasets yielding forty prediction sets that form the basis for the bulk of this report. See Appendix A.

[1] We used SQuAD instead of SQuAD 2 (Rajpurkar et al., 2018) because SQuAD 2 introduced questions without answers. Our task requires the derivation of questions *from* answers, so questions without answers are not useful to us.

[2] We used Sen and Saffari's QA dataset converter to convert datasets into SQuAD format (Sen and Saffari, 2020).

We trained T5 and BART models on the source datasets sequentially to explore the impact of how the blended dataset was assembled. In each case, we took a model trained on SQuAD, then fine-tuned that model on NQ, then QuAC, and finally TriviaQA. We discuss these models in Section 5.4.2 separately from the models referenced in the preceding paragraph.

4 Model Architectures, Datasets, and Evaluation Metrics

There are many model architectures, datasets, and evaluation metrics we might have selected. We discuss our choices and what motivated them in this section.

4.1 Model Architectures

We chose T5 version 1.1 and BART for our models because each is an encoder-decoder architecture well-suited to sequence-to-sequence tasks like question generation. Neither comes from its source trained on question generation.

We used T5 as a baseline because it is well-known for being particularly amenable to training on new tasks, with ample guidance included by the authors (Raffel et al., 2019). Unlike the original version of T5, version 1.1 must be fine-tuned before use on any downstream task.[3]

We were curious to see how BART would compare to T5, both for ease of fine-tuning and performance. Its design suggested to us that, like T5, it would work well on question generation (Lewis et al., 2019).

See Appendix B for additional details regarding model design and the length of context and target sequences as well as text generation hyperparameters.

4.2 Datasets

Our datasets were created for question answering and not question generation. We are not aware of any datasets created specifically for question generation.

Each sample in our training and validation data includes one target question for each context and answer pair. Since we work from answers to questions, we discard any questions identified as unanswerable. Table 1 shows the number of training and validation samples that remain in each of our datasets after we eliminate questions without answers.

	NQ	QuAC	SQuAD	TriviaQA
training	74,218	69,109	87,599	77,085
validation	2,356	5,868	10,570	9,835

Table 1: Dataset Sample Counts

SQuAD consists of questions posed by crowd workers on a set of Wikipedia articles, where the answer to each question is a segment of text from the corresponding reading passage. It is almost universally used for question answering research. We chose it as our first dataset for that reason (Rajpurkar et al., 2016).

NQ consists of aggregated queries issued to the Google search engine. An annotator is presented with a question and a Wikipedia page from the top 5 search results. The annotator chooses long and short answers or marks the question as not being answerable. It was created to study the impact of annotation methodologies, which is not the focus of this project. Nevertheless, the dataset is a good complement to SQuAD. We included it for that reason (Kwiatkowski et al., 2019).

TriviaQA consists of question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents, six per question on average. The authors contend that, in comparison to other datasets, TriviaQA has compositionally complex questions, considerable syntactic and lexical variability between questions and corresponding answers, and requires more cross-sentence reasoning to find answers. We included TriviaQA primarily because it was constructed to be more challenging than SQuAD (Joshi et al., 2017).

QuAC stands in stark contrast to the other datasets. Its prompts are taken from dialogues and contain questions that may only be meaningful within the dialogue context. We included QuAC without attempting to adjust for the dialogue context effect. We did so to see how our models would perform on the unadjusted dataset. It is an outlier among our datasets and one that we see as a useful model stress test (Choi et al., 2018).

Blended Data is what we used to explore the impact of training models on more than one dataset at a time. As a general rule, neural networks improve when exposed to more data. We believed that the same would be true for our question generation models. That is why we introduced the blended dataset described in Section 3.

We also believed that shuffling the data in the blended dataset would improve model performance by introducing more variation into the training process and avoiding potential local minima reflective of consistent patterns found in the source datasets.

4.3 Metrics

We chose evaluation metrics currently popular in evaluating sequence-to-sequence text generation tasks. These metrics evaluate the alignment of textual features such as word choice, word order, or semantics between reference and candidate sequences. We selected metrics that measure similarity in different ways to explore which features contribute most to faithfulness, fluidity, and coherence.

Our metrics fall into two categories: measures of lexical similarity and measures of semantic similarity. We use BLEU, ROUGE-L, and METEOR to evaluate the lexical similarity between pairs of generated and reference questions. We use BERTScore and USE to

[3] https://huggingface.co/google/t5-v1_1-base.

evaluate semantic similarity.

BLEU is a precision-based measure of overlapping n-grams, including n-grams up to an order of four. It is used to indicate how many words in common there are between pairs of sequences. We include it for its historical importance and as a simple measure of matching words (Papineni et al., 2002).

ROUGE-L is a measure of longest common subsequence between a pair of sentences. Of the ROUGE variants, we use ROUGE-L because it automatically considers longest subsequences without needing a pre-defined n-gram input and naturally accounts for sentence-level structure. Similar to BLEU, we included ROUGE-L for its historical importance and simplicity (Lin, 2004).

METEOR is a recall-oriented measure of unigram matching that builds upon BLEU's formulation to improve evaluation. METEOR adds matching heuristics so that word stems and synonyms are included in matches. Some have argued that, in the family of simple metrics, it correlates the best with human judgement (Denkowski and Lavie, 2014).

BERTScore evaluates semantic equivalence by computing the similarity of two sentences as the sum of the cosine similarities of their tokens' BERT contextual embeddings.[4] BERTScore was designed to be task-agnostic and correlates well with human evaluation. We include it for its robustness and its use of a popular pre-trained transformer language model for machine comprehension (Zhang et al., 2019).

USE (Universal Sentence Encoder) encodes sentences into embedding vectors using a pre-trained encoder. Its authors demonstrate that the cosine similarity of USE embeddings can be used to evaluate semantic similarity between sentences. We include it for its creation of sentence-level embeddings using a non-transformer architecture[5] (Cer et al., 2018).

5 Results and Discussion

We conducted extensive experiments and have four primary observations.

1. Both T5 and BART can be trained for question generation.

2. We did not find meaningful consistent differences between T5 and BART when fine-tuning for question generation.

3. Metrics that account for semantic similarity produce scores more reflective of successful question generation than those based on lexical similarity. Lexical metrics can be used together with semantic metrics for a more comprehensive evaluation that is especially well-suited to identifying higher-quality questions.

4. A model's performance is impacted by the dataset it is trained on. When looking at predictions made in the aggregate across all datasets, models trained on a blended dataset outperformed those trained on a single dataset. Models trained on shuffled blended data outperformed models trained on the source datasets sequentially.

5.1 Question Generation Success

Our models can produce fluent and high-quality questions. Table 2 shows the performance scores for our best and baseline models (each of which is based on T5). The scores reflect each model's aggregate performance across all of our datasets.[6]

Training Dataset	BLEU	ROUGE-L	METEOR	BERTScore	USE
Shuffled Blended	10.05	39.17	38.43	83.29	58.57
SQuAD	6.45	33.60	32.10	81.20	52.81

Table 2: Baseline and Best Model Results

At times, our models match the target precisely (*i.e*, they generate questions with the correct words in the correct order and without extraneous text). In these cases, we effectively have perfect scores on all metrics. While this illustrates our success with question generation, we were especially interested in exploring situations where our models produced high-quality predictions using different words than in the target. Table 4 includes examples. We were also very interested in results with high semantic metric scores but low lexical metric scores and *vice versa*. We discuss these topics in Section 5.3.3.

Finally, we gave our models examples from outside our source datasets. See Appendices C, D and E for examples and related discussion. The models do not always succeed. For example, as we describe in Appendix C, when working with "Mary Had a Little Lamb" some of our models tended to confuse Mary from the story with the biblical character.

5.2 Comparing T5 and BART for Question Generation

As described in Section 3, we created ten models by fine-tuning T5 and BART on the four source datasets, as well as a blended dataset. We ran each of those ten models on the validation data for each source dataset and grouped the results by model. After considering only those instances where differences in the results between T5 and BART models were statistically significant, we concluded that on balance, the two architectures were

[4]We chose the `distilbert-base-uncased` checkpoint for its manageable file size and recommendation by Hugging Face (see https://huggingface.co/spaces/evaluate-metric/bertscore). Evaluation quality would be expected to improve with larger model checkpoints.

[5]Of the two encoding models introduced in this paper, we use the Deep Averaging Network model published on TensorFlow Hub at https://tfhub.dev/google/universal-sentence-encoder/4.

[6]We did not attempt to optimize results from any of our models. As we note in Section 6, performance optimization remains a topic for future study.

indistinguishable for our project, with a slight edge favoring T5 or BART depending on the circumstances. See Appendix F for additional detail.

5.3 Metrics Evaluation

Selecting a metric for question generation evaluation is challenging because metric scores often contrast with human judgment (Nema and Khapra, 2018). We considered each of the metrics introduced in Section 4.3 in an effort to identify our preferred metrics.

5.3.1 Assessing Lexical Similarity Metrics

We evaluated each of the lexical similarity metrics referenced in Section 4.3 using the questions generated by our models against our test data. We examined statistical trends and specific examples. We found that predictions that scored well with METEOR better aligned with our expectations of well-generated questions compared to BLEU and ROUGE-L. METEOR was particularly robust against variations in word choice and order. Consequently, we believe that for our task, METEOR is the best of the lexical similarity metrics we considered. See Appendix G for additional detail.

5.3.2 Assessing Semantic Similarity Metrics

Our evaluation metrics that measure semantic similarity are BERTScore and USE. We examined target-prediction pairs with varying combinations of score ranges to determine the accuracy and usefulness of each. We found higher scores for each of BERTScore and USE to reflect higher quality results. While each is associated with high-quality questions, the best questions scored well with both. A high score in one metric but not the other often indicated incomplete semantic alignment. So, instead of adopting one or the other as the best semantic similarity metric for question generation, we think it best to use both. See Appendix G for additional detail.

5.3.3 Considering Lexical and Semantic Scores Together

As mentioned in Section 5.1, we were interested in studying examples of questions generated with high semantic metric scores but low lexical metric scores and *vice versa*.

We start with the case of predictions with high semantic metric scores but low lexical metric scores. We found many of the questions generated had slightly different words than the target, *but the essence was the same*. Therefore, we believe semantic metrics are more robust indicators of success at question generation than lexical metrics. See Table 3 and, for additional detail, Appendix H.

Next, we consider the case of predictions with high lexical metric scores but low semantic metric scores. For these, even though the coincidental use of words may result in a strong lexical metric score, the predicted questions typically *do not adequately capture the essence* of the target question. See Appendix H, Table 17 for additional detail.

Finally, we consider the case of predictions with high METEOR, BERTScore and USE scores. These were generally of the highest quality, combining the robustness of semantic scores with the lexical faithfulness of METEOR. See Table 4 and, for additional detail Appendix G.

Target	Prediction
who is beauty and the beast written by	Who wrote La Belle et la Bête?
what is the first basic process in the light reaction of photosynthesis	How does photosynthesis begin?
What was the nationality of composer Frederic Chopin?	where did frédéric chopin come from
How large are Cytoplasmic ribosomes?	what is the size of a chloroplast ribosome
From which common English tree are cricket stumps traditionally made?	what kind of wood is used for stumps in cricket

Table 3: High Semantic / Low Lexical Scores

Target	Prediction
What parts of plants have chloroplasts?	Chloroplasts are found in which part of a plant?
what was the name of the other HD channel Virgin media could carry in the future?	What HD channel did Virgin Media have an option to carry in the future?
Approximately how many adherents does the United Methodist Church have across the world?	How many members are there in the United Methodist Church?
Singer Dan Reynolds is the frontman for which US rock band?	Dan Reynolds is the lead singer with which band?
Carbon dioxide in solid form is called what?	What is the frozen solid form of carbon dioxide known as?
Who's home runs record did Mark McGwire break in the 1998 season?	In 1998, Big Mac McGwire broke the Major League Baseball home run record by hitting 70 home runs. Who did he beat?

Table 4: High Lexical / High Semantic Scores

5.4 Training Dataset Impact on Model Performance

We explored several hypotheses to better understand the impact training datasets have on our models' performance.

5.4.1 Training and Testing on Different Datasets and the Impact of Signature Dataset Features

Our first hypothesis was that the best-performing model for a dataset would be a model trained on that dataset. See Appendix A for a table showing model performance on each source dataset. Our findings are consistent with that hypothesis. This is not surprising as each dataset has distinctive attributes, such as a predominant context length and characteristic syntactic and lexical structure. See Section 4.2. Some datasets require more cross-sentence reasoning to find answers, and the span between the pieces of relevant text varies.

As we note in Section 4.2, QuAC is very different from the other datasets. Its prompts are taken from dialogues and contain questions that may only be meaningful within the dialogue context. We did not adjust

for the additional dialogue. We expected that not adjusting would deprive QuAC-trained models of important information and degrade their performance.

As we expected, the best results for QuAC came from a model trained on QuAC. However, those results are still low compared to the best performance on other datasets. Moreover, our models struggled with the QuAC dataset in general. See Appendices A and I.

5.4.2 Blended Data

We expected that training our models on more than one dataset at a time would enhance performance. That turned out to be the case.

When we grouped predictions by model and ranked them for each performance metric, the top two models are always the models trained on the shuffled blended dataset followed by the SQuAD-trained model. See Appendix I, Table 18.

We also believed that shuffling the data in the blended dataset would improve model performance by introducing more variation into the training process and avoiding potential local minima reflective of patterns or tendencies found in a source dataset. To test our hypothesis, we trained BART and T5 models on the four source datasets in a sequential fashion. In each case, we took a model that was trained on SQuAD, then fine-tuned that model on NQ, then QuAC, and finally TriviaQA. The sequentially trained models did not perform as well as the models trained on the shuffled blended dataset. This is consistent with our hypothesis. See Table 5 and Figure 1.[7]

Training Dataset	METEOR	BERTScore	USE
Shuffled Blended	38.43	83.29	58.57
Sequential Blended	33.56	80.80	53.54

Table 5: T5 Blended Model Ranked Performance

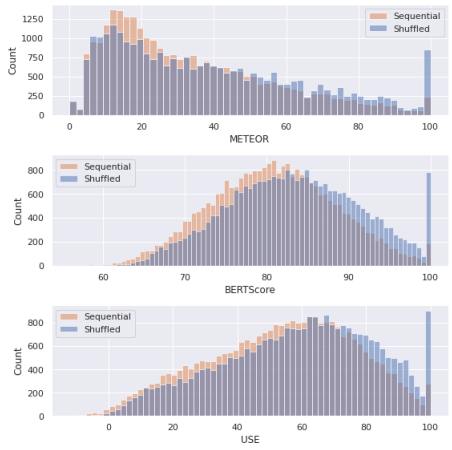

Figure 1: T5 Blended Model Performance Histograms

[7]The results with the BART sequentially trained model were not meaningfully different than the T5 results, so we only show the T5 results.

5.4.3 Context Length

We expected that context length would impact model performance and that longer context sequences would tend to confound models more than shorter context sequences. Our models perform better on SQuAD and NQ compared to TriviaQA. TriviaQA has substantially longer context sequences than either. See Appendix B, Figure 2.

The TriviaQA authors claim their dataset is more challenging because it has compositionally complex questions, considerable syntactic and lexical variability between questions and corresponding answers, and requires more cross-sentence reasoning to find answers (Joshi et al., 2017). We wondered if our models' relative underperformance on TriviaQA stemmed primarily from the length of its context sequences or those other factors.

To test this, we partitioned the TriviaQA validation set on input sequence length and created test data comprised of the shortest samples. We ran our models on the "short sample" test data. We reviewed the results for the four models that had the strongest performance on the full TriviaQA validation set. See Appendix J.

We found statistically significant improvement in METEOR and BERTScore. However, we did not observe a significant change for USE. These results supported our hypothesis that context length impacts question generation model performance and is at least part of the challenge posed by TriviaQA. However, even with the improved scores, the best results on TriviaQA still fall below the equivalent SQuAD and NQ results. This finding supports the contention that TriviaQA is inherently a more complex dataset and that only part of that complexity is attributable to context length.

6 Conclusion

Our models can produce fluent and high-quality questions. The differences between T5 and BART were, on balance, not meaningful. Our models perform better with shorter context sequences. We built our best-performing models by training on a randomly shuffled blended dataset. Metrics that account for semantic similarity produce scores more reflective of successful question generation than those based on lexical similarity. Lexical metrics can be used together with semantic metrics for a more comprehensive evaluation that is especially well-suited to identifying higher-quality questions.

We make several observations about datasets and training strategies concerning question generation. Future research could focus on model improvement by hyperparameter tuning and training for more than a single epoch. Additional datasets would allow for further fine-tuning and investigation of more sequential training combinations. Our datasets were constructed around the task of question answering. Therefore, we see an opportunity to create a dataset specific to question generation.

References

Associated-Press. 2022. Bears notch stirring victory in 125th big game.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with A new question type ontology. *CoRR*, abs/2107.00152.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Ying-Hong Chan and Yao-Chung Fan. 2019. BERT for question generation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 173–177, Tokyo, Japan. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. *CoRR*, abs/1705.00106.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*, abs/1705.03551.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9:3698.

Preksha Nema and Mitesh M. Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, Brussels, Belgium. Association for Computational Linguistics.

Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. Recent advances in neural question generation. *CoRR*, abs/1905.08949.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. *CoRR*, abs/1910.10683.

Thilina Rajapakse. 2020. Asking the right questions: Training a t5 transformer model on a new task.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2429–2438, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Validation Data Performance Comparison

Table 6 shows model results by test dataset, sorted by USE scores.

Test Dataset	Base Model	Training Dataset	BLEU	ROUGE-L	METEOR	BERTScore	USE
NQ	BART	NQ	20.31	55.04	52.88	86.74	73.67
	BART	Shuffled Blended	19.67	54.50	52.35	86.43	72.93
	T5	Shuffled Blended	16.97	52.78	49.79	85.80	71.99
	T5	NQ	17.30	52.77	49.73	85.95	71.89
	T5	SQuAD	1.20	41.54	35.78	79.97	63.39
	BART	SQuAD	1.18	41.06	35.53	79.86	63.03
	BART	TriviaQA	1.14	36.58	36.76	79.56	62.30
	T5	TriviaQA	0.84	30.90	31.92	78.13	57.25
	BART	QuAC	0.50	32.61	22.44	75.05	51.66
	T5	QuAC	0.00	6.85	4.07	66.28	18.07
QuAC	BART	QuAC	5.72	28.27	28.81	79.28	41.88
	T5	Shuffled Blended	5.37	24.90	26.24	77.81	37.80
	BART	Shuffled Blended	2.74	23.37	25.02	77.05	36.38
	T5	SQuAD	0.71	18.52	20.80	75.37	32.05
	BART	SQuAD	0.71	18.42	20.59	75.40	31.85
	T5	QuAC	2.35	17.35	18.99	75.60	30.31
	T5	NQ	0.15	15.19	11.16	69.69	28.77
	BART	NQ	0.11	14.96	11.26	69.72	28.66
	T5	TriviaQA	0.17	11.06	16.06	71.56	24.46
	BART	TriviaQA	0.15	11.36	16.30	70.93	24.32
SQuAD	BART	SQuAD	15.12	46.92	47.53	86.95	65.70
	T5	Shuffled Blended	14.29	47.08	46.98	86.70	65.64
	T5	SQuAD	14.73	46.77	46.77	86.88	65.07
	BART	Shuffled Blended	13.58	45.86	46.01	86.28	65.05
	T5	NQ	1.74	34.62	27.35	79.17	55.47
	BART	NQ	1.44	32.26	25.20	78.38	54.07
	BART	TriviaQA	4.70	31.06	33.79	81.66	53.51
	T5	TriviaQA	4.26	28.06	32.80	80.89	49.98
	BART	QuAC	2.77	29.09	24.56	80.29	44.56
	T5	QuAC	0.04	9.67	9.76	71.42	16.62
TriviaQA	BART	Shuffled Blended	7.80	36.87	35.46	82.64	60.92
	BART	TriviaQA	7.60	36.78	35.10	82.59	60.78
	T5	Shuffled Blended	6.64	35.91	33.79	82.28	60.16
	T5	TriviaQA	5.58	33.70	32.31	81.63	59.14
	BART	NQ	0.80	28.64	20.08	77.08	53.43
	T5	NQ	0.75	27.84	18.98	76.66	51.84
	T5	SQuAD	2.24	26.55	22.19	78.88	49.50
	BART	SQuAD	2.19	26.03	21.91	78.81	49.21
	BART	QuAC	0.69	19.78	15.09	75.55	39.68
	T5	QuAC	0.00	4.32	5.67	67.62	11.92

Table 6: Model Results By Test Dataset

B Model Design

We reviewed whether our models would accommodate the input sequences presented or if we would need to split or otherwise reduce lengthy input sequences. We also reviewed the length of target sequences to ensure that our models were capable of producing outputs at least as long as the target sequences.

We built our models with a maximum token length of 1024 to handle the overwhelming majority of the input sequences in the datasets. When generating predictions we set the maximum length of the generated tokens to be 50 (suitable in view of the bulk of target responses), the no repeat ngram size to be 3, and utilized a 4 beam strategy. See Figure 2.

In our initial work, we failed to recognize that the models made available on Hugging Face do not have uniform defaults for text generation hyperparameters. Before we made the text generation hyperparameters consistent we saw that BART consistently outperformed T5. When we applied consistent text generation hyperparameters, those differences dissipated.

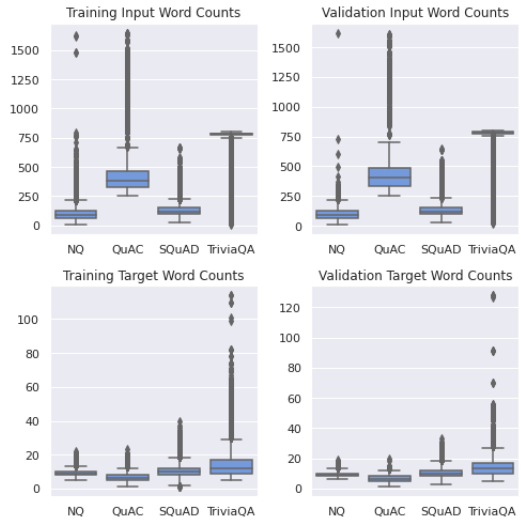

Figure 2: Input and Target Sequence Word Counts

C Mary Had a Little Lamb and Lesser Known Stories

"Mary Had a Little Lamb" is a very simple story, one that is well-suited for preliminary testing of our models. We start with the children's classic and then make several small changes to the story to see how our models react. We limit our review to our four highest-performing models.

"**Mary had a little lamb, its fleece was white as snow. Everywhere the child went, the little lamb was sure to go.**"

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	Mary	who had a little lamb that was white as snow
T5	SQuAD	Mary	Who had a little lamb?
BART	Shuffled Blended	Mary	who had a little lamb in the bible
BART	SQuAD	Mary	Who had a little lamb?
T5	Shuffled Blended	little	what was the size of mary's lamb
T5	SQuAD	little	What size was Mary's lamb?
BART	Shuffled Blended	little	how many lamb did mary have in the bible
BART	SQuAD	little	How big was Mary's lamb?
T5	Shuffled Blended	white	what color was the fleece of mary's lamb
T5	SQuAD	white	What color was Mary's lamb's fleece?
BART	Shuffled Blended	white	what color was the fleece of mary's lamb
BART	SQuAD	white	What color was the fleece of the little lamb?

Table 7: Mary Had a Little Lamb

Our models misconstrue references to Mary in the context sequence with the biblical character. It is easy to find web-based material reviewing the meaning of references to lambs in the bible and also discussions about whether "Mary Had a Little Lamb" is, in itself, about Mary from the bible. Given the vast amount of internet-sourced material used to train T5 and BART, what we see in our models is not surprising. It's also likely that T5 and BART have seen "Mary Had a Little Lamb" before. So, we substitute Fossie, the puppy, for Mary.

"**Fossie had a little lamb, its fleece was white as snow. Everywhere the puppy went, the little lamb was sure to go.**"

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	Fossie	who had a little lamb in snow
T5	SQuAD	Fossie	Who had a little lamb?
BART	Shuffled Blended	Fossie	what was the name of the little lamb
BART	SQuAD	Fossie	Who had a little lamb?
T5	Shuffled Blended	little	what was the size of fossie's lamb
T5	SQuAD	little	What was the size of Fossie's lamb?
BART	Shuffled Blended	little	how many lamb does a Fossie have
BART	SQuAD	little	How big was Fossie's lamb?
T5	Shuffled Blended	white	what color was the fleece of fossie's lamb
T5	SQuAD	white	What color was Fossie's lamb's fleece?
BART	Shuffled Blended	white	what color was Fossie's fleece
BART	SQuAD	white	What color was Fossie's fleece?

Table 8: Fossie Had a Little Lamb

Removing the reference to Mary got rid of the bible references. Fossie is not a common name and the models struggled a bit. Fossie was not "in snow" and the phrase "how many lamb does a Fossie have" is awkward at best. We understand that the phrase "its fleece was white as snow" is intended to refer to the lamb and not Fossie, but "it" is ambiguous. So while BART's references to the color of Fossie's fleece are technically correct, that is not how a person would be expected to respond. Next, we substitute Irving for Fossie.

"**Irving had a little lamb, its fleece was white as snow. Everywhere the child went, the little lamb was sure to go.**"

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	Irving	who had a little lamb that was white as snow
T5	SQuAD	Irving	Who had a little lamb?
BART	Shuffled Blended	Irving	who was the father of the little lamb
BART	SQuAD	Irving	Who had a little lamb?
T5	Shuffled Blended	little	What was Irving's lamb's size?
T5	SQuAD	little	What was Irving's lamb's size?
BART	Shuffled Blended	little	how many lamb did roberts have in the bible
BART	SQuAD	little	How big was Irving's lamb?
T5	Shuffled Blended	white	What color was Irving's fleece?
T5	SQuAD	white	What color was Irving's lamb's fleece?
BART	Shuffled Blended	white	what color was the fleece of the lamb
BART	SQuAD	white	What color was Irving's lamb's fleece?

Table 9: Irving Had a Little Lamb

The phrase "Irving had a little lamb" is ambiguous and can be interpreted both as "Irving possessed a little lamb" and "Irving fathered a little lamb" – so BART's unconventional response is not wrong. Surprisingly, another bible reference has emerged and it is not at all obvious who "roberts" is. According to https://www.meaningofname-baby.com/irving/ the name Irving does not appear in the Bible. We see another unconventional reference involving fleece color, although this time in one of the T5 models instead of BART. Moreover, in this case, it seems more likely that Irving is person instead of an animal and we don't often talk about a person's fleece. Finally, we replaced the lamb with a Labradoodle named Fossie. (Irving prefers Labradoodles to lambs).

"**Irving had a Labradoodle named Fossie, its fur was white as snow. Everywhere that child went, the little dog was sure to go.**"

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	Irving	who had a labradoodle named fossie
T5	SQuAD	Irving	Who had a labradoodle named Fossie?
BART	Shuffled Blended	Irving	who had a dog called Fossie
BART	SQuAD	Irving	Who had a dog named Fossie?
T5	Shuffled Blended	Fossie	what was the name of irving's dog
T5	SQuAD	Fossie	What was Irving's labradoodle?
BART	Shuffled Blended	Fossie	what was the name of the dog in the movie
BART	SQuAD	Fossie	What was Irving's labradoodle named?
T5	Shuffled Blended	little	What was Fossie's size?
T5	SQuAD	little	What was Fossie's size?
BART	Shuffled Blended	little	how old was the dog that went to the dog show
BART	SQuAD	little	How big was Fossie?
T5	Shuffled Blended	white	What color was Fossie's fur?
T5	SQuAD	white	What color was Fossie's fur?
BART	Shuffled Blended	white	What color was Fossie's fur?
BART	SQuAD	white	What color was Fossie's fur?
T5	Shuffled Blended	labradoodle	what kind of dog did irving have
T5	SQuAD	labradoodle	What type of dog did Irving have?
BART	Shuffled Blended	labradoodle	what kind of dog was fossie in the movie
BART	SQuAD	labradoodle	What breed of dog did Irving have?

Table 10: Irving Had a Little Labradoodle

"What was Irving's labradoodle?" is an awkward question given the target "Fossie". The references to "show" and "movie" are not appropriate given the context paragraph. On the other hand, the models text accurately reflects that Labradoodle is a kind or breed of dog.

D Goldilocks and the Three Bears

Once upon a time there were three bears who lived in a house in the forest.
There was a great big father bear, a middle-sized mother bear and a tiny baby bear.
One morning, their breakfast porridge was too hot to eat, so they decided to go for a walk in the forest.
While they were out, a little girl called Goldilocks came through the trees and found their house.
She knocked on the door and, as there was no answer, she pushed it open and went inside.
In front of her was a table with three chairs, one large chair, one middle-sized chair and one small chair.
On the table were three bowls of porridge, one large bowl, one middle-sized bowl and one small bowl – and three spoons.
Goldilocks was hungry and the porridge looked good, so she sat in the great big chair, picked up the large spoon and tried some of the porridge from the big bowl.
But the chair was very big and very hard, the spoon was heavy and the porridge too hot.
Goldilocks jumped off quickly and went over to the middle-sized chair.
But this chair was far too soft, and when she tried the porridge from the middle-sized bowl it was too cold.
So she went over to the little chair and picked up the smallest spoon and tried some of the porridge from the tiny bowl.
This time it was neither too hot nor too cold. It was just right and so delicious that she ate it all up.
But she was too heavy for the little chair and it broke in pieces under her weight.
Next Goldilocks went upstairs, where she found three beds.
There was a great big bed, a middle-sized bed and a tiny little bed.
By now she was feeling rather tired. So she climbed into the big bed and lay down.
The big bed was very hard and far too big. Then she tried the middle-sized bed, but that was far too soft.
So she climbed into the tiny little bed. It was neither too hard nor too soft.
In fact, it felt just right, all cosy and warm. And in no time at all Goldilocks fell fast asleep.
In a little while, the three bears came back from their walk in the forest.
They saw at once that someone had pushed open the door of their house and Father Bear looked around.
Then roared with a growly voice.
Mother Bear said in a quiet gentle voice. “Somebody has been sitting in my chair.”
Then Little Bear said in small squeaky baby voice. “Somebody has been sitting in my chair and has broken it!”
Then Father Bear looked at his bowl of porridge and saw the spoon in it and he said in his great big growly voice, “SOMEBODY HAS BEEN EATING MY PORRIDGE”
Then Mother Bear saw that her bowl had a spoon in it, and said in her quiet voice. “Somebody has been eating my porridge.”
Little Bear looked at his porridge bowl and said in his small squeaky baby voice, “Somebody has been eating my porridge, and has eaten it all up:-”
Then the three bears went upstairs, and Father Bear saw at once that his bed was untidy, and he said in his great big growly voice, “SOMEBODY HAS BEEN SLEEPING IN MY BED!”
Mother Bear saw that her bed, too, had the bedclothes turned back, and she said in her quiet gentle voice, “Somebody has been sleeping in my bed!”
Then Little Bear looked at his bed and said in his small squeaky baby voice, “Somebody is sleeping in my bed!”
He squeaked so loudly that Goldilocks woke up with a start.
She jumped out of bed, and away she ran, down the stairs and out into the forest. And the three bears never saw her again.

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	the forest	Where did the three bears live?
T5	SQuAD	the forest	Where did the three bears go for a walk?
BART	Shuffled Blended	the forest	Where did Goldilocks live?
BART	SQuAD	the forest	Where did the three bears live?
T5	Shuffled Blended	it broke	What happened to the little chair?
T5	SQuAD	it broke	What happened to the little chair?
BART	Shuffled Blended	it broke	What happened to the little chair?
BART	SQuAD	it broke	What happened to the little chair?
T5	Shuffled Blended	it was too hot	What was the problem with the porridge?
T5	SQuAD	it was too hot	What was the problem with the porridge?
BART	Shuffled Blended	it was too hot	Why did Goldilocks want to eat porridge?
BART	SQuAD	it was too hot	Why did Goldilocks go for a walk in the forest?
T5	Shuffled Blended	tiny	What was the size of the baby bear?
T5	SQuAD	tiny	What was the size of the little bed that Goldilocks climbed into?
BART	Shuffled Blended	tiny	What was the size of Goldilocks’s bed?
BART	SQuAD	tiny	How big was the little bed?

Table 11: Goldilocks and the Three Bears Results

We do not know from the story that Goldilocks lived in the forest, although it might be a fair inference. Then again, she could just be visiting. BART’s predictions to “it was too hot” are not correct. On the other hand, there are several references to tiny things in the story, each predicted answer is reasonable.

E The Big Game

Figure 3: Go Bears!!!

BERKELEY - Jackson Sirmon scored on a 37-yard return following a wild play with two fumbles and the Cal football team rallied from a double-digit deficit in the fourth quarter to beat Stanford 27-20 in the 125th edition of the Big Game on Saturday.

"It's really neat... as a defensive player, you don't get a lot of opportunities to score touchdowns and you have to enjoy them when you get them," said Sirmon, who played in his first Big Game on Saturday. "It was very neat especially considering the timing of it all."

The Golden Bears (4-7, 2-6 Pac-12) trailed 17-6 in the fourth quarter when Jack Plummer started the comeback with a 1-yard touchdown pass to Monroe Young.

Then things really went crazy on the ensuing possession for Stanford (3-8, 1-8).

Backup quarterback Ashton Daniels took a direct snap and ran to his left. He was stripped of the ball by Daniel Scott and Cal's Jeremiah Earby recovered. But Daniels then knocked the ball loose from Earby for a second fumble on the play only to have Sirmon scoop it up and run in to give Cal a 20-17 lead with 9:54 to play.

"It bounced right to me," Sirmon said. "I was in the right place at the right time. All I did that play is I didn't mess it up. The ball came right to me and I ran with it."

Stanford had three more drives to rally. But the Cardinal punted twice and Scott came up with an interception with just more than two minutes to play, setting up Jaydn Ott's 1-yard run that sealed the win in front of a sellout crowd of 51,892.

"We know how important this game is to us as a team, our administration and our support staff," Travers Family Head Football Coach Justin Wilcox said. "Our fans and students were just incredible tonight. What an awesome environment." (Associated-Press, 2022)

Base Model	Training Dataset	Answer	Prediction
T5	Shuffled Blended	Jakson Sirmon	Who scored on a 37-yard return in the Big Game?
T5	SQuAD	Jakson Sirmon	Who scored on a 37-yard return?
T5	Shuffled Blended	Stanford	Who did Cal beat in the Big Game?
T5	SQuAD	Stanford	Who did the Golden Bears beat in the 125th edition of the Big Game?
T5	Shuffled Blended	Jaydn Ott	Who scored a 1-yard run in the final quarter of the Big Game?
T5	SQuAD	Jaydn Ott	Who scored a 1-yard run in the fourth quarter?
T5	Shuffled Blended	51,892	How many people attended the game?
T5	SQuAD	51,892	How many people attended the game?
T5	Shuffled Blended	17-6	How many points did Cal trail in the fourth quarter?
T5	SQuAD	17-6	How many points did the Golden Bears trail in the fourth quarter?
T5	Shuffled Blended	20-17	What was the final score for Cal in the Big Game?
T5	SQuAD	20-17	What was the score with 9:54 to play?
T5	Shuffled Blended	27-20	How many points did Cal score in the 125th Big Game?
T5	SQuAD	27-20	What was the final score of the 125th edition of the Big Game?
T5	Shuffled Blended	three-drives	How many drives did Stanford have to rally?
T5	SQuAD	three-drives	How many drives did Stanford have to rally?

Table 12: Big Game Story Results

The T5 model trained on the shuffled blended dataset mistook 20-17 as the final score when it was the score with 9:54 left to play, the SQuAD-trained model got that right. The model trained on the shuffled blended dataset also attributed the entire score (27-20) that reflects both teams' scores to Cal. The SQuad-trained version had the correct answer.

F Comparing T5 and BART for Question Generation

As described in Section 3, we created ten models by fine-tuning T5 and BART on the four source datasets, as well as on a blended dataset. We ran each of those ten models on the validation data for each source dataset and grouped the results by model. See Table 13.

We ran Welch's T-Test on each metric for each similarly trained pair of T5 and BART models. The differences in the distribution of performance data were not statistically significant in 13 of the 25 comparisons. Of the 12 instances where the differences in the distribution were statistically significant, BART prevailed ten times and T5 twice.

- BART outperformed T5 for each metric on QuAC and Trivia QA-trained models. The relevant BART scores are shown in Table 13 in blue. Those models are also among the weaker-performing models.
- T5 outperformed BART on the Shuffled Blended trained models for BERTScore and on the NQ-trained models for ROUGE-L. The relevant T5 scores are shown in Table 13 in red. Our best-performing models are the models trained on the Shuffled Blended dataset (followed next by the SQuAD-trained models). See Appendix I Table 18.
- Where both numbers for a metric comparison are black, the difference between them is not statistically significant.

After considering only those instances where differences in the results between T5 and BART models were statistically significant we concluded that on balance, the two architectures were indistinguishable for our project with a slight edge favoring one or the other depending on the circumstances.

Training Dataset	Base Model	BLEU	ROUGE-L	METEOR	BERTScore	USE
Shuffled Blended	T5	10.05	39.17	38.43	83.29	58.57
	BART	9.87	38.87	38.60	83.15	58.40
NQ	T5	2.35	29.80	23.00	76.92	50.10
	BART	2.50	29.35	22.86	76.85	50.25
QuAC	T5	0.50	9.17	9.78	70.55	17.93
	BART	2.47	26.01	22.00	78.02	42.92
SQuAD	T5	6.45	33.60	32.10	81.20	52.81
	BART	6.58	33.42	32.22	81.20	52.88
TriviaQA	T5	3.59	26.75	29.13	79.01	48.49
	BART	4.47	29.44	30.90	79.61	50.75

Table 13: Comparing T5 and BART for Question Generation

G Evaluation of Lexical and Semantic Metrics

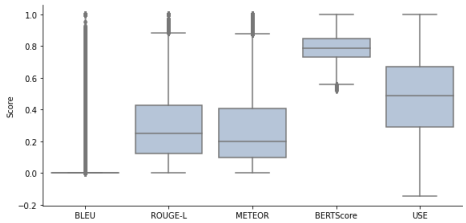

Figure 4: Evaluation Metric Distributions

We chose BLEU, ROUGE-L, METEOR, BERTScore, and USE as evaluation metrics. Figure 4 shows the distribution of each for predictions generated across all models. We examined statistical trends and individual text examples to consider how these metrics could be used to assess the quality of a generated question.

As we describe in Section 5.3.1, we chose METEOR as our preferred lexical metric for its robustness against variations in word choice and order. BLEU was the least useful metric, having a value of zero for 88% of predictions across all models. Since BLEU uses a precision-based geometric mean of n-gram matches up to 4-grams, a lack of matches at any of the four levels results in a score of zero. Predictions with top BLEU scores and bottom METEOR scores had many overlapping words but missed key question topics.[8] Similarly, predictions with higher METEOR scores and lower ROUGE-L scores generally aligned better with target questions.

We found BERTScore and USE to be effective when assessing the quality of generated questions. See Section 5.3.2. Table 14 shows three target-prediction pairs with their respective BERTScore and USE scores and whether those scores are high or low. For the first two pairs, the generated questions have observable partial semantic alignment and only score well on one of the metrics. The third pair is the highest quality question and, of the three, is the only one that scores well on both metrics. Though each metric can be used individually, predictions that perform well on both are typically higher-quality questions.

Target	Prediction	BERTScore	USE	BERTScore Strength	USE Strength
What is colonialism's core meaning?	What is the relation between imperialism and colonialism?	88.97	21.12	high	low
What vehicle did Doctor Who use for his escape from Gallifrey?	what is the TARDIS in doctor who	72.60	78.87	low	high
Persistent unemployment has what effect on long-term economic growth?	: what is the effect of high and persistent unemployment on economic growth?	88.86	85.02	high	high

Table 14: BERTScore and USE Examples with Varying Agreement

The highest quality predictions generally had high METEOR and semantic scores simultaneously. Table 15 consists of high-quality predictions with high METEOR, BERTScore, and USE scores. The examples showcase the ability of these metrics to identify a high quality question despite lexical variation from the target.

Target	Prediction	METEOR	BERTScore	USE
What parts of plants have chloroplasts?	Chloroplasts are found in which part of a plant?	50.96	90.69	86.98
what was the name of the other HD channel Virgin media could carry in the future?	What HD channel did Virgin Media have an option to carry in the future?	57.62	90.87	84.99
Matter particles are shown as what kind of lines in a Feynman diagram?	In a Feynman diagram, what line is each matter particle represented as?	63.71	90.05	84.99
Approximately how many adherents does the United Methodist Church have across the world?	How many members are there in the United Methodist Church?	46.33	89.14	84.99
who played the beast in the movie beauty and the beast	Who plays the title role in the film Beauty and the Beast?	78.87	89.78	84.98
Where did Tesla go to feed the pigeons daily?	In which park did Tesla walk to feed the pigeons?	66.58	90.94	84.97
What percentage of Newcastle's population is it believed the Bolivians account for?	What percentage of Newcastle's population are Bolivians?	57.70	92.74	84.94
Singer Dan Reynolds is the frontman for which US rock band?	Dan Reynolds is the lead singer with which band?	63.56	89.04	84.94
Carbon dioxide in solid form is called what?	What is the frozen solid form of carbon dioxide known as?	61.55	88.03	77.82
"For what television series was the theme tune ""I'll be there for you"" ?"	"Which American TV series had a theme song called ""I'll Be There For You""?"	70.42	94.97	77.75
Who's home runs record did Mark McGwire break in the 1998 season?	In 1998, Big Mac McGwire broke the Major League Baseball home run record by hitting 70 home runs. Who did he beat?	52.90	85.85	73.04
Used to describe US citizens who attempted to foment insurrections in Latin America in the mid-19th century, what is the term for someone who engages in an unauthorized military expedition into a foreign country to support a revolution?	What is the name given to an unauthorized military expedition into a foreign country or territory to foment or support a revolution?	46.13	87.11	72.49

Table 15: High METEOR and High Semantic Scores

[8] We define "top" or "high" as scores in the metric's 75th percentile across all of our experiments. We similarly define "bottom" or "low" as scores in the metric's 25th percentile.

H Conflicting Semantic and Lexical Metric Scores

Table 16 consists of predictions with at least one high semantic score but a low lexical metric score. These have different words than the target, *but the essence is the same*.

Target	Prediction	BLEU	ROUGE-L	METEOR	BERTScore	USE
who is beauty and the beast written by	Who wrote La Belle et la Bête?	0.0	12.50	6.25	71.57	82.44
what is the first basic process in the light reaction of photosynthesis	How does photosynthesis begin?	0.0	12.50	4.42	80.56	77.28
How large are Cytoplasmic ribosomes?	what is the size of a chloroplast ribosome	0.0	0.00	8.06	84.11	76.81
What was the nationality of composer Frederic Chopin?	where did frédéric chopin come from	0.0	12.50	5.75	80.00	76.61
From which common English tree are cricket stumps traditionally made?	what kind of wood is used for stumps in cricket	0.0	10.00	9.17	79.89	73.04

Table 16: High Semantic and Low Lexical Scores

Table 17 consists of predictions with high lexical metric scores but low semantic metric scores. For these, even though the coincidental use of words may result in a strong lexical metric score, the predicted questions *do not adequately capture the essence* of the target question. We note that the BERTScores shown in that table are not all especially low, but the USE scores are. Since we believe that it is best to look at both semantic scores and not one or the other, we take the low USE scores as a sufficient indicator of semantic similarity weakness.

Target	Prediction	BLEU	ROUGE-L	METEOR	BERTScore	USE
What happened after the mass?	what happened to the president of the philippines	0.0	46.15	41.22	72.16	27.77
Who was at the conference?	who was the prime minister of australia at the time of the war	0.0	44.44	47.11	67.51	25.36
What is the exam at the end of Form Four?	what is the end of primary education in kenya	0.0	52.63	44.81	72.96	25.14
What was the reaction to the event?	what was the name of the man in tiananmen square	0.0	47.06	45.73	72.47	20.84
what was the concert for?	what was the theme song for pirates of the caribbean	0.0	53.33	49.32	71.09	20.51

Table 17: Low Semantic and High Lexical Scores

I Model Comparison

Table 18 shows model performance for all inferences, sorted by USE scores.

Training Dataset	Base Model	BLEU	ROUGE-L	METEOR	BERTScore	USE
Shuffled Blended	T5	10.05	39.17	38.43	83.29	58.57
Shuffled Blended	BART	9.87	38.87	38.60	83.15	58.40
SQuAD	BART	6.58	33.42	32.22	81.20	52.88
SQuAD	T5	6.45	33.60	32.10	81.20	52.81
TriviaQA	BART	4.47	29.44	30.90	79.61	50.75
NQ	BART	2.50	29.35	22.86	76.85	50.25
NQ	T5	2.35	29.80	23.00	76.92	50.10
TriviaQA	T5	3.59	26.75	29.13	79.01	48.49
QuAC	BART	2.47	26.01	22.00	78.02	42.92
QuAC	T5	0.50	9.17	9.78	70.55	17.93

Table 18: Model Comparison

J Testing Context Length Impact with TriviaQA

We expected that context length would impact model performance and that longer context sequences would tend to confound models more than shorter context sequences. Our models perform better on SQuAD and NQ compared to TriviaQA. TriviaQA has substantially longer context sequences than either. See Appendix B, Figure 2.

To test our hypothesis, we partitioned the TriviaQA validation set into 20 quantiles (each a ventile) based on the length of the input sequences. The first ventile has sequences of up to 379 words, and the second has sequences of 379 to 611 words. The first ventile includes 493 samples, and the second 494. We created test sets from each. Additionally, we created a combined set comprised of the data from both ventiles. We ran our models on the "short sample" test data. We reviewed the results for the four models that had the strongest performance on the full TriviaQA validation set (BART trained on TriviaQA and BART trained on the shuffled blended dataset). We did not use T5 because BART performed better than T5 on TriviaQA across the board.

We found statistically significant improvement in METEOR and BERTScore results. See Tables 19 and 20 below. In each of those cases, the performance differences among the "short samples" data sets is not statistically significant, but the improvement over the full validation set performance is statistically significant.

However, we did not observe a statistically significant change for USE. See 21. There is no statistically significant difference among any of the results reported in that table.

These results supported our hypothesis that context length impacts question generation model performance and is at least part of the challenge posed by TriviaQA. However, even with the improved scores, the best results on TriviaQA still fall below the equivalent SQuAD and NQ results. This finding supports the contention that TriviaQA is inherently a more complex dataset and that only part of that complexity is attributable to context length.

Test Dataset	Base Model	Training Dataset	METEOR
ventile 1	BART	TriviaQA	38.51
ventile 2	BART	Shuffled Blended	38.44
decile 1	BART	Shuffled Blended	38.40
ventile 1	BART	Shuffled Blended	38.36
decile 1	BART	TriviaQA	38.20
ventile 2	BART	TriviaQA	37.90
TriviaQA	BART	Shuffled Blended	35.46
TriviaQA	BART	TriviaQA	35.10

Table 19: METEOR

Test Dataset	Base Model	Training Dataset	BERTScore
ventile 1	BART	TriviaQA	84.02
ventile 1	BART	Shuffled Blended	83.94
decile 1	BART	Shuffled Blended	83.90
decile 1	BART	TriviaQA	83.88
ventile 2	BART	Shuffled Blended	83.85
ventile 2	BART	TriviaQA	83.74
TriviaQA	BART	Shuffled Blended	82.64
TriviaQA	BART	TriviaQA	82.59

Table 20: BERTScore

Test Dataset	Base Model	Training Dataset	USE
ventile 2	BART	TriviaQA	61.63
ventile 2	BART	Shuffled Blended	61.41
decile 1	BART	TriviaQA	60.99
TriviaQA	BART	Shuffled Blended	60.92
decile 1	BART	Shuffled Blended	60.89
TriviaQA	BART	TriviaQA	60.78
ventile 1	BART	Shuffled Blended	60.37
ventile 1	BART	TriviaQA	60.36

Table 21: USE